FAKULTÄT FÜR INFORMATIK

DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Efficient Lifting Methods for Variational Problems

Thomas Möllenhoff

Technische Universität München

Fakultät für Informatik
Lehrstuhl für Bildverarbeitung und Künstliche Intelligenz

Efficient Lifting Methods for Variational Problems

Thomas Möllenhoff

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften
(Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Rüdiger Westermann

Prüfer der Dissertation:
1. Prof. Dr. Daniel Cremers
2. Prof. Dr. Bastian Goldlücke
Universität Konstanz

Die Dissertation wurde am 29.05.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 23.07.2020 angenommen.

Abstract

Variational methods have a long and colorful history in image analysis and computer vision. Unfortunately, most of the variational problems appearing in practical applications are nonconvex. One idea to tackle such nonconvex problems is by lifting the energy functional to a higher dimensional space and then performing a convex relaxation. The increased dimensionality due to the lifting procedure relies on efficient implementations, which are the primary concern of this thesis.

Inspired by similar constructions in the context of discrete-continuous Markov random fields, we propose a relaxation which assigns meaningful costs to fractional labelings in a spatially continuous multilabeling setup. In experiments, we show that this formulation leads to substantial savings in memory and runtime and allows one to interpolate between a naive convexification and tighter relaxations more gracefully. Additionally, we show that these sublabel-accurate labeling approaches correspond to a dual finite-element approximation of a relaxation for problems with infinite label-spaces. This viewpoint allows one to extend the previous results to more general regularizations. Also, it provides a new and principled way to implement existing functional lifting formulations and suggests the possibility of even more accurate discretizations.

In this thesis, we also contribute to functional lifting methods for vectorial variational problems. A difficulty for vectorial problems arises from the representation of minimal surfaces with dimension and codimension larger than one. Previous approaches resorted to discrete multilabeling approximations or representations, where one instead considers multiple surfaces of codimension one. Using tools from geometric measure theory, we present a lifting which considers a single surface in higher codimension. A convex formulation is obtained by relaxing the search space from surfaces to currents. Unlike previous relaxations, it allows one to tackle vectorial problems with general polyconvex regularizations. The proposed discretization of the resulting minimal surface problem with Whitney forms includes previous sublabel-accurate formulations as special cases.

Finally, we demonstrate that the introduced formalisms from geometric measure theory have applications beyond the convexification of variational problems. Specifically, we propose to represent high-dimensional data with currents. This representation can be interpreted as a first-order approximation to the data manifold. Based on the flat norm, we propose FlatGAN, which is a formulation in the spirit of generative adversarial networks (GANs), but generalized to currents. In experiments, we show that the formulation with currents allows one to learn interpretable and disentangled latent representations.

Zusammenfassung

Variationsmethoden haben eine lange Geschichte in der mathematischen Bildverarbeitung und dem maschinellen Sehen. Leider sind die meisten Variationsprobleme, die in praktischen Anwendungen auftreten, nicht konvex. Eine Möglichkeit um solche Probleme anzugehen, besteht darin, das Energiefunktional in einem allgemeinerem Raum einzubetten um dort dann eine gute konvexe Relaxierung zu berechnen. Die erhöhte Dimensionalität erfordert ein besonderes Augenmerk auf einer effizienten Darstellung, welche das Hauptanliegen dieser Arbeit sein wird.

Inspiriert durch vergleichbare Ansätze im Kontext von diskret-kontinuierlichen *Markov random fields*, schlagen wir ein räumlich kontinuierliches Multilabeling-Verfahren vor, welches auch nicht-ganzzahligen Labelzuständen sinnvolle Kosten zuweist. In Experimenten zeigen wir, dass diese Formulierung zu erheblichen Einsparungen bezüglich des Speicherverbrauchs und der Laufzeit gegenüber vorherigen Methoden führt. Zusätzlich zeigen wir, dass diese Formulierung auch alternativ durch die Appoximation einer gewissen Dualformulierung mittels stückweise linearen finiten Elementen hergeleitet werden kann. Diese Sichtweise ermöglicht es, die vorherigen Ergebnisse auf allgemeinere Kostenfunktionale zu erweitern. Außerdem bietet sie eine prinzipientreue Möglichkeit, bestehende Relaxierungen für Variationsprobleme zu implementieren, und suggeriert die Möglichkeit noch genauerer Diskretisierungen.

Weiterhin betrachten wir auch konvexe Relaxierungen für allgemeine Variationsprobleme mit vektoriellem Wertebereich. Bestehende Ansätze basieren meistens auf Darstellungen durch diskrete Zustände oder Formulierungen bei denen mehrere Flächen in Kodimension eins betrachtet werden. Mittels Einsichten aus der geometrischen Maßtheorie schlagen wir eine neue Methode vor, bei der nur eine einzige Fläche in höherer Kodimension verwendet wird. Ein konvexe Formulierung erhalten wir, indem der Suchraum von Flächen auf Ströme (engl. *currents*) erweitert wird. Im Gegensatz zu früheren Arbeiten, können wir nichtkonvexe Variationsprobleme mit allgemeiner polykonvexer Regularisierung konvex darstellen.

Als Ausblick zeigen wir schließlich noch, dass die eingeführten Methoden der geometrischen Maßtheorie auch Anwendungen im maschinellen Lernen haben. Insbesondere schlagen wir die neue Sichtweise vor, hochdimensionale Datensätze als Ströme zu verstehen. Basierend auf der *flat norm* verallgemeinern wir die kürzlich vorgestellten *generative adversarial networks* (GANs) von Wahrscheinlichkeitsverteilungen auf Ströme. In Experimenten zeigen wir dass es durch diese Verallgemeinerung möglich wird, mittels GANs eine interpretierbare und entflochtene Repräsentationen der hochdimensionalen Datenverteilung zu lernen.

Acknowledgments

First and foremost, I would like to thank my doctoral advisor Prof. Daniel Cremers for his consistent support and guidance over the years. Thank you, Daniel, for many fruitful and inspiring discussions, and for giving me the freedom to independently explore different topics and directions. I would also like to thank Prof. Bastian Goldlücke for his interest in this work and for agreeing to be the second referee.

During my time at TU Munich, I had the privilege to meet many amazing people. In particular, I would like to thank Eno Töppe, Claudia Nieuwenhuis, and Evgeny Strekalovskiy for all their advice and for introducing me to the world of research when I had just started as a master student in the lab. Emanuel Laude, Jan Lellmann, and Michael Möller for the great collaboration on sublabel-accurate multilabeling, which a substantial part of this thesis is based on. Mohamed Souiai and Thomas Windheuser for all the helpful advice and inspiring discussions. Thanks, Thomas, also for introducing me to the fascinating topic of geometric measure theory, which eventually lead to the papers on which the second half of this thesis is based on. Thomas Frerix, Björn Häfner, Yvain Quéau, Tao Wu, Zhenzhang Ye, and Pierre Bréchet for the fruitful collaborations. Hartmut Bauermeister, Florian Bernard, Peter Ochs, Sarah Sachs and Thomas Vogt for all the interesting conversations.

Additionally, I want to thank Prof. Christopher Zach for running experiments and sharing his results on the discrete-continuous graphical models, and Prof. Kevin R. Vixie for his valuable feedback related to geometric measure theory.

I would like to thank Sabine and Quirin, who helped me out so much with all the administrative and technical issues. Thanks also to everyone else from the lab, who made my time over the last years very enjoyable. John and Tao for the many after-work dinners. Csaba and Zhenzhang for being great office mates. The optimization guys (you know who you are!) for all the nice coffee and sandwich breaks. Also, all the other wonderful people I've had the chance to meet along this journey: Martin, Frank, Jan, Youngwook, Maria, Julia, Erik, Vladi, Hubert, Christian, Jakob, Matthias, Ema, Frank, Zorah, Rudi, Virginia, Lingni, Vladyslav, Robert, Georg, Songyou, Yuesong, Rui, Nan, Christiane, David and Niko!

Finally, I would like to thank my family for always supporting me over the years.

Contents

List of Symbols

General Notations

$\mathcal{X}, \Omega$	Domain
$\mathcal{Y}, \Gamma$	Range, label space
$\mathcal{G}_f$	Graph of the function f
$\mathbf{1}_f$	Indicator function of the subgraph of f
ρ, η	Data term ρ and regularizer η
$\|\cdot\|_{S^1}, \|\cdot\|_{S^\infty}$	Schatten–1 and Schatten–∞ norms
Δ^ℓ	The $(\ell-1)$-dimensional probability simplex

Geometric Measure Theory

$BV(\mathcal{X}; R)$	Functions of bounded variation with domain $\mathcal{X}$ and range $\mathbf{R}$
$\mathcal{C}_c^\infty(\mathcal{X}; \mathbf{R}^n)$	Smooth vector-fields with compact support
$\mathcal{L}^n$, dx	n-dimensional Lebesgue measure
$\mathcal{H}^m$	m-dimensional Hausdorff measure
$\mu \llcorner S$	Restriction of the measure μ to the set S, $(\mu \llcorner S)(A) = \mu(A \cap S)$
Du	Distributional derivative of the function u
δ_x	Dirac distribution at x
$d\omega$	Exterior derivative of the differential form ω
∂T	Boundary of the current T
$g^\sharp \omega$	Pullback of the differential form ω
$g_\sharp T$	Pushforward of the current T
spt T	Support of the current T
$\mathbb{M}$	Mass norm
$\mathbb{F}$	Flat norm
$[\![\mathcal{M}]\!]$	Lifts the oriented manifold $\mathcal{M}$ to a current
$\mathcal{D}^k(\mathbf{R}^n)$	Smooth, compactly supported differential k-forms in $\mathbf{R}^n$
$\mathcal{D}_k(\mathbf{R}^n)$	k-currents in $\mathbf{R}^n$, dual space of differential forms
$\mathbf{M}_k(\mathbf{R}^n)$	Finite mass k-currents in $\mathbf{R}^n$
$\boldsymbol{N}_{k,\mathcal{X}}(\mathbf{R}^n)$	Normal k-currents in $\mathbf{R}^n$ with support in $\mathcal{X}$

Convex Analysis and Optimization

epi$f$	Epigraph of the function f
δ_C	Indicator function of set $C \subset \mathbf{R}^n$
f^*, f^{**}	Convex conjugate and biconjugate of f
prox$_{\tau,f}$	Proximal operator of f with step size τ

Part I

Introduction and Foundations

Chapter 1

Introduction

1.1 Motivation and Overview

Since their adoption by the computer vision community nearly forty years ago [HS81; IH81], variational methods[1] enjoyed a long and colorful history. Over the years, many different variational problems have been proposed to tackle a diverse set of tasks and applications such as image restoration problems, reconstruction of 3D geometries from a set of images, or the estimation of motions, see for example [Sch+09; CP16a] for an overview.

The (often nonconvex) optimization of such variational problems has also received an equally or perhaps even more diverse treatment, both within the discrete [Kap+13] and the continuous [CP16a] optimization communities. This thesis belongs to the field of continuous optimization as our optimization problems will be integral functionals. The key motivation is that such formulations in function space are quite general and flexible as they admit a variety of different approximations and discretizations, some of which can be more faithful to underlying continuous phenomena. The resulting discrete formulations would often be difficult to reach if one would perform finite approximations at an too early stage. On the optimization side, the major part of this thesis will be concerned with global minimization approaches for variational problems based on a technique called *functional lifting*. The driving motivation behind global optimization approaches (both in the discrete and the continuous worlds) is to have a robust and transparent way to solve problems, which is also explainable and reliable in the sense that one is also clearly aware of the limitations. In contrast, for local or learning-based approaches, it can often be tricky to disentangle whether specific effects are due to the initialization, selection of tuning parameters, carefully curated training and

[1] Here and throughout this thesis, the term ”variational” strictly refers to its original meaning from the *calculus of variations*, i.e., the minimization of integral functionals. Note that in the optimization community, the word variational is now used in a broader context (”freed from its limitations of the past” as remarked in [RWW98]). In machine learning, ”variational” usually refers to optimization over families of probability distributions in the context of approximate Bayesian inference, see for example, [WJ+08; BKM17].

testing sets, the chosen optimization procedure or the mathematical modeling of the problem.

Despite a considerable amount of previous research efforts, global optimization methods for variational problems have not entered the ”mainstream” yet. We believe this is due to the following reasons.

1. Due to their global nature and a certain ”lifting” procedure (which increases the dimensionality) the methods traditionally come with excessive demands in runtime and memory.

2. They have limited applicability and are restricted to certain classes of variational energies.

3. The mathematical formalism of *geometric measure theory* lurking behind the lifting procedure is known to be technically involved and is usually not treated in the standard computer science curriculum.

Instead, local or discrete optimization approaches are still the preferred choice for most practitioners. The major part of this thesis presents a conscious effort to make substantial progress in all of the three directions above. Specifically, we address the points as follows:

1. **Efficiency in runtime and memory.** This thesis present a framework for sublabel-accurate multilabeling, which scales more gracefully in the dimension than previous functional lifting methods. Specifically, it allows one to control the amount of lifting more gradually. The limit case with minimal lifting is as efficient as the local optimization. Further, we interpret these sublabel-accurate multilabeling methods as dual finite-element approximations of relaxations with infinite label spaces. This insight yields more efficient formulations for a large class of previously proposed relaxations.

2. **Generality and applicability.** Essentially, the lifting procedure amounts to a reformulation of the original task as a minimal surface problem in a higher-dimensional space. Almost all previous lifting methods either employ one or more surfaces of codimension one. Instead, we consider minimal surface problems with codimension higher than one, and show that this allows one to introduce more general (polyconvex) regularizations.

3. **Technicality and accessibility.** We hope to convince the reader that most of the ideas which are eventually relevant for implementation are of an intuitive geometric nature and do not require the more technical subtleties of the subject. While the primary aim of this thesis is to make progress in the two points above, it also represents an effort to distill key ideas, concepts and intuitions from geometric measure theory which are useful for applications.

At the end of this thesis, we show that the notions from geometric measure theory can also be useful beyond the relaxation of variational problems. Specifically, we show that *currents*, which are a generalized and flexible notion of surface, can be used to represent high-dimensional data in machine learning. This rests upon a critical assumption, that is sometimes referred as the *manifold hypothesis*: the distribution of real-world data often concentrates nearby a low-dimensional manifold embedded in the high-dimensional data space [FMN16].

This will put the previous theory into a broader context: optimization problems over surfaces and currents in general codimension can have interesting applications, without necessarily stemming from the relaxation of a variational problem.

1.2 Variational Methods

In this section, we will introduce the basic idea of variational methods and also introduce some notations that we will keep throughout this thesis.

The main idea for variational approaches is to model the problem at hand with an *energy functional*. For each candidate solution, which in our setting is a map $f : \mathcal{X} \to \mathcal{Y}$ between spaces $\mathcal{X}$ and $\mathcal{Y}$, an energy functional E outputs a real number or perhaps plus infinity[2]. The structure of the sets $\mathcal{X}$ and $\mathcal{Y}$ and the energy E depend on the application, but the main idea is to formulate E in such a way that low energy values correspond to desirable solutions.

The aim of an optimization procedure is to find a candidate $f : \mathcal{X} \to \mathcal{Y}$ with minimal energy. The prototypical problem is given as follows:

$$E^* = \inf \left\{ E(f) \mid f : \mathcal{X} \to \mathcal{Y} \right\}. \tag{1.1}$$

A central distinction is typically made between convex and nonconvex energy functionals E, as pointed out by R. T. Rockafellar in his 1993 SIAM review paper: " ... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity. ".

As illustrated in Fig. 1.1, an advantage of convex energies is, that every point which cannot be improved locally is also the best in a global sense. Therefore, one can apply algorithms which operate locally such as the method of steepest descent, and not worry too much about a suitable starting point.

Unfortunately, many interesting or relevant practical tasks are much more easily formulated as a nonconvex optimization problem. In this thesis, we will develop efficient convex formulations for a certain class of nonconvex energy functionals. We will introduce the specific energies we are concerned with in the following section.

[2] In Chapter 3, Chapter 4 and Chapter 5 we use the notation $u : \Omega \to \Gamma$ instead of $f : \mathcal{X} \to \mathcal{Y}$.

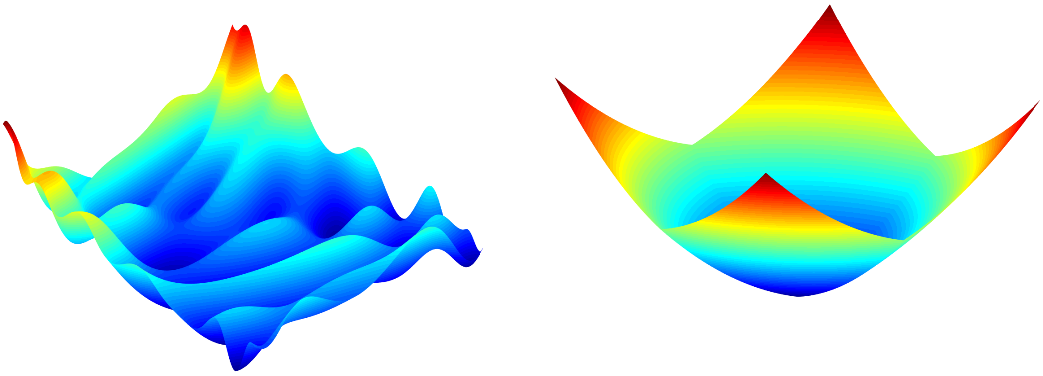


Figure 1.1: Example of a nonconvex energy function and a convex one.

1.2.1 A class of first-order energies

The energies we consider in the following are often denoted as ”first-order energies”, since they involve the first derivative of the function f. Specifically, for continuously differentiable maps $f : \mathcal{X} \to \mathcal{Y}$, we consider the following class of integral functionals:

$$E(f) = \int_{\mathcal{X}} c(x, f(x), \nabla f(x))\, \mathrm{d}x. \tag{1.2}$$

For $\mathcal{X} \subset \mathbf{R}^n$, $\mathcal{Y} \subset \mathbf{R}^N$, the cost integrand (sometimes referred to as a *Lagrangian density*) $c : \mathcal{X} \times \mathcal{Y} \times \mathbf{R}^{N \times n} \to \mathbf{R}$ is assumed to be nonnegative.

In some applications, this density (or cost) is given as a separable sum of two terms, often denoted as the data fidelity term and regularizer:

$$c(x, y, \xi) = \underbrace{\rho(x, y)}_{\text{data term}} + \underbrace{\eta(\xi)}_{\text{regularizer}}. \tag{1.3}$$

This splitting of the cost into a prior and a data model can also be motivated from the perspective of maximum a-posteriori (MAP) inference in a Bayesian setting, see for example [Mum94]. Given an observation z, the posterior probability of a candidate f is proportional to the product of a likelihood and a prior:

$$p(f \mid z) \propto p(z \mid f) \cdot p(f). \tag{1.4}$$

Instead of maximizing the probability, minimizing the negative logarithm of (1.4) and restricting the prior to only depend on the gradient of f leads to the above choice of Lagrangian density (1.3).

Functionals of the form (1.2), possibly subject to further side constraints, find a large variety of applications. In the following, we will give an overview of some common application examples.

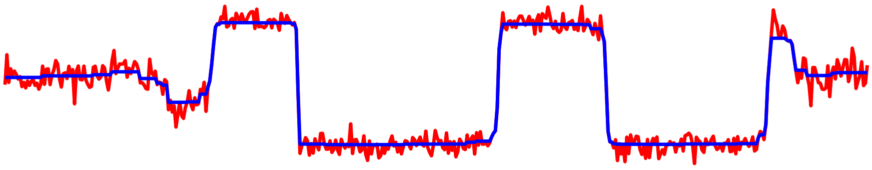


Figure 1.2: Discontinuity preserving denoising of a given signal z (shown in red) by total variation minimization.

1.2.2 Example applications

Denoising. As a simple example, let us consider the denoising of an observed signal $z : [0,1] \to \mathbf{R}$ that has been corrupted by Gaussian noise. Assuming independent identically distributed noise, the negative log-likelihood is given by

$$-\log p(z \mid f) = \frac{1}{2\sigma} \int_{\mathcal{X}} (f(x) - z(x))^2 \mathrm{d}x + \text{const.} \tag{1.5}$$

Note that in practical scenarios, this noise model can be unrealistic. Often the noise is known to instead follow a Cauchy distribution (see, e.g. [Mei+18]), which leads to a nonconvex energy model. The methods proposed in this thesis would allow also for the (near global) optimization of such energies.

A popular prior due to its edge-preserving property is the *total variation*, which for continuously differentiable f is given by

$$-\log p(f) = \int_{\mathcal{X}} |\nabla f(x)| \,\mathrm{d}x. \tag{1.6}$$

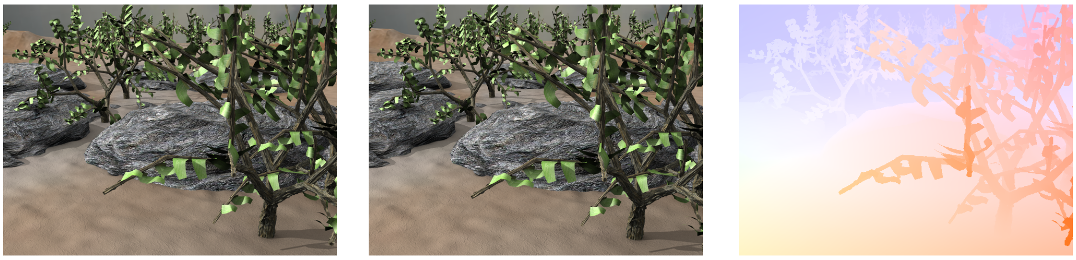
It has been extensively studied and applied in the image analysis and computer vision communities [Cha+10].

Putting everything together and minimizing the negative logarithm of (1.4) corresponds to selecting the Lagrangian density in Eq. (1.2) to be:

$$c(x, y, \xi) = \frac{1}{2\sigma} (y - z(x))^2 + |\xi|. \tag{1.7}$$

This type of cost was originally proposed by [ROF92], and is also referred to as the Rudin-Osher-Fatemi (ROF) model. In Fig. 1.2 we show in red a given signal corrupted by Gaussian noise and the minimizer of the variational energy (1.2) in blue.

Optical flow and correspondence estimation. Given two images $I_1, I_2 : \mathcal{X} \to \mathbf{R}^c$ of the same scene, but perhaps recorded from different camera positions or containing some kind of motion, a fundamental task in computer vision is to estimate correspondences between the images.



$I_1 : \mathcal{X} \to \mathbf{R}^3$ $I_2 : \mathcal{X} \to \mathbf{R}^3$ displacement $f : \mathcal{X} \to \mathbf{R}^2$

Figure 1.3: Illustration of optical flow between to images. The displacement field is color-coded to indicate direction and magnitude of the motion.

One approach is to estimate a displacement field $f : \mathcal{X} \to \mathbf{R}^2$ under a photoconsistency assumption, which in the simplest case leads to the following choice of data term:

$$\rho(x, y) = |\, I_1(x) - I_2(x + y)\,|. \tag{1.8}$$

Due to the nonconvexity of the image functions I_1 and I_2, this data term is also nonconvex. Further assuming that the displacement field exhibits spatial regularity, e.g., $\eta(\xi) = \lambda|\xi|$ one arrives at classical variational models for the determination of optical flow [HS81; Wed+09]. Despite the simplicity of such energies, global optimization followed by post-processing steps recently has been shown to yield competitive results [CK16].

Stereo matching. In case the motion between the two images is given by the change of viewpoint within a static scene, the correspondence search can be reduced to a search along a one-dimensional curve, which greatly simplifies the problem. A displacement along that curve can be associated with a depth value, which allows one to reformulate the problem as an estimation of a depth map $f : \mathcal{X} \to \mathbf{R}_{\geq 0}$. In that case, the data term is given as follows:

$$\rho(x, y) = |I_1(x) - I_2(x + W(x, y))|, \tag{1.9}$$

where for a depth value $y = f(x)$, the warping function W computes the correct offset along the epipolar line [HZ03]. More sophisticated models estimate depth and normal parameters at each point, i.e., $f : \mathcal{X} \to \mathbf{R}_{\geq 0} \times \mathbb{S}^2$. This allows one to devise more precise patch-based data terms, see for example [Hei+13]. A first-order regularization of the normal parameters can also induce desirable priors. For example, a piecewise constant regularization on the normal field encourages piecewise-affine depth values which can be a good prior for planar geometry, often present in indoor scenes.

1.2.3 Relaxation and regularization of the discontinuity set

Depending on the choice of the cost function $c(x, y, \xi)$, a sequence of solution candidates with decreasing energy in the sense of (1.2) might approach a function which is not differentiable or even continuous anymore. This is known to be the case in the aforementioned example of total variation regularization. Another situation where the solution is discontinuous by design are *multilabeling* problems. There, one has a finite set of labels, for example $\mathcal{Y} = \{\text{sky}, \text{mountain}, \text{water}, \dots\}$ and due to the discrete nature of $\mathcal{Y}$ any function will have jumps, unless it is completely constant. In such cases, the energy (1.2) only makes sense at points where f is differentiable.

Therefore, it is natural to consider a relaxation, which also assigns a sensible cost to jump discontinuities.

In case f is sufficiently regular (a special function of bounded variation, see [AFP00, Sec. 4.1]), one can define an energy on the $(n-1)$-dimensional discontinuity set in terms of the jump points $(f^-, f^+) \in \mathcal{Y} \times \mathcal{Y}$ and the normal of the interface $\nu_f \in \mathbb{S}^{n-1}$:

$$E_{\text{jump}}(f) = \int_{J_f} d(x, f^-, f^+, \nu_f) \, \mathrm{d}\mathcal{H}^{n-1}(x). \tag{1.10}$$

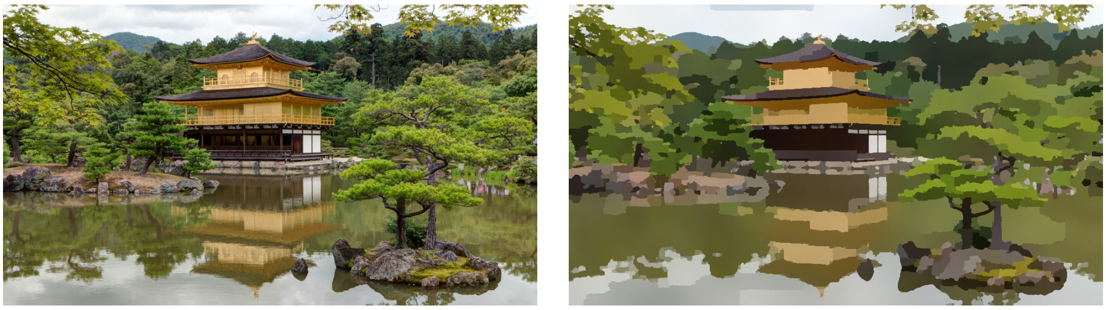
There are several technical assumptions to guarantee existence of minimizers for the model $E + E_{\text{jump}}$, see [AFP00, Theorem 5.22, Theorem 5.24].

A special case of the energy (1.2) together with (1.10) was suggested by Mumford and Shah [MS89] for the piecewise smooth approximation of functions, which amounts to the following choice:

$$c(x, y, \xi) = |y - z(x)|^2 + \alpha |\xi|^2, \quad d(x, f^-, f^+, \nu_f) = \lambda. \tag{1.11}$$

In Fig. 1.4, we show such an approximation of a natural image $z : \mathcal{X} \to [0, 1]^3$ obtained by minimization of the Mumford-Shah functional (1.11). We remark that the piecewise smooth assumption in RGB color space is not a very good prior for most natural images. For example, the prior does not account for textures in images. Still, one obtains interesting cartooning effects, as remarked in [Xu+11; SC14].

Nevertheless, discontinuities are ubiquitous in low-level vision due to object edges, occlusion boundaries, changes in albedo or (self-)shadowing. Therefore, for many other quantities such as (rigid) motion [CS05; Jai+15; GGK19; For+18], surface normals [QDA18], albedo [Hae+18; LB14], texture [Nie+14] the piecewise smoothness assumption can be reasonable. The piecewise constant case, which corresponds to passing α to infinity in (1.11) is also referred to as the Potts model or as ℓ_0-smoothing and also received considerable attention in recent applications [Xu+11; NB15; SW14].



(a) input image $z : \mathcal{X} \to [0,1]^3$ **(b)** piecewise smooth approximation

Figure 1.4: Minimization of the Mumford-Shah functional [MS89] leads to a piecewise smooth approximation of the given data. Assuming piecewise smoothness of the solution is not a particularly good prior for most natural images but produces an artistic cartooning effect.

1.3 The Lifting and Relaxation Principle

In the previous section, we have seen a class of nonconvex energies with applications in imaging or vision. Eventually, the aim is to find a convex formulation. In this section we give an introduction to the general lifting and relaxation principle which will be used to derive such convexifications.

1.3.1 Illustration of the lifting idea

Let us illustrate the principle first on a simple multilabeling task. Consider some fixed set $\mathcal{Y}$ with $|\mathcal{Y}| = \ell$ labels, i.e., $\mathcal{Y} = \{\text{sky, mountain, water}, \dots\}$ and an associated cost vector $c \in \mathbf{R}^\ell$. Determining the label with the minimal cost can be seen as a discrete optimization problem $\min_{1 \le i \le \ell} c_i$. It is common practice to instead consider "one-hot-encodings" $[0, 1, 0, \cdots] \in \Delta^{\ell-1} \subset \mathbf{R}^\ell$ in the $(\ell - 1)$-dimensional probability simplex. This leads to a convex relaxation as a continuous optimization problem $\min_{\mu \in \Delta^{\ell-1}} \sum_i c_i \cdot \mu_i$ which is amenable to gradient-based methods.

The above idea of discrete probabilities can be neatly generalized also to sets with infinitely many ”labels”, i.e., compact sets $\mathcal{Y} \subset \mathbf{R}^N$. The associated ”cost vector” is now a function $\rho : \mathcal{Y} \to \mathbf{R}$ and finding the label with minimal cost corresponds to a continuous (and possibly nonconvex) optimization problem

$$\min_{y \in \mathcal{Y}} \rho(y). \tag{1.12}$$

We can now state an equivalent reformulation of the problem by considering Dirac measures, which generalize the idea of the above encodings:

$$\rho(y) = \int \rho(y') \, \mathrm{d}\delta_y(y') =: \boldsymbol{\rho}(\delta_y). \tag{1.13}$$

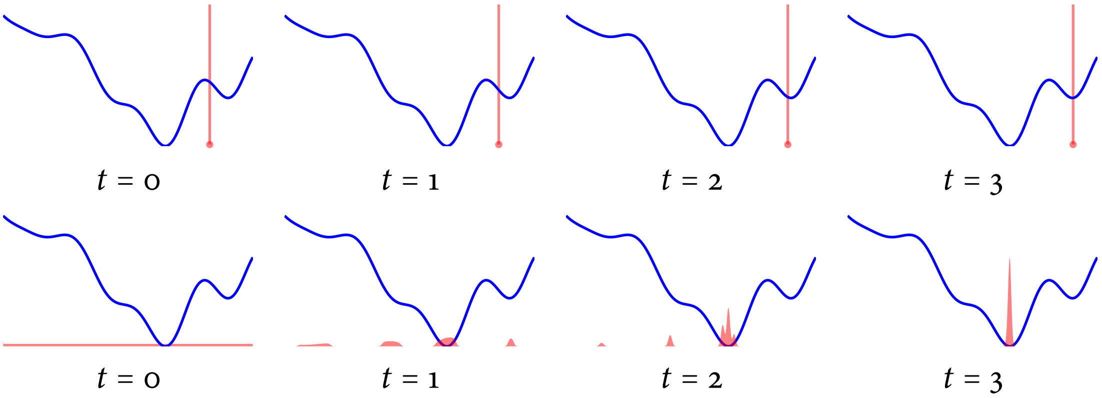


Figure 1.5: We compare local optimization (shown in the top row) against the global lifting approach (bottom row). Gradient descent optimization on the original problem can be interpreted by a Dirac distribution, whose center is adjusted locally to decrease the energy. Descent on the lifted and relaxed problem is a search in the convex set of all probability distributions. Due to the convexity, one is able to overcome local optima.

Here, δ_y is the Dirac delta distribution centered at the point $y \in \mathcal{Y}$. The new energy functional $\boldsymbol{\rho}$ is now defined on probability measures, and it is linear as it amounts to an integration operation. We have therefore reformulated a possibly nonconvex energy into a linear one, at the price of moving from finite to infinite dimensions.

The lifting $y \mapsto \delta_y$, from (1.12) to (1.13) still does not lead to a convex problem, as the search space of Dirac distributions is nonconvex. A relaxation is obtained by convexifying the search space. In this case, we relax to the convex hull of the Dirac distributions, which is given by the probability distributions on $\mathcal{Y}$ denoted throughout the thesis as $\mathcal{P}(\mathcal{Y})$. This leads to the relaxed problem

$$\min_{\mu \in \mathcal{P}(\mathcal{Y})} \boldsymbol{\rho}(\mu), \qquad \text{where} \qquad \boldsymbol{\rho}(\mu) := \int \rho(y) \, \mathrm{d}\mu(y). \tag{1.14}$$

Crucially, problem (1.14) is convex. A recurring question throughout the thesis will be whether solutions to the relaxed, convexified problem can be identified with solutions of the original problem. For the above problem (1.14) the answer is affirmative, as the support of the optimal probability distribution will be contained within the set of global minimizers to (1.12). Formally, this can be justified by applying Bauer's maximum principle [AB94, Theorem 7.69] as the objective functional is linear in μ and the feasible set $\mathcal{P}(\mathcal{Y})$ is compact in the weak*-topology. The principle further asserts that minimizers are attained at the extreme points of the set of probability measures. For $\mathcal{P}(\mathcal{Y})$ it can be shown that the extreme points correspond to the Dirac delta measures, see [Cho69, p. 112, Example 26.1].

Another interpretation of the above procedure is that one makes the optimization problem "artificially" probabilistic. Instead of considering just a single state, the idea is to allow a probabilistic superposition over all possible states. This is illustrated in

Fig. 1.5 on a simple one-dimensional energy. In that figure, we illustrate a gradient flow in distribution space, starting at the uniform density at $t = 0$. The evolving distribution is shown in red. While at intermediate points, the flow concentrates near local minima and saddle-points in the original nonconvex energy (shown in blue), it is eventually able to overcome them due to convexity of the lifted energy (1.14). The final distribution (shown on the right) concentrates at the global minimizers of the energy. Local optimization of a single ”Dirac” state starting from a bad initialization will get stuck at a suboptimal point.

While this lifting and relaxation idea seems perhaps intractable at first sight, we find it to be rather profound and of fundamental importance. Indeed, it is the starting point for many different popular formulations and considerations in literature which go far beyond the scope of variational problems. For example, it is the starting point of the Lasserre hierarchy in the global optimization of polynomials [Las00].

To obtain a finite problem, many methods consider a parametrized subset of measures $\{\mu_\theta : \theta \in \Theta\} \subset \mathcal{P}(\mathcal{Y})$. This yields a (typically nonconvex) upper bound to (1.14). Interestingly, it is possible to compute the gradient of the lifted energy with respect to the parameters by just sampling the original energy, which is referred to as the ”log-derivative” trick or REINFORCE estimator [Wil92]:

$$\nabla_\theta \boldsymbol{\rho}(\mu_\theta) = \mathbb{E}_{y \sim \mu_\theta} \left[\rho(y) \nabla_\theta \log \mu_\theta \right]. \tag{1.15}$$

Since only an *evaluation* of the original energy $\rho : \mathcal{Y} \to \mathbf{R}$ but no gradient information is required, one is able to deal with very difficult and irregular energies. This makes the relaxed formulation on probabilities popular in reinforcement learning [Kak02; PS08], evolutionary search and black box optimization [Wie+08]. The reformulation (1.14) is sometimes referred to as a stochastic relaxation of the optimization problem. Optimization over such parametrized families of probability distributions is typically approached via *natural gradient methods*, see [Ama16, Section 12.1.4] also for further references and applications.

For the variational problems considered in this thesis, we perform such a relaxation at every point $x \in \mathcal{X}$ in the domain. Similarly to the above situation, we will be able to handle arbitrary cost functions without requiring to evaluate their gradients. In that sense, the lifting approaches bear some similarity to black box or zeroth-order optimization strategies. However, instead of considering a parametrized family of distributions in the primal, our approach will rather be based on duality which leads to convex lower bounds on the objective.

1.3.2 Monge and Kantorovich problems in optimal transport

An early instance of the previous lifting and relaxation principle can be found in the Kantorovich formulation [Kan60] of Monge’s optimal transportation problem [Mon81]. Since the underlying principle is quite similar to the relaxations we consider in this thesis, let us briefly review them. We refer the interested reader to [San15; PC18] for more details.

The Monge problem in optimal transport is a variational problem which involves a search over maps $f : \mathcal{X} \to \mathcal{Y}$ that transport a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ to another probability measure $\nu \in \mathcal{P}(\mathcal{Y})$. The cost function penalizes the transportation of a point $x \in \mathcal{X}$ to another point $f(x) \in \mathcal{Y}$ by the cost $c(x, f(x))$.

$$\inf_{f:\mathcal{X}\to\mathcal{Y}} \int_{\mathcal{X}} c(x, f(x))\, \mathrm{d}\mu(x), \text{ s.t.} \quad f_{\sharp}\mu = \nu. \tag{1.16}$$

Due to potential nonconvexities in the cost and the nonlinear constraint, this is a nonconvex optimization problem. The Kantorovich relaxation [Kan60] is a linear programming formulation of (1.16).

$$\begin{aligned} \min_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} & \int c(x, y)\mathrm{d}\gamma(x, y), \\ \text{s.t.} \quad & \pi_{1\sharp}\gamma = \mu, \\ & \pi_{2\sharp}\gamma = \nu. \end{aligned} \tag{1.17}$$

This relaxation can be motivated in the following way. Every map $f : \mathcal{X} \to \mathcal{Y}$ induces a probability measure $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that is concentrated on the graph of f. The cost function in (1.17) is formulated in such a way that it coincides with (1.16) in that situation. A convex formulation (here even a linear program) is obtained by relaxing the search space from measures which concentrate on graphs to a more general set of probability measures.

1.3.3 Weighted minimal surfaces

As mentioned earlier, this thesis is concerned with variational problems for energies (1.2), which have the form

$$E(f) = \int_{\mathcal{X}} c(x, f(x), \nabla f(x))\, \mathrm{d}x. \tag{1.18}$$

Unlike in the Monge problem (1.16), the cost can additionally depend on the derivative of the function. Generalizations of the Monge problem to costs such as (1.18) have recently been studied in the context of optimal transport [Bre03; DLS14; Lou14]. Corresponding convex relaxations, such as the one studied in [Gho+20], are closely related to the ones we consider in this thesis.

In order to penalize the derivative in the ”lifted” formulation, a relaxation to vector-valued measures is considered in which the vector describes the tangent information to the graph. In codimension one, it is possible to represent such vector-valued measures as the derivative of an indicator function. This geometric view on first-order variational energies (in particular minimal surface problems) has been studied in great detail by Federer [Fed69; Fed74].

The main idea is, that energies of the form (1.18) can be written as certain anisotropic minimal surface energies with convex weighting Ψ:

$$\mathbf{E}(S) = \int_S \Psi := \int_S \Psi(x, y, \tau_S(x, y)) \, \mathrm{d}\mathcal{H}^n(x, y). \tag{1.19}$$

The meaning of the above integral will be made precise later on, but the intuition is that τ_S is an *orientation* of the surface S. For example, an energy penalizing the surface area is simply given by $\Psi = \| \cdot \|$. The general idea which connects minimal surface problems to the class of variational problems (1.18) is to consider weights Ψ such that

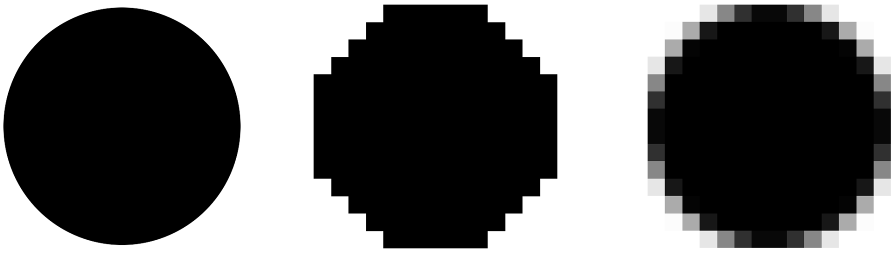
$$E(f) = \mathbf{E}(\mathcal{G}_f), \tag{1.20}$$

where $\mathcal{G}_f$ is the oriented graph of the function f. While such reformulations are well-known, we will give a self-contained proof of the equality (1.20) in the general setting in Chapter 6. A convex formulation is obtained in analogy to the previous section. Recall that in the simple setting we performed a relaxation from Dirac measures to all probability measures. Here, a similar principle can be applied: Instead of considering only graph surfaces, the idea is to relax to a larger convex set of vector-valued measures or *currents*. This principle will be revisited in more detail in Chapter 5 and Chapter 6.

1.4 Related Work: Discrete and Continuous

In this thesis, the underlying idea is to prolong an approximation or discretization of $\mathcal{X}$ and $\mathcal{Y}$ as long as possible. The motivation for such a strategy is that this eventually leads to discrete formulations which can be more faithful to the underlying continuous model. Such formulations are hard to understand or derive when discretizing at an too early stage. Due to the continuous optimization, such models often yield fractional solutions, as is illustrated in Fig. 1.6. These solutions can implicitly represent the true continuous object beyond the mesh size. In some sense, such ”subpixel-accurate” representations as shown in Fig. 1.6 can appear rather naturally, for example when one would record a picture of the continuous black disk with a camera.

This advantage of continuous variational approaches over their discrete counterparts has been noted since a while, see for example [Klo+08; Lel+13b]. In linear programming relaxations of integer programs in discrete optimization, fractional values appear due to exactness or tightness issues with the relaxation. As already remarked in [Lel+13b], the main idea here is rather different: *Fractional values rather appear as an effect of approximating the true continuous solution on a finite mesh.* An issue is, that it can become difficult to disentangle whether fractional solutions are due to an approximation of the continuous model or due tightness issues with the relaxation. A naive thresholding to a binary solution would destroy the subpixel-accuracy and more elaborate strategies are necessary.



(a) continuous (b) discrete (integer) (c) discrete (fractional)

Figure 1.6: On a fixed grid, it is possible to represent a continuous object (shown in (a)) more faithfully by fractional (non-integer) values than with binary values.

1.4.1 Discrete multilabeling

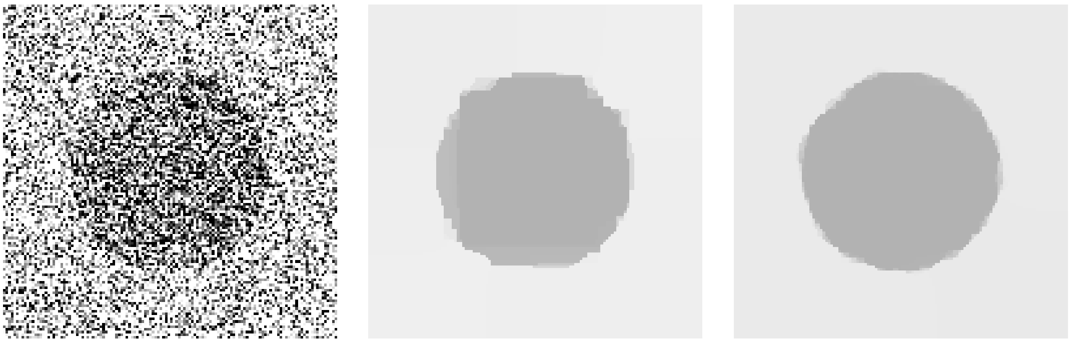
To arrive at a practical formulation that is implementable, one eventually has to discretize or approximate the continuous formulation in some way. Therefore, one could argue that one can directly start with a discrete formulation in which both $\mathcal{X}$ and $\mathcal{Y}$ are represented by finite sets. Sometimes, it is argued that in computer vision or imaging applications, this is naturally the case. For example the set $\mathcal{X}$ can represent pixels or superpixels in an image and $\mathcal{Y}$ is sometimes naturally a finite set, e.g., in segmentation tasks.

By doing so, one arrives at discrete multilabeling problems which we will briefly review in the following. They can be seen as a discrete analogue to the continuous first-order functional (1.2) and are given by the following energy:

$$E(f) = \sum_{i \in \mathcal{X}} E_i(f_i) + \sum_{(i,j) \in \mathcal{E}} E_{ij}(f_i, f_j). \tag{1.21}$$

The edge set $\mathcal{E} \subset \mathcal{X} \times \mathcal{X}$ in the graph $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ is used to encode pairwise dependence between the variables, in analogy to the derivative ∇f in (1.2). The energy (1.21) also corresponds to a Markov random field (MRF) with underlying probability distribution $p(f) \propto \exp(-E(f))$, see [Ish03].

In case $|\mathcal{Y}| = 2$, the energy (1.21) can be efficiently minimized under mild assumptions on the pairwise terms [BK04]. For $|\mathcal{Y}| > 2$, this is also true in case E_{ij} is *submodular* [Shl76; Wer07], for example given by $E_{ij}(f_i, f_j) = g(f_i - f_j)$ for an even convex function $g : \mathbf{R} \to \mathbf{R}_{\geq 0}$ as in [Ish03]. In general, when $|\mathcal{Y}| > 2$ and $\mathcal{Y}$ is an unordered set, finding the global optimum of (1.21) is known to be NP-hard [LSH16]. Another special case arises when the graph $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ has the structure of a one-dimensional chain or, more generally, of a tree. In that setting, the global optimum of (1.21) can also be efficiently determined [WF00; WJ+08; CK97; SC10].



noisy data $\sum_{(i,j)\in\mathcal{E}} |f_i - f_j|$ $\int_{\mathcal{X}} \|\nabla f(x)\| \mathrm{d}x$

Figure 1.7: Spatially continuous formulations (infinite $\mathcal{X}$) can rather naturally handle more isotropic regularizers such as total variation based on the ℓ_2-norm. Graph-based approaches (here with a 4-connected neighbourhood) tend to exhibit a certain bias towards the underlying grid.

1.4.2 Discrete-continuous Markov random fields

Closely related to the works presented in this thesis are *continuous* MRFs, in which the label set $\mathcal{Y} \subset \mathbf{R}^N$ is a continuum but $\mathcal{X}$ is still finite.

Based on an infinite-dimensional local marginal polytope relaxation and its dual program, the paper [Pen+11] proposes a local particle-based optimization scheme for such MRFs. In a subsequent work [Yam+12], that method is applied to the task of stereo matching. The tightness of this infinite-dimensional linear programming relaxation has been studied in [WG14; Ruo15].

The works [ZK12; Zac13] propose convex relaxations for continuous MRFs based on perspective functions. The paper [FA14] shows that these convex relaxation can actually be derived as a (discontinuous) piecewise-linear approximation of the infinite-dimensional dual linear program. Furthermore, a general hierarchy of piecewise-polynomial approximations to the dual is analyzed.

For continuous MRFs with polynomial potentials, specialized local optimization solvers based on ADMM [Sal13] and difference-of-convex programming strategies [WSU14] have been proposed. Since for $\mathcal{Y} = \mathbf{R}^N$ the energy (1.21) is a finite-dimensional continuous optimization problem, standard local solvers from continuous (nonconvex) optimization (such as gradient descent or L-BFGS) can also be applied.

For chain or tree-structured graphs, efficient dynamic programming algorithms for the global optimization have been proposed also in the continuous case. For example, the case of piecewise-linear potential functions was considered in [KPR16] and the Potts model on chains in [SW14].

1.4.3 Spatially continuous multilabeling

A different line of works considers multilabeling problems with finite $\mathcal{Y}$, but where the problem domain $\mathcal{X} \subset \mathbf{R}^n$ is modeled as a continuum. An advantage of such continuous formulations is, that they can rather naturally handle isotropic smoothness terms, which in turn allows for more isotropic discretizations. The difference between such anisotropic and isotropic regularization is illustrated on a simple ”denoising” example in Fig. 1.7.

Since the set $\mathcal{Y}$ is finite, the map $f : \mathcal{X} \to \mathcal{Y}$ will be piecewise constant, i.e., consist of constant regions. Since ∇f will be either zero (in the constant regions) or not well-defined (at the jump-parts), one considers functional of a form as (1.10):

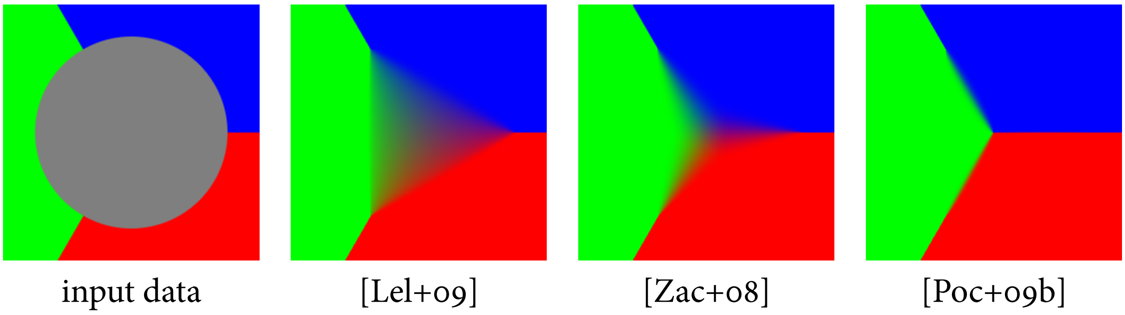
$$E(f) = \int_{\mathcal{X}} c(x, f(x)) \mathrm{d}x + \int_{J_f} d(x, f^-, f^+, \nu_f) \, \mathrm{d}\mathcal{H}^{n-1}(x). \tag{1.22}$$

Roughly speaking, the cost $d(x, f^-, f^+, \nu_f)$ plays the role of the pairwise potential in (1.21). Note that E_{ij} in (1.21) depends only on the two values of f at the jump along the edge. Due to the continuous nature of $\mathcal{X}$, in the above formulation an additional dependance on the jump direction $\nu_f \in \mathbb{S}^{n-1}$ is introduced which intuitively accounts for the fact that each point $x \in \mathcal{X}$ has infinitely many neighbours.

The energy (1.22) is nonconvex and there are three popular convex relaxations based on a lifting $\widehat{f} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. The two formulations [Zac+08; Lel+09] are simpler and therefore easier to optimize, while the relaxation [Poc+09b; CCP12] is tighter and more accurate in some situations. A comparison[3] between the relaxations on an ”inpainting” problem with three labels (red, green and blue) is shown in Fig. 1.8. The pairwise interaction term is chosen such that the perimeter of the interface is minimized. The relaxations [Zac+08; Lel+09] lead to a large amount of fractional labelings, which is visualized as a ”mixing” of the three colors. In contrast, the formulation from [Poc+09b; CCP12] yields an almost binary solution and the fractional solutions at the interface are likely due to the discretization. For more details, we refer to the paper [Lel+13b], which discusses the importance of distinguishing between fractional solutions which are due to the discretization and those which are caused by the relaxation.

Similarly to the discrete setting, in case $|\mathcal{Y}| = 2$, the global optimum to (1.10) can be found by a thresholding, see for example [BPV91; CEN06; Cha+10]. In further analogy to discrete multilabeling methods, extensions of maximum flow based approaches to the continuous setting have been proposed in [Yua+10; YBT10]. A numerical comparison between discrete and continuous multilabeling approaches was conducted in [NTC13]. The work [ZHP13] connects finite difference discretizations of continuous multilabeling relaxations to local marginal polytope relaxations for MRFs. It is sometimes argued as an advantage of discrete multilabeling formulations that they can also handle non-metric transition costs. As remarked

[3] Solutions computed using the optimization framework ”prost”, see Appendix C.



input data [Lel+09] [Zac+08] [Poc+09b]

Figure 1.8: Comparison of relaxations for spatially continuous multilabeling on an ”inpainting” problem. The relaxation [Poc+09b; CCP12] leads to an almost binary solution.

in [CCP12], the continuous model (1.10) would not admit a minimizer in such a non-metric case due to (1.10) not being lower-semicontinuous.

The study of continuous multilabeling problems in which the set $\mathcal{Y}$ contains *infinitely* many labels [Lel+13a] was the starting point for the considerations in this thesis. Specifically, in [Lel+13a] the set $\mathcal{Y}$ is assumed to be a manifold, which is eventually discretized into a discrete set of labels. Nevertheless, due to the continuous formulation, one expects fractional solutions which can be interpreted as a *sublabel-accurate* solution to the underlying continuous manifold. This allows one to arrive at smooth solutions despite a rather coarse approximation of the manifold $\mathcal{Y}$. This is especially desirable in high-dimensional settings, as due to the curse of dimensionality the number of labels grows *exponentially* with the dimension of the manifold.

As we will later see, the linear label-cost used in traditional multilabeling approaches and in [Lel+13a] does not take this sublabel-accurate structure into account appropriately. The main contribution of Chapter 3 and Chapter 4 will be a more faithful data term which assigns meaningful costs also to fractional labelings.

An issue with this multilabeling perspective is, that one is essentially restricted to regularizations on the jump part which has to be a metric. Therefore, it was initially considered unclear how to extend the sublabel-accurate multilabeling methods to more general convex regularizers such as Huber or Dirichlet-type energies. This issue is eventually resolved in Chapter 5, by interpreting the multilabeling methods as a certain discretization of a label-continuous relaxation for variational problems with general regularizations. As remarked before, these relaxations are essentially weighted anisotropic minimal surface problems in higher dimension.

1.4.4 Continuous variational problems

Finally, let us consider the case of continuous variational problems (1.2) where the sets $\mathcal{X}$ and $\mathcal{Y}$ both have infinite cardinality. In the following, we give an overview over existing methods to tackle such problems.

1.4.4.1 Local optimization methods

One approach is to solve the variational problem via direct optimization, by running gradient descent or fixed-point methods to find a zero of the Euler-Lagrange equations, see for example [ROF92]. In case the cost function is split into a nonconvex data term plus a convex regularizer, $c(x, y, \xi) = \rho(x, y) + \eta(\xi)$, more elaborate local optimization methods have been developed over the years. One popular approach, often referred to as quadratic relaxation or quadratic decoupling, introduces a splitting of the objective that is coupled with a quadratic penalty term:

$$\min_{f,g} \int_{\mathcal{X}} \rho(x, f(x)) + \eta(\nabla g(x)) + \frac{1}{2\theta}(f(x) - g(x))^2 \, \mathrm{d}x. \tag{1.23}$$

The subproblems appearing in alternating minimization on this split formulation are both easy. The f-subproblem is an independent point-wise search that can be solved globally while the g subproblem is a convex optimisation problem. For $\theta \to 0^+$, it can be expected that one approaches a (local) solution to the original problem. Such a splitting strategy has been successfully used for large-displacement optical flow [SPC09] or real-time dense 3D mapping [NLD11]. Introducing a Lagrange multiplier for the constraint, which corresponds to an inexact augmented Lagrangian or ADMM type method, has been shown to be beneficial [KC13].

Another popular line of works locally replaces the nonconvex data term with a convex model ρ_t at the current solution f^t for time step $t \geq 0$. In the simplest case, this model is based on a first or second order Taylor expansion. Then, one solves a sequence of convex problems of the form

$$f^{t+1} = \arg\min_{f} \int_{\mathcal{X}} \rho_t(x, f(x)) + \eta(\nabla f(x)) + \frac{1}{2\tau}(f(x) - f^t(x))^2 \, \mathrm{d}x \tag{1.24}$$

where the additional trust-region or proximal term keeps the solution in an area in which this model is considered to be accurate. Since the model is often only valid locally, this sequential convex programming strategy is usually embedded in a coarse-to-fine framework, see for example [ZPB07; Wed+09; SGC10]. In case the model is a first-order expansion, this algorithm is also known as forward-backward splitting or the proximal gradient method [CP11b].

1.4.4.2 Convex relaxation approaches

While the previous local optimization approaches are fast and can lead to good solutions, they require additional hyperparameters such as step sizes $\tau > 0$ or homotopy continuation parameters $\theta \to 0$. Furthermore, they cannot be guaranteed to always reach the global optimum and depend on the initialization. The focus of this thesis to advance global optimization approaches, where the aim is to find a single convex optimization problem.

The most direct way to obtain a convex relaxation is to replace the nonconvex function with a global convex surrogate. This has been considered by Bhusnurmath et al. [Bhu08; BT08], on the example of stereo matching. Specifically, the

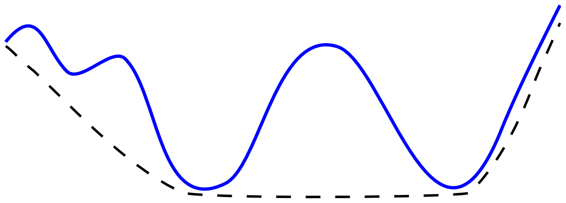


Figure 1.9: Computing a lower convex envelope of the nonconvex cost to obtain a convex formulation can lead to quite large ”gaps” between the envelope and actual function. Lifting strategies can overcome this problem by embedding the cost in a higher-dimensional space prior to the convexification.

nonconvex image matching cost is replaced by its convex hull. Due to convexity of the regularizer, this leads to an overall convex linear programming formulation which is solved using an interior point method. As illustrated in Fig. 1.9, an issue with this approach is that there can be quite a large gap between the convex and the nonconvex cost. Nevertheless, despite this seemingly poor approximation, the results obtained in [Bhu08; BT08] are surprisingly good.

We show in Chapter 3 and Chapter 4 that the proposed sublabel-accurate lifting approach includes this idea of direct convexification as a special case, when a minimal number of labels is chosen. As we will illustrate later on, the multilabeling approach embeds the nonconvex cost in a higher-dimensional space before taking the convex envelope which leads to a better approximation.

Minimal surface problems in codimension one. As we have seen in the previous section, variational problems involving first-order energies can be rephrased as certain weighted anisotropic minimal surface problems for which global optimization methods exist.

In his Ph.D. thesis, John Sullivan [Sul90] proposes to determine discrete minimal surfaces supported on a large set of polygonal surface elements using max-flow min-cut algorithms. Since one determines a certain polyhedral surface, this yields an upper bound to the true continuous minimal surface energy. While the algorithm is quite efficient, and there are convergence guarantees under the refinement of the grid, there is a certain bias towards the chosen underlying surface elements. The work [KG04] extends that approach to also handle Neumann boundary conditions and provides an actual numerical implementation. The paper [BK03] also proposes a discrete approach for computing minimal surfaces using minimum cuts in a graph. Using tools from integral geometry, the paper provides a formula for optimal edge weights in the graph which minimize metrication artifacts.

The paper [Par92] proposes to determine a minimal surface in codimension one by minimizing the gradient norm of a zero-one-indicator function represented in a piecewise linear basis. The method also finds an upper bound to the true minimal surface energy, but by representing the surface as the level-set of a piecewise linear function, a smoother solution which can go beyond the chosen mesh accuracy is obtained. This is in similar spirit as the proposed sublabel-accurate lifting methods, but we will consider a discretization of the dual problem instead.

The approach [AT05] is concerned with the determination of minimal surfaces using the theory of continuous maximum flows [Str83; Str09]. Unlike the methods proposed in this thesis, it still requires a fine discretization of the range into labels, as it is based on a simple finite differencing approach.

In contrast to the above works, what we will propose in Chapter 5 and Chapter 6 is a lower bound to the true minimum, as we will discretize a dual representation of the energy. Quite similarly, the work [Bra95] also considers a discretization of the dual problem using piecewise linear finite elements. However, we consider a more general setting and solve the resulting optimization problem using a first-order primal dual algorithm while [Bra95] considers an interior point method. It is noted in [Bra95] that the interior point method finds a superposition of all possible solutions. This can be attributed to the fact that the interior point method always stays in the interior of the feasible set. Despite lack of theoretical guarantees, it is mentioned in [Bra95] that this superposition of solutions always turns out to be close to the true continuous solution in practice. We observe a similar effect in the experiments in Chapter 6.

Another difference to all the above works is, that we eventually consider the setting of general dimension and codimension, and use a specialized first-order primal-dual algorithm on GPUs to solve very large scale problems. The works [Poc+08; Poc+10] also propose to solve the (anisotropic) minimal surface energy with the same algorithm. This thesis builds upon these works by using a more precise (sublabel-accurate) discretization, that can still be efficiently solved using the same primal-dual methods.

Beyond codimension one. It is a major research challenge to tackle vectorial variational problems or multilabeling problems where the labels do not form an ordered set. Under the minimal surface viewpoint, such problems corresponds to the task of finding a surface with codimension larger than one, for example a two-dimensional surface in four-dimensional space. It has already been noted in [KG04] that it would be "nice to solve variational problems of higher codimension" but deemed "unlikely that such a generalization exists", by which they refer to efficient min-cut max-flow techniques.

The works [GSC13; SCC12; SCC14] are among the first to tackle the difficult challenge of such vectorial variational problems. To that end, a collection of surfaces in codimension one is considered. This efficient representation avoids a discretization of $\mathcal{X} \times \mathcal{Y}$, but assumes a factorization $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N$ into one-dimensional spaces. Approaches which consider the full product space have also been proposed [GBO12; Lel+13a; WC16], but are restricted to certain types of surface energies.

In this thesis, we make two contributions to this line of works. In Chapter 5 we show that more accurate results can be obtained by using a piecewise linear discretization in the dual, which also applies to the above works. We illustrate the effect on the example of the work [SCC12]. Furthermore, we extend lifting approaches which operate in the full space $\mathcal{X} \times \mathcal{Y}$ to general polyconvex regularizations in Chapter 6.

1.5 Contributions

This cumulative thesis consists of five publications [Möl+16; Lau+16; MC17; MC19a; MC19b] which are the result of collaborations with Emanuel Laude, Prof. Michael Möller, Prof. Jan Lellmann and Prof. Daniel Cremers. All of these works were published in highly ranked and peer reviewed international conferences. The following table gives an overview of the publications included in this thesis.

[Möl+16] T. Möllenhoff, E. Laude, M. Moeller, J. Lellmann and D. Cremers. **Sublabel-accurate relaxation of nonconvex energies**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. Received the best paper honorable mention award at CVPR 2016. The first two authors contribued equally. (Cited on pages 22, 63, 77, 80, 82, 83, 85, 86, 88, 91, 95, 106, 156, 157, 170).

[Lau+16] E. Laude, T. Möllenhoff, M. Moeller, J. Lellmann and D. Cremers. **Sublabel-accurate convex relaxation of vectorial multilabel energies**. In: *European Conference on Computer Vision (ECCV)*. 2016. The first two authors contribued equally. (Cited on pages 22, 77, 95, 180).

[MC17] T. Möllenhoff and D. Cremers. **Sublabel-accurate discretization of nonconvex free-discontinuity problems**. In: *International Conference on Computer Vision (ICCV)*. 2017 (Cited on pages 22, 94, 95, 106, 197).

[MC19a] T. Möllenhoff and D. Cremers. **Lifting vectorial variational problems: A natural formulation based on geometric measure theory and discrete exterior calculus**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (Cited on pages 22, 113, 207).

[MC19b] T. Möllenhoff and D. Cremers. **Flat metric minimization with applications in generative modeling**. In: *International Conference on Machine Learning (ICML)*. 2019 (Cited on pages 22, 218).

During the course of the master and doctoral studies a couple of additional papers not directly related to the topic of this thesis have been prepared. They are not included in this thesis and simply listed here for completeness.

[Hae+18] B. Haefner, T. Möllenhoff, Y. Queau and D. Cremers. **Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (Cited on page 9).

[Möl+13] T. Möllenhoff, C. Nieuwenhuis, E. Toeppe and D. Cremers. **Efficient convex optimization for minimal partition problems with volume constraints**. In: *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*. 2013.

[Möl+15a] T. Möllenhoff, E. Strekalovskiy, M. Moeller and D. Cremers. **Low rank priors for color image regularization**. In: *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*. 2015.

[Möl+15b] T. Möllenhoff, E. Strekalovskiy, M. Moeller and D. Cremers. **The primal-dual hybrid gradient method for semiconvex splittings**. In: *SIAM J. Imaging Sci.* 8.2 (2015), pp. 827–857.

[Fre+18] T. Frerix, T. Möllenhoff, M. Moeller and D. Cremers. **Proximal backpropagation**. In: *International Conference on Learning Representations (ICLR)*. 2018. The first three authors contribued equally.

[Möl+18] T. Möllenhoff, Z. Ye, T. Wu and D. Cremers. **Combinatorial preconditioners for proximal algorithms on graphs**. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2018.

[MMC19] M. Moeller, T. Möllenhoff and D. Cremers. **Controlling neural networks via energy dissipation**. In: *International Conference on Computer Vision (ICCV)*. 2019.

[Bré+19] P. Bréchet, T. Wu, T. Möllenhoff and D. Cremers. **Informative GANs via structured regularization of optimal transport**. In: *Optimal Transport and Machine Learning (NeurIPS Workshop)*. 2019.

[Ye+20] Z. Ye, T. Möllenhoff, T. Wu and D. Cremers. **Optimization of Graph Total Variation via Active-Set-based Combinatorial Reconditioning**. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2020.

This thesis makes several contributions to advance the state-of-the-art in spatially continuous multilabeling and convex relaxation methods, which are listed in the following.

1.5.1 Sublabel-accurate multilabeling

The first contribution this thesis makes is a novel formulation for spatially continuous multilabeling problems, in which fractional labelings are assigned a more meaningful cost. This allows one to tackle labeling problems which originally have huge or even infinite label spaces. The method is first developed for scalar-valued labels in Chapter 3 and then extended to the vectorial setting in Chapter 4. We show that the resulting convex-concave saddle point problems can be implemented with little overhead using standard primal dual algorithms and epigraphical projections. In contrast to discrete-continuous MRFs [ZK12], the formulation allows a particularly efficient implementation of isotropic regularizations.

1.5.2 Interpretation as a discretization

In Chapter 5 we show that the previous sublabel-accurate multilabeling approaches can be derived as a particular discretization to a known convex relaxation for continuous variational problems which are based on the theory of calibrations. Specifically, we show that the sublabel-accurate representation corresponds to an approximation of the calibration (dual variable) with continuous piecewise linear functions. Using this insight, we can extend these methods from total variation or convex one-homogeneous regularizers to more general convex regularizations. Furthermore, we show that concave penalizations of the jump set can be implemented with finitely many constraints. However, we see the main contribution of this chapter as a new principled way to obtain sublabel-accurate solutions to continuous convex relaxations. Furthermore, it highlights the generality and usefulness of the infinite-dimensional treatment and gives further justification of the function space formulations. We expect the insights developed in this chapter to inspire and drive further developments and analyses of finite element approaches applied to dual formulations.

1.5.3 Liftings in arbitrary dimension and codimension

In Chapter 5, we have seen that sublabel-accurate multilabeling approaches are discretizations of continuous function space formulations. Therefore, in Chapter 6, the aim is to extend the scope of such formulations. In particular, we focus on vectorial variational problems. The main contribution we make over previous works is a novel formulation which can handle general polyconvex regularizations. To achieve this, we view the lifting as an anisotropic minimal surface problem in general dimension and codimension. This requires to introduce various notions from geometric measure theory, which are not previously used in computer vision. While previous liftings for vectorial or even manifold-valued problems support general convex regularizers [SCC14; Vog+19], in our framework they correspond to a certain trivial polyconvex extension of the convex cost. The presented framework potentially allows for tighter relaxations by taking the geometric structure of the underlying problem into account. Furthermore, it allows to introduce bijectivity constraints, which we demonstrate on the example of image registration.

1.5.4 Outlook: Geometric measure theory in machine learning

In Chapter 7, we show that the previously introduced notions from geometric measure theory can be useful beyond the convex relaxation of variational problems. The main contribution is the novel perspective to view high-dimensional data not as a probability distribution but rather as a current. This is motivated by the manifold hypothesis, which states that the true data distribution concentrates near a low-dimensional manifold with high codimension. Our insight is that the generalized surfaces developed in geometric measure theory to solve variational problems in general dimension and codimension can be also useful to represent high-dimensional data. This viewpoint also allows one to equip the data with a local *orientation*, which can be useful for the learning of equivariant representations. Our second contribution in this chapter is a generalization of the recent Wasserstein generative adversarial networks [ACB17] to this perspective. The generalization is based on the flat norm, which serves as a distance between currents. In our theoretical contribution, we show Lipschitz continuity of the flat norm with respect to the parameters which ensures well-posedness of the optimization problem. Finally, we demonstrate that the formulation can actually be implemented, and that it leads to disentangled and equivariant latent representations.

1.6 Outline of the Thesis

Following this introduction in Part I is Chapter 2 which revisits some mathematical foundations. The three theoretical cornerstones of this thesis are *(1) convex analysis*, *(2) (geometric) measure theory and functions of bounded variation* and *(3) mathematical optimization*. Correspondingly, this chapter consists of three sections which review some of the required backgrounds from these disciplines. Each of these are vast fields and the purpose is not to repeat the standard textbooks on the topic. The aim is rather to focus on the concepts which are of specific importance to this thesis. We will present the concepts in a rather non-technical fashion, with focus on intuitions and examples rather than on proofs. Nevertheless, we will refer to the points in the literature where more rigorous treatments can be found. As the later chapters contain some background as well, we keep the presentation rather brief.

In Part II, Chapters 3 – 7 correspond to one the aforementioned full-length publications.

Finally, in Part III we conclude the thesis. Chapter 8 summarizes the results and contributions made in this thesis. In Chapter 9, some recent related work that has been carried out concurrently or builds upon the results of this thesis is discussed. Finally, we point out some promising open directions for future research.

Chapter 2

Theoretical Foundations

In this chapter we provide a short introduction to the mathematical concepts used in this thesis. The first section is concerned with convex analysis, which lies at the heart of the convexification approaches presented in the remainder of this thesis. We will put particular emphasis on the notions of convex envelope and convex duality as they form the central tools in obtaining convex reformulations of nonconvex problems.

The second section reviews notions from (geometric) measure theory and functions of bounded variation. These technicalities are required as the lifting principle is essentially based on a reformulation of the optimization problem to spaces of certain (geometric) measures.

The third section briefly introduces proximal algorithms for convex optimization. These are used to solve the nonsmooth convex optimization problems arising from the discretization of the lifted and convexified formulations. There is a vast body of literature on first-order methods and their relations [Ess10; PB13]. We will discuss a popular and simple primal-dual algorithm [CP11a] that was used in most experiments of this thesis. We further discuss some tweaks that can be used to speed up the convergence such as (adaptive) step size selection and preconditioning.

2.1 Convex Analysis

We will present most of the results in the setting of finite-dimensional real vector spaces $\mathbf{V}$ equipped with a norm $|\cdot|$ and inner product $\langle\cdot,\cdot\rangle$. This chapter will state many results without proof but we will refer to the according results in the literature. As a general introduction to convex analysis we recommend the book by [HL12], on which this section is mostly based on. The comprehensive treatise of [Roc96] is often considered the definite reference to the subject. Many of the concepts also carry over to the more general setting of infinite-dimensional topological vector spaces; see for example [Roc74; AB94].

2.1.1 Basic definitions and notation

The extended reals are denoted by $\overline{\mathbf{R}} = \mathbf{R} \cup \{+\infty\}$, with the laws of addition, multiplication and comparison as customary in convex analysis, see [HL12, pp. 5–6]. We consider functions mapping into the extended reals, as it conveniently allows us to formulate constrained optimization problems in an unified "unconstrained" notation. For $f : \mathbf{V} \to R$ and $C \subset \mathbf{V}$ compare

$$\min_{x \in C} \; f(x), \tag{2.1}$$

with the problem over $g : \mathbf{V} \to \overline{\mathbf{R}}$:

$$\min_{x \in \mathbf{V}} \; g(x) := f(x) + \delta_C(x). \tag{2.2}$$

The function $\delta_C : \mathbf{V} \to \overline{\mathbf{R}}$ is the *indicator function* of the set C and defined by

$$\delta_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.3}$$

When it is clear from the context it will be convenient to write $\delta\{x_1 \geq x_2 + 5\}$ instead of $\delta_C(x)$ as it avoids to explicitly define a set $C = \{x \in \mathbf{R}^2 \; : \; x_1 \geq x_2 + 5\}$.

Recall that a set $C \subset \mathbf{V}$ is called *convex* if for any two points $x, y \in C$ the line segment $\alpha x + (1 - \alpha)y$, $\alpha \in [0, 1]$ is also contained in the set. Often, one can apply results for sets to functions by considering their *epigraph*

$$\text{epi} f = \{(x, t) \in \mathbf{V} \times \mathbf{R} \; : \; f(x) \leq t\} \subset \mathbf{V} \times \mathbf{R}. \tag{2.4}$$

A function $f : \mathbf{V} \to \overline{\mathbf{R}}$ is *convex* if and only if its epigraph is a convex set. For a lower-semicontinuous function $f : \mathbf{V} \to \overline{\mathbf{R}}$

$$f(x_0) \leq \liminf_{x \to x_0} f(x), \tag{2.5}$$

the epigraph is a closed set and we will also refer to the function as *closed*.

The *domain* of a function $f : \mathbf{V} \to \overline{\mathbf{R}}$ is given by

$$\text{dom} f = \{x \in \mathbf{V} \; : \; f(x) \neq \infty\}. \tag{2.6}$$

A function $f : \mathbf{V} \to \overline{\mathbf{R}}$ is called proper if its domain is nonempty. The set of proper, convex and lower-semicontinuous functions mapping from $\mathbf{V}$ to $\overline{\mathbf{R}}$ is denoted as $\Gamma_0(\mathbf{V})$. The *subdifferential* of a function $f \in \Gamma_0(\mathbf{V})$ is defined by

$$\partial f(x) = \{p \in \mathbf{V}^* \; : \; f(y) \geq f(x) + \langle p, y - x \rangle \; \text{ for all } \; y \in \mathbf{V}\}. \tag{2.7}$$

It is empty if $x \notin \text{dom} f$.

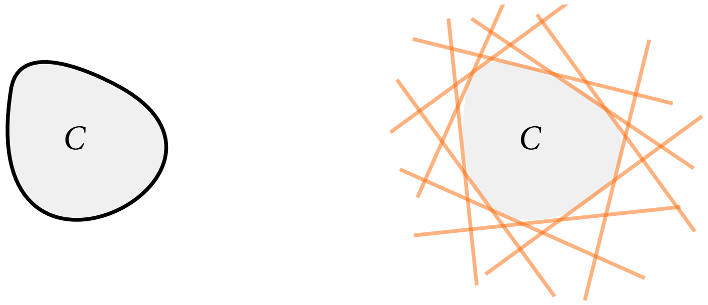


Figure 2.1: Illustration of the geometric intuition behind convex duality. A closed, convex set $C \subset \mathbf{V}$ can be viewed as a collection of points but at the same time also as an intersection of half-spaces.

2.1.2 Convex duality

A fundamental observation (see e.g. [HL12, Chapter A, Corollary 4.2.4]) which leads to the concept of convex duality is that a closed convex set $C \subset V$ can either be viewed as a collection of points but also as the intersection of closed half-spaces which contain C, see Fig. 2.1. The proof of this observation relies on a separation theorem, which for a point $x \notin C$ ensures the existence of a hyperplane which strictly separates x from C. This is also referred to as the geometric version of the Hahn-Banach theorem, see [HL12, Chapter A, Remark 4.1.2]

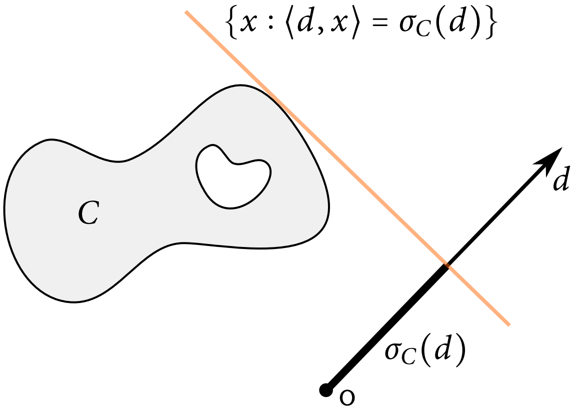


Figure 2.2: Support function of a (nonconvex) set C for a direction d with $\|d\| = 1$.

For a nonempty set $C \subset \mathbf{V}$, an important notion (dual to the previous indicator functions δ_C) is that of a *support function* $\sigma_C : \mathbf{V}^* \to \overline{\mathbf{R}}$:

$$\sigma_C(d) = \sup_{x \in C} \langle d, x \rangle. \tag{2.8}$$

Support functions are *sublinear*, i.e., convex and positively one-homogeneous. A function is positively one-homogeneous if $\sigma(cx) = c\sigma(x)$ holds for all $c \geq 0$. For such functions, the norm of the argument is not too important. Therefore, the argument itself can be interpreted as an oriented direction. This intuition will be helpful to keep in mind. As illustrated in Fig. 2.2, for $\|d\| = 1$, the support function (2.8) represents the distance of a supporting hyperplane with normal d from the origin.

Another important notion is the *convex conjugate* (also called Legendre-Fenchel conjugate) of a function $f : \mathbf{V} \to \overline{\mathbf{R}}$, defined as follows:

$$f^*(p) = \sup_{x \in \mathbf{V}} \langle p, x \rangle - f(x). \tag{2.9}$$

By [HL12, Chapter E, Prop. 1.2.1], it can be written as the support function of the epigraph and therefore seen as a dual representation of the function:

$$f^*(p) = \sup_{(x,t) \in \mathrm{epi} f} \langle p, x \rangle - t = \sigma_{\mathrm{epi} f}(p, -1). \tag{2.10}$$

With Fig. 2.2 in mind, this gives an interpretation of $f^*(p)$ as a (scaled) distance from the origin to a supporting hyperplane of the epigraph.

Another useful interpretation is due to [HL12, Chapter E, Sec. 1.2]. Suppose we are given a family of hyperplanes in the ”graph space” $\mathbf{V} \times \mathbf{R}$ with normal vector $(p, -1)$ and parametrized by $\langle p, x \rangle - r$. Now, we wish to push the hyperplane as close as possible to the function f from below by varying the parameter r:

$$\langle p, x \rangle - r \leq f(x), \text{ for all } x \in \mathbf{V}. \tag{2.11}$$

The constraints in the above equation can be reduced to a single one by taking the supremum:

$$\sup_{x \in \mathbf{V}} \langle p, x \rangle - f(x) = f^*(p) \leq r. \tag{2.12}$$

The optimal parameter is given by $r = f^*(p)$ and the best affine lower bound with slope p is $l(x) = \langle p, x \rangle - f^*(p)$.

Applying the convex conjugate operation (2.9) twice, one obtains the biconjugate $f^{**} = (f^*)^*$. Using the definition, it is given as

$$f^{**}(x) = \sup_{p \in \mathbf{V}^*} \langle x, p \rangle - f^*(p). \tag{2.13}$$

With the previous geometrical interpretations, one can observe that the biconjugate represents the function as a supremum over all supporting hyperplanes to the epigraph, see Fig. 2.3 for an illustration.

Indeed, it can also be shown that the epigraph of the biconjugate f^{**} is the closure of the convex hull of epi$f$, see [HL12, Theorem 1.3.5]. It follows that for $f \in \Gamma_0(\mathbf{V})$ we have the desirable representation

$$f(x) = \sup_{p \in \mathbf{V}^*} \langle x, p \rangle - f^*(p) = f^{**}(x). \tag{2.14}$$

In the following chapters, we will often use the definition of the convex biconjugate as a useful technical tool to compute the largest closed convex underapproximation to a nonconvex function.

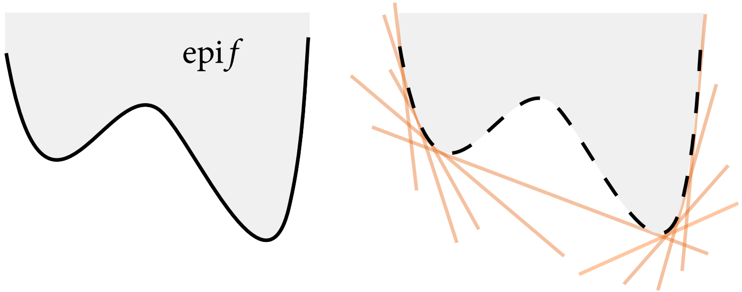


Figure 2.3: Visualization of the Legendre-Fenchel biconjugate f^{**} as the supremum over a collection over supporting hyperplanes. The biconjugate f^{**} is the largest convex function below f, the intuition being that the supporting hyperplanes cannot reach into the nonconvex dents of f.

2.1.3 Perspective functions

Perspective functions are a central ingredient for the functional lifting procedures presented in this thesis. They offer a general way of turning a convex function into a sublinear function in one dimension higher. The lifting procedure is essentially based on a reformulation using perspective functions. Furthermore, in another context, perspective functions will appear in the discretization of the infinite-dimensional objective using piecewise-linear approximations.

Formally, the perspective of a function $f \in \Gamma_0(V)$ is given as:

$$f^{\oslash}(x,t) = \begin{cases} tf(x/t), & \text{if } t > 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.15}$$

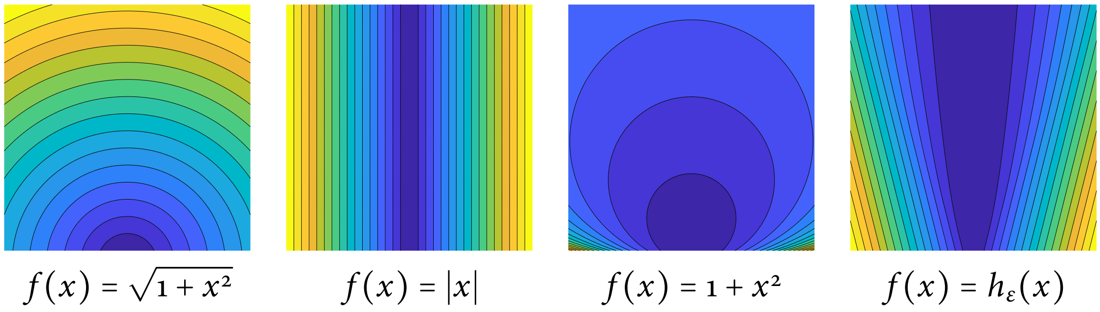
At $t = 1$ it coincides with the original function, i.e., $f^{\oslash}(x,1) = f(x)$.

It turns out that the perspective of a function is convex if and only if the function is convex. To gain an intuition, we show the perspective of a few selected functions in Fig. 2.4. For $f(x) = \sqrt{1+x^2}$ and $t > 0$ let us calculate

$$f^{\oslash}(x,t) = t\sqrt{1 + x^2/t^2} = \sqrt{t^2 + x^2} = \|(x,t)\|_2. \tag{2.16}$$

Due to its construction the perspective is not closed at the point $t = 0$ where it is perhaps a bit harshly set to $+\infty$. For optimization purposes later on it is convenient to rather work with the lower-semicontinuous envelope, which from now on we will also refer to as $f^{\oslash}$. It extends the function at $t = 0$ in a lower-semicontinuous fashion and is given as follows [HL12]:

$$f^{\oslash}(x,t) = \sigma_{\text{epi} f^*}(x,-t) = \begin{cases} tf^{**}(x/t), & \text{if } t > 0, \\ \sigma_{\text{dom} f^*}(x), & \text{if } t = 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.17}$$



$f(x) = \sqrt{1+x^2}$ $f(x) = |x|$ $f(x) = 1 + x^2$ $f(x) = h_\varepsilon(x)$

Figure 2.4: The perspective function is a canonical way to turn a convex function $f : \mathbf{R} \to \mathbf{R}$ into a convex one-homogeneous function $f^{\oslash} : \mathbf{R} \times \mathbf{R}_{>0} \to \mathbf{R}$.

The quantity $f^{\infty}(x) = \sigma_{\mathrm{dom} f^*}(x)$ is referred to as the *recession function* of f.

To get an intuition about (2.17) recall that by definition of the convex conjugate (2.10) we have the representation

$$f^{**}(x) = \sigma_{\mathrm{epi} f^*}(x, -1) = f^{\oslash}(x, 1). \tag{2.18}$$

By relaxing the second argument in the support function from 1 to t we naturally arrive at (2.17).

Perspective functions and convex (bi)conjugates are the main concepts from convex analysis which are required to understand Chapter 3 – Chapter 6. Additional results will be introduced in the chapters themselves.

2.2 Notions from Geometric Measure Theory

In this section, we recall some basic notions from geometric measure theory. More specific concepts such as exterior algebra, currents and differential forms will be introduced in Chapter 6 and Chapter 7. For this section, we follow the presentation in [AFP00] and [Mor16].

2.2.1 Basic measure theory

Let $\mathcal{X} \subset \mathbf{R}^n$ be a compact set. We denote the space of finite, $\mathbf{R}^m$-valued Radon measures by $\mathcal{M}(\mathcal{X}; \mathbf{R}^m)$, and we abbreviate $\mathcal{M}(\mathcal{X}; \mathbf{R}) = \mathcal{M}(\mathcal{X})$. Positive Radon measures are denoted by $\mathcal{M}_+(\mathcal{X})$ and probability measures by $\mathcal{P}(\mathcal{X})$.

For a finite $\mathbf{R}^m$-valued Radon measure $\mu \in \mathcal{M}(\mathcal{X}; \mathbf{R}^m)$, its *total variation* is a positive measure $\|\mu\| \in \mathcal{M}_+(\mathcal{X})$ given by [AFP00, Prop. 1.47]:

$$\|\mu\|(A) = \sup \left\{ \sum_{i=1}^{m} \int \varphi_i \, \mathrm{d}\mu_i : \varphi \in \mathcal{C}_c(A; \mathbf{R}^m), \|\varphi\|_\infty \leq 1 \right\}. \tag{2.19}$$

For a discrete measure $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ with coefficients $a \in \mathbf{R}^N$ the total variation reduces to the usual finite-dimensional ℓ_1-norm, i.e., $\|\mu\|(\mathcal{X}) = \sum_i |a_i| = \|a\|_1$. For a measure $\mu = f\,\mathrm{d}x$ which is absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x$, we have $\|\mu\|(\mathcal{X}) = \int_{\mathcal{X}} |f(x)|\mathrm{d}x$.

Another measure that we will frequently use is the m-dimensional Hausdorff measure, see [Mor16, Section 2.3]:

$$\mathcal{H}^m(A) = \lim_{\delta \to 0} \inf \left\{ \sum_j \alpha_m \left(\frac{\operatorname{diam}(S_j)}{2} \right)^m : A \subset \bigcup S_j, \operatorname{diam}(S_j) < \delta \right\}. \tag{2.20}$$

Here, α_m is the volume of the unit ball in $\mathbf{R}^m$ and the infimum is taken over all countable covers $\{S_j\}$ of A with diameter less than δ. It measures the k-dimensional volume of sets in $\mathbf{R}^n$.

The restriction of a measure $\mu \llcorner S$ to some set S is another measure given by $(\mu \llcorner S)(A) = \mu(S \cap A)$. The Hausdorff measure (2.20) is very useful, as it allows one to define m-dimensional geometric structures in a purely intrinsic way, i.e., without specifying a certain parametrization or topology. For example, this can be done by restricting the m-dimensional Hausdorff measure to some m-dimensional embedded manifold $S \subset \mathbf{R}^n$, i.e., $\mu = \mathcal{H}^m \llcorner S \in \mathcal{M}(\mathcal{X})$. Measures of this type (and more general less regular ones) are called *rectifiable measures*, see [AFP00, Definition 2.59]. In particular, if $S \subset \mathbf{R}^n$ is a $\mathcal{H}^m$-measurable set with $\mathcal{H}^m(S) < \infty$, then $\mathcal{H}^m \llcorner S \in \mathcal{M}(\mathcal{X})$, see [Mat99, p. 57].

The attractiveness of the measure theoretic viewpoint for practical and optimization purposes comes from the fact that it allows one to treat geometric structures (e.g. embedded manifolds) as elements of a *linear (vector) space*: $\mathcal{M}(\mathcal{X})$! This linear structure is important, as it makes the geometric theory compatible with vector space optimization, convex analysis and duality techniques. This fundamental fact has been exploited already many times in the past to solve different shape optimization problems in computer vision with the help of convex optimization techniques. Importantly, no restrictions on the topology of the shape or explicit parametrizations are necessary!

2.2.2 Convex functionals on measures

In Chapter 3, Chapter 4 and Chapter 6 we will make use of the *polar decomposition* of a measure, [AFP00, Corollary 1.29]. For a vector-valued measure $\mu \in \mathcal{M}(\mathcal{X};\mathbf{R}^m)$, the polar decomposition guarantees the existence of a unique $\mathbb{S}^{m-1}$-valued function $\vec{\mu} \in L^1(\mathcal{X}, \|\mu\|, \mathbb{S}^{m-1})$ such that $\mu = \vec{\mu} \cdot \|\mu\|$.

Given a Borel function $\Psi : \mathcal{X} \times \mathbf{R}^n \to \mathbf{R}$ which is convex and positively one-homogeneous in the second argument, the above decomposition can be used to define a functional on measures $\mu \in \mathcal{M}(\mathcal{X};\mathbf{R}^n)$ as follows:

$$I(\mu) = \int \Psi(x, \mu) := \int \Psi(x, \vec{\mu}(x))\, \mathrm{d}\|\mu\|(x). \tag{2.21}$$

For example, if the vector-measure μ describes the normal field of an embedded surface in codimension one, the above functional could for example penalize the weighted area of that surface under the choice $\Psi(x, n) = w(x)\|n\|$.

The functional $I : \mathcal{M}(\mathcal{X};\mathbf{R}^n) \to \mathbf{R}$ is again convex and positively 1-homogeneous [AFP00, Prop. 2.37]. Lower-semicontinuity can be guaranteed, if Ψ is lower-semicontinuous in its first argument [AFP00, Theorem 2.38].

Recall from the previous section, that for convex one-homogeneous functions Ψ, the argument can be interpreted as a direction. This fits nicely together with the fact that in the polar-decomposition we have $\vec{\mu}(x) \in \mathbb{S}^{n-1}$. For the lifting procedures presented in this thesis, this argument will carry the interpretation of a unit jump normal along an interface between two regions or more generally in Chapter 6 a multivector which describes an oriented tangent space of a manifold.

2.2.3 Functions of bounded variation

In the following chapters, we will often not directly work with vector-valued measures but rather with functions whose distributional derivative is a measure:

$$BV(\mathcal{X};\mathbf{R}) = \{f \in L^1(\mathcal{X};\mathbf{R}) \;:\; Df \in \mathcal{M}(\mathcal{X};\mathbf{R}^n)\}\,. \tag{2.22}$$

For such a function of bounded variation, its total variation is given by the total variation of its distributional derivative in the sense of (2.19), i.e.,

$$TV(f) = \|Df\|(\mathcal{X}) = \sup\left\{\int \operatorname{div}\varphi \cdot f \mathrm{d}x : \varphi \in C_c^1(\mathcal{X};\mathbf{R}^n), \|\varphi\|_\infty \leq 1\right\}. \tag{2.23}$$

Functions $f \in BV(\mathcal{X};\mathbf{R})$ have finite total variation, i.e., $TV(f) < \infty$. For more information, we refer the reader to [AFP00, Chapter 3].

2.2.4 Lifting, area and coarea formulas

In this section, we give short review of theoretical results from [Poc+10], as we build upon their formulation for the liftings considered in Chapter 5 and Chapter 6. For that, we introduce two more tools. The first one is the area formula [KP08, Corollary 5.1.13], which is essentially a change of variables.

Theorem 1 (Area formula). *Let $M \leq N$ and $\phi : \mathbf{R}^M \to R^N$ be a Lipschitz function. For $g : A \to \mathbf{R}$ and $A \subset \mathbf{R}^M$ both $\mathcal{H}^M$-measurable, we have:*

$$\int_A g(x) J_M\phi(x)\,\mathrm{d}x = \int_{\mathbf{R}^N} \sum_{x \in A \cap \phi^{-1}(y)} g(x)\,\mathrm{d}\mathcal{H}^M(y). \tag{2.24}$$

In the above theorem, $J_M\phi$ denotes the Jacobian of ϕ, see [KP08, Definition 5.1.3]. For differentiable ϕ, it is given by $J_M\phi(x) = \sqrt{\det(\nabla\phi(x)^\top \nabla\phi(x))}$.

Consider the perspective function (cf. [Poc+10, Eq. 3.4]):

$$\Psi(x,y,v) = c^{\oslash}(x,y,v_x,-v_t) = \begin{cases} -v_t c(x,y,v_x/-v_t), & \text{if } v_t < 0, \\ c^\infty(x,y,v_x), & \text{if } v_t = 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.25}$$

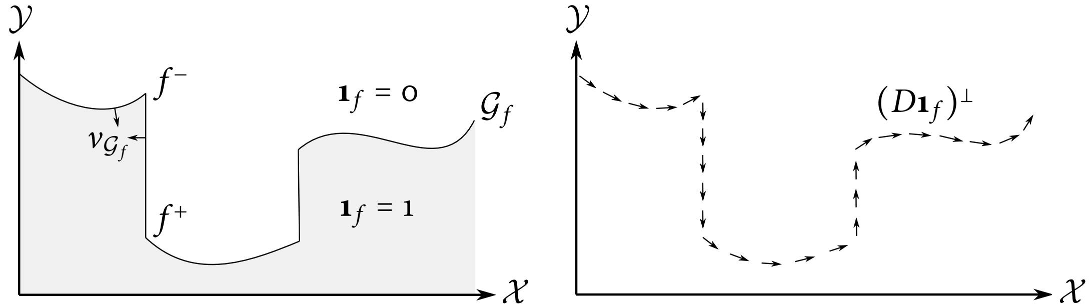


Figure 2.5: In codimension one it is possible to represent the graph $\mathcal{G}_f$ as the derivative of an indicator function of the subgraph. The derivative $D\mathbf{1}_f$ is a vector-valued measure concentrating on the graph. On the right we show a 90 degree rotated version $(D\mathbf{1}_f)^\perp$ for better visualization. In general, the relation between tangent and normal vectors is given by the Hodge star operator.

Assuming that f is differentiable, by using the area formula we can reparametrize our variational problem (1.18) to the graph of the function, see [Poc+10, pg. 6]:

$$\begin{aligned} E(f) &= \int_{\mathcal{X}} c(x, f(x), \nabla f(x))\,\mathrm{d}x = \int_{\mathcal{X}} c^{\oslash}(x, f(x), \nabla f(x), 1)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \Psi\Big(x, f(x), \frac{(\nabla f(x), -1)}{\sqrt{1 + \|\nabla f(x)\|^2}}\Big)\sqrt{1 + \|\nabla f(x)\|^2}\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \Psi(x, f(x), \nu_{\mathcal{G}_f}(x, f(x)))\sqrt{1 + \|\nabla f(x)\|^2}\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \Psi(x, f(x), \nu_{\mathcal{G}_f}(x, f(x)))J_n f(x)\,\mathrm{d}x \\ &= \int_{\mathcal{G}_f} \Psi(x, y, \nu_{\mathcal{G}_f}(x, y))\,\mathrm{d}\mathcal{H}^n(x, y) \\ &=: \mathbf{E}(\nu_{\mathcal{G}_f} \cdot \mathcal{H}^n \llcorner \mathcal{G}_f). \end{aligned} \tag{2.26}$$

To arrive at the final result, we applied Theorem 1 with $\phi : \mathbf{R}^n \to \mathbf{R}^{n+1}$, $\phi(x) = (x, f(x))$ and $g(x) = \Psi\left(x, f(x), \nu_{\mathcal{G}_f}(x, f(x))\right)$. The lifted energy $\mathbf{E} : \mathcal{M}(\mathcal{X} \times \mathcal{Y}; \mathbf{R}^{n+1}) \to \overline{\mathbf{R}}$ is a convex functional on measures as defined in the previous section. The above calculation shows that if $\mu = \nu_{\mathcal{G}_f} \cdot \mathcal{H}^n \llcorner \mathcal{G}_f$, we have $\mathbf{E}(\mu) = E(f)$! In Chapter 6, this calculation will be carried out in general codimension, see also [AG91; Mor02; GMS98].

The set of measures of type $\nu_{\mathcal{G}_f} \cdot \mathcal{H}^n \llcorner \mathcal{G}_f$, which concentrate on the graph of a function, is a nonconvex subset of $\mathcal{M}(\mathcal{X}; \mathbf{R}^{n+1})$. By representing these measures as the derivative of a subgraph function, we can define a relaxation without requiring the general notions of currents. The idea is illustrated in Fig. 2.5. The subgraph

$\mathbf{1}_f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ of the function f is given as follows:

$$\mathbf{1}_f(x, y) = \begin{cases} 0, & \text{if } f(x) < y, \\ 1, & \text{if } f(x) \geq y. \end{cases} \tag{2.27}$$

It turns out that the (distributional) derivative of $\mathbf{1}_f$ is exactly a vector-valued measure that concentrates on the graph of f:

$$D\mathbf{1}_f = \nu_{\mathcal{G}_f} \cdot \mathcal{H}^n \llcorner \mathcal{G}_f, \tag{2.28}$$

see for example [ABD03, Eq. 2.3]. Additionally, $\nu_{\mathcal{G}_f} : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}^{n+1}$ is the inward pointing normal on $\mathcal{G}_f$. In case f has discontinuities, the measure $D\mathbf{1}_f$ will have support on the *complete graph*, where the jumps have been "stitched". Note that the vertical parts in the graph correspond to the "limiting" case in the perspective function (2.25).

To obtain a convex problem, the idea in [Poc+10] is to relax from subgraph functions to the larger set of nonincreasing functions given by:

$$\begin{aligned} \mathcal{C} = \Big\{ \boldsymbol{v} \in BV(\mathcal{X} \times \mathcal{Y}; [0, 1]) : & \lim_{y \to y^-} \boldsymbol{v}(x, y) = 1, \\ & \lim_{y \to y^+} \boldsymbol{v}(x, y) = 0, \\ & \boldsymbol{v}(x, \cdot) \text{ is nonincreasing.} \Big\}. \end{aligned} \tag{2.29}$$

The relaxed optimization problem is then given by

$$\min_{\boldsymbol{v} \in \mathcal{C}} \mathbf{E}(D\boldsymbol{v}) = \int \Psi(D\boldsymbol{v}) := \int \Psi(x, y, \vec{D\boldsymbol{v}}(x, y)) \, \mathrm{d}\|D\boldsymbol{v}\|(x, y). \tag{2.30}$$

The main idea of the generalization we present in Chapter 6 is the following: instead of optimizing over a function $\boldsymbol{v} \in \mathcal{C}$ and then defining an energy on the measure $D\boldsymbol{v}$ we directly optimize over a vector-measure (equivalently a *finite mass current*) μ! Instead of a normal vector $\nu_{\mathcal{G}_f}$ the measure is multivector-valued, where the multivectors span the (oriented) tangent plane of the graph.

Returning to the relaxed optimization problem (2.30), an important question is whether this relaxation to a larger set of admissible candidates decreases the minimum, i.e., if $\min_{\boldsymbol{v} \in \mathcal{C}} \mathbf{E}(D\boldsymbol{v}) < \inf_f \, E(f)$. It turns out that this is not the case, as was proven in [Poc+10, Theorem 3.1] using a (generalized) coarea formula for BV functions [AFP00, Theorem 3.40].

Theorem 2 (Coarea formula)**.** *For $\boldsymbol{v} \in BV(\mathcal{X} \times \mathcal{Y}; \mathbf{R})$ and Ψ convex one-homogeneous one has the following representation*

$$\int \Psi(D\boldsymbol{v}) = \int_{-\infty}^{\infty} \int \Psi(D\mathbf{1}_{\boldsymbol{v}>s}) \, \mathrm{d}s, \tag{2.31}$$

where $\mathbf{1}_{\boldsymbol{v}>s} \in BV(\mathcal{X} \times \mathcal{Y}; \{0, 1\})$ is a zero-one thresholding of the function $\boldsymbol{v}$ at s.

For sake of completeness, we repeat the proof from [Poc+10, Theorem 3.1]. Assume that $\boldsymbol{v}^*$ is a minimizer of (2.30). Due to the minimality we know that $\mathbf{E}(D\boldsymbol{v}^*) \leq \mathbf{E}(D\mathbf{1}_{\boldsymbol{v}^*>s})$ for any s. Therefore, we have that

$$\mathbf{E}(D\boldsymbol{v}^*) = \int_0^1 \mathbf{E}(D\boldsymbol{v}^*)\mathrm{d}s \leq \int_0^1 \mathbf{E}(D\mathbf{1}_{\boldsymbol{v}^*>s})\mathrm{d}s. \tag{2.32}$$

Due to the above coarea formula, equality holds and it follows that for almost every $s \in (0,1)$ the thresholded solution $D\mathbf{1}_{\boldsymbol{v}^*>s}$ is also a minimizer since otherwise we would have a contradiction to the minimality of $\boldsymbol{v}^*$.

Essentially, the coarea formula can be interpreted in the following way: it is possible to write any ”diffuse” graph surface $D\boldsymbol{v}$ as an integral over binary ones $D\mathbf{1}_{\boldsymbol{v}>s}$ without changing the energy. In higher dimension and codimension it is unclear under which conditions such a result holds. Such ”foliations” of normal currents into an integral over integral currents has recently been studied in [AM17; AMS19] for general dimension and codimension. However, counterexamples exist and it seems that a nonconvex involutivity constraint is required for such a foliation to exist.

Central to the numerical implementation in [Poc+10] is the following dual representation of the convex one-homogeneous energy as a support functional of a certain set $\mathcal{K} \subset \mathcal{C}_0(\mathcal{X} \times \mathcal{Y}; \mathbf{R}^{n+1})$:

$$\mathbf{E}(D\boldsymbol{v}) = \sup_{\varphi \in \mathcal{K}} \int \varphi \cdot D\boldsymbol{v}. \tag{2.33}$$

In the seminal work [ABD03], additional nonlocal constraints are introduced on the dual variable $\mathcal{K}$. These constraints allow a more refined penalization of the jump discontinuities and will be considered in Chapter 5.

2.3 Proximal Splitting Methods

The convex relaxations considered in this thesis will eventually lead to large-scale (nonsmooth) convex optimization problems. While in principle there are many ways to solve such optimization problems, we will consider proximal splitting methods. This is mainly due to their low memory requirement (which scales linearly in the problem dimension) and their ability to exploit the sparsity present in the problem structure. Furthermore, the methods are attractive due to their flexibility, parallelizability and simplicity of implementation.

Perhaps because of the above reasons, proximal splitting methods have become the standard way to solve large-scale nonsmooth convex problems in imaging and vision [CP11a]. The drawback is, that these methods typically have a sublinear worst-case convergence rate, i.e., they require $\mathcal{O}(1/\varepsilon)$ iterations to find an ε-accurate solution. Even to reach a modest accuracy of $\varepsilon = 10^{-6}$, millions of iterations can be required in the worst case. In practice, a much faster convergence is observed, and locally, a linear convergence behaviour is often observed. Nevertheless, an efficient implementation on GPUs is important.

2.3.1 Projection and proximal mappings

The *proximal mapping* forms the basic building block for the presented algorithms. For $f \in \Gamma_0(\mathbf{V})$ it is given as follows:

$$\operatorname{prox}_{\tau,f}(v) = (\mathrm{id} + \tau\partial f)^{-1}(v) = \arg\min_{x \in \mathbf{V}} \ f(x) + \frac{1}{2\tau}\|x - v\|^2. \tag{2.34}$$

Due to the quadratic term, the minimization problem is strongly convex and therefore has a unique minimizer. The parameter $\tau > 0$ can be interpreted as a step-size. It determines the trade-off between minimizing f and staying close to the input argument v.

In case $f(x) = \delta_C(x)$ is the indicator function of a convex set $C \subset \mathbf{V}$ the proximal mapping

$$\operatorname{prox}_{\tau,f}(v) = \arg\min_{x \in C} \|x - v\|^2 = \operatorname{proj}_C(v). \tag{2.35}$$

reduces to the orthogonal projection onto the set C.

A useful tool is Moreau's identity, which relates the proximal operator of f to the one of f^*:

$$\operatorname{prox}_{\tau,f}(x) = x - \tau \operatorname{prox}_{\tau^{-1},f^*}(x/\tau). \tag{2.36}$$

We will also consider proximal operators for matrix-valued step-sizes. Given a symmetric positive definite matrix T, we define

$$\operatorname{prox}_{T,f}(v) = (\mathrm{id} + T\partial f)^{-1}(v) = \arg\min_{x \in \mathbf{V}} \ f(x) + \frac{1}{2}\|x - v\|^2_{T^{-1}}, \tag{2.37}$$

where the scaled norm is given by

$$\|x\|^2_{T^{-1}} = \|T^{-1/2}x\|^2 = \langle x, T^{-1}x\rangle. \tag{2.38}$$

Usually, the matrix T is chosen to be diagonal in order for the proximal mapping to still be computable in closed form.

2.3.2 A first-order primal-dual algorithm

We consider the primal-dual algorithm [Poc+09a; CP11a]. It is applicable to the following class of structured convex optimization problems:

$$\min_{x \in \mathbf{R}^n} \ G(x) + F(Kx), \tag{2.39}$$

which can also be written in the form of a saddle-point problem:

$$\min_{x \in \mathbf{R}^n} \max_{y \in \mathbf{R}^m} G(x) - F^*(y) + \langle Kx, y\rangle. \tag{2.40}$$

The map $K : \mathbf{R}^n \to \mathbf{R}^m$ is assumed to be linear and furthermore we assume that $G \in \Gamma_0(\mathbf{R}^n)$, $F \in \Gamma_0(\mathbf{R}^m)$. The iterative algorithm is given as follows:

$$\begin{aligned} x^{k+1} &= (I + \tau\partial G)^{-1}(x^k - \tau K^T y^k), \\ y^{k+1} &= (I + \sigma\partial F^*)^{-1}(y^k + \sigma K(x^{k+1} + \theta(x^{k+1} - x^k))). \end{aligned} \tag{2.41}$$

One way to interpret this algorithm is as an alternating gradient descent / gradient ascent scheme on the saddle point energy (2.40). For extrapolation parameter $\theta = 1$, the algorithm can also be understood as a *proximal point* method to find a zero of the maximally monotone operator describing the optimality conditions of (2.40), see [Roc76; HY12].

The main convergence result in [CP11a] establishes a $\mathcal{O}(1/\varepsilon)$ (ergodic) convergence rate for the partial primal-dual gap, in case $\tau\sigma\|K\|^2 < 1$. After convergence, we take the (non-ergodic) last iterate as it tends to perform better in practice [CP16b].

2.3.3 Extensions and implementation

All optimization problems in this thesis were solved using the `prost` framework, see Appendix C. It is based on a C++/CUDA implementation of the above primal-dual algorithm (2.41). The framework consists of a collection of commonly used proximal operators and abstract linear maps implemented in a matrix-free fashion. The framework includes bindings to MATLAB exist to allow for rapid prototyping.

Due to the generic structure of the implementation there is a large loss in efficiency over a completely optimized and hand-tailored CUDA implementation, often by a factor of more than 10-100 in runtime. Nevertheless, existing convex optimization libraries such as CVX [DB16] did not scale to the required problem size.

The framework uses several well-known tricks and heuristics that sometimes can improve the convergence speed of (2.41) in practice. What follows is mainly an adaptation of the heuristics used for the graph-form ADMM solver in [FB18] to the primal-dual algorithm (2.41).

2.3.3.1 Problem scaling and preconditioning

The first such trick is to introduce two symmetric positive definite matrices $\Sigma \in \mathbf{R}^{m \times m}$, $\mathrm{T} \in \mathbf{R}^{n \times n}$ into the problem formulation.

$$\min_{x \in \mathbf{R}^n, z \in \mathbf{R}^m} g(\mathrm{T}^{\frac{1}{2}} x) + f(\Sigma^{-\frac{1}{2}} z), \quad \text{s.t.} \quad z = \Sigma^{\frac{1}{2}} K \mathrm{T}^{\frac{1}{2}} x. \tag{2.42}$$

Under an invertible change of variables, problem (2.42) and (2.39) are equivalent. Similarly, one can derive a rescaled dual problem:

$$\max_{y \in \mathbf{R}^m, w \in \mathbf{R}^n} - g^*(\mathrm{T}^{-\frac{1}{2}} w) - f^*(\Sigma^{\frac{1}{2}} y), \quad \text{s.t.} \quad w = -\mathrm{T}^{\frac{1}{2}} K^T \Sigma^{\frac{1}{2}} y. \tag{2.43}$$

Formally applying the algorithm (2.41) to the modified problem (2.42) yields the iterations of the preconditioned primal dual method [PC11]:

$$\begin{aligned} x^{k+1} &= (I + \tau \mathrm{T} \partial g)^{-1}(x^k - \tau \mathrm{T} K^T y^k), \\ y^{k+1} &= (I + \sigma \Sigma \partial f^*)^{-1}(y^k + \sigma \Sigma K(x^{k+1} + \theta(x^{k+1} - x^k))). \end{aligned} \tag{2.44}$$

The choice of Σ and T can significantly influence the convergence speed in practice. Due to the nonlinear update dynamics of the algorithm, the effect of preconditioning on the theoretical convergence speed is however, far less understood as in the case of linear systems. We adopt the diagonal preconditioners proposed in [PC11], which typically yield a reliable performance boost in comparison to the method without preconditioning (i.e. $\Sigma = I,\ \mathrm{T} = I$).

2.3.3.2 Stopping criterion

As a stopping criterion, we consider the feasibility residuals of the linear constraint in the scaled problem (2.42) and its dual problem (2.43).

$$\begin{aligned} r_p^{k+1} &= \|Kx^{k+1} - z^{k+1}\|_\Sigma, \\ r_d^{k+1} &= \|K^T y^{k+1} + w^{k+1}\|_\mathrm{T}. \end{aligned} \tag{2.45}$$

Following [FB18], the auxiliary variables w and z can be recovered from the following relations:

$$\begin{aligned} w^{k+1} &= \frac{\mathrm{T}^{-1}(x^k - x^{k+1})}{\tau} - K^T y^k, \\ z^{k+1} &= \frac{\Sigma^{-1}(y^k - y^{k+1})}{\sigma} + K(x^{k+1} + \theta(x^{k+1} - x^k)). \end{aligned} \tag{2.46}$$

If both residuals in (2.45) are zero, then we have found an optimal primal-dual solution pair. In practice, the algorithm is stopped once the residuals $r_p < \varepsilon_a + \varepsilon_r \|z^{k+1}\|_\Sigma$ and $r_d < \varepsilon_a + \varepsilon_r \|w^{k+1}\|_\mathrm{T}$ are below some given absolute and relative tolerances $\varepsilon_a, \varepsilon_r > 0$ as suggested in [FB18].

2.3.3.3 Residual balancing

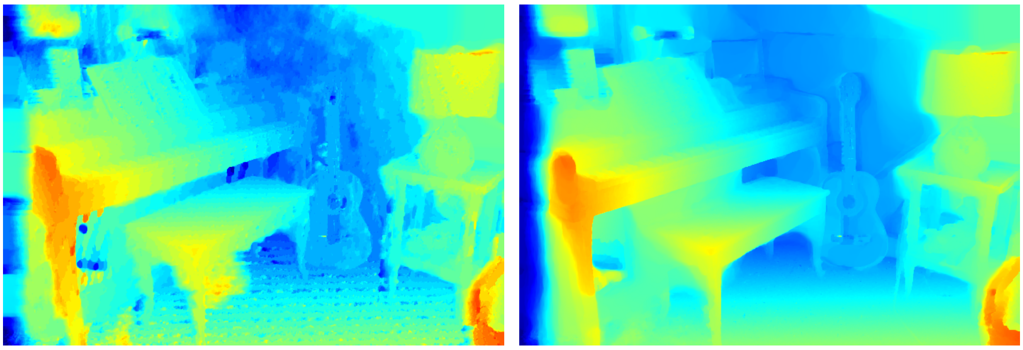
A heuristic which can speed up the convergence in practice is to adjust τ and σ during the iterations to balance the residuals (2.45). In [FB18] it is proposed to use fixed step sizes, until one of the residuals hits the desired tolerance and then adjust the step sizes in a way that the converged residual remains constant while the other one decreases. Another heuristic suggested in [Gol+13] is to adapt τ and σ such that during the process of optimization both residuals are in the same order of magnitude. We found the first heuristic to yield faster convergence in all of our experiments.

Part II

Own Publications

Chapter 3

Sublabel-Accurate Relaxation of Nonconvex Energies



[Poc+10], 48 labels, 1.49 GB, 52*s*.

Proposed, 8 labels, 0.49 GB, 30*s*.

Figure 3.1: We propose a convex relaxation for the variational model (3.1), which opposed to existing functional lifting methods [Poc+10; Poc+08] allows continuous label spaces *even after* discretization. Our method (here applied to stereo matching) avoids label space discretization artifacts, while saving on memory and runtime.

3.1 Introduction

Energy minimization methods have become the central paradigm for solving practical problems in computer vision. The energy functional can often be written as the sum of a data fidelity and a regularization term. One of the most popular regularizers is the total variation (TV) due to its many favorable properties [Cha+10]. Hence, an important class of optimization problems is given as

$$\min_{u:\Omega\to\Gamma} \int_\Omega \rho(x, u(x))\,\mathrm{d}x + \lambda\,\mathrm{TV}(u), \tag{3.1}$$

defined for functions u with finite total variation, arbitrary, possibly nonconvex dataterms $\rho : \Omega \times \Gamma \to \mathbf{R}$, label spaces Γ which are closed intervals in $\mathbf{R}$, $\Omega \subset \mathbf{R}^d$, and $\lambda \in \mathbf{R}^+$. The multilabel interpretation of the dataterm is that $\rho(x, u(x))$ represents the costs of assigning label $u(x)$ to point x. For (weakly) differentiable functions $TV(u)$ equals the integral over the norm of the derivative, and therefore favors a spatially coherent label configuration. The difficultly of minimizing the nonconvex energy (3.1) has motivated researchers to develop convex reformulations.

Convex representations of (3.1) and more general related energies have been studied in the context of the calibration method for the Mumford-Shah functional [ABD03]. Based on these works, relaxations for the piecewise constant [Poc+09b] and piecewise smooth Mumford-Shah functional [Poc+09a] have been proposed. Inspired by Ishikawa's graph-theoretic globally optimal solution to discrete variants of (3.1), continuous analogues have been considered by Pock et al.in [Poc+10; Poc+08]. Continuous relaxations for multilabeling problems with finite label spaces Γ have also been studied in [LS11].

Interestingly, the discretization of the aforementioned continuous relaxations is very similar to the linear programming relaxations proposed for MAP inference in the Markov Random Field (MRF) community [Ish03; Sch76; Wer07; ZHP13]. Both approaches ultimately discretize the range Γ into a finite set of labels. A closer analysis of these relaxations reveals, however, that they are not well-suited to represent the continuous valued range that we face in most computer vision problems such as stereo matching or optical flow. More specifically, the above relaxations are not designed to assign meaningful cost values to non-integral configurations. As a result, a large number of labels is required to achieve a faithful approximation. Solving real-world vision problems therefore entails large optimization problems with high memory and runtime requirement. To address this problem, Zach and Kohli [ZK12], Zach [Zac13] and Fix and Agarwal [FA14] introduced MRF-based approaches which retain continuous label spaces after discretization. For manifold-valued labels, this issue was addressed by Lellmann et al.[Lel+13a], however with the sole focus on the regularizer.

3.1.1 Contributions

We propose the first sublabel–accurate convex relaxation of nonconvex problems in a spatially continuous setting. It exhibits several favorable properties:

- In contrast to existing spatially continuous lifting approaches [Poc+10; Poc+08], the proposed method provides substantially better solutions with far fewer labels – see Fig. 3.1. This provides savings in runtime and memory.
- In Sec. 3.3 we show that the functional lifting methods [Poc+10; Poc+08] are a special case of the proposed framework.

- In Sec. 3.3 we show that, in a local sense, our formulation is the tightest convex relaxation which takes dataterm and regularizer into account separately. It is unknown whether this "local convex envelope" property also holds for the discrete approach [ZK12].

- Our formulation is compact and requires only half the amount of variables for the dataterm than the formulation in [ZK12]. We prove that the sublabel–accurate total variation can be represented in a very simple way, introducing no overhead compared to [Poc+10; Poc+08]. In contrast, the regularizer in [ZK12] is much more involved.

- Since our method is derived in a spatially continuous setting, the proposed approach easily allows different gradient discretizations. In contrast to [Zac13; ZK12] the regularizer is isotropic leading to noticeably less grid bias.

3.2 Notation and Mathematical Preliminaries

We make heavy use of the convex conjugate, which is given as $f^*(y) = \sup_{x \in \mathbf{R}^n} \langle y, x \rangle - f(x)$ for functions $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$. The biconjugate f^{**} denotes its *convex envelope*, i.e. the largest lower-semicontinuous convex under-approximation of f. For a set C we denote by δ_C the function which maps any element from C to 0 and is ∞ otherwise. For a comprehensive introduction to convex analysis, we refer the reader to [Roc96]. Vector valued functions $\boldsymbol{u} : \Omega \to \mathbf{R}^k$ are written in bold symbols. If it is clear from the context, we will drop the $x \in \Omega$ inside the functions, e.g., we write $\rho(u)$ for $\rho(x, u(x))$, or α for $\alpha(x)$.

3.3 Functional Lifting

To derive a convex representation of (3.1), we rely on the framework of functional lifting. The idea is to reformulate the optimization problem in a higher dimensional space. We numerically show in Sec. 3.5 that considering the convex envelope of the dataterm and regularizer in this higher dimensional space leads to a better approximation of the original nonconvex energy.

We start by sampling the range Γ at $L = k + 1$ labels $\gamma_1 < \cdots < \gamma_L \in \Gamma$. This partitions the range into k intervals $\Gamma_i = [\gamma_i, \gamma_{i+1}]$ so that $\Gamma = \Gamma_1 \cup \cdots \cup \Gamma_k$. For any value in the range of $u : \Omega \to \Gamma$ there exist a label index $1 \leq i \leq k$ and $\alpha \in [0, 1]$ such that

$$u(x) = \gamma_i^\alpha := \gamma_i + \alpha(\gamma_{i+1} - \gamma_i). \tag{3.2}$$

We represent a value in the range Γ by a vector in $\mathbf{R}^k$

$$\boldsymbol{u}(x) = \mathbf{1}_i^\alpha := \alpha \mathbf{1}_i + (1 - \alpha)\mathbf{1}_{i-1}, \tag{3.3}$$

where $\mathbf{1}_i$ denotes a vector starting with i ones followed by $k - i$ zeros.

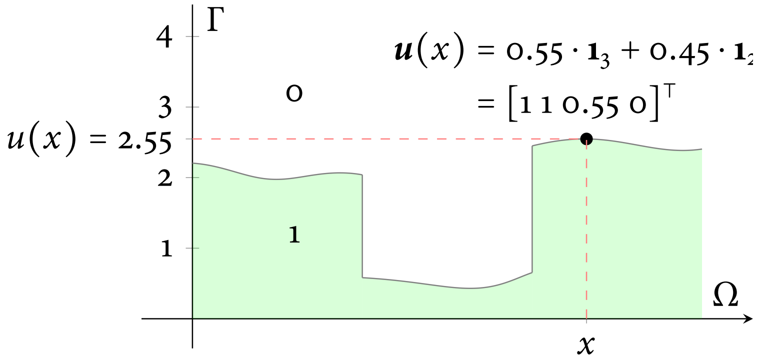

Γ
4
3
2
1
$u(x) = 2.55$
0
1
$\boldsymbol{u}(x) = 0.55 \cdot \mathbf{1}_3 + 0.45 \cdot \mathbf{1}_2$
$= [1\ 1\ 0.55\ 0]^\top$
Ω
x

Figure 3.2: Lifted representation. Instead of optimizing over the function $u : \Omega \to \Gamma$, we optimize over all possible graph functions (here shaded in green) on $\Omega \times \Gamma$. The main idea behind our approach is the finite dimensional representation of the graph at every $x \in \Omega$ by means of $\boldsymbol{u} : \Omega \to \mathbf{R}^k$ (here $k = 4$).

We call $\boldsymbol{u} : \Omega \to \mathbf{R}^k$ the *lifted* representation of u, representing the graph of u. This notation is depicted in Fig. 3.2 for $k = 4$. Back-projecting the lifted $\boldsymbol{u}(x)$ to the range of u using the layer cake formula yields a one-to-one correspondence between $u(x) = \gamma_i^\alpha$ and $\boldsymbol{u}(x) = \mathbf{1}_i^\alpha$ via

$$u(x) = \gamma_1 + \sum_{i=1}^{k} \boldsymbol{u}_i(x)(\gamma_{i+1} - \gamma_i). \tag{3.4}$$

We write problem (3.1) in terms of such graph functions, a technique that is used in the theory of Cartesian currents [GMS98].

3.3.1 Convexification of the dataterm

For now, we consider a fixed $x \in \Omega$. Then the dataterm from (3.1) is a possibly nonconvex real-valued function (cf. Fig. 3.3) that we seek to minimize over a compact interval Γ:

$$\min_{u \in \Gamma} \rho(u). \tag{3.5}$$

Due to the one-to-one correspondence between γ_i^α and $\mathbf{1}_i^\alpha$ it is clear that solving problem (3.5) is equivalent to finding a minimizer of the lifted energy:

$$\boldsymbol{\rho}(\boldsymbol{u}) = \min_{1 \le i \le k} \boldsymbol{\rho}_i(\boldsymbol{u}), \tag{3.6}$$

$$\boldsymbol{\rho}_i(\boldsymbol{u}) = \begin{cases} \rho(\gamma_i^\alpha), & \text{if } \boldsymbol{u} = \mathbf{1}_i^\alpha,\ \alpha \in [0,1], \\ \infty, & \text{else.} \end{cases} \tag{3.7}$$

Note that the constraint in (3.7) is essentially the nonconvex special ordered set of type 2 (SOS2) constraint [BT70]. More precisely, we demand that the "derivative"

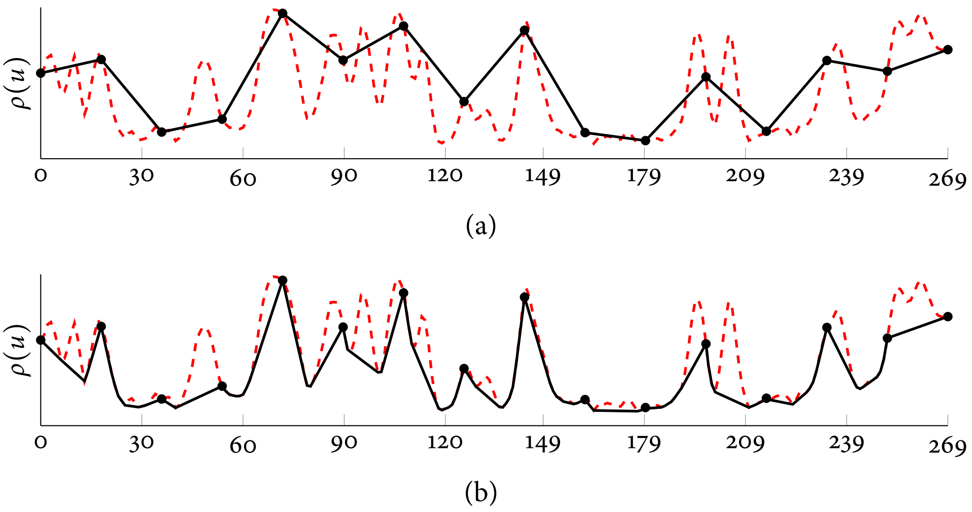


Figure 3.3: We show the nonconvex energy $\rho(u)$ at a fixed point $x \in \Omega$ (red dashed line in both plots) from the stereo matching experiment in Fig. 3.9 over the full range of 270 disparities. The black dots indicate the positions of the labels and the black curves show the approximations used by the respective methods. Fig. 3.3a: The baseline lifting method [Poc+10] uses a piecewise linear approximation with labels as nodes. Fig. 3.3b: The proposed method uses an optimal piecewise convex approximation. As we can see, the piecewise convex approximation is closer to the original nonconvex energy and therefore more accurate.

in label direction $\left(\partial_\gamma \boldsymbol{u}\right)_i := \boldsymbol{u}_{i+1} - \boldsymbol{u}_i$ is zero, except for two neighboring elements, which add up to one. In the following proposition, we derive the tightest convex relaxation of $\boldsymbol{\rho}$.

Proposition 1. *The convex envelope of* (3.6) *is given as:*

$$\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \sup_{\boldsymbol{v} \in \mathbf{R}^k} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \max_{1 \le i \le k} \boldsymbol{\rho}_i^*(\boldsymbol{v}), \tag{3.8}$$

where the conjugate of the individual $\boldsymbol{\rho}_i$ is

$$\boldsymbol{\rho}_i^*(\boldsymbol{v}) = c_i(\boldsymbol{v}) + \rho_i^*\left(\frac{\boldsymbol{v}_i}{\gamma_{i+1} - \gamma_i}\right), \tag{3.9}$$

with $c_i(\boldsymbol{v}) = \langle \mathbf{1}_{i-1}, \boldsymbol{v} \rangle - \frac{\gamma_i}{\gamma_{i+1} - \gamma_i} \boldsymbol{v}_i$ and $\rho_i = \rho + \delta_{\Gamma_i}$.

Proof. See Appendix A.1. □

The above proposition reveals that the convex relaxation implicitly convexifies the dataterm ρ on each interval Γ_i. The equality $\rho_i^* = \rho_i^{***}$ implies that starting with ρ_i yields exactly the same convex relaxation as starting with ρ_i^{**}.

Corollary 1. *If ρ is linear on each Γ_i, then the convex envelopes of $\boldsymbol{\rho}(\boldsymbol{u})$ and $\boldsymbol{\sigma}(\boldsymbol{u})$ coincide, where the latter is:*

$$\boldsymbol{\sigma}(\boldsymbol{u}) = \begin{cases} \rho(\gamma_i^\alpha), & \text{if } \exists i : \boldsymbol{u} = \mathbf{1}_i^\alpha,\ \alpha \in \{0, 1\}, \\ \infty, & \text{else.} \end{cases} \tag{3.10}$$

Proof. Consider an additional constraint $\delta_{\{\gamma_i, \gamma_{i+1}\}}$ for each ρ_i, which corresponds to selecting $\alpha \in \{0, 1\}$ in (3.7). The fact that our relaxation is independent of whether we choose ρ_i or ρ_i^{**}, along with the fact that the convex hull of two points is a line, yields the assertion. □

For the piecewise linear case, it is possible to find an explicit form of the biconjugate.

Proposition 2. *Let us denote by $\boldsymbol{r} \in \mathbf{R}^k$ the vector with*

$$\boldsymbol{r}_i = \rho(\gamma_{i+1}) - \rho(\gamma_i), \quad 1 \le i \le k. \tag{3.11}$$

Under the assumptions of Prop. 1, one obtains:

$$\boldsymbol{\sigma}^{**}(\boldsymbol{u}) = \begin{cases} \rho(\gamma_1) + \langle \boldsymbol{u}, \boldsymbol{r} \rangle, & \text{if } \boldsymbol{u}_i \ge \boldsymbol{u}_{i+1}, \boldsymbol{u}_i \in [0, 1], \\ \infty, & \text{else.} \end{cases} \tag{3.12}$$

Proof. See Appendix A.1. □

Up to an offset (which is irrelevant for the optimization), one can see that (3.12) coincides with the dataterm of [Poc+09b], the discretizations of [Poc+10; Poc+08], and – after a change of variable – with [LS11]. This not only proves that the latter is optimizing a convex envelope, but also shows that our method naturally generalizes the work from piecewise linear to arbitrary piecewise convex energies. Fig. 3.3a and Fig. 3.3b illustrate the difference of $\boldsymbol{\sigma}^{**}$ and $\boldsymbol{\rho}^{**}$ on the example of a nonconvex stereo matching cost.

Because our method allows arbitrary convex functions on each Γ_i, we can prove that, for the two label case, our approach optimizes the convex envelope of the dataterm.

Proposition 3. *In the case of binary labeling, i.e., $L = 2$, the convex envelope of* (3.6) *reduces to*

$$\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \rho^{**}\left(\gamma_1 + \boldsymbol{u}(\gamma_2 - \gamma_1)\right), \text{ with } \boldsymbol{u} \in [0, 1]. \tag{3.13}$$

Proof. See Appendix A.1. □

3.3.2 A lifted representation of the total variation

We now want to find a lifted convex formulation that emulates the total variation regularization in (3.1). We follow [CCP12] and define an appropriate integrand of the functional

$$TV(\boldsymbol{u}) = \int_{\Omega} \boldsymbol{\Phi}(x, D\boldsymbol{u}), \tag{3.14}$$

where the distributional derivative $D\boldsymbol{u}$ is a finite $\mathbf{R}^{k \times d}$-valued Radon measure [AFP00]. We define

$$\boldsymbol{\Phi}(\boldsymbol{g}) = \min_{1 \le i \le j \le k} \boldsymbol{\Phi}_{i,j}(\boldsymbol{g}). \tag{3.15}$$

The individual $\boldsymbol{\Phi}_{i,j} : \mathbf{R}^{k \times d} \to \mathbf{R} \cup \{\infty\}$ are given by:

$$\boldsymbol{\Phi}_{i,j}(\boldsymbol{g}) = \begin{cases} \left|\gamma_i^\alpha - \gamma_j^\beta\right| \cdot |\nu|_2, & \text{if } \boldsymbol{g} = \left(\mathbf{1}_i^\alpha - \mathbf{1}_j^\beta\right) \nu^\top, \\ \infty, & \text{else}, \end{cases} \tag{3.16}$$

for some $\alpha, \beta \in [0,1]$ and $\nu \in \mathbf{R}^d$. The intuition is that $\Phi_{i,j}$ penalizes a jump from γ_i^α to γ_j^β in the direction of ν. Since $\boldsymbol{\Phi}$ is nonconvex we compute the convex envelope.

Proposition 4. *The convex envelope of* (3.15) *is*

$$\boldsymbol{\Phi}^{**}(\boldsymbol{g}) = \sup_{\boldsymbol{p} \in \mathcal{K}} \langle \boldsymbol{p}, \boldsymbol{g} \rangle, \tag{3.17}$$

where $\mathcal{K} \subset \mathbf{R}^{k \times d}$ *is given as:*

$$\begin{aligned} \mathcal{K} = \Big\{ \boldsymbol{p} \in \mathbf{R}^{k \times d} \;\Big|\; \left|\boldsymbol{p}^\top \left(\mathbf{1}_i^\alpha - \mathbf{1}_j^\beta\right)\right|_2 \le \left|\gamma_i^\alpha - \gamma_j^\beta\right|, \\ \forall\, 1 \le i \le j \le k,\ \forall \alpha, \beta \in [0,1] \Big\}. \end{aligned} \tag{3.18}$$

Proof. See Appendix A.1. □

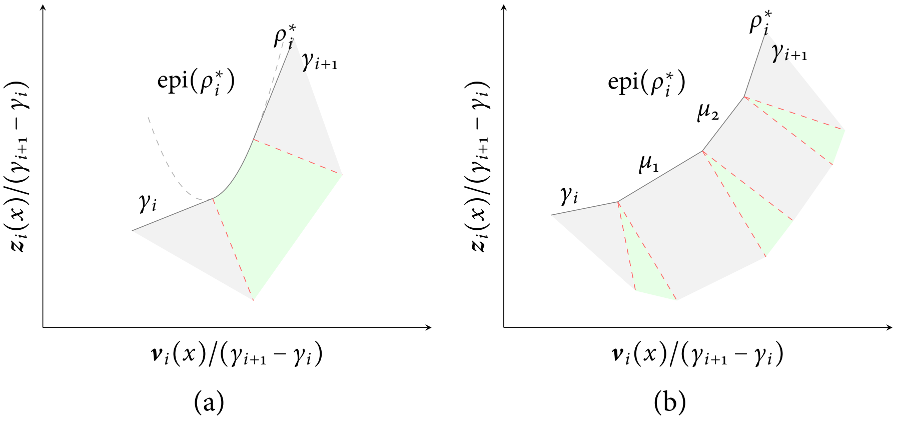
The set $\mathcal{K}$ from Eq. (3.18) involves infinitely many constraints which makes numerical optimization difficult. As the following proposition reveals, the infinite number of constraints can be reduced to only linearly many, allowing to enforce the constraint $\boldsymbol{p} \in \mathcal{K}$ exactly.

Proposition 5. *If the labels are ordered ($\gamma_1 < \gamma_2 < \cdots < \gamma_L$) then the constraint set $\mathcal{K}$ from Eq.* (3.18) *is equal to*

$$\mathcal{K} = \{ \boldsymbol{p} \in \mathbf{R}^{k \times d} \mid |\boldsymbol{p}_i|_2 \le \gamma_{i+1} - \gamma_i,\ \forall i \}. \tag{3.19}$$

Proof. See Appendix A.1. □

This shows that the proposed regularizer coincides with the total variation from [CCP12], where it has been derived based on (3.16) for α and β restricted to $\{0,1\}$. Prop. 5 together with Prop. 3 show that for $k = 1$ our formulation amounts to unlifted TV optimization with a convexified dataterm.



(a)

(b)

Figure 3.4: In the left subfigure the projection onto the epigraph of the conjugate of a convex quadratic ρ_i is shown. In the right subfigure the piecewise linear case is illustrated. In the both cases all points that lie in the gray sets are orthogonally projected onto the respective linear parts whereas the points that lie in the green sets are projected onto the parabolic part (in the quadratic case) respectively the kinks (in the piecewise linear case). In the piecewise linear case the green sets are normal cones. The red dashed lines correspond to the boundary cases. γ_i, γ_{i+1}, μ_1, μ_2 are the slopes of the segments of ρ_i^* respectively the (sub-)label positions of ρ_i.

3.4 Numerical Optimization

Discretizing $\Omega \subset \mathbf{R}^d$ as a d-dimensional Cartesian grid, the relaxed energy minimization problem becomes

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^k} \sum_{x\in\Omega} \boldsymbol{\rho}^{**}(x, \boldsymbol{u}(x)) + \boldsymbol{\Phi}^{**}(x, \nabla \boldsymbol{u}(x)), \tag{3.20}$$

where ∇ denotes a forward-difference operator with $\nabla \boldsymbol{u} : \Omega \to \mathbb{R}^{k\times d}$. We rewrite the dataterm given in equation (3.8) by replacing the pointwise maximum over the conjugates $\boldsymbol{\rho}_i^*$ with a maximum over a real number $q \in \mathbb{R}$ and obtain the following saddle point formulation of problem (3.20):

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^k} \max_{\substack{(\boldsymbol{v},q)\in\mathcal{C}\\ \boldsymbol{p}:\Omega\to\mathcal{K}}} \langle \boldsymbol{u}, \boldsymbol{v}\rangle - \sum_{x\in\Omega} q(x) + \langle \boldsymbol{p}, \nabla \boldsymbol{u}\rangle, \tag{3.21}$$

$$\mathcal{C} = \{(\boldsymbol{v}, q) : \Omega \to \mathbf{R}^k \times \mathbf{R} \mid q(x) \geq \boldsymbol{\rho}_i^*(\boldsymbol{v}(x)),\ \forall x, \forall i\}. \tag{3.22}$$

We numerically compute a minimizer of problem (3.21) using a first-order primal-dual method [EZC10; Poc+09a] with diagonal preconditioning [PC11] and adaptive steps [Gol+13]. It alternates between a gradient descent step in the primal variable and a gradient ascent step in the dual variable. Subsequently the dual variables are orthogonally projected onto the sets $\mathcal{C}$ respectively $\mathcal{K}$. In the following we give some hints on the implementation of the individual steps. For a detailed discussion we refer to [Gol+13]. The projection onto the set $\mathcal{K}$ is a simple ℓ_2-ball projection. To simplify the projection onto $\mathcal{C}$, we transform the k-dimensional epigraph constraints in (3.22) into 1-dimensional scaled epigraph constraints by introducing an additional variable $z : \Omega \to \mathbb{R}^k$ with:

$$\boldsymbol{z}_i(x) = \left[q(x) - c_i\left(\boldsymbol{v}(x)\right)\right]\left(\gamma_{i+1} - \gamma_i\right). \tag{3.23}$$

Using equation (3.9) we can write the constraints in (3.22) as

$$\frac{\boldsymbol{z}_i(x)}{\gamma_{i+1} - \gamma_i} \geq \rho_i^*\left(\frac{\boldsymbol{v}_i(x)}{\gamma_{i+1} - \gamma_i}\right). \tag{3.24}$$

We implement the newly introduced equality constraints (3.23) introducing a Lagrange multiplier $\boldsymbol{s} : \Omega \to \mathbb{R}^k$. It remains to discuss the orthogonal projections onto the epigraphs of the conjugates ρ_i^*. Currently we support quadratic and piecewise linear convex pieces ρ_i. For the piecewise linear case, the conjugate ρ_i^* is a piecewise linear function with domain $\mathbf{R}$. The slopes correspond to the x-positions of the sublabels and the intercepts correspond to the function values at the sublabel positions. The conjugates as well as the epigraph projections of both, a quadratic and a piecewise linear piece are depicted in Fig. 3.4. For the quadratic case, the projection onto the epigraph of a parabola is computed using [SCC14, App. B.2].

3.5 Experiments

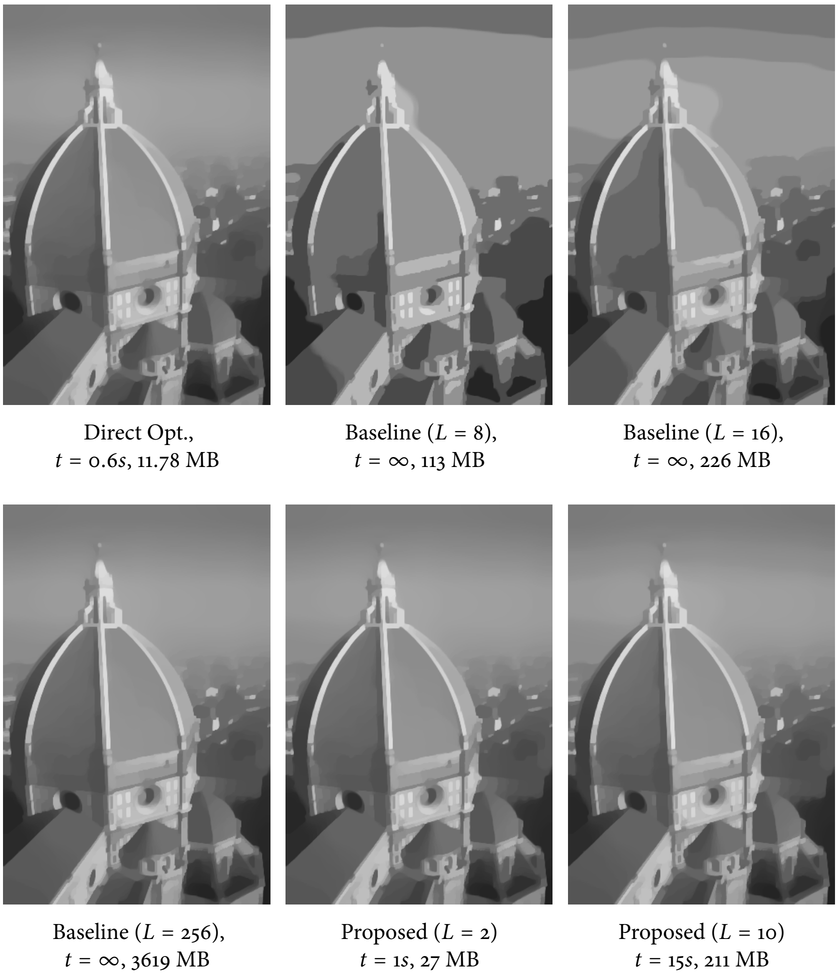
We implemented the primal-dual algorithm in CUDA to run on GPUs. For $d = 2$, our implementation of the functional lifting framework [Poc+10], which will serve as a baseline method, requires $4N(L-1)$ optimization variables, while the proposed method requires $6N(L-1) + N$ variables, where N is the number of points used to discretize the domain $\Omega \subset \mathbf{R}^d$. As we will show, our method requires much fewer labels to yield comparable results, thus, leading to an improvement in accuracy, memory usage, and speed.

3.5.1 Rudin-Osher-Fatemi model

As a proof of concept, we first evaluate the novel relaxation on the well-known Rudin-Osher-Fatemi (ROF) model [ROF92]. It corresponds to (3.1) with the following dataterm:

$$\rho(x, u(x)) = \left(u(x) - f(x)\right)^2, \tag{3.25}$$

where $f : \Omega \to \mathbf{R}$ denotes the input data. While there is no practical use in applying convex relaxation methods to an already convex problem such as the ROF model,



Direct Opt.,
$t = 0.6s$, 11.78 MB

Baseline ($L = 8$),
$t = \infty$, 113 MB

Baseline ($L = 16$),
$t = \infty$, 226 MB

Baseline ($L = 256$),
$t = \infty$, 3619 MB

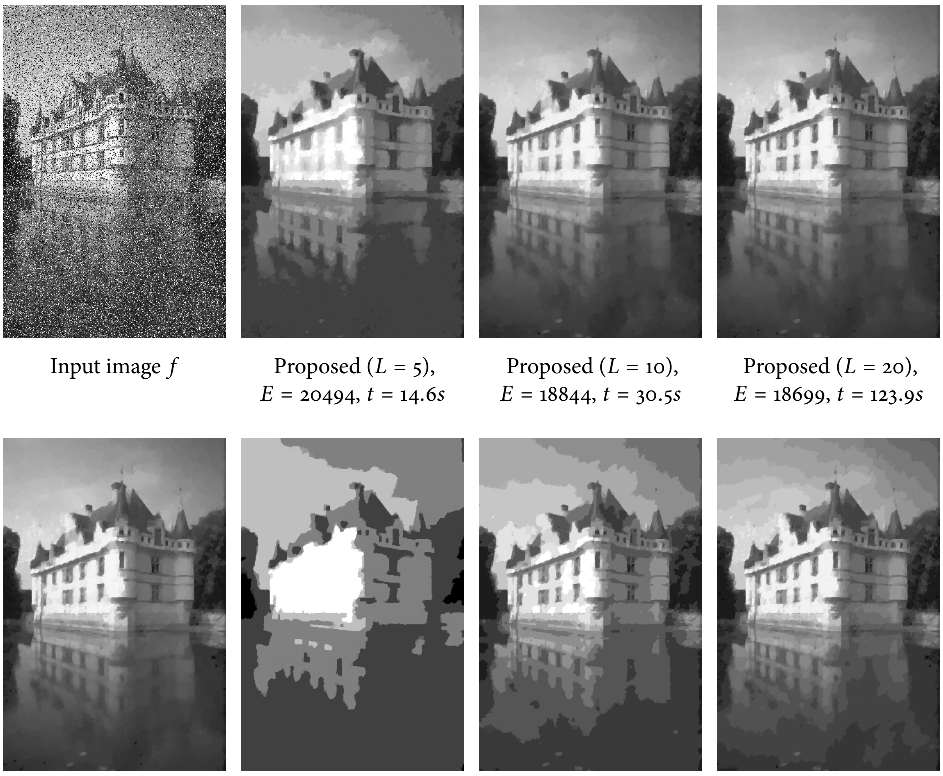
Proposed ($L = 2$)
$t = 1s$, 27 MB

Proposed ($L = 10$)
$t = 15s$, 211 MB

Figure 3.5: Denoising comparison. We compare the proposed method to the baseline method [Poc+10] on the convex ROF problem. We show the time in seconds required for each method to produce a solution within a certain energy gap to the optimal solution. As the baseline method optimizes a piecewise linear approximation of the quadratic dataterm, it fails to reach that optimality gap even for $L = 256$ (indicated by $t = \infty$). While the proposed lifting method can solve a large class of non-convex problems, it is almost as efficient as direct methods on convex problems.

the purpose of this is two-fold. Firstly, it allows us to measure the overhead introduced by our method by comparing it to standard convex optimization methods which do not rely on functional lifting. Secondly, we can experimentally verify that the relaxation is tight for a convex dataterm.

In Fig. 3.5 we solve (3.25) directly using the primal-dual algorithm [Gol+13], using the baseline functional lifting method [Poc+10] and using our proposed algorithm. First, the globally optimal energy was computed using the direct method with a very high number of iterations. Then we measure how long each method took to



Input image f

Proposed ($L = 5$), $E = 20494$, $t = 14.6s$

Proposed ($L = 10$), $E = 18844$, $t = 30.5s$

Proposed ($L = 20$), $E = 18699$, $t = 123.9s$

Baseline ($L = 256$), $E = 18660$, $t = 1001s$

Baseline ($L = 5$), $E = 23864$, $t = 4.7s$

Baseline ($L = 10$), $E = 19802$, $t = 6.3s$

Baseline ($L = 20$), $E = 18876$, $t = 12.8s$

Figure 3.6: Denoising using a robust truncated quadratic dataterm. The top row shows the input image along with the result obtained by our approach for a varying number of labels L. The bottom row illustrates the results obtained by the baseline method [Poc+10]. The energy of the final solution as well as the total runtime are given below each image.

reach this global optimum to a fixed tolerance.

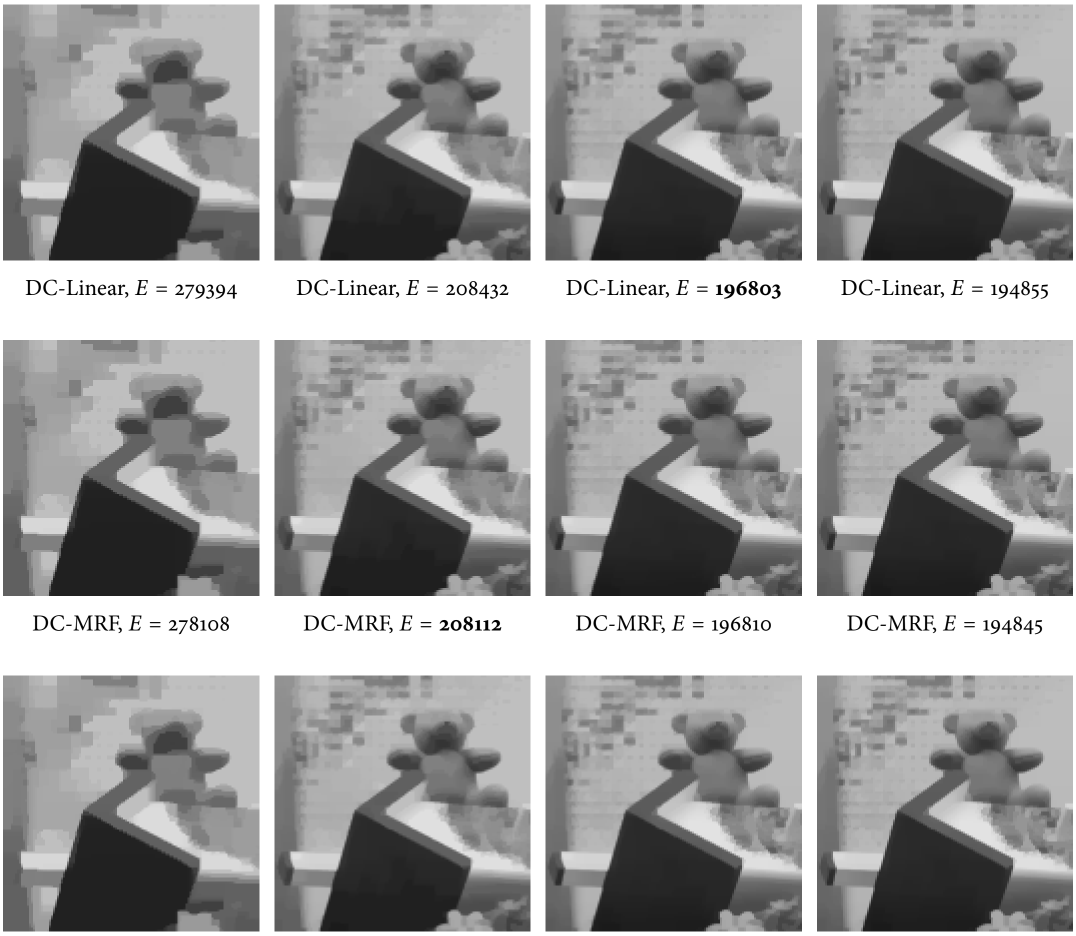
The baseline method fails to reach the global optimum even for 256 labels. While the lifting framework introduces a certain overhead, the proposed method finds the same globally optimal energy as the direct unlifted optimization approach and generalizes to nonconvex energies.

3.5.2 Robust truncated quadratic dataterm

The quadratic dataterm in (3.25) is often not well suited for real-world data as it comes from a pure Gaussian noise assumption and does not model outliers. We now consider a robust truncated quadratic dataterm:

$$\rho(x, u(x)) = \frac{\alpha}{2} \min \left\{ (u(x) - f(x))^2, \nu \right\}. \tag{3.26}$$

To implement (3.26), we use a piecewise polynomial approximation of the dataterm. In Fig. 3.6 we degraded the input image with additive Gaussian and salt and pepper

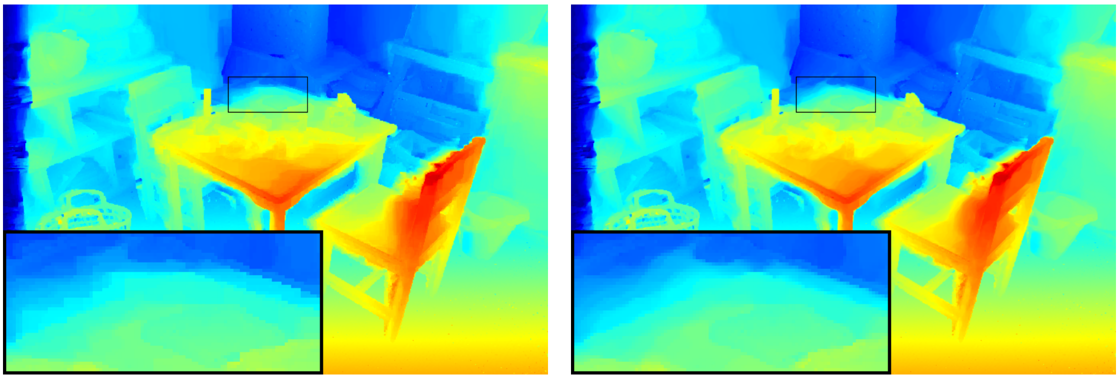


DC-Linear, E = 279394 DC-Linear, E = 208432 DC-Linear, E = **196803** DC-Linear, E = 194855

DC-MRF, E = 278108 DC-MRF, E = **208112** DC-MRF, E = 196810 DC-MRF, E = 194845

Proposed, E = **277970** **Proposed**, E = 208493 **Proposed**, E = 196979 **Proposed**, E = **194836**

Figure 3.7: Comparison to the MRF approach presented in [ZK12]. The first row shows DC-Linear, second row DC-MRF and third row our results for 4, 8, 16 and 32 convex pieces on the truncated quadratic energy (3.26). Below the figures we show the final nonconvex energy. We achieve competitive results while using a more compact representation and generalizing to isotropic regularizers.


(a) Anisotropic Regularization

(b) Isotropic Regularization

Figure 3.8: We compare the proposed relaxation with anistropic regularizer to isotropic regularization on the stereo matching example. Using an anisotropic formulation as in [ZK12] leads to grid bias.

noise. The parameters in (3.26) were chosen as $\alpha = 25$, $\nu = 0.025$ and $\lambda = 1$. It can be seen that the proposed method requires fewer labels to find lower energies than the baseline.

3.5.3 Comparison to the method of Zach and Kohli

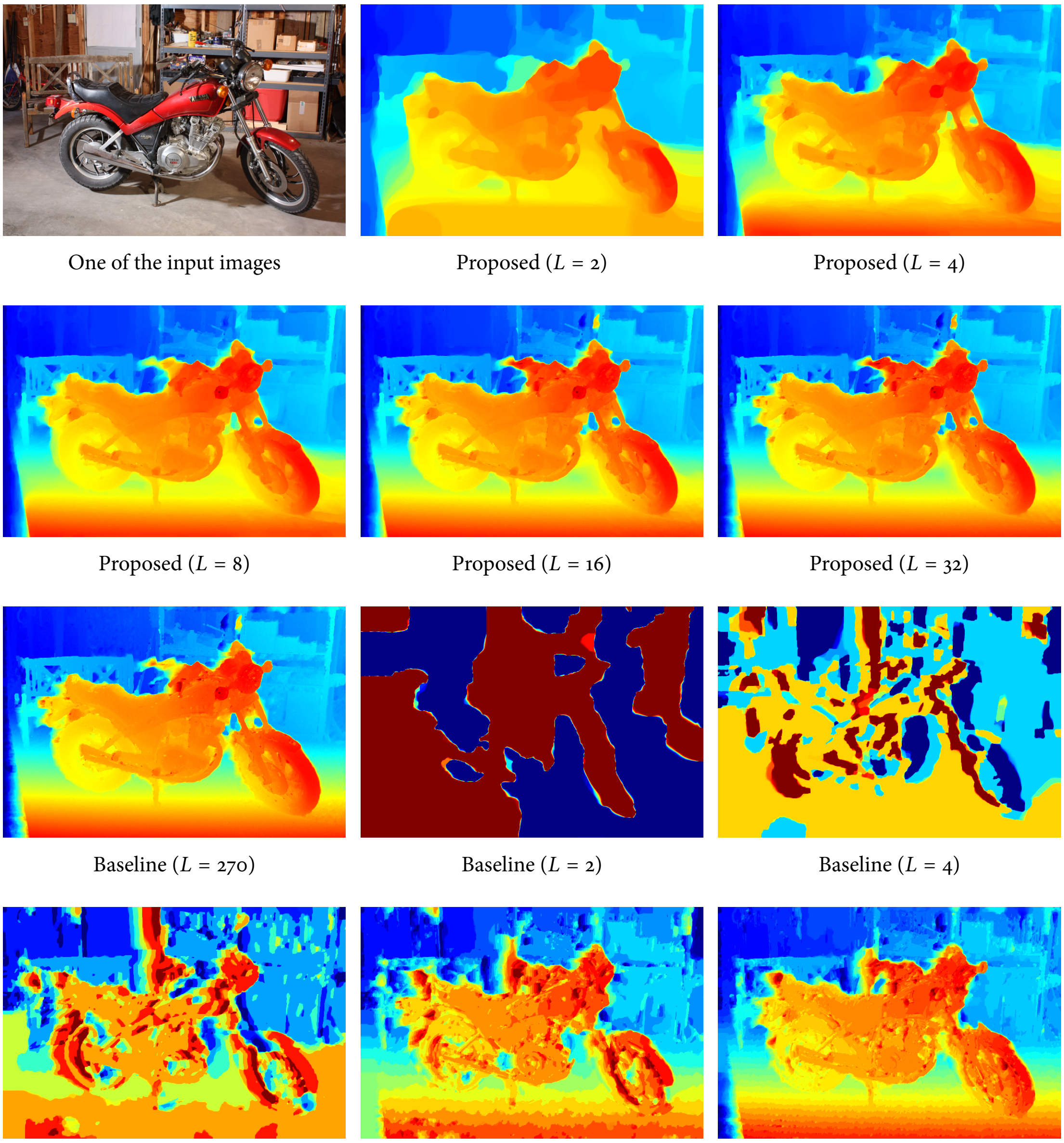
We remark that Prop. 4 and Prop. 5 hold for arbitrary convex one-homogeneous functionals $\phi(v)$ instead of $|v|_2$ in equation (3.16). In particular, they hold for the anisotropic total variation $\phi(v) = |v|_1$. This generalization allows us to directly compare our convex relaxation to the MRF approach of Zach and Kohli [ZK12].

In Fig. 3.7 we show the results of optimizing the two models entitled "DC-Linear" and "DC-MRF" proposed in [ZK12], and of our proposed method with anisotropic regularization on the robust truncated denoising energy (3.26). We picked the parameters as $\alpha = 0.2$, $\nu = 500$, and $\lambda = 1$. The label space is also chosen as $\Gamma = [0, 256]$ as described in [ZK12].

Note that overall, all the energies are better than the ones reported in [ZK12]. It can be seen from Fig. 3.7 that the proposed relaxation is competitive to the one proposed by Zach and Kohli. In addition, the proposed relaxation uses a more compact representation and extends to isotropic and convex one-homogeneous regularizers. To illustrate the advantages of isotropic regularizations, Fig. 3.8a and Fig. 3.8b show a comparison of our proposed method for isotropic and anisotropic regularization for the example of stereo matching discussed in the next section.

3.5.4 Stereo matching

Given a pair of rectified images, the task of finding a correspondence between the two images can be formulated as an optimization problem over a scalar field



One of the input images

Proposed ($L = 2$)

Proposed ($L = 4$)

Proposed ($L = 8$)

Proposed ($L = 16$)

Proposed ($L = 32$)

Baseline ($L = 270$)

Baseline ($L = 2$)

Baseline ($L = 4$)

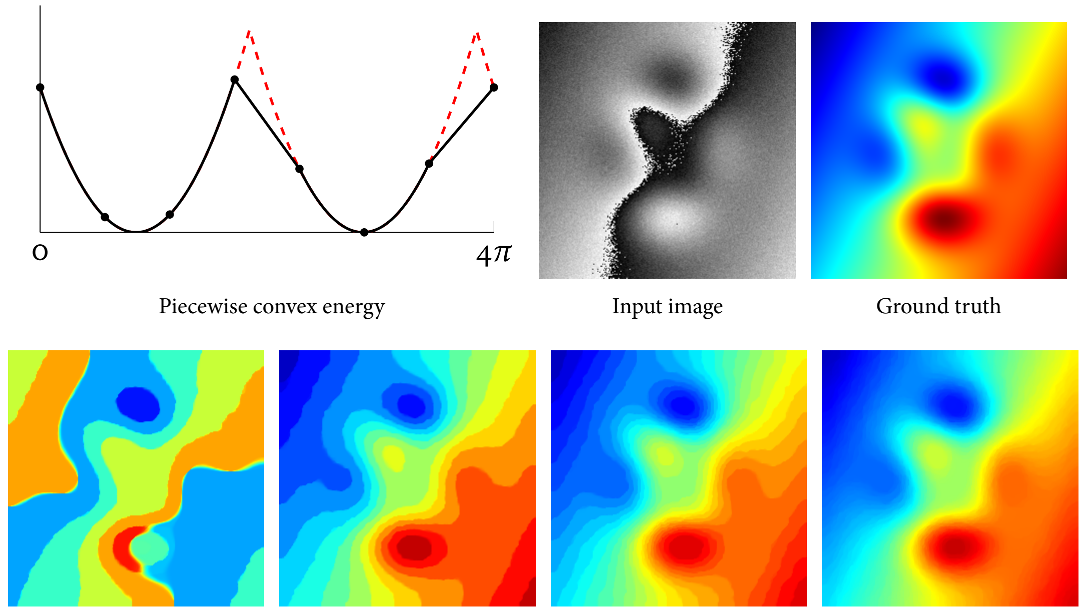
Baseline ($L = 8$)

Baseline ($L = 16$)

Baseline ($L = 32$)

Figure 3.9: Stereo comparison. We compare the proposed method to the baseline method on the example of stereo matching. The first column shows one of the two input images and below the baseline method with the full number of labels. The proposed relaxation requires much fewer labels to reach a smooth depth map. Even for $L = 32$, the label space discretization of the baseline method is strongly visible, while the proposed method yields a smooth result already for $L = 8$.

$u : \Omega \to \Gamma$ where each point $u(x) \in \Gamma$ denotes the displacement along the epipolar line associated with each $x \in \Omega$. The overall cost functional fits Eq. (3.1). In our experiments, we computed $\rho(x, u(x))$ for 270 disparities on the Middlebury stereo benchmark [Sch+14] in a 4 × 4 patch using a truncated sum of absolute gradient



Piecewise convex energy

Input image

Ground truth

Baseline ($L = 8$)

Baseline ($L = 16$)

Baseline ($L = 32$)

Proposed ($L = 8$)

Figure 3.10: We show the piecewise convex approximation of the phase unwrapping energy, followed by the cyclic input image and the unwrapped ground truth. With only 8 labels, the proposed method already yields a smooth reconstruction. The baseline method fails to unwrap the heightmap correctly using 8 labels, and for 16 and 32 labels, the discretization is still noticable.

differences. We convexify the matching cost ρ in each range Γ_i by numerically computing the convex envelope using the gift wrapping algorithm.

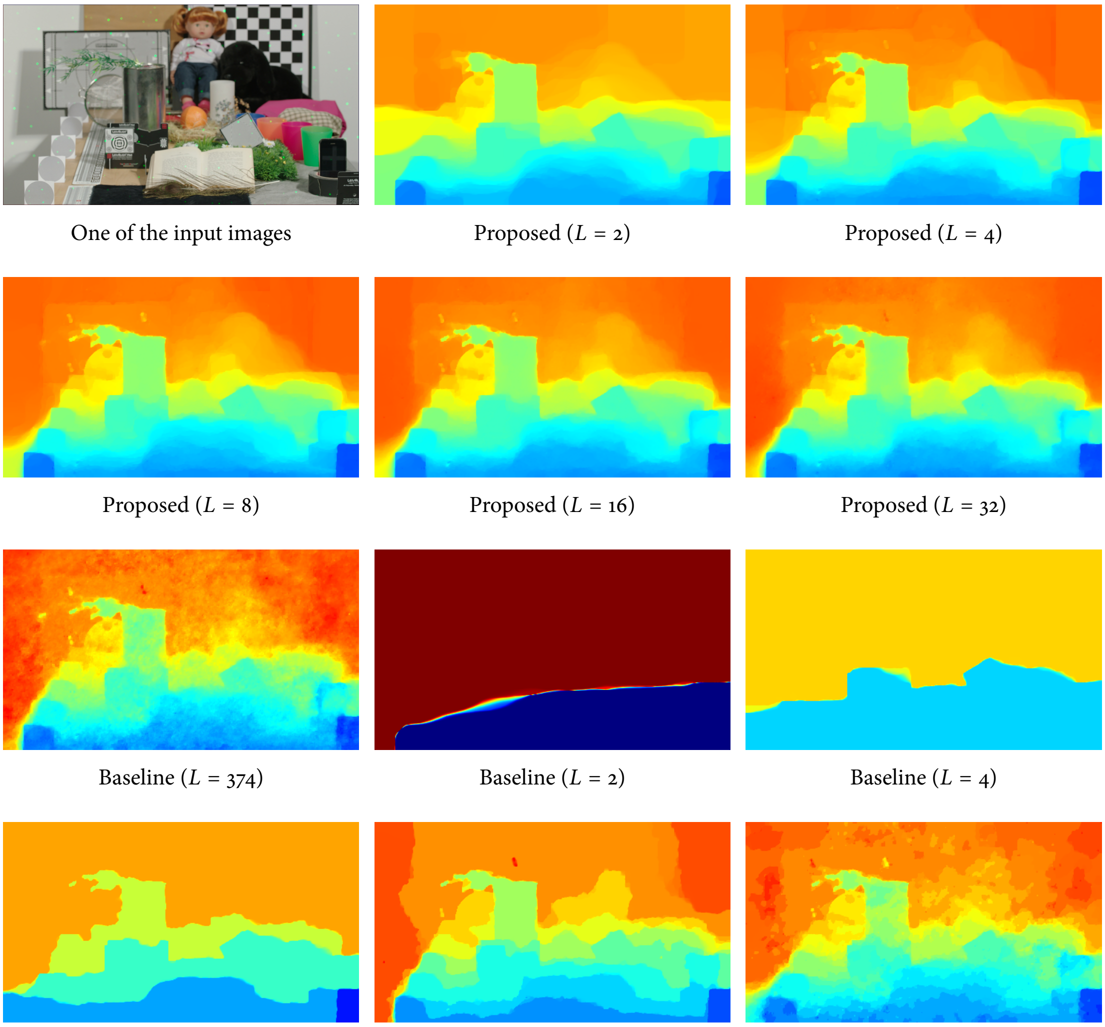
The first row in Fig. 3.9 shows the result of the proposed relaxation using the convexified energy between two labels. The second row shows the baseline approach using the same amount of labels. Even for $L = 2$, the proposed method produces a reasonable depth map while the baseline approach basically corresponds to a two region segmentation.

3.5.5 Phase unwrapping

Many sensors such as time-of-flight cameras or interferometric synthetic aperture radar (SAR) yield cyclic data lying on the circle $\mathcal{S}^1$. Here we consider the task of total variation regularized unwrapping. As is shown on the left in Fig. 3.10, the dataterm is a nonconvex function where each minimum corresponds to a phase shift by 2π:

$$\rho\left(x, u(x)\right) = d_{\mathcal{S}^1}\left(u(x), f(x)\right)^2. \tag{3.27}$$

For the experiments, we approximated the nonconvex energy by quadratic pieces as depicted in Fig. 3.10. The label space is chosen as $\Gamma = [0, 4\pi]$ and the regularization



One of the input images

Proposed (L = 2)

Proposed (L = 4)

Proposed (L = 8)

Proposed (L = 16)

Proposed (L = 32)

Baseline (L = 374)

Baseline (L = 2)

Baseline (L = 4)

Baseline (L = 8)

Baseline (L = 16)

Baseline (L = 32)

Figure 3.11: Depth from focus comparison. We compare our method to the baseline approach on the problem of depth from focus. First column: one of the 374 differently focused input images and the baseline method for full number of labels. Following columns: proposed relaxation (top row) vs. baseline (bottom row) for 2, 4, 8, 16 and 32 labels each.

parameter was set to $\lambda = 0.005$. Again, it is visible in Fig. 3.10 that the baseline method shows label space discretization and fails to unwrap the depth map correctly if the number of labels is chosen too low. The proposed method yields a smooth unwrapped result using only 8 labels.

3.5.6 Depth from focus

In depth from focus the task is to recover the depth of a scene, given a stack of images each taken from a constant position but in a different focal setting, so that in each image only the objects of a certain depth are sharp. images. We compute the dataterm cost ρ by using the modified Laplacian function [NN94] as a contrast measure.

Similar to the stereo experiments, we convexify the cost on each label range by computing the convex hull. The results are shown in Fig. 3.11. While the baseline method clearly shows the label space discretization, the proposed approach yields a smooth depth map. Since the proposed method uses a convex lower bound of the lifted energy, the regularizer has slightly more influence on the final result. This explains why the resulting depth maps in Fig. 3.11 and Fig. 3.9 look overall less noisy.

3.6 Conclusion

In this work we proposed a tight convex relaxation that can be interpreted as a sublabel–accurate formulation of classical multilabel problems. The final formulation is a simple saddle-point problem that admits fast primal-dual optimization. Our method maintains sublabel accuracy even after discretization and for that reason outperforms existing spatially continuous methods. Interesting directions for future work include higher dimensional label spaces, manifold valued data and more general regularizers.

Chapter 4

Sublabel-Accurate Convex Relaxation of Vectorial Multilabel Energies

4.1 Introduction

4.1.1 Nonconvex vectorial problems

In this paper, we derive a sublabel-accurate convex relaxation for vectorial optimization problems of the form

$$\min_{u:\Omega\to\Gamma} \int_{\Omega} \rho\big(x, u(x)\big)\,\mathrm{d}x + \lambda\, TV(u), \tag{4.1}$$

where $\Omega \subset \mathbf{R}^d$, $\Gamma \subset \mathbf{R}^n$ and $\rho : \Omega \times \Gamma \to \mathbf{R}$ denotes a generally nonconvex pointwise dataterm. As regularization we focus on the *total variation* defined as:

$$TV(u) = \sup_{q \in C_c^\infty(\Omega, \mathbf{R}^{n\times d}), \|q(x)\|_{S^\infty} \leq 1} \int_{\Omega} \langle u, \operatorname{div} q\rangle\,\mathrm{d}x, \tag{4.2}$$

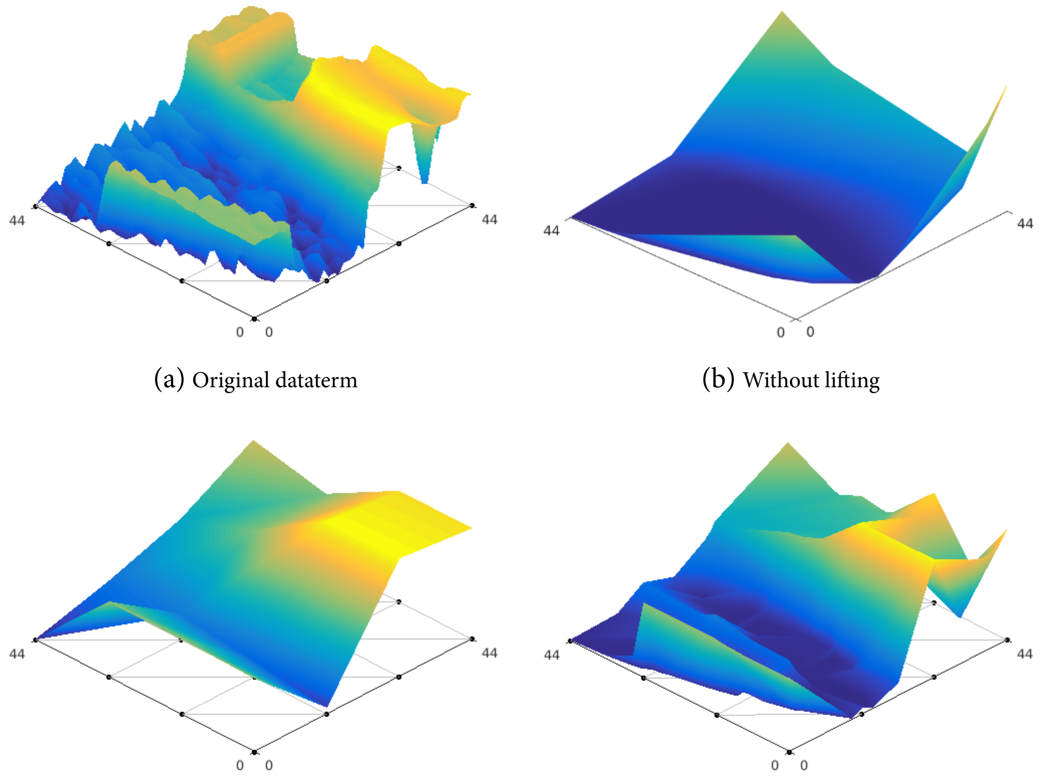
where $\|\cdot\|_{S^\infty}$ is the Schatten-∞ norm on $\mathbf{R}^{n\times d}$, i.e., the largest singular value. For differentiable functions u we can integrate (4.2) by parts to find

$$TV(u) = \int_{\Omega} \|\nabla u(x)\|_{S^1}\,\mathrm{d}x, \tag{4.3}$$

where the dual norm $\|\cdot\|_{S^1}$ essentially penalizes Jacobians ∇u which have high rank, i.e., the individual components of u jump in a different direction. This type of regularization is part of the framework of Sapiro and Ringach [SR96].

4.1.2 Related work

Due to its nonconvexity the optimization of (4.1) is challenging. For the scalar case ($n = 1$), Ishikawa [Ish03] proposed a pioneering technique to obtain globally



(a) Original dataterm

(b) Without lifting

(c) Classical lifting

(d) Proposed lifting

Figure 4.1: In (a) we show a nonconvex dataterm. Convexification without lifting would result in the energy (b). Classical lifting methods [Lel+13a] (c), approximate the energy piecewise linearly between the labels, whereas the proposed method results in an approximation that is convex on each triangle (d). Therefore, we are able to capture the structure of the nonconvex energy much more accurately.

optimal solutions in a spatially discrete setting, given by the minimum s-t-cut of a graph representing the space $\Omega \times \Gamma$. A continuous formulation was introduced by Pock et al. [Poc+08] exhibiting several advantages such as less grid bias and parallelizability.

In a series of papers [Poc+10; Poc+09a], connections of the above approaches were made to the mathematical theory of *cartesian currents* [GMS98] and the calibration method for the Mumford-Shah functional [ABD03], leading to a generalization of the convex relaxation framework [Poc+08] to more general (in particular nonconvex) regularizers.

In the following, researchers have strived to generalize the concept of functional lifting and convex relaxation to the vectorial setting ($n > 1$). If the dataterm and the regularizer are both separable in the label dimension, one can simply apply the above convex relaxation approach in a channel-wise manner to each component separately. But when either the dataterm or the regularizer couple the label components, the situation becomes more complex [GSC13; SCC14].

The approach which is most closely related to our work, and which we consider as a baseline method, is the one by Lellmann et al. [Lel+13a]. They consider coupled dataterms with coupled total variation regularization of the form (4.2).

A drawback shared by all mentioned papers is that ultimately one has to discretize the label space. While Lellmann et al. [Lel+13a] propose a sublabel-accurate regularizer, we show that their dataterm leads to solutions which still have a strong bias towards the label grid. For the scalar-valued setting, continuous label spaces have been considered in the MRF community by Zach et al. [ZK12] and Fix et al. [FA14]. The paper [Zac13] proposes a method for mixed continuous and discrete vectorial label spaces, where everything is derived in the spatially discrete MRF setting. Möllenhoff et al. [Möl+16] recently proposed a novel formulation of the scalar-valued case which retains fully continuous label spaces even after discretization. The contribution of this work is to extend [Möl+16] to vectorial label spaces, thereby complementing [Lel+13a] with a sublabel-accurate dataterm.

4.1.3 Contribution

In this work we propose the first sublabel-accurate convex formulation of vectorial labeling problems. It generalizes the formulation for scalar-valued labeling problems [Möl+16] and thus includes important applications such as optical flow estimation or color image denoising. We show that our method, derived in a spatially continuous setting, has a variety of interesting theoretical properties as well as practical advantages over the existing labeling approaches:

- We generalize existing functional lifting approaches (see Sec. 4.2.2).
- We show that our method is the best convex under-approximation (in a local sense), see Prop. 6 and Prop. 7.
- Due to its sublabel-accuracy our method requires only a small amount of labels to produce good results which leads to a drastic reduction in memory. We believe that this is a vital step towards the real-time capability of lifting and convex relaxation methods. Moreover, our method eliminates the label bias, that previous lifting methods suffer from, even for many labels.
- In Sec. 4.2.3 we propose a regularizer that couples the different label components by enforcing a joint jump normal. This is in contrast to [GSC13], where the components are regularized separately.
- For convex dataterms, our method is equivalent to the unlifted problem – see Prop. 9. Therefore, it allows a seamless transition between direct optimization and convex relaxation approaches.

4.1.4 Notation

We write $\langle x, y\rangle = \sum_i x_i y_i$ for the standard inner product on $\mathbf{R}^n$ or the Frobenius product if x, y are matrices. Similarly $\|\cdot\|$ without any subscript denotes the usual Euclidean norm, respectively the Frobenius norm for matrices.

We denote the convex conjugate of a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ by $f^*(y) = \sup_{x\in\mathbf{R}^n} \langle y, x\rangle - f(x)$. It is an important tool for devising convex relaxations, as the biconjugate f^{**} is the largest lower-semicontinuous (lsc.) convex function below f. For the indicator function of a set C we write δ_C, i.e., $\delta_C(x) = 0$ if $x \in C$ and ∞ otherwise. $\Delta_n^U \subset \mathbf{R}^n$ stands for the unit n-simplex.

4.2 Convex Formulation

4.2.1 Lifted representation

Motivated by Fig. 4.1, we construct an equivalent representation of (4.1) in a higher dimensional space, before taking the convex envelope.

Let $\Gamma \subset \mathbf{R}^n$ be a compact and convex set. We partition Γ into a set $\mathcal{T}$ of n-simplices Δ_i so that Γ is a disjoint union of Δ_i up to a set of measure zero. Let t^{i_j} be the j-th vertex of Δ_i and denote by $\mathcal{V} = \{t^1, \dots, t^{|\mathcal{V}|}\}$ the union of all vertices, referred to as labels, with $1 \le i \le |\mathcal{T}|$, $1 \le j \le n+1$ and $1 \le i_j \le |\mathcal{V}|$. For $u : \Omega \to \Gamma$, we refer to $u(x)$ as a *sublabel*. Any sublabel can be written as a convex combination of the vertices of a simplex Δ_i with $1 \le i \le |\mathcal{T}|$ for appropriate barycentric coordinates $\alpha \in \Delta_n^U$:

$$u(x) = T_i\alpha := \sum_{j=1}^{n+1} \alpha_j t^{i_j}, \; T_i := (t^{i_1},\ t^{i_2},\ \dots,\ t^{i_{n+1}}) \in \mathbf{R}^{n\times n+1}. \tag{4.4}$$

By encoding the vertices $t^k \in \mathcal{V}$ using a one-of-$|\mathcal{V}|$ representation e^k we can identify any $u(x) \in \Gamma$ with a sparse vector $\boldsymbol{u}(x)$ containing at least $|\mathcal{V}| - n$ many zeros and vice versa:

$$\begin{aligned} \boldsymbol{u}(x) &= E_i\alpha := \sum_{j=1}^{n+1} \alpha_j e^{i_j}, \; E_i := (e^{i_1},\ e^{i_2}, \dots,\ e^{i_{n+1}}) \in \mathbf{R}^{|\mathcal{V}|\times n+1}, \\ u(x) &= \sum_{k=1}^{|\mathcal{V}|} t^k \boldsymbol{u}_k(x), \; \alpha \in \Delta_n^U,\ 1 \le i \le |\mathcal{T}|. \end{aligned} \tag{4.5}$$

The entries of the vector e^{i_j} are zero except for the (i_j)-th entry, which is equal to one. We refer to $\boldsymbol{u} : \Omega \to \mathbf{R}^{|\mathcal{V}|}$ as the *lifted* representation of u. This one-to-one-correspondence between $u(x) = T_i\alpha$ and $\boldsymbol{u}(x) = E_i\alpha$ is shown in Fig. 4.2. Note that both, α and i depend on x. However, for notational convenience we drop the dependence on x whenever we consider a fixed point $x \in \Omega$.

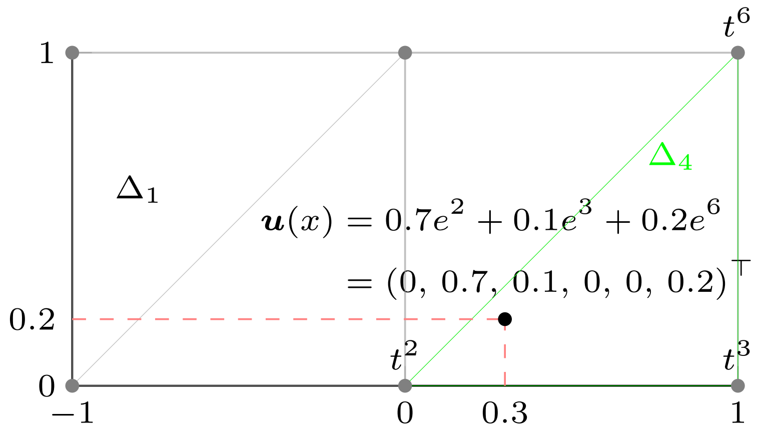

t^6
1
Δ_4
Δ_1
$\boldsymbol{u}(x) = 0.7e^2 + 0.1e^3 + 0.2e^6$
$= (0,\ 0.7,\ 0.1,\ 0,\ 0,\ 0.2)^\top$
0.2
t^2
t^3
0
-1
0
0.3
1

Figure 4.2: This figure illustrates our notation and the one-to-one correspondence between $u(x) = (0.3, 0.2)^\top$ and the lifted $\boldsymbol{u}(x)$ containing the barycentric coordinates $\alpha = (0.7, 0.1, 0.2)^\top$ of the sublabel $u(x) \in \Delta_4 = \text{conv}\{t^2, t^3, t^6\}$. The triangulation $(\mathcal{V}, \mathcal{T})$ of $\Gamma = [-1; 1] \times [0; 1]$ is visualized via the gray lines, corresponding to the triangles and the gray dots, corresponding to the vertices $\mathcal{V} = \{(-1, 0)^\top, (0, 0)^\top, \dots, (1, 1)^\top\}$, that we refer to as the labels.

4.2.2 Convexifying the dataterm

Let for now the weight of the regularizer in (4.1) be zero. Then, at each point $x \in \Omega$ we minimize a generally nonconvex energy over a compact set $\Gamma \subset \mathbf{R}^n$:

$$\min_{u \in \Gamma} \rho(u). \tag{4.6}$$

We set up the lifted energy so that it attains finite values if and only if the argument $\boldsymbol{u}$ is a sparse representation $\boldsymbol{u} = E_i \alpha$ of a sublabel $u \in \Gamma$:

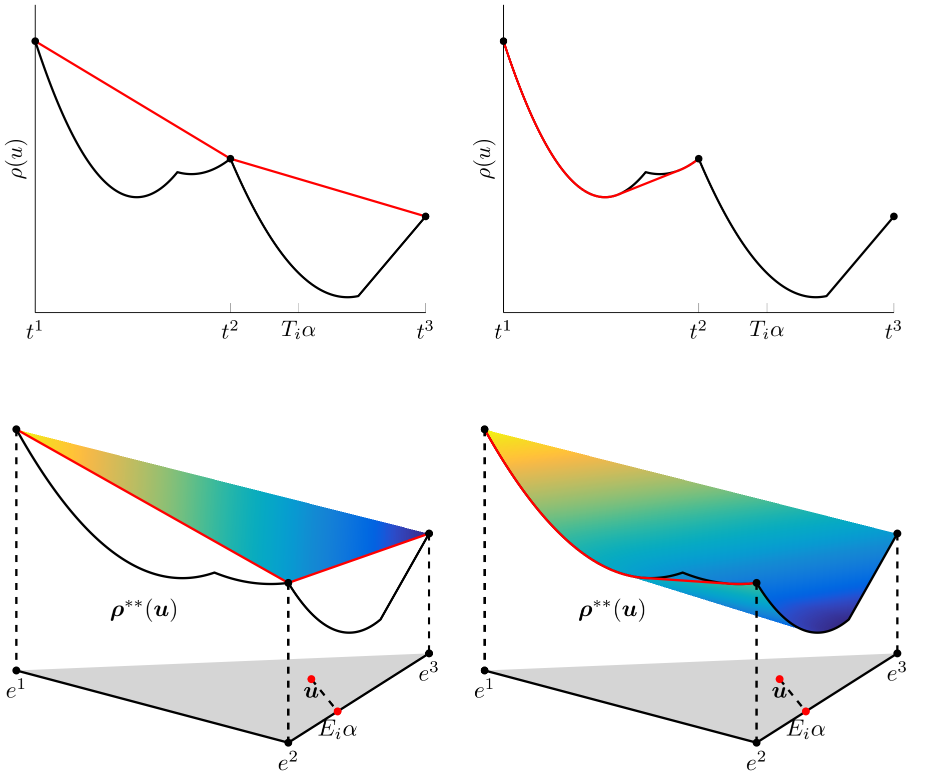
$$\boldsymbol{\rho}(\boldsymbol{u}) = \min_{1 \le i \le |\mathcal{T}|} \boldsymbol{\rho}_i(\boldsymbol{u}), \qquad \boldsymbol{\rho}_i(\boldsymbol{u}) = \begin{cases} \rho(T_i \alpha), & \text{if } \boldsymbol{u} = E_i \alpha,\ \alpha \in \Delta_n^U, \\ \infty, & \text{otherwise.} \end{cases} \tag{4.7}$$

Problems (4.6) and (4.7) are equivalent due to the one-to-one correspondence of $u = T_i \alpha$ and $\boldsymbol{u} = E_i \alpha$. However, energy (4.7) is finite on a nonconvex set only. In order to make optimization tractable, we minimize its convex envelope.

Proposition 6. *The convex envelope of* (4.7) *is given as:*

$$\begin{aligned} \boldsymbol{\rho}^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{v} \in \mathbf{R}^{|\mathcal{V}|}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \max_{1 \le i \le |\mathcal{T}|} \boldsymbol{\rho}_i^*(\boldsymbol{v}), \\ \boldsymbol{\rho}_i^*(\boldsymbol{v}) &= \langle E_i b_i, \boldsymbol{v} \rangle + \rho_i^*(A_i^\top E_i^\top \boldsymbol{v}), \quad \rho_i := \rho + \delta_{\Delta_i}. \end{aligned} \tag{4.8}$$

b_i and A_i are given as $b_i := M_i^{n+1}$, $A_i := \left(M_i^1,\ M_i^2,\ \dots,\ M_i^n\right)$, where M_i^j are the columns of the matrix $M_i := (T_i^\top, \mathbf{1})^{-\top} \in \mathbf{R}^{n+1 \times n+1}$.



Standard lifting [Lel+13a]

Proposed lifting

Figure 4.3: Geometrical intuition for the proposed lifting and standard lifting [Lel+13a] for the special case of 1-dimensional range $\Gamma = [a, b]$ and 3 labels $\{t^1, t^2, t^3\}$. The standard lifting correponds to a linear interpolation of the original cost in between the locations t^1, t^2, t^3, which are associated with the vertices e^1, e^2, e^3 in the lifted energy (lower left). The proposed method extends the cost to the relaxed set in a more precise way: The original cost is preserved on the connecting lines between adjacent e^i (black lines on the bottom right) up to concave parts (red graphs and lower surface on the right). This information, which may influence the exact location of the minimizer, is lost in the standard formulation. If the solution of the lifted formulation $\boldsymbol{u}$ is in the interior (gray area) an approximate solution to the original problem can still be obtained via Eq. (4.5).

Proof. Follows from a calculation starting at the definition of $\boldsymbol{\rho}^{**}$. See Appendix A.2 for a detailed derivation. □

The geometric intuition of this construction is depicted in Fig. 4.3. Note that if one prescribes the value of $\boldsymbol{\rho}_i$ in (4.7) only on the *vertices* of the unit simplices Δ_n^U, i.e., $\boldsymbol{\rho}(\boldsymbol{u}) = \rho(t^k)$ if $\boldsymbol{u} = e^k$ and $+\infty$ otherwise, one obtains the linear biconjugate $\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{s} \rangle$, $\boldsymbol{s} = (\rho(t^i), \ldots, \rho(t^L))$ on the feasible set. This coincides with the standard relaxation of the dataterm used in [Poc+10; LS11; CCP12; Lel+13a]. In that sense, our approach can be seen as a relaxing the dataterm in a more precise way, by incorporating the true value of ρ not only on the finite set of labels $\mathcal{V}$, but also everywhere in between, i.e., on every *sublabel.*

4.2.3 Lifting the vectorial total variation

We define the lifted vectorial total variation as

$$\boldsymbol{TV}(\boldsymbol{u}) = \int_{\Omega} \boldsymbol{\Psi}(D\boldsymbol{u}), \tag{4.9}$$

where $D\boldsymbol{u}$ denotes the distributional derivative of $\boldsymbol{u}$ and $\boldsymbol{\Psi}$ is positively one-homogeneous, i.e., $\boldsymbol{\Psi}(c\boldsymbol{u}) = c\,\boldsymbol{\Psi}(\boldsymbol{u}), c \geqslant 0$. For such functions, the meaning of (4.9) can be made fully precise using the polar decomposition of the Radon measure $D\boldsymbol{u}$ [AFP00, Cor. 1.29, Thm. 2.38]. However, in the following we restrict ourselves to an intuitive motivation for the derivation of $\boldsymbol{\Psi}$ for smooth functions.

Our goal is to find $\boldsymbol{\Psi}$ so that $\boldsymbol{TV}(\boldsymbol{u}) = \mathrm{TV}(u)$ whenever $\boldsymbol{u} : \Omega \to \mathbf{R}^{|\mathcal{V}|}$ corresponds to some $u : \Omega \to \Gamma$, in the sense that $\boldsymbol{u}(x) = E_i\alpha$ whenever $u(x) = T_i\alpha$. In order for the equality to hold, it must in particular hold for all u that are classically differentiable, i.e., $Du = \nabla u$, and whose Jacobian $\nabla u(x)$ is of rank 1, i.e., $\nabla u(x) = (T_i\alpha - T_j\beta) \otimes v(x)$ for some $v(x) \in \mathbb{R}^d$. This rank 1 constraint enforces the different components of u to have the same jump normal, which is desirable in many applications. In that case, we observe

$$TV(u) = \int_{\Omega} \|T_i\alpha - T_j\beta\| \cdot \|v(x)\| \,\mathrm{d}x. \tag{4.10}$$

For the corresponding lifted representation $\boldsymbol{u}$, we have $\nabla\boldsymbol{u}(x) = (E_i\alpha - E_j\beta) \otimes v(x)$. Therefore it is natural to require $\boldsymbol{\Psi}(\nabla\boldsymbol{u}(x)) = \boldsymbol{\Psi}\left((E_i\alpha - E_j\beta) \otimes v(x)\right) := \|T_i\alpha - T_j\beta\| \cdot \|v(x)\|$ in order to achieve the goal $\boldsymbol{TV}(\boldsymbol{u}) = \mathrm{TV}(u)$. Motivated by these observations, we define

$$\boldsymbol{\Psi}(\boldsymbol{p}) := \begin{cases} \|T_i\alpha - T_j\beta\| \cdot \|v\| & \text{if } \boldsymbol{p} = (E_i\alpha - E_j\beta) \otimes v, \\ \infty & \text{otherwise,} \end{cases} \tag{4.11}$$

where $\alpha, \beta \in \Delta^U_{n+1}$, $v \in \mathbf{R}^d$ and $1 \leq i, j \leq |\mathcal{T}|$. Since the convex envelope of (4.9) is intractable, we derive a "locally" tight convex underapproximation:

$$\boldsymbol{R}(\boldsymbol{u}) = \sup_{\boldsymbol{q}:\Omega \to \mathbf{R}^{d\times|\mathcal{V}|}} \int_{\Omega} \langle \boldsymbol{u}, \operatorname{div} \boldsymbol{q}\rangle - \boldsymbol{\Psi}^*(\boldsymbol{q}) \,\mathrm{d}x. \tag{4.12}$$

Proposition 7. *The convex conjugate of* $\boldsymbol{\Psi}$ *is*

$$\boldsymbol{\Psi}^*(\boldsymbol{q}) = \delta_{\mathcal{K}}(\boldsymbol{q}) \tag{4.13}$$

with convex set

$$\mathcal{K} = \bigcap_{1 \leq i,j \leq |\mathcal{T}|} \left\{ \boldsymbol{q} \in \mathbf{R}^{d\times|\mathcal{V}|} \;\middle|\; \|Q_i\alpha - Q_j\beta\| \leq \|T_i\alpha - T_j\beta\|,\ \alpha, \beta \in \Delta^U_{n+1} \right\}, \tag{4.14}$$

and $Q_i = (\boldsymbol{q}^{i_1}, \boldsymbol{q}^{i_2}, \ldots, \boldsymbol{q}^{i_{n+1}}) \in \mathbf{R}^{d\times n+1}$. $\boldsymbol{q}^j \in \mathbb{R}^d$ *are the columns of* $\boldsymbol{q}$.

Proof. Follows from a calculation starting at the definition of the convex conjugate $\boldsymbol{\Psi}^*$. See Appendix A.2. □

Interestingly, although in its original formulation (4.14) the set $\mathcal{K}$ has infinitely many constraints, one can equivalently represent $\mathcal{K}$ by finitely many.

Proposition 8. *The set $\mathcal{K}$ in equation* (4.14) *is the same as*

$$\mathcal{K} = \left\{ \boldsymbol{q} \in \mathbf{R}^{d\times|\mathcal{V}|} \mid \left\| D^i_{\boldsymbol{q}} \right\|_{S^\infty} \le 1,\ 1 \le i \le |\mathcal{T}| \right\},\ D^i_{\boldsymbol{q}} = Q_i D \left(T_i D \right)^{-1}, \tag{4.15}$$

where the matrices $Q_i D \in \mathbb{R}^{d\times n}$ and $T_i D \in \mathbb{R}^{n\times n}$ are given as

$$Q_i D := \left(\boldsymbol{q}^{i_1} - \boldsymbol{q}^{i_{n+1}},\ \ldots,\ \boldsymbol{q}^{i_n} - \boldsymbol{q}^{i_{n+1}} \right),\ T_i D := \left(t^{i_1} - t^{i_{n+1}},\ \ldots,\ t^{i_n} - t^{i_{n+1}} \right).$$

Proof. Similar to the analysis in [Lel+13a], equation (4.14) basically states the Lipschitz continuity of a piecewise linear function defined by the matrices $\boldsymbol{q} \in \mathbf{R}^{d\times|\mathcal{V}|}$. Therefore, one can expect that the Lipschitz constraint is equivalent to a bound on the derivative. For the complete proof, see Appendix A.2. □

4.2.4 Lifting the overall optimization problem

Combining dataterm and regularizer, the overall optimization problem is given

$$\min_{\boldsymbol{u}:\Omega\to\mathbf{R}^{|\mathcal{V}|}} \sup_{\boldsymbol{q}:\Omega\to\mathcal{K}} \int_\Omega \boldsymbol{\rho}^{**}(\boldsymbol{u}) + \langle \boldsymbol{u}, \operatorname{div} \boldsymbol{q} \rangle \, \mathrm{d}x. \tag{4.16}$$

A highly desirable property is that, opposed to any other vectorial lifting approach from the literature, our method with just one simplex applied to a convex problem yields the same solution as the unlifted problem.

Proposition 9. *If the triangulation contains only 1 simplex, $\mathcal{T} = \{\Delta\}$, i.e., $|\mathcal{V}| = n+1$, then the proposed optimization problem* (4.16) *is equivalent to*

$$\min_{u:\Omega\to\Delta} \int_\Omega (\rho + \delta_\Delta)^{**}(x, u(x)) \, \mathrm{d}x + \lambda TV(u), \tag{4.17}$$

which is (4.1) *with a globally convexified dataterm on* Δ.

Proof. For $u = t^{n+1} + TD\tilde{u}$ the substitution $\boldsymbol{u} = \left(\tilde{u}_1, \cdots, \tilde{u}_n, 1 - \sum_{j=1}^n \tilde{u}_j \right)$ into $\boldsymbol{\rho}^{**}$ and $\boldsymbol{R}$ yields the result. For a complete proof, see see Appendix A.2. □

4.3 Numerical Optimization

4.3.1 Discretization

For now assume that $\Omega \subset \mathbf{R}^d$ is a d-dimensional Cartesian grid and let div denote a finite-difference divergence operator with $\operatorname{div} \boldsymbol{q} : \Omega \to \mathbb{R}^{|\mathcal{V}|}$. Then the relaxed energy minimization problem becomes

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^{|\mathcal{V}|}} \max_{\boldsymbol{q}:\Omega\to\mathcal{K}} \sum_{x\in\Omega} \boldsymbol{\rho}^{**}(x, \boldsymbol{u}(x)) + \langle \operatorname{div} \boldsymbol{q}, \boldsymbol{u} \rangle. \tag{4.18}$$

In order to get rid of the pointwise maximum over $\boldsymbol{\rho}_i^*(\boldsymbol{v})$ in Eq. (4.8), we introduce additional variables $w(x) \in \mathbb{R}$ and additional constraints $(\boldsymbol{v}(x), w(x)) \in \mathcal{C}$, $x \in \Omega$ so that $w(x)$ attains the value of the pointwise maximum:

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^{|\mathcal{V}|}} \max_{\substack{(\boldsymbol{v},w):\Omega\to\mathcal{C} \\ \boldsymbol{q}:\Omega\to\mathcal{K}}} \sum_{x\in\Omega} \langle \boldsymbol{u}(x), \boldsymbol{v}(x)\rangle - w(x) + \langle \operatorname{div} \boldsymbol{q}, \boldsymbol{u}\rangle, \tag{4.19}$$

where the set $\mathcal{C}$ is given as

$$\mathcal{C} = \bigcap_{1\le i\le|\mathcal{T}|} \mathcal{C}_i, \quad \mathcal{C}_i := \left\{(x,y) \in \mathbf{R}^{|\mathcal{V}|+1} \mid \boldsymbol{\rho}_i^*(x) \le y\right\}. \tag{4.20}$$

For numerical optimization we use a GPU-based implementation[1] of a first-order primal-dual method [Poc+09a]. The algorithm requires the orthogonal projections of the dual variables onto the sets $\mathcal{C}$ respectively $\mathcal{K}$ in every iteration. However, the projection onto an epigraph of dimension $|\mathcal{V}|+1$ is difficult for large values of $|\mathcal{V}|$. We rewrite the constraints $(\boldsymbol{v}(x), w(x)) \in \mathcal{C}_i$, $1 \le i \le |\mathcal{T}|$, $x \in \Omega$ as $(n+1)$-dimensional epigraph constraints introducing variables $r^i(x) \in \mathbf{R}^n$, $s_i(x) \in \mathbf{R}$:

$$\rho_i^*\left(r^i(x)\right) \le s_i(x), \quad r^i(x) = A_i^\top E_i^\top \boldsymbol{v}(x), \quad s_i(x) = w(x) - \langle E_i b_i, \boldsymbol{v}(x)\rangle. \tag{4.21}$$

These equality constraints can be implemented using Lagrange multipliers. For the projection onto the set $\mathcal{K}$ we use an approach similar to [GSC12, Figure 7].

4.3.2 Epigraphical projections

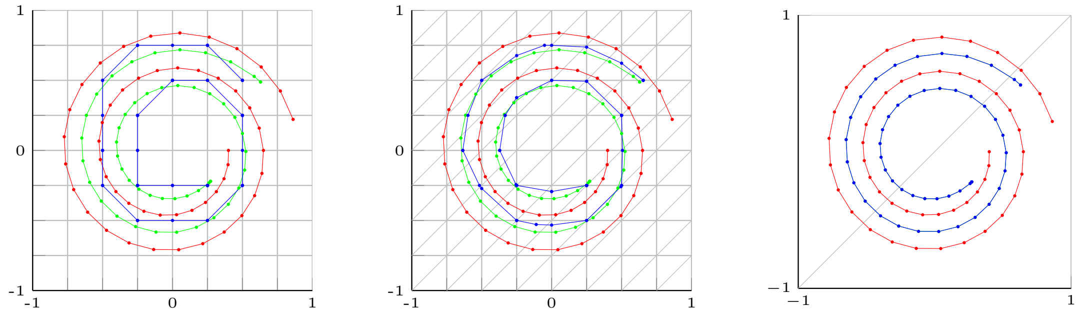
Computing the Euclidean projection onto the epigraph of ρ_i^* is a central part of the numerical implementation of the presented method. However, for $n > 1$ this is nontrivial. Therefore we provide a detailed explanation of the projection methods used for different classes of ρ_i. We will consider quadratic, truncated quadratic and piecewise linear ρ.

Quadratic case. Let ρ be of the form $\rho(u) = \frac{a}{2} u^\top u + b^\top u + c$. A direct projection onto the epigraph of $\rho_i^* = (\rho + \delta_{\Delta_i})^*$ for $n > 1$ is difficult. However, the epigraph can be decomposed into separate epigraphs for which it is easier to project onto: For proper, convex, lsc. functions f, g the epigraph of $(f+g)^*$ is the Minkowski sum of the epigraphs of f^* and g^* (cf. [RWW98, Exercise 1.28, Theorem 11.23a]). This means that it suffices to compute the projections onto the epigraphs of a quadratic function $f^* = \rho^*$ and a convex, piecewise linear function $g^*(v) = \max_{1\le j\le n+1} \langle t^{i_j}, v\rangle$ by rewriting constraint (4.21) as

$$\rho^*(r_f) \le s_f, \; {\delta_{\Delta_i}}^*(c_g) \le d_g \;\text{ s.t. } (r,s) = (r_f, s_f) + (c_g, d_g). \tag{4.22}$$

For the projection onto the epigraph of a n-dimensional quadratic function we use the method described in [SCC14, Appendix B.2]. The projection onto a piecewise linear function is described in the last paragraph of this section.

[1] `https://github.com/tum-vision/sublabel_relax`



Naive, 81 labels.   [Lel+13a], 81 labels.   Ours, **4 labels**.

Figure 4.4: ROF denoising of a vector-valued signal $f : [0,1] \to [-1,1]^2$, discretized on 50 points (shown in red). We compare the proposed approach (right) with two alternative techniques introduced in [Lel+13a] (left and middle). The labels are visualized by the gray grid. While the naive (standard) multilabel approach from [Lel+13a] (left) provides solutions that are constrained to the chosen set of labels, the sublabel accurate regularizer from [Lel+13a] (middle) does allow sublabel solutions, yet – due to the dataterm bias – these still exhibit a strong preference for the grid points. In contrast, the proposed approach does not exhibit any visible grid bias providing fully sublabel-accurate solutions: With only 4 labels, the computed solutions (shown in blue) coincide with the "unlifted" problem (green).

Truncated quadratic case. Let ρ be of the form $\rho(u) = \min\{\, \nu, \, \frac{a}{2} u^\top u + b^\top u + c \,\}$ as it is the case for the nonconvex robust ROF with a truncated quadratic dataterm in Sec. 4.4.2. Again, a direct projection onto the epigraph of ρ_i^* is difficult. However, a decomposition of the epigraph into simpler epigraphs is possible as the epigraph of $\min\{f, g\}^*$ is the intersection of the epigraphs of f^* and g^*. Hence, one can separately project onto the epigraphs of $(\nu + \delta_{\Delta_i})^*$ and $(\frac{a}{2} u^\top u + b^\top u + c + \delta_{\Delta_i})^*$. Both of these projections can be handled using the methods from the other paragraphs.

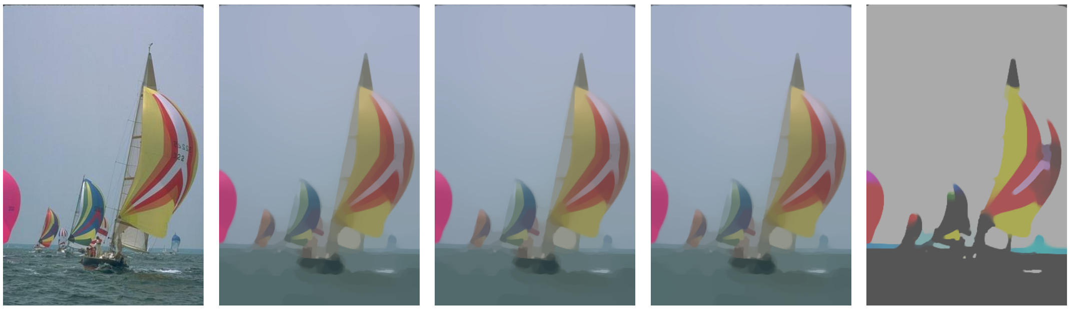
Piecewise linear case. In case ρ is piecewise linear on each Δ_i, i.e., ρ attains finite values at a discrete set of sampled sublabels $\mathcal{V}_i \subset \Delta_i$ and interpolates linearly between them, we have that

$$(\rho + \delta_{\Delta_i})^*(v) = \max_{\tau \in \mathcal{V}_i} \langle \tau, v \rangle - \rho(\tau). \tag{4.23}$$

Again this is a convex, piecewise linear function. For the projection onto the epigraph of such a function, a quadratic program of the form

$$\min_{(x,y) \in \mathbf{R}^{n+1}} \frac{1}{2}\|x - c\|^2 + \frac{1}{2}\|y - d\|^2 \;\text{ s.t. } \langle \tau, x \rangle - \rho(\tau) \le y, \forall \tau \in \mathcal{V}_i \tag{4.24}$$

needs to be solved.



Input image

Unlifted Problem, $E = 992.50$

Ours, $|\mathcal{T}| = 1$, $|\mathcal{V}| = 4$, $E = 992.51$

Ours, $|\mathcal{T}| = 6$ $|\mathcal{V}| = 2 \times 2 \times 2$ $E = 993.52$

Baseline, $|\mathcal{V}| = 4 \times 4 \times 4$, $E = 2255.81$

Figure 4.5: Convex ROF with vectorial TV. Direct optimization and proposed method yield the same result. In contrast to the baseline method [Lel+13a] the proposed approach has no discretization artefacts and yields a lower energy. The regularization parameter is chosen as $\lambda = 0.3$.

We implemented the primal active-set method described in [NW06, Algorithm 16.3], and found it solves the program in a few (usually $2 - 10$) iterations for a moderate number of constraints.

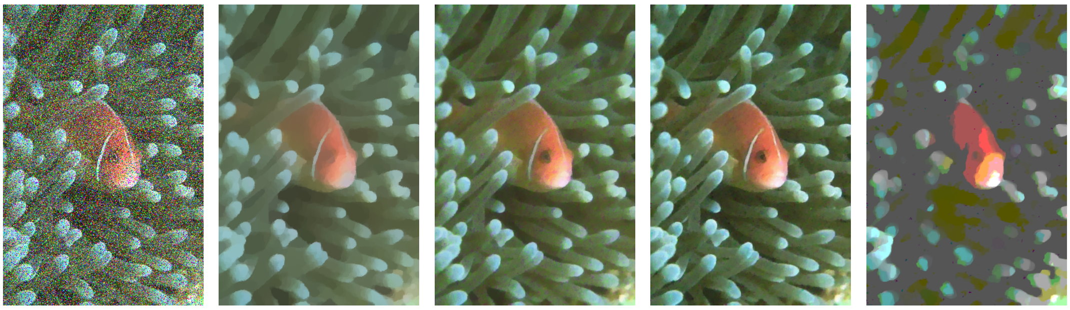
4.4 Experiments

4.4.1 Vectorial ROF denoising

In order to validate experimentally, that our model is exact for convex dataterms, we evaluate it on the Rudin-Osher-Fatemi [ROF92] (ROF) model with vectorial TV (4.2). In our model this corresponds to defining $\rho(x, u(x)) = \frac{1}{2}\|u(x) - I(x)\|^2$. As expected based on Prop. 9 the energy of the solution of the unlifted problem is equal to the energy of the projected solution of our method for $|\mathcal{V}| = 4$ up to machine precision, as can be seen in Fig. 4.4 and Fig. 4.5. We point out, that the sole purpose of this experiment is a proof of concept as our method introduces an overhead and convex problems can be solved via direct optimization. It can be seen in Fig. 4.4 and Fig. 4.5, that the baseline method [Lel+13a] has a strong label bias.

4.4.2 Denoising with truncated quadratic dataterm

For images degraded with both, Gaussian and salt-and-pepper noise we define the dataterm as $\rho(x, u(x)) = \min\left\{\frac{1}{2}\|u(x) - I(x)\|^2, \nu\right\}$. We solve the problem using the epigraph decomposition described in the second paragraph of Sec. 4.3.2. It can be seen, that increasing the number of labels $|\mathcal{V}|$ leads to lower energies and at



Noisy input

Ours, $|\mathcal{T}| = 1$, $|\mathcal{V}| = 4$, $E = 2849.52$

Ours, $|\mathcal{T}| = 6$, $|\mathcal{V}| = 2 \times 2 \times 2$, $E = 2806.18$

Ours, $|\mathcal{T}| = 48$, $|\mathcal{V}| = 3 \times 3 \times 3$, $E = 2633.83$

Baseline, $|\mathcal{V}| = 4 \times 4 \times 4$, $E = 3151.80$

Figure 4.6: ROF with a truncated quadratic dataterm ($\lambda = 0.03$ and $\nu = 0.025$). Compared to the baseline method [Lel+13a] the proposed approach yields much better results, already with a very small number of 4 labels.

the same time to a reduced effect of the TV. This occurs as we always compute a piecewise convex underapproximation of the original nonconvex dataterm, that gets tighter the larger the number of labels. The baseline method [Lel+13a] again produces strong discretization artefacts even for a large number of labels $|\mathcal{V}| = 4 \times 4 \times 4 = 64$.

4.4.3 Optical flow

We compute the optical flow $v : \Omega \to \mathbf{R}^2$ between two input images I_1, I_2. The label space $\Gamma = [-d, d]^2$ is chosen according to the estimated maximum displacement $d \in \mathbf{R}$ between the images. The dataterm is $\rho(x, v(x)) = \|I_2(x) - I_1(x + v(x))\|$, and $\lambda(x)$ is based on the norm of the image gradient $\nabla I_1(x)$.

In Fig. 4.7 we compare the proposed method to the product space approach [GSC13]. Note that we implemented the product space dataterm using Lagrange multipliers, also referred to as the *global* approach in [GSC13]. While this increases the memory consumption, it comes with lower computation time and guaranteed convergence. For our method, we sample the label space $\Gamma = [-15, 15]^2$ on 150×150 sublabels and subsequently convexify the energy on each triangle using the quickhull algorithm [BDH96]. For the product space approach we sample the label space at equidistant labels, from 5×5 to 27×27. As the regularizer from the product space approach is different from the proposed one, we chose μ differently for each method. For the proposed method, we set $\mu = 0.5$ and for the product space and baseline approach $\mu = 3$. We can see in Fig. 4.7, our method outperforms the product space approach w.r.t. the average end-point error. Our method outperforms previous lifting approaches: In Fig. 4.8 we compare our method on large displacement optical flow to the baseline [Lel+13a]. To obtain competitive results on the Middlebury benchmark, one would need to engineer a better dataterm.

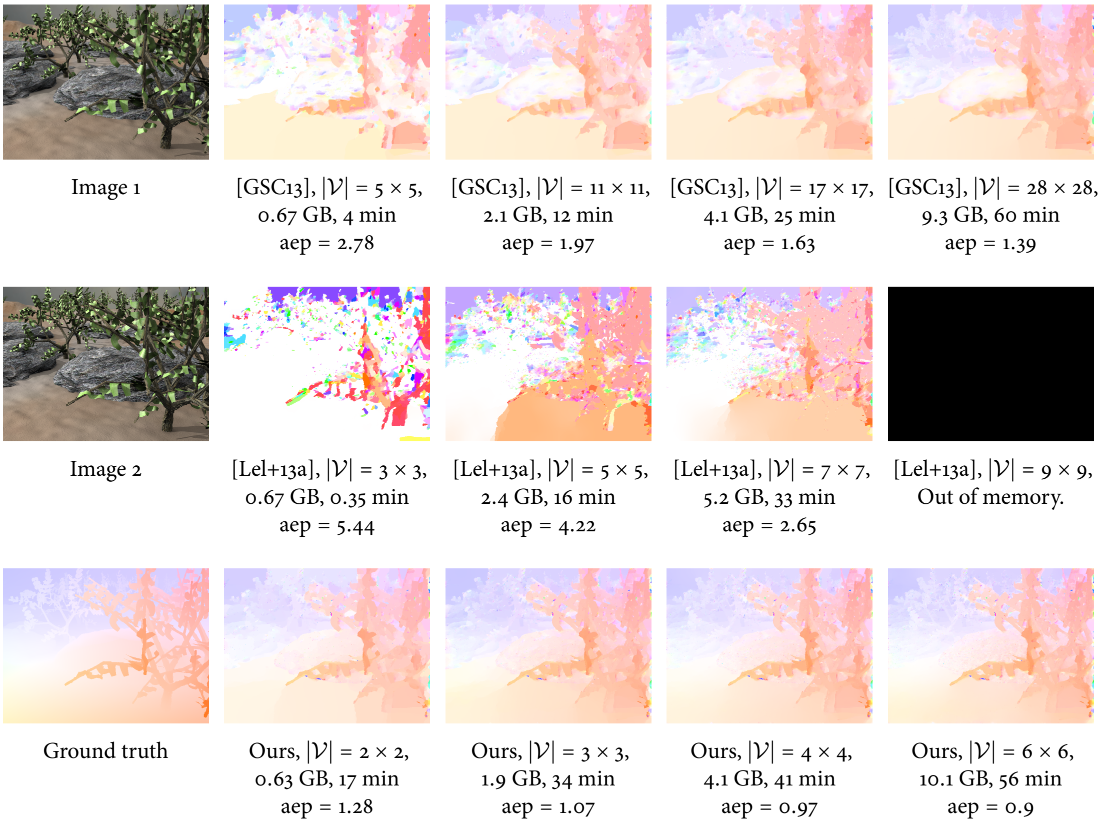


Image 1

[GSC13], $|\mathcal{V}| = 5 \times 5$, 0.67 GB, 4 min aep = 2.78

[GSC13], $|\mathcal{V}| = 11 \times 11$, 2.1 GB, 12 min aep = 1.97

[GSC13], $|\mathcal{V}| = 17 \times 17$, 4.1 GB, 25 min aep = 1.63

[GSC13], $|\mathcal{V}| = 28 \times 28$, 9.3 GB, 60 min aep = 1.39

Image 2

[Lel+13a], $|\mathcal{V}| = 3 \times 3$, 0.67 GB, 0.35 min aep = 5.44

[Lel+13a], $|\mathcal{V}| = 5 \times 5$, 2.4 GB, 16 min aep = 4.22

[Lel+13a], $|\mathcal{V}| = 7 \times 7$, 5.2 GB, 33 min aep = 2.65

[Lel+13a], $|\mathcal{V}| = 9 \times 9$, Out of memory.

Ground truth

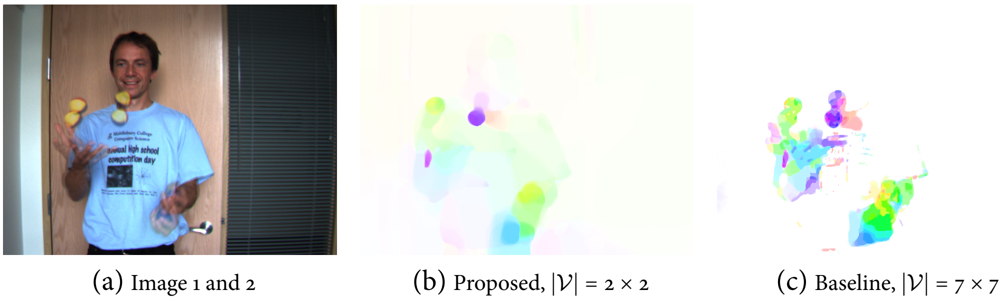
Ours, $|\mathcal{V}| = 2 \times 2$, 0.63 GB, 17 min aep = 1.28

Ours, $|\mathcal{V}| = 3 \times 3$, 1.9 GB, 34 min aep = 1.07

Ours, $|\mathcal{V}| = 4 \times 4$, 4.1 GB, 41 min aep = 0.97

Ours, $|\mathcal{V}| = 6 \times 6$, 10.1 GB, 56 min aep = 0.9

Figure 4.7: We compute the optical flow using our method, the product space approach [GSC13] and the baseline method [Lel+13a] for a varying amount of labels and compare the average endpoint error (aep). The product space method clearly outperforms the baseline, but our approach finds the overall best result already with 2 × 2 labels. To achieve a similarly precise result as the product space method, we require 150 times fewer labels, 10 times less memory and 3 times less time. For the same number of labels, the proposed approach requires more memory as it has to store a convex approximation of the energy instead of a linear one.



(a) Image 1 and 2

(b) Proposed, $|\mathcal{V}| = 2 \times 2$

(c) Baseline, $|\mathcal{V}| = 7 \times 7$

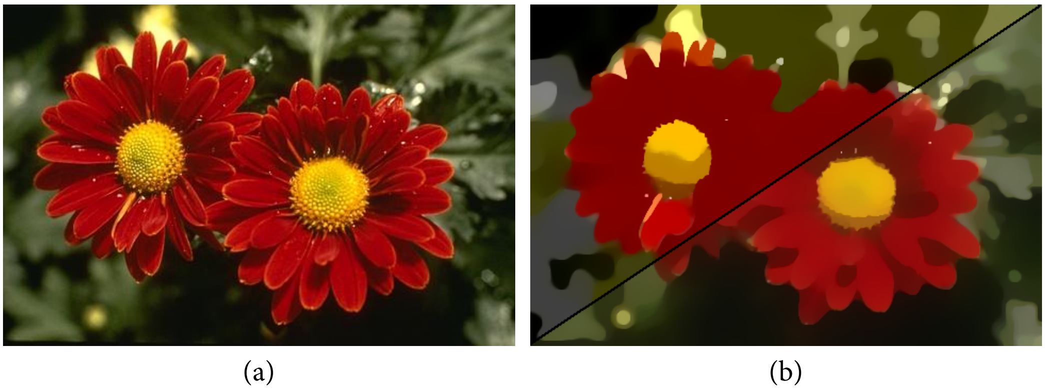
Figure 4.8: Large displacement flow between two 640 × 480 images (a) using a 81 × 81 search window. The result of our method with 4 labels is shown in (b), the baseline [Lel+13a] in (c).

4.5 Conclusions

We proposed the first sublabel-accurate convex relaxation of vectorial multilabel problems. To this end, we approximate the generally nonconvex dataterm in a piecewise convex manner as opposed to the piecewise linear approximation done in the traditional functional lifting approaches. This assures a more faithful approximation of the original cost function and provides a meaningful interpretation for the non-integral solutions of the relaxed convex problem. In experimental validations on large-displacement optical flow estimation and color image denoising, we show that the computed solutions have superior quality to the traditional convex relaxation methods while requiring substantially less memory and runtime.

Chapter 5

Sublabel-Accurate Discretization of Nonconvex Free-Discontinuity Problems


(a)

(b)

Figure 5.1: The classical way to discretize continuous convex relaxations such as the vectorial Mumford-Shah functional [SCC12] leads to solutions (**b**), top-left) with a strong bias towards the chosen labels (here an equidistant $5 \times 5 \times 5$ sampling of the RGB space). This can be seen in the bottom left part of the image, where the green color is truncated to the nearest label which is gray. The proposed sublabel-accurate approximation of the continuous relaxation leads to bias-free solutions (**b**), bottom-right).

5.1 Introduction

5.1.1 A class of continuous optimization problems

Many tasks particularly in low-level computer vision can be formulated as optimization problems over mappings $u : \Omega \to \Gamma$ between sets Ω and Γ. The energy functional is usually designed in such a way that the minimizing argument corresponds to a mapping with the desired solution properties. In classical discrete Markov random field (MRF) approaches, which we refer to as *fully discrete optimization*, Ω is typically a set of nodes (e.g., pixels or superpixels) and Γ a set of labels $\{1, \cdots, \ell\}$.

However, in many problems such as image denoising, stereo matching or optical flow where $\Gamma \subset \mathbf{R}^d$ is naturally modeled as a continuum, this discretization into *labels* can entail unreasonably high demands in memory when using a fine sampling, or it leads to a strong label bias when using a coarser sampling, see Figure 5.1. Furthermore, as jump discontinuities are ubiquitous in low-level vision (e.g., caused by object edges, occlusion boundaries, changes in albedo, shadows, etc.), it is important to model them in a meaningful manner. By restricting either Ω or Γ to a discrete set, one loses the ability to mathematically distinguish between continuous and discontinuous mappings.

Motivated by these two points we consider *fully-continuous* optimization approaches, where the idea is to postpone the discretization of $\Omega \subset \mathbf{R}^n$ and $\Gamma \subset \mathbf{R}$ as long as possible. The prototypical class of continuous optimization problems which we consider in this work are nonconvex free-discontinuity problems, inspired by the celebrated Mumford-Shah functional [BZ87; MS89]:

$$\begin{aligned} E(u) = &\int_{\Omega \setminus J_u} f\left(x, u(x), \nabla u(x)\right) \mathrm{d}x \\ &+ \int_{J_u} d\left(x, u^-(x), u^+(x), \nu_u(x)\right) \mathrm{d}\mathcal{H}^{n-1}(x). \end{aligned} \tag{5.1}$$

The first integral is defined on the region $\Omega \setminus J_u$ where u is continuous. The integrand $f : \Omega \times \Gamma \times \mathbf{R}^n \to [0, \infty]$ can be thought of as a combined data term and regularizer, where the regularizer can penalize variations in terms of the (weak) gradient ∇u. The second integral is defined on the $(n-1)$-dimensional discontinuity set $J_u \subset \Omega$ and $d : \Omega \times \Gamma \times \Gamma \times \mathcal{S}^{n-1} \to [0, \infty]$ penalizes jumps from u^- to u^+ in unit direction ν_u. The appropriate function space for (5.1) are the *special functions of bounded variation*. These are functions of bounded variation (cf. Section 5.2 for a defintion) whose distributional derivative Du can be decomposed into a continuous part and a jump part in the spirit of (5.1):

$$Du = \nabla u \cdot \mathcal{L}^n + (u^+ - u^-)\, \nu_u \cdot \mathcal{H}^{n-1} \llcorner J_u, \tag{5.2}$$

where $\mathcal{L}^n$ denotes the n-dimensional Lebesgue measure and $\mathcal{H}^{n-1} \llcorner J_u$ the $(n-1)$-dimensional Hausdorff measure restricted to the jump set J_u. For an introduction to functions of bounded variation and the study of existence of minimizers to (5.1) we refer the interested reader to [AFP00].

Note that due to the possible nonconvexity of f in the first two variables a surprisingly large class of low-level vision problems fits the general framework of (5.1). While (5.1) is a difficult nonconvex optimization problem, the state-of-the-art are convex relaxations [ABD03; Bou98; Cha01]. We give an overview of the idea behind the convex reformulation in Section 5.3.

Extensions to the vectorial setting, i.e., $\dim(\Gamma) > 1$, have been studied by Strekalovskiy et al.in various works [GSC13; SCC12; SCC14] and recently using the theory of currents by Windheuser and Cremers [WC16]. The case when Γ is a manifold has been considered by Lellmann et al.[Lel+13a]. These advances have allowed for a wide range of difficult vectorial and joint optimization problems to be solved within a convex framework.

5.1.2 Related work

The first practical implementation of (5.1) was proposed by Pock et al.[Poc+09a], using a simple finite differencing scheme in both Ω and Γ which has remained the standard way to discretize convex relaxations. This leads to a strong label bias (see Figure 5.1b), top-left) *despite* the initially label-continuous formulation.

In the MRF community, a related approach to overcome this label-bias are *discrete-continuous* models (discrete Ω and continuous Γ), pioneered by Zach et al.[Zac13; ZK12]. Most similar to the present work is the approach of Fix and Agarwal [FA14]. They derive the discrete-continuous approaches as a discretization of an infinite dimensional dual linear program. Their approach differs from ours, as we start from a different (nonlinear) infinite-dimensional optimization problem and consider a representation of the dual variables which enforces continuity. The recent work of Bach [Bac19] extends the concept of submodularity from discrete to continuous Γ along with complexity estimates.

There are also *continuous-discrete* models, i.e. the range Γ is discretized into labels but Ω is kept continuous [CCP12; LS11]. Recently, these spatially continuous multilabeling models have been extended to allow for so-called *sublabel accurate* solutions [Lau+16; Möl+16], i.e., solutions which lie between two labels. These are, however, limited to total variation regularization, due to the separate convexification of data term and regularizer. We show in this work that for general regularizers a joint convex relaxation is crucial.

Finally, while not focus of this work, there are of course also *fully-discrete* approaches, among many [Ish03; Sch76; Wer07], which inspired some of the continuous formulations.

5.1.3 Contribution

In this work, we propose an approximation strategy for *fully-continuous* relaxations which retains continuous Γ even after discretization (see Figure 5.1b), bottom-right). We summarize our contributions as:

- We generalize the work [Möl+16] from total variation to general convex and nonconvex regularization.
- We prove (see Prop. 11 and Prop. 13) that different approximations to a convex relaxation of (5.1) give rise to existing relaxations [Poc+09a] and [Möl+16].

We investigate the relationship to discrete-continuous MRFs in Prop. 14.

- On the example of the vectorial Mumford-Shah functional [SCC12] we show that our framework yields also sublabel-accurate formulations of extensions to (5.1).

5.2 Notation and Preliminaries

We denote the Iverson bracket as $[\![\cdot]\!]$. Indicator functions from convex analysis which take on values 0 and ∞ are denoted by $\delta\{\cdot\}$. We denote by f^* the convex conjugate of $f : \mathbf{R}^n \to \overline{\mathbf{R}}$. Let $\Omega \subset \mathbf{R}^n$ be a bounded open set. For a function $u \in L^1(\Omega; \mathbf{R})$ its total variation is defined by

$$TV(u) = \sup \left\{ \int_\Omega u \operatorname{div} \varphi \, \mathrm{d}x : \varphi \in C_c^1(\Omega; \mathbf{R}^n) \right\}. \tag{5.3}$$

The space of functions of bounded variation, i.e., for which $\mathrm{TV}(u) < \infty$ (or equivalently for which the distributional derivative Du is a finite Radon measure) is denoted by $BV(\Omega; \mathbf{R})$ [AFP00]. We write $u \in SBV(\Omega; \mathbf{R})$ for functions $u \in BV(\Omega; \mathbf{R})$ whose distributional derivative admits the decomposition (5.2). For the rest of this work, we will make the following simplifying assumptions:

- The Lagrangian f in (5.1) is separable, i.e.,

  $$f(x, t, g) = \rho(x, t) + \eta(x, g), \tag{5.4}$$

  for possibly nonconvex $\rho : \Omega \times \Gamma \to \mathbf{R}$ and regularizers $\eta : \Omega \times \mathbf{R}^n \to \mathbf{R}$ which are convex in g.

- The jump regularizer in (5.1) is isotropic and induced by a concave function $\kappa : \mathbf{R}_{\geq 0} \to \mathbf{R}$:

  $$d(x, u^-, u^+, \nu_u) = \kappa(|u^- - u^+|) \|\nu_u\|_2, \tag{5.5}$$

  with $\kappa(a) = 0 \Leftrightarrow a = 0$.

- The range $\Gamma = [\gamma_1, \gamma_\ell] \subset \mathbf{R}$ is a compact interval.

5.3 The Convex Relaxation

In [ABD03; Bou98; Cha01] the authors propose a convex relaxation for the problem (5.1). Their basic idea is to reformulate the energy (5.1) in terms of the *complete graph* of u, i.e. lifting the problem to one dimension higher as illustrated in Figure 5.2. The complete graph $G_u \subset \Omega \times \Gamma$ is defined as the (measure-theoretic) boundary of the characteristic function of the subgraph $\mathbf{1}_u : \Omega \times \mathbf{R} \to \{0, 1\}$ given by:

$$\mathbf{1}_u(x, t) = [\![t < u(x)]\!]. \tag{5.6}$$

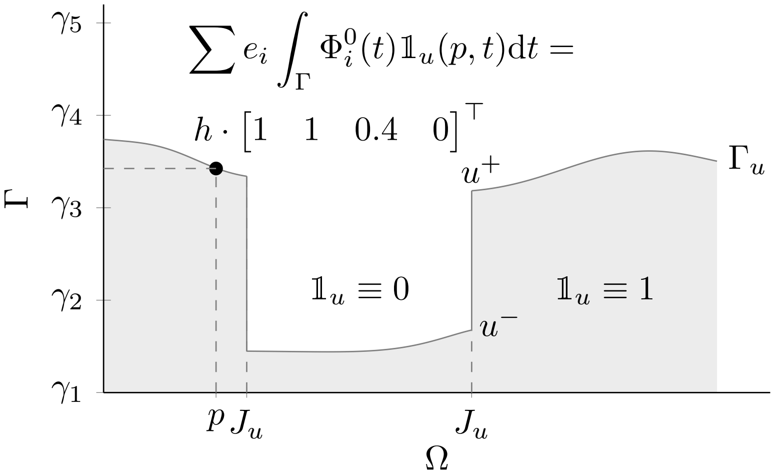


Figure 5.2: The central idea behind the convex relaxation for problem (5.1) is to reformulate the functional in terms of the complete graph $G_u \subset \Omega \times \Gamma$ of $u : \Omega \to \Gamma$ in the product space. This procedure is often referred to as "lifting", as one lifts the dimensionality of the problem.

Furthermore we denote the inner unit normal to $\mathbf{1}_u$ with ν_{G_u}. It is shown in [ABD03] that for $u \in SBV(\Omega;\mathbf{R})$ one has

$$E(u) = F(\mathbf{1}_u) = \sup_{\varphi \in \mathcal{K}} \int_{G_u} \langle \varphi, \nu_{G_u} \rangle \, \mathrm{d}\mathcal{H}^n, \tag{5.7}$$

with constraints on the dual variables $\varphi \in \mathcal{K}$ given by

$$\begin{aligned} \mathcal{K} = \Big\{ &(\varphi_x, \varphi_t) \in C_c^1(\Omega \times \mathbf{R}; \mathbf{R}^n \times \mathbf{R}) : \\ &\varphi_t(x,t) + \rho(x,t) \geq \eta^*(x, \varphi_x(x,t)), && (5.8) \\ &\Big\| \int_t^{t'} \varphi_x(x,t) \mathrm{d}t \Big\|_2 \leq \kappa(|t-t'|), \forall t, t', \forall x \Big\}. && (5.9) \end{aligned}$$

The functional (5.7) can be interpreted as the maximum flux of admissible vector fields $\varphi \in \mathcal{K}$ through the cut given by the complete graph G_u. The set $\mathcal{K}$ can be seen as capacity constraints on the flux field φ. This is reminiscent to constructions from the discrete optimization community [Ish03]. The constraints (5.8) correspond to the first integral in (5.1) and the non-local constraints (5.9) to the jump penalization.

Using the fact that the distributional derivative of the subgraph indicator function $\mathbf{1}_u$ can be written as

$$D\mathbf{1}_u = \nu_{G_u} \cdot \mathcal{H}^m \llcorner G_u, \tag{5.10}$$

one can rewrite the energy (5.7) as

$$F(\mathbf{1}_u) = \sup_{\varphi \in \mathcal{K}} \int_{\Omega \times \Gamma} \langle \varphi, D\mathbf{1}_u \rangle. \tag{5.11}$$

A convex formulation is then obtained by relaxing the set of admissible primal variables to a convex set:

$$\begin{aligned}\mathcal{C} = \Big\{ & v \in BV_{\text{loc}}(\Omega \times \mathbf{R}; [0,1]) : \\ & v(x,t) = 1 \ \ \forall t \leq \gamma_1, v(x,t) = 0 \ \ \forall t > \gamma_\ell, \\ & v(x,\cdot) \text{ non-increasing} \Big\}.\end{aligned} \tag{5.12}$$

This set can be thought of as the convex hull of the subgraph functions $\mathbf{1}_u$. The final optimization problem is then a convex-concave saddle point problem given by:

$$\inf_{v \in \mathcal{C}} \sup_{\varphi \in \mathcal{K}} \int_{\Omega \times \mathbf{R}} \langle \varphi, Dv \rangle. \tag{5.13}$$

In dimension one ($n = 1$), this convex relaxation is tight [Car16; Cha01]. For $n > 1$ global optimality can be guaranteed by means of a thresholding theorem in case $\kappa \equiv \infty$ [BF15; Poc+10]. If the primal solution $\widehat{v} \in \mathcal{C}$ to (5.13) is binary, the global optimum u^* of (5.1) can be recovered simply by pointwise thresholding $\widehat{u}(x) = \sup\{t : \widehat{v}(x,t) > \frac{1}{2}\}$. If $\widehat{v}$ is not binary, in the general setting it is not clear how to obtain the global optimal solution from the relaxed solution. An a posteriori optimality bound to the global optimum $E(u^*)$ of (5.1) for the thresholded solution $\widehat{u}$ can be computed by:

$$|E(\widehat{u}) - E(u^*)| \leq |F(\mathbf{1}_{\widehat{u}}) - F(\widehat{v})|. \tag{5.14}$$

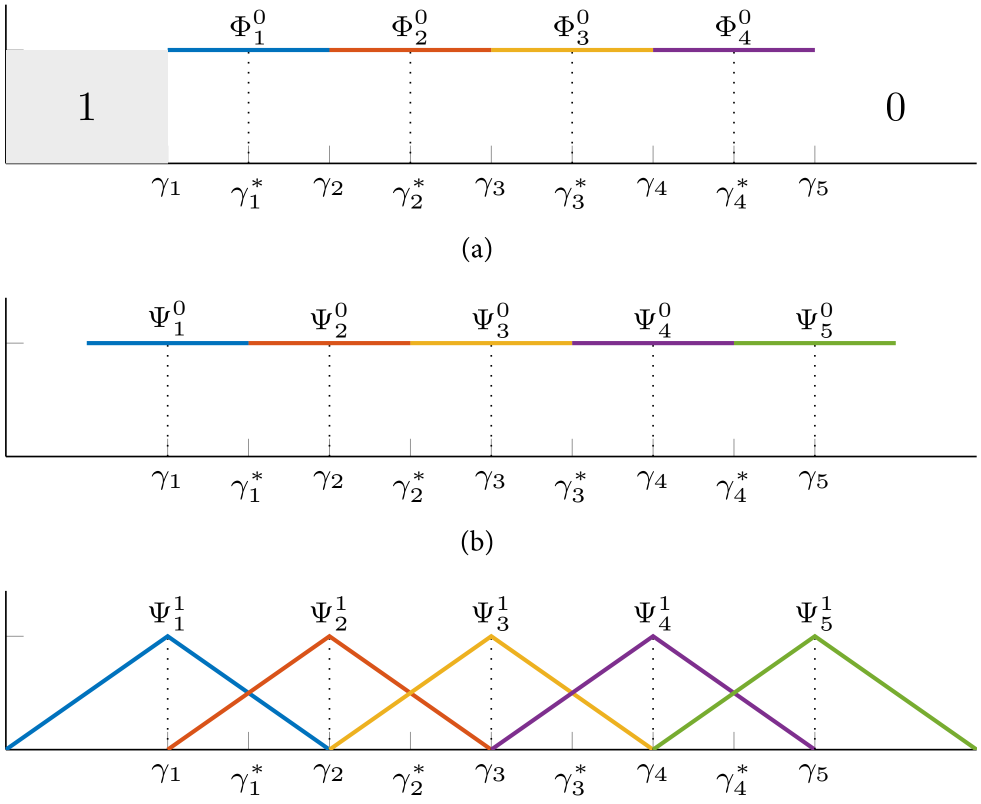
Using that bound, it has been observed that solutions are usually near globally optimal [SCC12]. In the following section, we show how different discretizations of the continuous problem (5.13) lead to various existing lifting approaches and to generalizations of the recent sublabel-accurate continuous multilabeling approach [Möl+16].

5.4 Sublabel-Accurate Discretization

5.4.1 Choice of primal and dual mesh

In order to discretize the relaxation (5.13), we partition the range $\Gamma = [\gamma_1, \gamma_\ell]$ into $k = \ell - 1$ intervals. The individual intervals $\Gamma_i = [\gamma_i, \gamma_{i+1}]$ form a one dimensional *simplicial complex* (see e.g., [Hir03]), and we have $\Gamma = \Gamma_1 \cup \cdots \cup \Gamma_k$. The points $\gamma_i \in \Gamma$ are also referred to as *labels*. We assume that the labels are equidistantly spaced with label distance $h = \gamma_{i+1} - \gamma_i$. The theory generalizes also to non-uniformly spaced labels, as long as the spacing is homogeneous in Ω. Furthermore, we define $\gamma_0 = \gamma_1 - h$ and $\gamma_{\ell+1} = \gamma_\ell + h$.

The mesh for dual variables is given by *dual complex*, which is formed by the intervals $\Gamma_i^* = [\gamma_{i-1}^*, \gamma_i^*]$ with nodes $\gamma_i^* = \frac{\gamma_i + \gamma_{i+1}}{2}$. An overview of the notation and the considered finite dimensional approximations is given in Figure 5.3.


Φ^0_1
Φ^0_2
Φ^0_3
Φ^0_4
1
0
γ_1 γ^*_1 γ_2 γ^*_2 γ_3 γ^*_3 γ_4 γ^*_4 γ_5
(a)
Ψ^0_1
Ψ^0_2
Ψ^0_3
Ψ^0_4
Ψ^0_5
γ_1 γ^*_1 γ_2 γ^*_2 γ_3 γ^*_3 γ_4 γ^*_4 γ_5
(b)
Ψ^1_1
Ψ^1_2
Ψ^1_3
Ψ^1_4
Ψ^1_5
γ_1 γ^*_1 γ_2 γ^*_2 γ_3 γ^*_3 γ_4 γ^*_4 γ_5

(c)

Figure 5.3: Overview of the notation and proposed finite dimensional approximation spaces.

5.4.2 Representation of the primal variable

As $\mathbf{1}_u$ is a discontinuous jump function, we consider a piecewise constant approximation for $v \in \mathcal{C}$,

$$\Phi^0_i(t) = [\![t \in \Gamma_i]\!], \ 1 \le i \le k, \tag{5.15}$$

see Figure 5.3a). Due to the boundary conditions in Eq. (5.12), we set v outside of Γ to 1 on the left and 0 on the right. Note that the non-decreasing constraint in $\mathcal{C}$ is implicitly realized as $\varphi_t \in \mathcal{K}$ can be arbitrarily large.

For coefficients $\hat{v} : \Omega \times \{1, \cdots, k\} \to \mathbf{R}$ we have

$$v(x,t) = \sum_{i=1}^{k} \hat{v}(x,i)\Phi^0_i(t). \tag{5.16}$$

As an example of this representation, consider the approximation of $\mathbf{1}_u$ at point p shown in Figure 5.2:

$$\begin{aligned} \widehat{v}(p,\cdot) &= \sum_{i=1}^{k} e_i \int_\Gamma \Phi^0_i(t)\mathbf{1}_u(p,t)\mathrm{d}t \\ &= h \cdot \begin{bmatrix} 1 & 1 & 0.4 & 0 \end{bmatrix}^\top . \end{aligned} \tag{5.17}$$

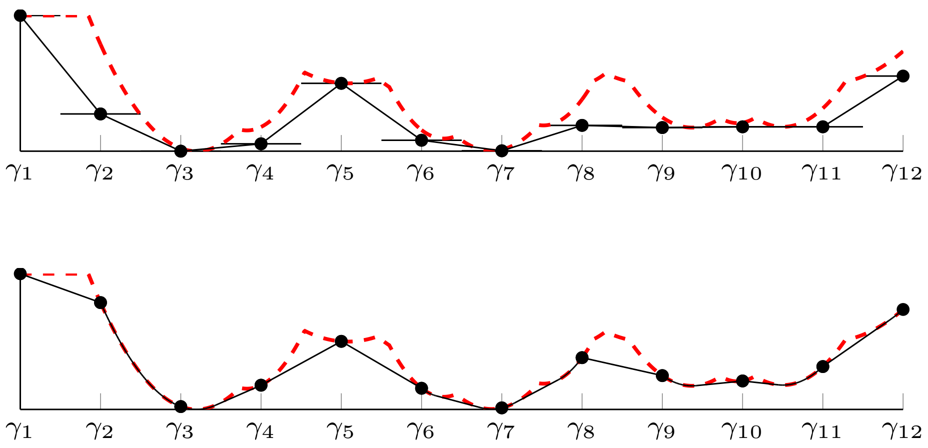


Figure 5.4: **Left:** piecewise constant dual variables φ_t lead to a linear approximation (shown in black) to the original cost function (shown in red). The unaries are determined through min-pooling of the continuous cost in the Voronoi cells around the labels. **Right:** continuous piecewise linear dual variables φ_t convexify the costs on each interval.

This leads to the sublabel-accurate representation also considered in [Möl+16]. In that work, the representation from the above example (5.17) encodes a convex combination between the labels γ_3 and γ_4 with interpolation factor 0.4. Here it is motivated from a different perspective: we take a finite dimensional subspace approximation of the infinite dimensional optimization problem (5.13).

5.4.3 Representation of the dual variables

5.4.3.1 Piecewise constant φ_t

The simplest discretization of the dual variable φ_t is to pick a piecewise constant approximation on the dual intervals Γ_i^* as shown in Figure 5.3b): The functions are given by

$$\Psi_i^0(t) = [\![t \in \Gamma_i^*]\!], \; 1 \leq i \leq \ell, \tag{5.18}$$

As φ is a vector field in C_c^1, the functions Ψ vanish outside of Γ. For coefficient functions $\hat{\varphi}_t : \Omega \times \{1, \cdots, \ell\} \to \mathbf{R}$ and $\hat{\varphi}_x : \Omega \times \{1, \cdots, k\} \to \mathbf{R}^n$ we have:

$$\varphi_t(t) = \sum_{i=1}^{\ell} \hat{\varphi}_t(i) \Psi_i^0(t), \; \varphi_x(t) = \sum_{i=1}^{k} \hat{\varphi}_x(i) \Phi_i^0(t). \tag{5.19}$$

To avoid notational clutter, we dropped $x \in \Omega$ in (5.19) and will do so also in the following derivations. Note that for φ_x we chose the same piecewise constant approximation as for v, as we keep the model continuous in Ω, and ultimately discretize it using finite differences in x.

Discretization of the constraints. In the following, we will plug in the finite dimensional approximations into the constraints from the set $\mathcal{K}$. We start by reformulating the constraints in (5.8). Taking the infimum over $t \in \Gamma_i$ they can be equivalently written as:

$$\inf_{t \in \Gamma_i} \varphi_t(t) + \rho(t) - \eta^* \left(\varphi_x(t)\right) \geq 0, \; 1 \leq i \leq \ell. \tag{5.20}$$

Plugging in the approximation (5.19) into the above leads to the following constraints for $1 \leq i \leq k$:

$$\begin{aligned} \hat{\varphi}_t(i) + \inf_{t \in [\gamma_i, \gamma_i^*]} \rho(t) &\geq \eta^*(\hat{\varphi}_x(i)), \\ \hat{\varphi}_t(i+1) + \underbrace{\inf_{t \in [\gamma_i^*, \gamma_{i+1}]}}_{\text{min-pooling}} \rho(t) &\geq \eta^*(\hat{\varphi}_x(i)). \end{aligned} \tag{5.21}$$

These constraints can be seen as min-pooling of the continuous unary potentials in a symmetric region centered on the label γ_i. To see that more easily, assume one-homogeneous regularization so that $\eta^* \equiv 0$ on its domain. Then two consecutive constraints from (5.21) can be combined into one where the infimum of ρ is taken over $\Gamma_i^* = [\gamma_i^*, \gamma_{i+1}^*]$ centered the label γ_i. This leads to capacity constraints for the flow in vertical direction $-\hat{\varphi}_t(i)$ of the form

$$- \hat{\varphi}_t(i) \leq \inf_{t \in \Gamma_i^*} \rho(t), \; 2 \leq i \leq \ell - 1, \tag{5.22}$$

as well as similar constraints on $\hat{\varphi}_t(1)$ and $\hat{\varphi}_t(\ell)$. The effect of this on a nonconvex energy is shown in Figure 5.4 on the left. The constraints (5.21) are convex inequality constraints, which can be implemented using standard proximal optimization methods and orthogonal projections onto the epigraph $\text{epi}(\eta^*)$ as described in [Poc+10, Section 5.3].

For the second part of the constraint set (5.9) we insert again the finite-dimensional representation (5.19) to arrive at:

$$\begin{aligned} &\left\| (1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l) + \beta \hat{\varphi}_x(j) \right\| \\ &\leq \frac{\kappa(\gamma_j^\beta - \gamma_i^\alpha)}{h}, \; \forall\, 1 \leq i \leq j \leq k, \alpha, \beta \in [0,1], \end{aligned} \tag{5.23}$$

where $\gamma_i^\alpha := (1-\alpha)\gamma_i + \alpha\gamma_{i+1}$. These are infinitely many constraints, but similar to [Möl+16] these can be implemented with finitely many constraints.

Proposition 10. *For concave $\kappa : \mathbf{R}_0^+ \to \mathbf{R}$ with $\kappa(a) = 0 \Leftrightarrow a = 0$, the constraints* (5.23) *are equivalent to*

$$\left\| \sum_{l=i}^{j} \hat{\varphi}_x(l) \right\| \leq \frac{\kappa(\gamma_{j+1} - \gamma_i)}{h}, \; \forall 1 \leq i \leq j \leq k. \tag{5.24}$$

Proof. Proofs are given in Appendix A.3. □

This proposition reveals that only information from the labels γ_i enters into the jump regularizer κ. For $\ell = 2$ we expect all regularizers to behave like the total variation.

Discretization of the energy. For the discretization of the saddle point energy (5.13) we apply the divergence theorem

$$\int_{\Omega\times\mathbf{R}} \langle \varphi, Dv \rangle = \int_{\Omega\times\mathbf{R}} -\operatorname{div}\varphi \cdot v \,\mathrm{d}t\,\mathrm{d}x, \tag{5.25}$$

and then discretize the divergence by inserting the piecewise constant representations of φ_t and v:

$$\int_{\mathbf{R}} -\partial_t \varphi_t(t) v(t)\,\mathrm{d}t = -\hat{\varphi}_t(1) - \sum_{i=1}^{k} \hat{v}(i)\left[\hat{\varphi}_t(i+1) - \hat{\varphi}_t(i)\right]. \tag{5.26}$$

The discretization of the other parts of the divergence are given as the following:

$$\int_{\mathbf{R}} -\partial_{x_j}\varphi_x(t) v(t)\,\mathrm{d}t = -h\sum_{i=1}^{k} \partial_{x_j}\hat{\varphi}_x(i)\hat{v}(i), \tag{5.27}$$

where the spatial derivatives ∂_{x_j} are ultimately discretized using standard finite differences. It turns out that the above discretization can be related to the one from [Poc+09a]:

Proposition 11. *For convex one-homogeneous η the discretization with piecewise constant φ_t and φ_x leads to the traditional discretization as proposed in [Poc+09a], except with min-pooled instead of sampled unaries.*

5.4.3.2 Piecewise linear φ_t

As the dual variables in $\mathcal{K}$ are continuous vector fields, a more faithful approximation is given by a continuous piecewise linear approximation, given for $1 \le i \le \ell$ as:

$$\Psi_i^1(t) = \begin{cases} \frac{t-\gamma_{i-1}}{h}, & \text{if } t \in [\gamma_{i-1}, \gamma_i], \\ \frac{\gamma_{i+1}-t}{h}, & \text{if } t \in [\gamma_i, \gamma_{i+1}], \\ 0 & \text{otherwise.} \end{cases} \tag{5.28}$$

They are shown in Figure 5.3c), and we set:

$$\varphi_t(t) = \sum_{i=1}^{\ell} \hat{\varphi}_t(i)\Psi_i^1(t). \tag{5.29}$$

Note that the piecewise linear dual representation considered by Fix et al.in [FA14] differs in this point, as they do not ensure a continuous representation. Unlike the proposed approach their approximation does not take a true subspace of the original infinite dimensional function space.

Discretization of the constraints. We start from the reformulation (5.20) of the original constraints (5.8). With (5.29) for φ_t and (5.19) for φ_x, we have for $1 \leq i \leq k$:

$$\inf_{t \in \Gamma_i} \hat{\varphi}_t(i) \frac{\gamma_{i+1} - t}{h} + \hat{\varphi}_t(i+1) \frac{t - \gamma_i}{h} + \rho(t) \geq \eta^*(\hat{\varphi}_x(i)). \tag{5.30}$$

While the constraints (5.30) seem difficult to implement, they can be reformulated in a simpler way involving ρ^*.

Proposition 12. *The constraints* (5.30) *can be equivalently reformulated by introducing additional variables $a \in \mathbf{R}^k$, $b \in \mathbf{R}^k$, where $\forall i \in \{1, \cdots, k\}$:*

$$\begin{aligned} & r(i) = (\hat{\varphi}_t(i) - \hat{\varphi}_t(i+1))/h, \\ & a(i) + b(i) - (\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(x, i+1)\gamma_i)/h = 0, \\ & r(i) \geq \rho_i^* \left(a(i)\right), \hat{\varphi}_x(i) \geq \eta^* \left(b(i)\right), \end{aligned} \tag{5.31}$$

with $\rho_i(x,t) = \rho(x,t) + \delta\{t \in \Gamma_i\}$.

The constraints (5.31) are implemented by projections onto the epigraphs of η^* and ρ_i^*, as they can be written as:

$$(r(i), a(i)) \in \text{epi}(\rho_i^*), \; (\hat{\varphi}_x(i), b(i)) \in \text{epi}(\eta^*). \tag{5.32}$$

Epigraphical projections for quadratic and piecewise linear ρ_i are described in [Möl+16]. In Section 5.5.1 we describe how to implement piecewise quadratic ρ_i. As the convex conjugate of ρ_i enters into the constraints, it becomes clear that this discretization only sees the *convexified* unaries on each interval, see also the right part of Figure 5.4.

Discretization of the energy. It turns out that the piecewise linear representation of φ_t leads to the same discrete bilinear saddle point term as (5.26). The other term remains unchanged, as we pick the same representation of φ_x.

Relation to existing approaches. In the following we point out the relationship of the approximation with piecewise linear φ_t to the sublabel-accurate multilabeling approaches [Möl+16] and the discrete-continuous MRFs [ZK12].

Proposition 13. *The discretization with piecewise linear φ_t and piecewise constant φ_x, together with the choice $\eta(g) = \|g\|$ and $\kappa(a) = a$ is equivalent to the relaxation [Möl+16].*

Thus we extend the relaxation proposed in [Möl+16] to more general regularizations. The relaxation [Möl+16] was derived starting from a discrete label space and involved a separate relaxation of data term and regularizer. To see this, first note that the convex conjugate of a convex one-homogeneous function is the indicator function of a convex set [Roc96, Corollary 13.2.1]. Then the constraints (5.8) can be written as

$$-\varphi_t(x,t) \leq \rho(x,t), \tag{5.33}$$

$$\varphi_x(x,t) \in \operatorname{dom}\{\eta^*\}, \tag{5.34}$$

where (5.33) is the data term and (5.34) the regularizer. This provides an intuition why the separate convex relaxation of data term and regularizer in [Möl+16] worked well. However, for general choices of η a joint relaxation of data term and regularizer as in (5.30) is crucial. The next proposition establishes the relationship between the data term from [ZK12] and the one from [Möl+16].

Proposition 14. *The data term from [Möl+16] (which is in turn a special case of the discretization with piecewise linear φ_t) can be (pointwise) brought into the primal form*

$$\mathcal{D}(\widehat{v}) = \inf_{\substack{x_i \geq 0, \sum_i x_i = 1 \\ \widehat{v} = y/h + I^\top x}} \sum_{i=1}^{k} x_i \rho_i^{**}\left(\frac{y_i}{x_i}\right), \tag{5.35}$$

where $I \in \mathbf{R}^{k \times k}$ is a discretized integration operator.

The data term of Zach and Kohli [ZK12] is precisely given by (5.35) except that the optimization is directly performed on $x, y \in \mathbf{R}^k$. The variable x can be interpreted as 1-sparse indicator of the interval Γ_i and $y \in \mathbf{R}^k$ as a sublabel offset. The constraint $\widehat{v} = y/h + I^\top x$ connects this representation to the subgraph representation $\widehat{v}$ via the operator $I \in \mathbf{R}^{k \times k}$ (see supplementary material for the definition). For general regularizers η, the discretization with piecewise linear φ_t differs from [Möl+16] as we perform a *joint convexification* of data term and regularizer and from [ZK12] as we consider the spatially continuous setting. Another important question to ask is which primal formulation is actually optimized after discretization with piecewise linear φ_t. In particular the distinction between jump and smooth regularization only makes sense for continuous label spaces, so it is interesting to see what is optimized after discretizing the label space.

Proposition 15. *Let $\gamma = \kappa(\gamma_2 - \gamma_1)$ and $\ell = 2$. The approximation with piecewise linear φ_t and piecewise constant φ_x of the continuous optimization problem* (5.13) *is equivalent to*

$$\inf_{u:\Omega\to\Gamma} \int_\Omega \rho^{**}(x, u(x)) + (\eta^{**} \mathbin{\square} \gamma\|\cdot\|)(\nabla u(x)) \, \mathrm{d}x, \tag{5.36}$$

where $(\eta \mathbin{\square} \gamma\|\cdot\|)(x) = \inf_y \eta(x-y) + \gamma\|y\|$ denotes the infimal convolution (cf. [Roc96, Section 5]).

From Proposition 15 we see that the minimal discretization with $\ell = 2$ amounts to approximating problem (5.1) by globally convexifying the data term. Furthermore, we can see that Mumford-Shah (truncated quadratic) regularization ($\eta(g) = \alpha\|g\|^2$, $\kappa(a) \equiv \lambda[\![a > 0]\!]$) is approximated by a convex Huber regularizer in case $\ell = 2$. This is because the infimal convolution between x^2 and $|x|$ corresponds to the Huber function. While even for $\ell = 2$ this is a reasonable approximation to the original model (5.1), we can gradually increase the number of labels to get an increasingly faithful approximation of the original nonconvex problem.

5.4.3.3 Piecewise quadratic φ_t

For piecewise quadratic φ_t the main difficulty are the constraints in (5.20). For piecewise linear φ_t the infimum over a linear function plus ρ_i lead to (minus) the convex conjugate of ρ_i. Quadratic dual variables lead to so called generalized Φ-conjugates [RWW98, Chapter 11L*, Example 11.66]. Such conjugates were also theoretically considered in the recent work [FA14] for discrete-continuous MRFs, however an efficient implementation seems challenging. The advantage of this representation would be that one can avoid convexification of the unaries on each interval Γ_i and thus obtain a tighter approximation. While in principle the resulting constraints could be implemented using techniques from convex algebraic geometry and semi-definite programming [BPT12] we leave this direction open to future work.

5.5 Implementation and Extensions

5.5.1 Piecewise quadratic unaries ρ_i

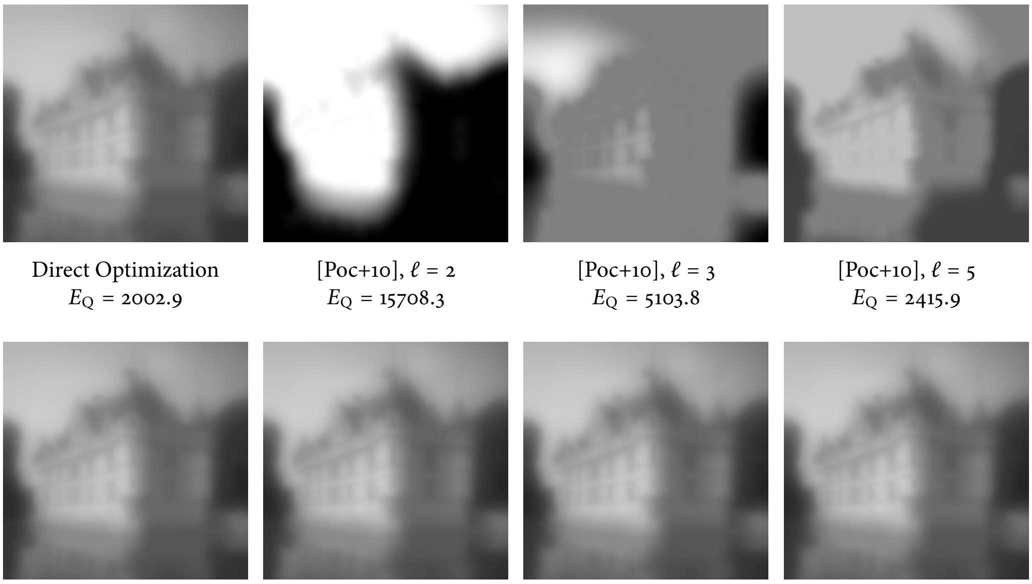
In some applications such as robust fusion of depth maps, the data term ρ has a piecewise quadratic form:

$$\rho(u) = \sum_{m=1}^{M} \min\left\{v_m, \alpha_m (u - f_m)^2\right\}. \tag{5.37}$$

The intervals on which the above function is a quadratic are formed by the breakpoints $f_m \pm \sqrt{v_m/\alpha_m}$. In order to optimize this within our framework, we need to compute the convex conjugate of ρ on the intervals Γ_i, see Eq. (5.31). We can write the data term (5.37) on each Γ_i as

$$\min_{1 \le j \le n_i} \underbrace{a_{i,j}u^2 + b_{i,j}u + c_{i,j} + \delta\{u \in I_{i,j}\}}_{=:\rho_{i,j}(u)}, \tag{5.38}$$

where n_i denotes the number of pieces and the intervals $I_{i,j}$ are given by the breakpoints and Γ_i. The convex conjugate is then given by $\rho_i^*(v) = \max_{1 \le j \le n_i} \rho_{i,j}^*(v)$. As the epigraph of the maximum is the intersection of the epigraphs, $\text{epi}(\rho_i^*) =$



Direct Optimization
E_Q = 2002.9

[Poc+10], $\ell = 2$
E_Q = 15708.3

[Poc+10], $\ell = 3$
E_Q = 5103.8

[Poc+10], $\ell = 5$
E_Q = 2415.9

[Poc+10], $\ell = 16$
E_Q = 2016.5

Proposed, $\ell = 2$
E_Q = 2002.9

Proposed, $\ell = 3$
E_Q = 2002.9

Proposed, $\ell = 5$
E_Q = 2002.9

Figure 5.5: To verify the tightness of the approximation, we optimize a convex problem (quadratic data term with quadratic regularization). The discretization with piecewise linear φ_t recovers the exact solution with 2 labels and remains tight (numerically) for all $\ell > 2$, while the traditional discretization from [Poc+10] leads to a strong label bias.

$\bigcap_{j=1}^{n_j} \operatorname{epi}\left(\rho^*_{i,j}\right)$, the constraints for the data term $(r^i, a^i) \in \operatorname{epi}(\rho^*_i)$, can be broken down:

$$\left(r^{i,j}, a^{i,j}\right) \in \operatorname{epi}\left(\rho^*_{i,j}\right), r^i = r^{i,j}, a^i = a^{i,j}, \forall j. \tag{5.39}$$

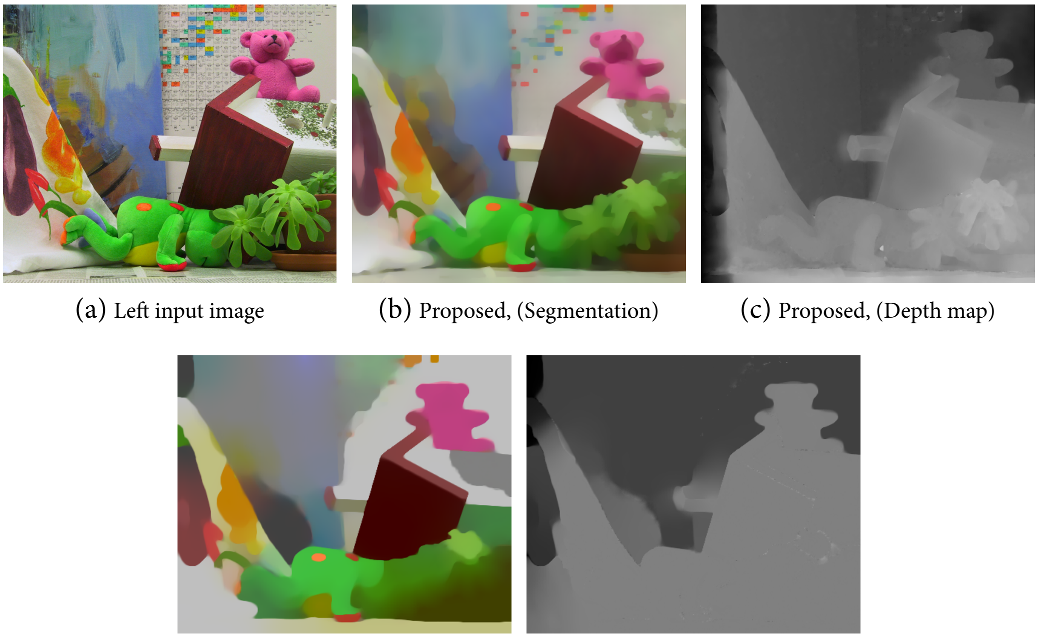
The projection onto the epigraphs of the $\rho^*_{i,j}$ are carried out as described in [Möl+16]. Such a convexified piecewise quadratic function is shown on the right in Figure 5.4.

5.5.2 The vectorial Mumford-Shah functional

Recently, the free-discontinuity problem (5.1) has been generalized to vectorial functions $u : \Omega \to \mathbf{R}^{n_c}$ by Strekalovskiy et al. [SCC12]. The model they propose is

$$\sum_{c=1}^{n_c} \int_{\Omega \setminus J_u} f_c(x, u_c(x), \nabla_x u_c(x))\, \mathrm{d}x + \lambda \mathcal{H}^{n-1}(J_u), \tag{5.40}$$

which consists of a separable data term and separable regularization on the continuous part. The individual channels are coupled through the jump part regularizer $\mathcal{H}^{n-1}(J_u)$ of the joint jump set across all channels. Using the same strategy as in Section 5.4, applied to the relaxation described in [SCC12, Section 3], a sublabel-accurate representation of the vectorial Mumford-Shah functional can be obtained.



(a) Left input image (b) Proposed, (Segmentation) (c) Proposed, (Depth map)

(d) [SCC12], (Segmentation) (e) [SCC12], (Depth map)

Figure 5.6: Joint segmentation and stereo matching. **b), c)** Using the proposed discretization we can arrive at smooth solutions using a moderate (5 × 5 × 5 × 5) discretization of the 4-dimensional RGB-D label space. **d), e)** When using such a coarse sampling of the label space, the classical discretization used in [SCC12] leads to a strong label bias. Note that with the proposed approach, a piecewise constant segmentation as in **d)** could also be obtained by increasing the smoothness parameter.

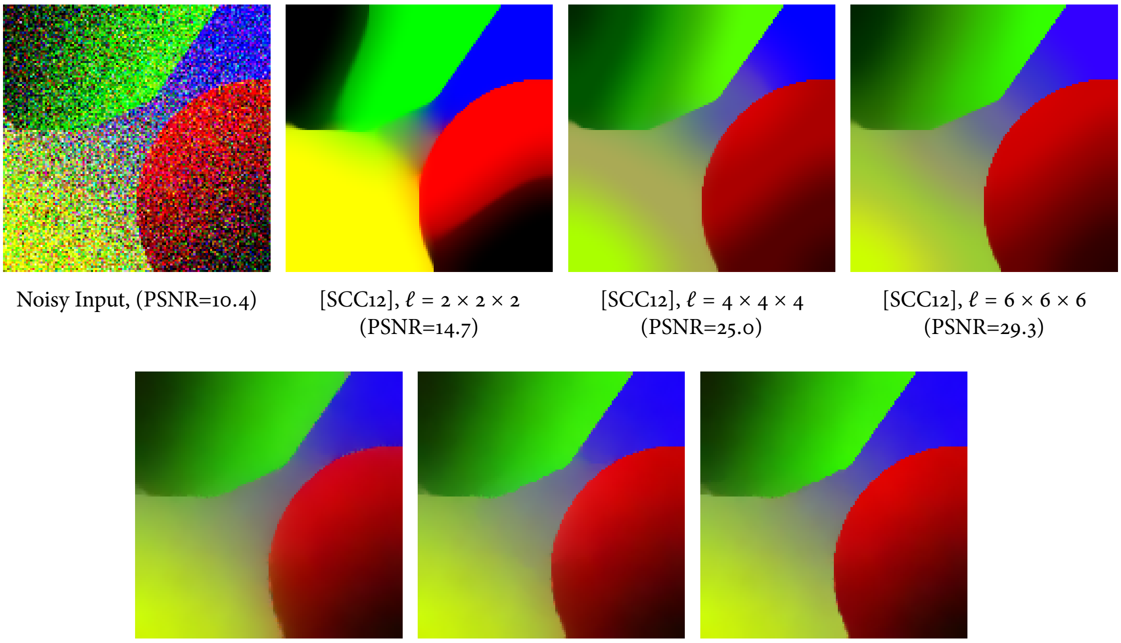
5.5.3 Numerical solution

We solve the final finite dimensional optimization problem after finite-difference discretization in spatial direction using the primal-dual algorithm [Poc+09a].

5.6 Experiments

5.6.1 Exactness in the convex case

We validate our discretization in Figure 5.5 on the convex problem $\rho(u) = (u - f)^2$, $\eta(\nabla u) = \lambda |\nabla u|^2$. The global minimizer of the problem is obtained by solving $(I - \lambda \Delta) u = f$. For piecewise linear φ_t we recover the exact solution using only 2 labels, and remain (experimentally) exact as we increase the number of labels. The discretization from [Poc+10] shows a strong label bias due to the piecewise constant dual variable φ_t. Even with 16 labels their solution is different from the ground truth energy.



Noisy Input, (PSNR=10.4)

[SCC12], $\ell = 2 \times 2 \times 2$ (PSNR=14.7)

[SCC12], $\ell = 4 \times 4 \times 4$ (PSNR=25.0)

[SCC12], $\ell = 6 \times 6 \times 6$ (PSNR=29.3)

Ours, $\ell = 2 \times 2 \times 2$, (PSNR=24.8)

Ours, $\ell = 4 \times 4 \times 4$, (PSNR=28.0)

Ours, $\ell = 6 \times 6 \times 6$, (PSNR=**30.0**)

Figure 5.7: Denoising of a synthetic piecewise smooth image degraded with 30% Gaussian noise. The standard discretization of the vectorial Mumford-Shah functional shows a strong bias towards the chosen labels (see also Figure 5.8), while the proposed discretization has no bias and leads to the highest overall peak signal to noise ratio (PSNR).

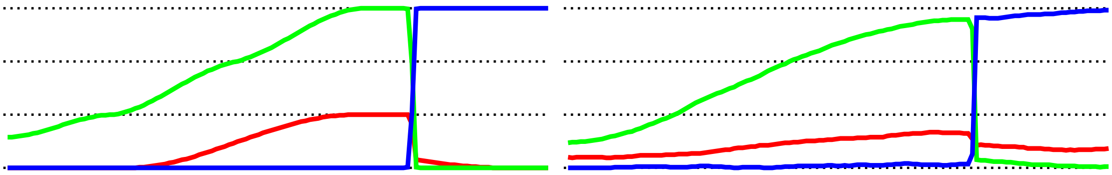


Figure 5.8: We show a 1D-slice through the resulting image in Figure 5.7 (with $\ell = 4 \times 4 \times 4$). The discretization [SCC12] (left) shows a strong bias towards the labels, while the proposed discretization (right) yields a sublabel-accurate solution.

5.6.2 The vectorial Mumford-Shah functional

Joint depth fusion and segmentation. We consider the problem of joint image segmentation and robust depth fusion from [PZB07] using the vectorial Mumford-Shah functional from Section 5.5.2. The data term for the depth channel is given by (5.37), where f_m are the input depth hypotheses, α_m is a depth confidence and ν_m is a truncation parameter to be robust towards outliers. For the segmentation, we use a quadratic difference dataterm in RGB space. For Figure 5.6 we computed multiple depth hypotheses f_m on a stereo pair using different matching costs (sum of absolute (gradient) differences, and normalized cross correlation) with varying patch radii (0 to 2). Even for a moderate label space of $5 \times 5 \times 5 \times 5$ we have no label discretization artifacts.

The piecewise linear approximation of the unaries in [SCC12] leads to an almost piecewise constant segmentation of the image. To highlight the sublabel-accuracy of the proposed approach we chose a small smoothness parameter which leads to a piecewise smooth segmentation, but with a higher smoothness term or different choice of unaries a piecewise constant segmentation could also be obtained.

Piecewise-smooth approximations. In Figure 5.7 we compare the discretizations for the vectorial Mumford-Shah functional. We see that the approach [SCC12] shows strong label bias (see also Figure 5.8 and 5.1) while the discretiziation with piecewise linear duals leads to a sublabel-accurate result.

5.7 Conclusion

We proposed a framework to numerically solve *fully-continuous* convex relaxations in a sublabel-accurate fashion. The key idea is to implement the dual variables using a piecewise linear approximation. We prove that different choices of approximations for the dual variables give rise to various existing relaxations: in particular piecewise constant duals lead to the traditional lifting [Poc+09a] (with min-pooling of the unary costs), whereas piecewise linear duals lead to the sublabel lifting that was recently proposed for total variation regularized problems [Möl+16]. While the latter method is not easily generalized to other regularizers due to the separate convexification of data term and regularizer, the proposed representation generalizes to arbitrary convex and non-convex regularizers such as the scalar and the vectorial Mumford-Shah problem. The proposed approach provides a systematic technique to derive sublabel-accurate discretizations for continuous convex relaxation approaches, thereby boosting their memory and runtime efficiency for challenging large-scale applications.

Chapter 6

Lifting Vectorial Variational Problems: A Natural Formulation based on Geometric Measure Theory and Discrete Exterior Calculus

6.1 Introduction

We consider functionals of C^1-mappings $f : \mathcal{X} \to \mathcal{Y}$

$$E(f) = \int_{\mathcal{X}} c\left(x, f(x), \nabla f(x)\right) \mathrm{d}x, \tag{6.1}$$

where $\mathcal{X} \subset \mathbf{R}^n$, $\mathcal{Y} \subset \mathbf{R}^N$ are bounded and open. The cost function $c \equiv c(x, y, \xi)$ is assumed to be a nonnegative (possibly nonconvex) continuous function on $\mathcal{X} \times \mathcal{Y} \times \mathbf{R}^{N \times n}$ that is *polyconvex* (see Def. 2) in the Jacobian matrix ξ.

This work is concerned with relaxation and global optimization of (6.1) when, both, dimension and codimension are possibly larger than one ($n > 1$, $N > 1$). This is expected to be difficult: In the discrete setting problems with $n = 1$ or $N = 1$ typically correspond to polynomial-time solvable shortest path ($n = 1$) or graph cut ($N = 1$) problems [CK97; Tsi95; Ish03; SC10], whereas for $n, N > 1$, the arising multilabel problems with unordered label spaces are known to be NP-hard - see [LSH16]. Nevertheless, heuristic strategies have been shown to yield excellent results in tasks such as optical flow [CK16] or shape matching [SKH08; CK15]. In contrast to such well-established Markov random field (MRF) works [Kol06; KR07; Koh+08; SKH08; MHG15; CK15; CK16; DSC18] we consider the way less explored continuous (infinite-dimensional) setting.

Our motivation partly stems from the fact that formulations in function space are very general and admit a variety of discretizations. Finite difference discretizations of continuous relaxations often lead to models that are reminiscent of MRFs [ZHP13], while piecewise-linear approximations are related to discrete-continuous MRFs [ZK12], see [FA14; MC17]. More recently, for the Kantorovich relaxation in optimal transport, approximations with deep neural networks were considered and achieved promising performance, for example in generative modeling [ACB17; Seg+18].

We further argue that fractional (non-integer) solutions to a careful discretization of the continuous model can implicitly approximate an “integer” continuous solution. Therefore one can achieve accuracies that go substantially beyond the mesh size. The resulting models would be difficult to interpret and derive from a finite-dimensional viewpoint such that the continuous considerations are required for the final implementation. Also, formulations arising from continuous relaxations allow one to introduce isotropic smoothness potentials without reverting to higher-order terms in the cost, and, as we show in this work, one can impose general polyconvex regularizations using only local constraints. An example of a polyconvex function (which is in general nonconvex) is the surface area of the graph, sometimes referred to as “Beltrami regularization” in the image processing community, see e.g., [KMS00].

In contrast to the discrete multi-labeling setting, an important question is whether variational problems involving the energy (6.1) admit a minimizer. A fruitful approach to address this question is to suitably relax the notion of solution, thereby enlarging the search space of admissible candidates (“lifting the problem to a larger space”). The origins of this idea can be traced back[1] to the turn of the century, see Hilbert’s twentieth problem [Hil00]. An example of that principle is the celebrated Kantorovich relaxation [Kan60] of Monge’s transportation problem [Mon81]. There, the search over maps $f : \mathcal{X} \to \mathcal{Y}$ is relaxed to one over probability measures on the product space $\mathcal{X} \times \mathcal{Y}$. Each map can be identified in that extended space with a measure concentrated on its graph. Existence of optimal transportation plans follows directly due to good compactness properties of the larger space. Furthermore, the nonlinearly constrained and nonconvex optimization problem is transformed into one of linear programming, leading to rich duality theories and fast numerical algorithms [PC18].

One may ask whether the relaxed solution in the extended space has certain regularity properties, for example whether it is the graph of a (sufficiently regular) map and thus can be considered a solution to the original (“unlifted”) problem. In the case of optimal transport, such regularity theory can be guaranteed under some assumptions [Vil08; San15]. Establishing existence and regularity for problems in which the cost additionally depends on the Jacobian (for example minimal surface problems) has been a driving factor in the development of geometric measure theory, see [Mor16] for an introduction. In this work, we will use ideas from geometric measure theory to pursue the above relaxation and lifting principle for the energy (6.1). The main idea is to reformulate the original variational problem as a shape optimization problem over oriented manifolds representing the graph of the map $f : \mathcal{X} \to \mathcal{Y}$ in the product space $\mathcal{X} \times \mathcal{Y}$. To obtain a convex formulation

[1] We refer the interested reader to the historical remarks in L. C. Young’s book on the calculus of variations [You80, pp. 122–123].

we enlarge the search space from oriented manifolds to currents.

6.1.1 Related work

A common strategy to solve problems involving (6.1) is to revert to local gradient descent minimization based on the Euler-Lagrange equations. But for nonconvex problems solutions might depend on the initialization and the computed stationary points may be quite suboptimal. Therefore, we pursue the aforementioned lifting of the energy (6.1) to currents. This lifting has been previously considered in geometric measure theory to establish the aforementioned existence and regularity theory for vectorial variational problems in a very broad setting, see e.g., [Fed69; Fed74; AG91]. In contrast to such impressive theoretical achievements, this paper is concerned with a discretization and implementation.

There is also a variety of related applied works. The paper [Win+11] tackles the problem of bijective and smooth shape matching using linear programming. Similar to the present work, the authors also look for graph surfaces in $\mathcal{X} \times \mathcal{Y}$ but they consider the discrete setting and use a different notion of boundary operator. We study the continuous setting, but also our discrete formulation is quite different.

For $N = 1$, the proposed continuous formulation specializes to [ABD03; Poc+10]. To tackle the setting of $N > 1$ in a memory efficient manner, Strekalovskiy et al. [SCC12; GSC13; SCC14] keep a collection of N surfaces with codimension one under the factorization assumption that $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N$. In contrast, we consider only one surface of codimension N, we do not require an assumption on $\mathcal{Y}$, our approach is applicable to a larger class of functionals and we expect it to yield a tighter relaxation. The lifting approaches [Lel+13a; GBO12] also tackle vectorial problems by considering the full product space, but are limited to total variation regularization (with the former allowing $\mathcal{Y}$ to be a manifold). The recent work [WC16] is most related to the present one, however their relaxation considers a specific instance of (6.1). Moreover, the above works are based on finite difference discretizations of the continuous model. In contrast, the proposed discretization using discrete exterior calculus yields solutions beyond the mesh accuracy as in recent sublabel-accurate approaches. The latter are restricted to $N = 1$ [Möl+16; MC17] or total variation regularization [Lau+16]. Recent works also include extensions to total generalized variation or Laplacian regularization [SG18; VL19; LL18].

Recent approaches in shape analysis [Sol+16; Ves+17b; Ves+17a] also operate in the product space $\mathcal{X} \times \mathcal{Y}$. However, these are based on local minimizations of the Gromov-Wasserstein distance [Mém07] and spectral variants thereof [Mém09] which leads to (nonconvex) quadratic assignment problems. While the goal to find a smooth (possibly bijective) map is similar, the formulations appear to be quite different. To alleviate the increased cost of the product space formulation, computationally efficient representations of densities in $\mathcal{X} \times \mathcal{Y}$ have been studied in the context of functional maps [Ovs+12; Rod+19].

6.2 Notation and Preliminaries

Throughout this paper we will introduce notions from geometric measure theory, as they are not commonly used in the vision community. While the subject is rather technical, our aim is to keep the presentation light and to focus on the geometric intuition and aspects which are important for a practical implementation. We invite the reader to consult chapter 4 in the book [Mor16] and the chapter on exterior calculus in [Cra19], which both contain many illuminating illustrations. For a more technical treatment we refer to [Fed69; KP08].

In the following, we denote a basis in $\mathbf{R}^d$ as $\{e_1, \cdots, e_d\}$ with dual basis $\{\mathrm{d}x_1, \cdots, \mathrm{d}x_d\}$ where $\mathrm{d}x_i : \mathbf{R}^d \to \mathbf{R}$ is the linear functional that maps every $x = (x_1, \cdots, x_d)$ to the i-th component x_i. Given an integer $k \leq d$, $I(d,k)$ are the multi-indices $\mathbf{i} = (i_1, \cdots, i_k)$ with $1 \leq i_1 < \cdots < i_k \leq d$.

As we will consider n-surfaces in $\mathcal{X} \times \mathcal{Y} \subset \mathbf{R}^{n+N}$, most of the time we set $d = n+N$ and $k = n$. To further simplify notation, we denote the basis vectors $\{e_{n+1}, \cdots, e_{n+N}\}$ by $\{\varepsilon_1, \cdots, \varepsilon_N\}$ and similarly refer to the dual basis as $\{\mathrm{d}x_1, \cdots \mathrm{d}x_n, \mathrm{d}y_1, \cdots, \mathrm{d}y_N\}$. When it is clear from the context, we treat vectors $e_i \in \mathbf{R}^n$ and $\varepsilon_i \in \mathbf{R}^N$ in the sense that $e_i \simeq (e_i, \mathbf{0}_N) \in \mathbf{R}^{n+N}$, $\varepsilon_i \simeq (\mathbf{0}_n, \varepsilon_i) \in \mathbf{R}^{n+N}$. As an example, for $\nabla f(x) \in \mathbf{R}^{N \times n}$ we can define the expression $e_i + \nabla f(x) e_i$ and read it as $(e_i, \nabla f(x) e_i) \in \mathbf{R}^{n+N}$.

6.2.1 Convex analysis

The extended reals are denoted by $\overline{\mathbf{R}} = \mathbf{R} \cup \{+\infty\}$. For a finite-dimensional real vector space V and $\Psi : V \to \overline{\mathbf{R}}$ we denote the convex conjugate as $\Psi^* : V^* \to \overline{\mathbf{R}}$ and the biconjugate as $\Psi^{**} : V \to \overline{\mathbf{R}}$. Ψ^{**} is the largest lower-semicontinuous convex function below Ψ. In our notation, for functions with several arguments, the conjugate is always taken only in the last argument. As a general reference to convex analysis, we refer the reader to the books [HL12; Roc96].

6.2.2 Multilinear algebra

The formalism of multi-vectors we introduce in this section is central to this work, as the idea of the relaxation is to represent the oriented graph of f by a k-vectorfield (more precisely: a k-current) in the product space $\mathcal{X} \times \mathcal{Y}$. Basically, one can multiply $v_i \in \mathbf{R}^d$ to obtain an object

$$v = v_1 \wedge \cdots \wedge v_k, \tag{6.2}$$

called a *simple* k-vector in $\mathbf{R}^d$. The geometric intuition of simple k-vectors is, that they describe the k-dimensional space spanned by the $\{v_i\}$, together with an orientation and the area of the parallelotope given by the $\{v_i\}$. Thus, simple k-vectors can be thought of oriented parallelotopes as shown in orange in Fig. 6.1. In general, k-vectors are defined to be formal sums

$$v = \sum_{\mathbf{i} \in I(d,k)} v^{\mathbf{i}} \cdot e_{\mathbf{i}_1} \wedge \cdots \wedge e_{\mathbf{i}_k} = \sum_{\mathbf{i} \in I(d,k)} v^{\mathbf{i}} \cdot e_{\mathbf{i}}, \tag{6.3}$$

for coefficients $v^{\mathbf{i}} \in \mathbf{R}$. They form the vector space $\Lambda_k \mathbf{R}^d$, which has dimension $\binom{d}{k}$.

The dual space $\Lambda^k \mathbf{R}^d$ of k-covectors is defined analogously, with $\langle dx_{\mathbf{i}}, e_{\mathbf{j}} \rangle = \delta_{\mathbf{ij}}$. We define for two k-vectors (and also for k-covectors) $v = \sum_{\mathbf{i}} v_{\mathbf{i}} e_{\mathbf{i}}$, $w = \sum_{\mathbf{i}} w_{\mathbf{i}} e_{\mathbf{i}}$ an inner product $\langle v, w \rangle = \sum_{\mathbf{i}} v_{\mathbf{i}} w_{\mathbf{i}}$ and norm $|v| = \sqrt{\langle v, v \rangle}$.

k-vectors (elements of $\Lambda_k \mathbf{R}^d$) are called *simple*, if they can be written as the *wedge product* of 1-vectors as in (6.2). Unfortunately, for $1 < k < d - 1$, not all k-vectors are simple and the set of simple k-vectors is a nonconvex cone in $\Lambda_k \mathbf{R}^d$, called the Grassmann cone [BES63]. This is one aspect why the setting of $n > 1$ and $N > 1$ is more challenging.

Later on, we will consider a relaxation from the nonconvex set of simple k-vectors to general k-vectors. Naturally, for the relaxation to be good, we want the convex energy to be *as large as possible* on non-simple k-vectors. For the Euclidean norm, a good convex extension is the *mass* norm

$$\|v\| = \inf \left\{ \sum_i |\xi_i| : \xi_i \text{ are simple}, v = \sum_i \xi_i \right\}. \tag{6.4}$$

The dual norm is the *comass* norm given by:

$$\|w\|^* = \sup \{ \langle w, v \rangle : v \text{ is simple }, |v| \leq 1 \}. \tag{6.5}$$

The mass norm can be understood as the largest norm that agrees with the Euclidean norm on simple k-vectors.

6.3 Lifting to Graphs in the Product Space

With the necessary preliminaries in mind, our goal is now to reparametrize the original energy (6.1) to the graph $\mathcal{G}_f \subset \mathcal{X} \times \mathcal{Y}$. As shown in Fig. 6.1, the graph is an oriented n-dimensional manifold in the product space with global parametrization $u(x) = (x, f(x))$.

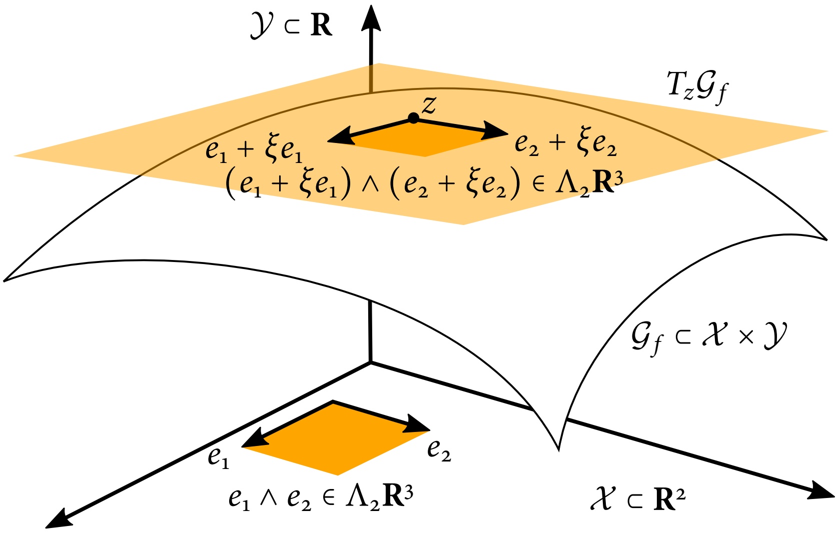
Definition 1 (Orientation). *If $\mathcal{M} \subset \mathbf{R}^d$ is a k-dimensional smooth manifold in $\mathbf{R}^d$ (possibly with boundary), an **orientation** of $\mathcal{M}$ is a continuous map $\tau_{\mathcal{M}} : \mathcal{M} \to \Lambda_k \mathbf{R}^d$ such that $\tau_{\mathcal{M}}(z)$ is a simple k-vector with unit norm that spans the tangent space $T_z \mathcal{M}$ at every point $z \in \mathcal{M}$.*

From differential geometry we know that the tangent space $T_z \mathcal{G}_f$ at $z = (x, f(x))$ is spanned by $\partial_i u(u^{-1}(z)) = e_i + \nabla f(x) e_i$. Therefore, an orientation of $\mathcal{G}_f$ is given by

$$\tau_{\mathcal{G}_f}(z) = \frac{M(\nabla f(\pi_1 z))}{|M(\nabla f(\pi_1 z))|}, \tag{6.6}$$

where the map $M : \mathbf{R}^{N \times n} \to \Lambda_n \mathbf{R}^{n+N}$ is given by

$$M(\xi) = (e_1 + \xi e_1) \wedge \cdots \wedge (e_n + \xi e_n), \tag{6.7}$$



$\mathcal{Y} \subset \mathbf{R}$
$T_z\mathcal{G}_f$
z
$e_1 + \xi e_1$
$e_2 + \xi e_2$
$(e_1 + \xi e_1) \wedge (e_2 + \xi e_2) \in \Lambda_2\mathbf{R}^3$
$\mathcal{G}_f \subset \mathcal{X} \times \mathcal{Y}$
e_1
e_2
$e_1 \wedge e_2 \in \Lambda_2\mathbf{R}^3$
$\mathcal{X} \subset \mathbf{R}^2$

Figure 6.1: Illustration for the setting of $n = 2$, $N = 1$. The graph $\mathcal{G}_f$ of the C^1-map $f : \mathcal{X} \to \mathcal{Y}$ is a smooth oriented manifold embedded in the product space $\mathcal{X} \times \mathcal{Y}$. The tangent space at $z = (x, f(x))$ is spanned by the simple n-vector $(e_1 + \xi e_1) \wedge \cdots \wedge (e_n + \xi e_n) \in \Lambda_n\mathbf{R}^{n+N}$, where $\xi = \nabla f(x) \in \mathbf{R}^{N \times n}$ is the Jacobian.

and $\pi_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ is the canonical projection onto the first argument. In order to derive the reparametrization, we have to connect a simple n-vector (representing an oriented tangent plane of the graph) with the Jacobian of the original energy. For that, we need an inverse of the map given in (6.7).

To derive such an inverse, we first introduce further helpful notations. For $\mathbf{i} \in I(m, l)$ we denote by $\bar{\mathbf{i}} \in I(m, m - l)$ the element which complements $\mathbf{i}$ in $\{1, 2, \cdots, m\}$ in increasing order, denote $\bar{0} = \{1, \cdots, m\}$ and 0 as the empty multi-index. Every $v \in \Lambda_n\mathbf{R}^{n+N}$ can be written as

$$v = \sum_{|\mathbf{i}|+|\mathbf{j}|=n} v^{\mathbf{i},\mathbf{j}} e_{\mathbf{i}} \wedge \varepsilon_{\mathbf{j}}, \tag{6.8}$$

where $\mathbf{i} \in I(n, l)$, $\mathbf{j} \in I(N, l')$, $l + l' = n$. To give an example, the $\binom{5}{2} = 10$ coefficients of a 2-vector $v \in \Lambda_2\mathbf{R}^5$ according to the notation (6.8) are:

$$\begin{array}{llll} v^{\bar{0},0} & & & \\ v^{1,1} & v^{2,1} & & \\ v^{1,2} & v^{2,2} & v^{0,(1,2)} & \\ v^{1,3} & v^{2,3} & v^{0,(1,3)} & v^{0,(2,3)}, \end{array} \tag{6.9}$$

where we highlighted the $N \times n$ coefficients with $|\mathbf{j}| = 1$. Now note that the vector $v = M(\xi)$ is by construction a simple n-vector with first component $v^{\bar{0},0} = 1$. To any $v \in \Lambda_n\mathbf{R}^{n+N}$ with $v^{\bar{0},0} = 1$ we associate $\xi(v) \in \mathbf{R}^{N \times n}$ given by

$$[\xi(v)]_{j,i} = (-1)^{n-i} v^{\bar{i},j}. \tag{6.10}$$

If and only if $v \in \Lambda_n \mathbf{R}^{n+N}$ is simple with first component $v^{\bar{0},0} = 1$ then $v = M(\xi(v))$. A proof is given in [GMS98, Vol. I, Ch. 2.1, Prop. 1]. Thus, on the set of simple n-vectors with first component $v^{\bar{0},0} = 1$,

$$\Sigma_1 = \{v \in \Lambda_n \mathbf{R}^{n+N} : v = M(\xi) \text{ for } \xi \in \mathbf{R}^{N \times n}\}, \tag{6.11}$$

the inverse of the map (6.7) is given by (6.10).

Using the above notations, we can define a generalized notion of convexity, which essentially states that there is a convex reformulation on k-vectors.

Definition 2 (Polyconvexity). *A map $c : \mathbf{R}^{N \times n} \to \overline{\mathbf{R}}$ is* ***polyconvex*** *if there is a convex function $\bar{c} : \Lambda_n \mathbf{R}^{n+N} \to \overline{\mathbf{R}}$ such that we have*

$$c(\xi) = \bar{c}(M(\xi)) \quad \textit{for all} \quad \xi \in \mathbf{R}^{N \times n}. \tag{6.12}$$

Equivalently one has that $c(\xi(v)) = \bar{c}(v)$ for all $v \in \Sigma_1$. We also refer to the convex function $\bar{c}$ as a ***polyconvex extension****.*

In general, the polyconvex extension is not unique. Any convex function has an obvious polyconvex extension by (6.10), but as discussed in the previous section we would like the convex extension to be as large as possible for $v \notin \Sigma_1$. The largest polyconvex extension which agrees with the original function on Σ_1 can be formally defined using the convex biconjugate, but is often hard to explicitly compute. The mass norm (6.4) corresponds to such a construction.

Nevertheless, given any polyconvex extension, we can now reparametrize the original energy (6.1) on the oriented graph $\mathcal{G}_f$, as we show in the following central proposition.

Proposition 16. *Let $\bar{c} : \mathcal{X} \times \mathcal{Y} \times \Lambda_n \mathbf{R}^{n+N} \to \overline{\mathbf{R}}$ be a polyconvex extension of the original cost c in the last argument. Define the function $\Psi : \mathcal{X} \times \mathcal{Y} \times \Lambda_n \mathbf{R}^{n+N} \to \overline{\mathbf{R}}$,*

$$\Psi(z, v) = \begin{cases} v^{\bar{0},0} \bar{c}(\pi_1 z, \pi_2 z, v / v^{\bar{0},0}), & \textit{if } v^{\bar{0},0} > 0, \\ +\infty, & \textit{otherwise,} \end{cases} \tag{6.13}$$

where $\pi_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and $\pi_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ are the canonical projections onto the first and second argument. Then we can reparametrize (6.1) *as follows:*

$$\begin{aligned} &\int_{\mathcal{X}} c(x, f(x), \nabla f(x)) \, \mathrm{d}\mathcal{L}^n(x) \\ &\quad = \int_{\mathcal{G}_f} \Psi(z, \tau_{\mathcal{G}_f}(z)) \, \mathrm{d}\mathcal{H}^n(z), \end{aligned} \tag{6.14}$$

where the second integral is the standard Lebesgue integral with respect to the n-dimensional Hausdorff measure on $\mathbf{R}^{n+N}$ restricted to the graph $\mathcal{G}_f$.

Proof. We directly calculate:

$$\begin{aligned}
&\int_{\mathcal{X}} c\,(x, f(x), \nabla f(x))\, \mathrm{d}\mathcal{L}^n(x) && (6.15)\\
&= \int_{\mathcal{X}} \Psi\,(x, f(x), M(\nabla f(x)))\, \mathrm{d}\mathcal{L}^n(x) && (6.16)\\
&= \int_{\mathcal{G}_f} \Psi\,(z, M(\nabla f(\pi_1 z)))\, \frac{1}{|M(\nabla f(\pi_1 z))|} \mathrm{d}\mathcal{H}^n(z) && (6.17)\\
&= \int_{\mathcal{G}_f} \Psi\left(z, \tau_{\mathcal{G}_f}(z)\right) \mathrm{d}\mathcal{H}^n(z). && (6.18)
\end{aligned}$$

The step from (6.15) to (6.16) uses that $\bar{c}$ is a polyconvex extension (so that we can apply (6.12)) and the fact that for $v = M(\nabla f(x))$ we have $v^{\bar{0},0} = 1$. To arrive at (6.17), an application of the area formula [KP08, Corollary 5.1.13] suffices and for (6.18) we used positive one-homogenity of Ψ and the definition of $\tau_{\mathcal{G}_f}$ in (6.6). □

Interestingly, the function (6.13) is convex and one-homogeneous in the last argument, as it is the *perspective* of a convex function. However, the search space of oriented graphs of C^1 mappings is nonconvex. Therefore we relax from oriented graphs to the larger set of currents, which we will introduce in the following section. Since currents form a vector space, we therefore obtain a convex functional over a convex domain.

6.4 From Oriented Graphs to Currents

Throughout this section, let $U \subset \mathbf{R}^d$ be an open set, which will later be a neighbourhood of $X \times Y \subset \mathbf{R}^{n+N}$, where $X = \mathrm{cl}(\mathcal{X})$, $Y = \mathrm{cl}(\mathcal{Y})$ are the closures of $\mathcal{X}, \mathcal{Y}$. The main idea of our relaxation and the geometric intuitions of *pushforward* and *boundary* operator we introduce in this section are summarized in the following Fig. 6.2. Currents are defined in duality with differential forms, which we will briefly introduce in the following section.

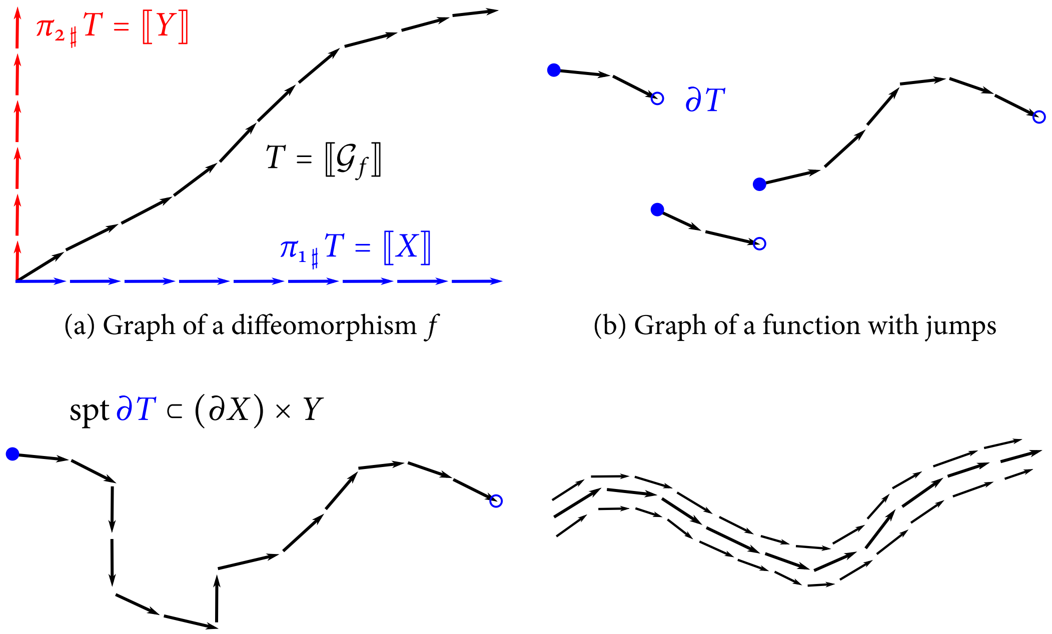
6.4.1 Differential forms

A differential form of order k (short: k-form) is a map $\omega : U \to \Lambda^k \mathbf{R}^d$. The *support* of a differential form $\operatorname{spt} \omega$ is defined as the closure of $\{z \in U : \omega(z) \neq 0\}$. Integration of a k-form over an oriented k-dimensional manifold is defined by

$$\int_{\mathcal{M}} \omega := \int_{\mathcal{M}} \langle \omega(z), \tau_{\mathcal{M}}(z) \rangle \, \mathrm{d}\mathcal{H}^k(z). \tag{6.19}$$

A notion of derivative for k-forms is the exterior derivative $d\omega$, which is a $(k+1)$-form given by:

$$\langle d\omega(z), v_1 \wedge \cdots \wedge v_{k+1} \rangle = \lim_{h \to 0} \frac{1}{h^{k+1}} \int_{\partial P} \omega, \tag{6.20}$$


$\pi_{2\sharp}T = [\![Y]\!]$
$T = [\![\mathcal{G}_f]\!]$
$\pi_{1\sharp}T = [\![X]\!]$
∂T
spt $\partial T \subset (\partial X) \times Y$

(a) Graph of a diffeomorphism f

(b) Graph of a function with jumps

(c) ”Stiched” graph (zero boundary)

(d) Current which is not a graph

Figure 6.2: The idea of our relaxation is to move from oriented graphs in the product space to the larger set of currents. These include oriented graphs as special cases, as shown in Fig. 6.2a. For a diffeomorphism, the push-forwards $\pi_{1\sharp}T$, $\pi_{2\sharp}T$ yield currents induced by domain and codomain, which will be a linear constraint in the relaxed problem. In Fig. 6.2b we show the current given by the graph of a discontinuous function. Since it has holes, the boundary operator ∂T has support inside the domain. We will constrain the support of the boundary to exclude such cases. Fig. 6.2c Stitching jumps yields a current with vertical parts at the jump points, which corresponds to the limiting case in the perspective function (6.13). To obtain an overall convex formulation, we will also allow currents Fig. 6.2d which don’t necessarily concentrate on the graph of a function.

where ∂P is the boundary of the parallelotope spanned by the $\{hv_i\}$ at point z.

To get an intuition, note that for $k = 0$ this reduces to the familiar directional derivative $\langle d\omega(x), v_1 \rangle = \lim_{h \to 0} \frac{1}{h} (\omega(x + hv_1) - \omega(x))$. With (6.19) and (6.20) in mind, one sees why Stokes’ theorem

$$\int_{\mathcal{M}} d\omega = \int_{\partial \mathcal{M}} \omega. \tag{6.21}$$

should hold intuitively. Given a map $\pi : \mathbf{R}^d \to \mathbf{R}^q$, the *pullback* $\pi^\sharp \omega$ of the k-form

ω is determined by

$$\langle \pi^\sharp \omega, v_1 \wedge .. \wedge v_k \rangle = \langle \omega \circ \pi, D_{v_1}\pi \wedge .. \wedge D_{v_k}\pi \rangle, \tag{6.22}$$

where $D_{v_i}\pi = \nabla\pi \cdot v_i$ and $\nabla\pi \in \mathbf{R}^{q\times d}$ is the Jacobian.

6.4.2 Currents

Denote the space of smooth k-forms with compact support on U as $\mathcal{D}^k(U)$. *Currents* are elements of the dual space $\mathcal{D}_k(U) = \mathcal{D}^k(U)'$, i.e., linear functionals acting on differential forms. As shown in Fig. 6.2a, an oriented k-dimensional manifold $\mathcal{M} \subset U$ induces a current by

$$[\![\mathcal{M}]\!](\omega) = \int_{\mathcal{M}} \omega. \tag{6.23}$$

However, since $\mathcal{D}_k(U)$ is a vector space, not all elements look like k-dimensional manifolds, see Fig. 6.2d. The *boundary* of the k-current $T \in \mathcal{D}_k(U)$ is the $(k-1)$-current $\partial T \in \mathcal{D}_{k-1}(U)$ defined via the exterior derivative:

$$\partial T(\omega) = T(d\omega), \text{ for all } \omega \in \mathcal{D}^{k-1}(U). \tag{6.24}$$

Stokes' theorem (6.21) ensures that for currents which are given by k-dimensional oriented manifolds, the boundary of the current agrees with the usual notion, see Fig. 6.2b.

The *support* of a current, denoted by spt T, is the complement of the biggest open set V such that

$$T(\omega) = 0 \text{ whenever } \operatorname{spt}(\omega) \subset V. \tag{6.25}$$

Given a map $\pi : \mathbf{R}^d \to \mathbf{R}^q$ the *pushforward* $\pi_\sharp T$ of the k-current $T \in \mathcal{D}_k(U)$ is given by

$$\pi_\sharp T(\omega) = T(\pi^\sharp \omega), \text{ for all } \omega \in \mathcal{D}^k(\mathbf{R}^q). \tag{6.26}$$

Intuitively, it transforms the current using the map π, as illustrated in Fig. 6.2a. The *mass* of a current $T \in \mathcal{D}_k(U)$ is

$$\mathbb{M}(T) = \sup\left\{ T(\omega) : \omega \in \mathcal{D}^k(U), \|\omega(z)\|^* \leq 1 \right\}, \tag{6.27}$$

and as expected $\mathbb{M}([\![\mathcal{M}]\!]) = \mathcal{H}^k(\mathcal{M})$. We denote the space of k-currents with finite mass and compact support by $\mathbf{M}_k(U)$. These are *representable by integration*, meaning there is a measure $\|T\|$ on U and a map $\vec{T} : U \to \Lambda_k \mathbf{R}^d$ such that $\|\vec{T}(z)\| = 1$ for $\|T\|$-almost all z such that

$$T(\omega) = \int \langle \omega(z), \vec{T}(z) \rangle \, \mathrm{d}\|T\|(z). \tag{6.28}$$

The decomposition (6.28) is crucial, and we will use it to define the relaxation in the next section.

6.4.3 The relaxed energy

We lift the original energy (6.1) to the space of finite mass currents $T \in \mathbf{M}_n(U)$ with $\operatorname{spt} T \subset X \times Y$ as follows:

$$\mathbf{E}(T) = \int \Psi^{**}\left(\pi_1 z, \pi_2 z, \vec{T}(z)\right) \mathrm{d}\|T\|(z). \tag{6.29}$$

Since for $T = [\![\mathcal{G}_f]\!]$ we have $\vec{T} = \tau_{\mathcal{G}_f}$, $\|T\| = \mathcal{H}^n \llcorner \mathcal{G}_f$ the desirable property $\mathbf{E}([\![\mathcal{G}_f]\!]) = E(f)$ holds due to Prop. 16.

Note that in (6.29) we use the lower-semicontinuous regularization Ψ^{**} which extends (6.13) at $v^{\bar{0},0} = 0$ with the correct value. Interestingly, this point corresponds to the situation when the graph has vertical parts, which cannot occur for C^1 functions but can happen for general currents, see Fig. 6.2c. In [Mor02] it was shown that one can penalize such jumps in a way depending on the jump distance and direction. We will not consider such additional regularization due to space limitations, but remark that they could be integrated by adding further constraints to the following dual representation, which is a consequence of [GMS98, Vol. II, Sec. 1.3.1, Thm. 2].

Proposition 17. *For $T \in \mathbf{M}_n(U)$ with* $\operatorname{spt} T \subset X \times Y$*, we have the dual representation*

$$\mathbf{E}(T) = \sup_{\omega \in \mathcal{K}} T(\omega), \tag{6.30}$$

where the constraint is the closed and convex set

$$\mathcal{K} = \left\{\omega \in C_c^0(U, \Lambda^n \mathbf{R}^{n+N}) : \Psi^*(\pi_1 z, \pi_2 z, \omega(z)) \leq 0, \forall z \in X \times Y\right\}. \tag{6.31}$$

The final relaxed optimization problem for (6.1) reads

$$\inf_{T \in \mathbf{M}_n(U)} \mathbf{E}(T), \text{ s.t. } \operatorname{spt} T \subset X \times Y,\ T \in \mathcal{C}. \tag{6.32}$$

Depending on the kind of problem one wishes to solve, a different convex constraint set $\mathcal{C}$ should be considered. For example, in the case of variational problems with Dirichlet boundary conditions, we set

$$\mathcal{C} = \left\{T \;:\; \pi_{1\sharp} T = [\![X]\!],\ \partial T = S\right\}, \tag{6.33}$$

where $S \in \mathbf{M}_{n-1}(U)$ is a given boundary datum. In case of Neumann boundary conditions, one constrains the support of the boundary to be zero inside the domain

$$\mathcal{C} = \left\{T \;:\; \pi_{1\sharp} T = [\![X]\!],\ \operatorname{spt} \partial T \subset (\partial X) \times Y\right\}, \tag{6.34}$$

to exclude surfaces with holes, but allow the boundary to be freely chosen on $(\partial X) \times Y$. In case $n = N$, one can also consider the constraint set

$$\mathcal{C} = \left\{T \;: \pi_{1\sharp} T = [\![X]\!], \pi_{2\sharp} T = [\![Y]\!], \operatorname{spt} \partial T \subset \partial(X \times Y)\right\}, \tag{6.35}$$

where the additional pushforward constraint encourages bijectivity. Notice also the similarity of (6.32) together with (6.35) to the Kantorovich relaxation in optimal transport.

Existence of minimizing currents to a similar problem as (6.32) in a certain space of currents (*real flat chains*) is shown in [Fed74, §3.9]. For dimension $n = 1$ or codimension $N = 1$, the infimum is actually realized by a surface (*integral flat chain*) [Fed74, §5.10, §5.12]. An adaptation of such theoretical considerations to our setting and conditions under which the relaxation is tight in the scenario $n > 1$, $N > 1$ is a major open challenge and left for future work.

6.5 Discrete Formulation

In this section we present an implementation of the continuous model (6.32) using discrete exterior calculus [Hir03]. We will base our discretization on cubes since they are easy to work with in high dimensions, but one could also use simplices. To define cubical meshes, we adopt some notations from computational homology [KMM06].

Definition 3 (Elementary interval and cube)**.** *An* ***elementary interval*** *is an interval $I \subset \mathbf{R}$ of the form $I = [l, l+1]$ or $I = \{l\}$ for $l \in \mathbf{Z}$. Intervals that consist of a single point are* ***degenerate****. An* ***elementary cube*** *is given by a product $\kappa = I_1 \times \cdots \times I_d$, where each I_i is an elementary interval. The set of elementary cubes in $\mathbf{R}^d$ is denoted by K^d.*

For $\kappa \in K^d$, denote by $\dim \kappa \in \{1, \cdots, d\}$ the number of nondegenerate intervals. We denote $\mathbf{i}(\kappa) \in I(d, \dim \kappa)$ as the multi-index referencing the nondegenerate intervals.

Definition 4 (Cubical set)**.** *A set $Q \subset \mathbf{R}^d$ is a* ***cubical set*** *if it can be written as a finite union of elementary cubes.*

Let $K_k^d(Q) = \{\kappa \in K^d : \kappa \subset Q, \dim \kappa = k\}$ be the set of k-dimensional cubes contained in $Q \subset \mathbf{R}^d$. A map $\phi : Q \to X \times Y$ will transform the cubical set to our domain. As we work with images, it will just be a mesh spacing, i.e., we set $\phi(z) = (h_1 z_1, \cdots, h_d z_d)$.

Definition 5 (k-chains, k-cochains)**.** *We denote the space of finite formal sums of elements in $K_k^d(Q)$ with real coefficients as $\mathcal{C}_k(Q)$, called (real) k-**chains**. We denote the dual as $\mathcal{C}_k(Q)^* = \mathcal{C}^k(Q)$ and call the elements k-**cochains**.*

Definition 6 (Boundary)**.** *For $\kappa \in K_k^d(Q)$, denote the primary faces obtained by collapsing the j-th non-degenerate interval to the lower respectively upper boundary as $\kappa_j^-, \kappa_j^+ \in K_{k-1}^d$. The* ***boundary*** *of an elementary cube $\kappa \in K_k^d(Q)$ is the $(k-1)$-chain,*

$$\partial \kappa = \sum_{j=1}^{k} (-1)^{j-1} (\kappa_j^+ - \kappa_j^-) \in \mathcal{C}_{k-1}(Q). \tag{6.36}$$

The ***boundary operator*** *is given by the extension to a linear map $\partial : \mathcal{C}_k(Q) \to \mathcal{C}_{k-1}(Q)$.*

A k-chain $T = \sum_\kappa T_\kappa \kappa \in \mathcal{C}_k(Q)$ can be identified with a k-current $T' \in \mathcal{D}_k(U)$ by $T' = \sum_\kappa T_\kappa \phi_\sharp [\![\kappa]\!]$. The above discrete notion of boundary is defined in analogy to the continuous definition (6.24).

In our discretization, we will use the dual representation of the lifted energy from Prop. 17. To implement differential forms, we introduce an interpolation operator.

Definition 7 (Whitney map)**.** *The* ***Whitney map*** *extends a k-cochain ω to a k-form $(\boldsymbol{W}\omega) : X \times Y \to \Lambda^k \mathbf{R}^d$,*

$$(\boldsymbol{W}\omega)(x) = \sum_{\kappa \in K_k^d(Q)} \omega_\kappa \widehat{\boldsymbol{W}}(\phi^{-1}(x), \kappa), \tag{6.37}$$

where $\omega_\kappa \in \mathbf{R}$ are the coefficients of the k-cochain,

$$\widehat{\boldsymbol{W}}(x, \kappa) = \mathrm{d}x_{\mathbf{i}(\kappa)} \prod_{i \in \overline{\mathbf{i}(\kappa)}} \max\{0, 1 - |x_i - I_i(\kappa)|\}, \tag{6.38}$$

and $I_i(\kappa) \in \mathbf{Z}$ is the element in the degenerate interval.

Interestingly, the Whitney map (for simplicial meshes) first appeared in [Whi57, Eq. 27.12] but specializes to lowest-order Raviart-Thomas [RT77] ($k = 2, d = 3$) and Nédélec [Néd80] elements (for $k = 1$, $d = 3$), see [AA14; AFW06]. Differential forms of type (6.37) are called Whitney forms.

We also define a weighted inner product $\langle \cdot, \cdot \rangle_\phi$ between chains and cochains by plugging the Whitney form associated to the k-cochain into the current corresponding to the k-chain. As both are constant on each k-cube, a quick calculation shows: $\langle T, \omega \rangle_\phi = \sum_\kappa T_\kappa \omega_\kappa \mathcal{H}^k(\phi(\kappa))$, where $\mathcal{H}^k(\phi(\kappa))$ is simply the volume of the k-cube under the mesh spacing ϕ.

Using the dual representation (6.30), and approximating the current by a k-chain and the differential forms with k-cochains we arrive at the following finite-dimensional convex-concave saddle-point problem on $Q \subset \mathbf{R}^{n+N}$:

$$\begin{aligned} &\min_{T \in \mathcal{C}_n(Q)} \max_{\substack{\omega \in \mathcal{C}^n(Q) \\ \varphi \in \mathcal{C}^{n-1}(Q)}} \langle T, \omega \rangle_\phi + \langle \partial T - S, \varphi \rangle_\phi, \\ &\text{subject to} \quad \pi_{1\sharp} T = \mathbf{1}, \boldsymbol{W}\omega \in \mathcal{K}, \\ &\text{potentially} \quad \pi_{2\sharp} T = \mathbf{1} \text{ in case } n = N. \end{aligned} \tag{6.39}$$

$S \in \mathcal{C}_{n-1}(Q)$ is a given boundary datum, for free boundary conditions we replace the inner product $\langle S, \varphi \rangle$ with an indicator function $S : \mathcal{C}^{n-1} \to \overline{\mathbf{R}}$ forcing φ to be zero on the boundary. The pushforwards $\pi_{1\sharp}$, $\pi_{2\sharp}$ are linear constraints on the coefficients of the k-chain T.

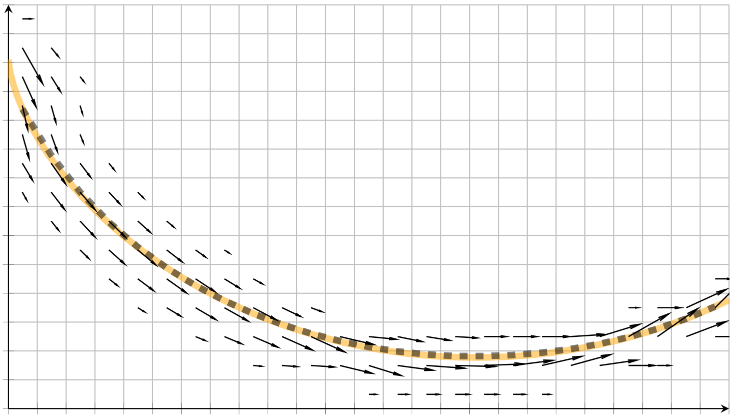


Figure 6.3: Minimization of the *Brachistochrone* energy on a 25 × 14 cubical set (gray squares). The proposed discretization yields a diffuse current (black vector field), whose center of mass (black, dashed) however is faithful to the analytical cycloid solution (orange) far beyond the mesh accuracy.

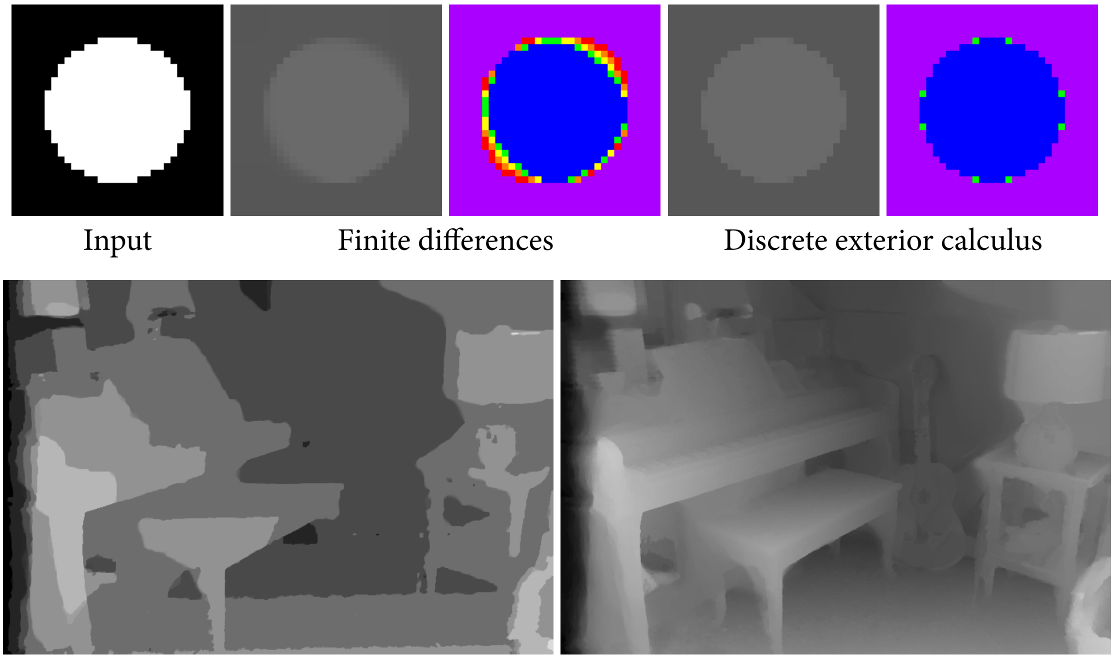
6.6 Numerical Implementation

In practice we solve (6.39) with the first order primal-dual algorithm [CP11a]. For the local constraints $\boldsymbol{W}\omega \in \mathcal{K}$ usually no closed form projection exists. In some situations ($N = 1$) they can be implemented exactly, see [Möl+16; MC17]. In the general setting, we resort to implementing them at sampled points. To enforce the constraint $\boldsymbol{W}\omega \in \mathcal{K}$ at samples $Z = \{z_1, z_2, \cdots\} \subset X \times Y$ we add another primal variable $\lambda : Z \to \Lambda_n \mathbf{R}^{n+N}$ and the additional term $\sum_{z \in Z} \Psi^{**}(z, \lambda(z)) - \langle \lambda(z), (\boldsymbol{W}\omega)(z) \rangle$ to the saddle-point formulation (6.39).

Finally, one requires the proximal operator of the perspective function Ψ^{**}. These can be implemented using epigraphical projections as in [Poc+10]. For an overview over proximal operators of perspective functions we refer to [CM18].

6.6.1 Properties of the discretization

As a first example we solve the Brachistochrone [Ber96], arguably the first variational approach. The cost is given by $c(x, y, \xi) = \sqrt{\frac{1+\xi^2}{2gy}}$ where $g > 0$ is the gravitational constant. Dirichlet boundary conditions are enforced using the boundary operator. In Fig. 6.3 we show the resulting current, which concentrates on the graph of the closed-form solution to the problem, which is a cycloid. The unlifted result is obtained by taking the center of mass of the first component $T^{\bar{0},0}$ of the current by summing over the horizontal edges in the 1-chain. The obtained result



Input

Finite differences

Discrete exterior calculus

Finite differences

Discrete exterior calculus

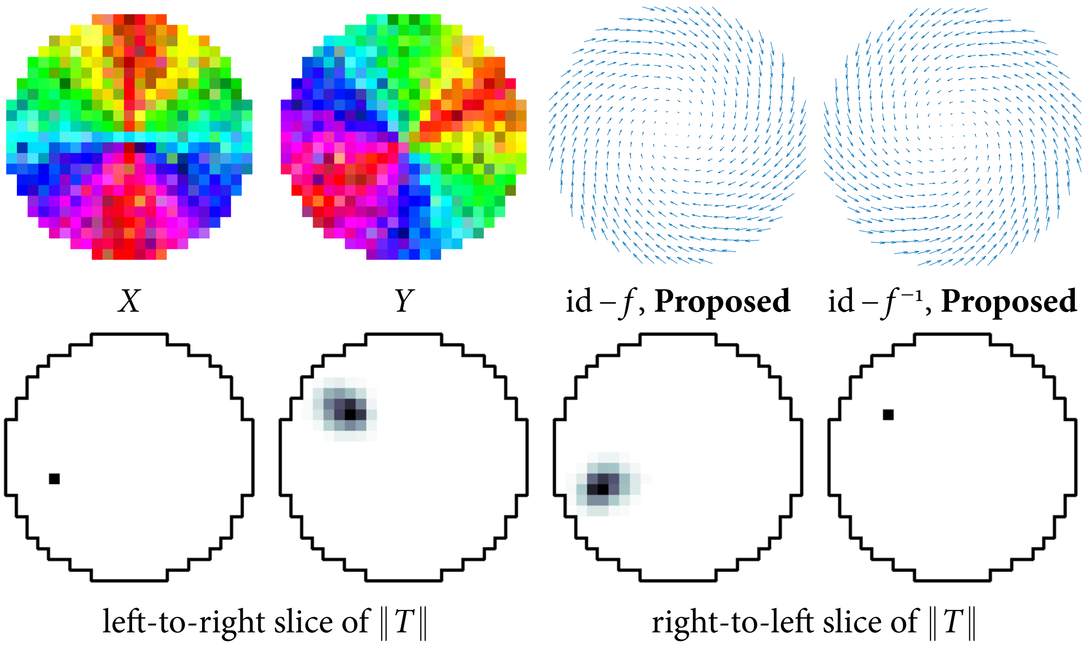
Figure 6.4: Total variation minimization. **Top:** The proposed DEC discretization yields solutions with better isotropy and sharper discontinuities. **Bottom:** In that stereo matching example, we enforce the continuous constraints $\boldsymbol{W}\omega \in \mathcal{K}$ between the discretization points (here 8 *labels*), which leads to more precise (sublabel-accurate) solutions compared to the naive finite-difference approach.

nearly coincides with the exact cycloid. Instead, solutions from MRF approaches would invariably be confined to the vertices or edges of the rather coarse grid.

In Fig. 6.4 we solve total variation regularized problems which corresponds to setting $c(x, y, \xi) = \rho(x, y) + |\xi|$ for some data ρ. The data is either a quadratic or a stereo matching cost in that example. The proposed approach based on discrete exterior calculus has better isotropy and leads to sharper discontinuities than the common forward difference approach used in literature. Furthermore, by enforcing the constraints $\boldsymbol{W}\omega \in \mathcal{K}$ also between the discretization points one can achieve "sublabel-accurate" results as demonstrated in the stereo matching example.

6.6.2 Global registration

As an example of $n > 1$, $N > 1$ with polyconvex regularization, we tackle the problem of orientation preserving diffeomorphic registration between two shapes $X, Y \subset \mathbf{R}^2$ with boundary. We use the cost $c(x, y, \xi) = (\rho(x, y) + \varepsilon)\sqrt{\det(I + \xi^\top \xi)}$, which penalizes the surface area in the product space and favors local isometry. The parameter $\varepsilon > 0$ models the trade-off between data and smoothness. In the example considered in Fig. 6.5 the data is given by $\rho(x, y) = \|I_1(x) - I_2(y)\|$, where I_1, I_2



X

Y

$\mathrm{id} - f$, **Proposed**

$\mathrm{id} - f^{-1}$, **Proposed**

left-to-right slice of $\|T\|$

right-to-left slice of $\|T\|$

Figure 6.5: Global registration of X and Y. **Top:** Our method yields dense pointwise correspondences that are smooth in both directions and correspond to the correct transformation. **Bottom:** 2-D slices through the 4-D density $\|T\|$ at the single black pixel. We empirically verify (also at the other points) that the current concentrated near a surface, therefore the recovered solution is near the global minimum of the original nonconvex problem.

are the shown color images. A polyconvex extension of the above cost, which is large for non-simple vectors is given by the (weighted) mass norm (6.4). The 4-D cubical set Q is the product space between the two shapes X and Y, which are given as quads (pixels). We impose the constraints $\boldsymbol{W}\omega \in \mathcal{K}$ at the 16 vertices of each four dimensional hypercube. The proximal operator of the mass norm is computed as in [WC16]. Note that the required 4×4 real Schur decomposition can be reduced to a 2×2 SVD using a few Givens rotations, see [WG76]. We further impose $T^{\bar{0},0} \geq 0$ and $T^{0,\bar{0}} \geq 0$, and boundary conditions ensure that ∂X is matched to ∂Y. Bijectivity of the matching is encouraged by the pushforward constraints $\pi_{1\sharp} T = \mathbf{1}$, $\pi_{2\sharp} T = \mathbf{1}$. After solving (6.39) we obtain the final pointwise correspondences $f : X \to Y$ from the 2-chain $T \in \mathcal{C}_2(Q)$ by taking its center of mass.

In Fig. 6.5 we visualize $f(x) = \sum_y y\,|(\boldsymbol{W}T)(x,y)|$, $f^{-1}(y) = \sum_x x\,|(\boldsymbol{W}T)(x,y)|$. As one can see, the maps f and f^{-1} are smooth and inverse to each other. Despite $n > 1$, $N > 1$, the current apparently concentrated near a surface (see bottom of Fig. 6.5) and the computed solutions are therefore near the *global optimum* of the original nonconvex problem.

6.7 Discussion and Limitations

In this work, we introduced a novel approach to vectorial variational problems based on geometric measure theory, along with a natural discretization using concepts from discrete exterior calculus. Though observed in practice, we do not have theoretical guarantees that the minimizing current will concentrate on a surface. In case of multiple global solutions, one might get a convex combination of minimizers. Some mechanism to select an extreme point of the convex solution set would therefore be desirable. The main drawback over MRFs, for which very efficient solvers exist [Kap+13], is that we had to resort to the generic algorithm [CP11a] with $\mathcal{O}(1/\varepsilon)$ convergence. Yet, solutions with high numerical accuracy are typically not required in practice and the algorithm parallelizes well on GPUs. To conclude, we believe that the present work is a step towards making continuous approaches an attractive alternative to MRFs, especially in scenarios where faithfulness to certain geometric properties of the underlying continuous model is desirable.

Chapter 7

Flat Metric Minimization with Applications in Generative Modeling

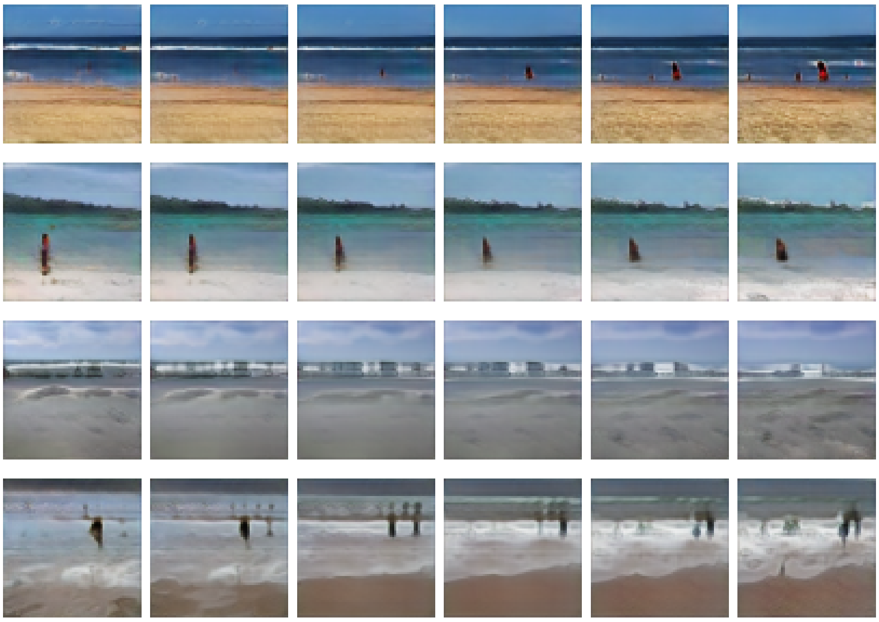


Figure 7.1: Discovering the arrow of time by training a generative model with the proposed formalism on the tinyvideos dataset [VPT16]. The approach we introduce allows one to learn latent representations which behave equivariantly to specified tangent vectors (here: difference of two successive video frames).

7.1 Introduction

This work is concerned with the problem of representation learning, which has important consequences for many tasks in artificial intelligence, cf. the work of [BCV13]. More specifically, our aim is to learn representations which behave equivariantly with respect to selected transformations of the data. Such variations are often known beforehand and could for example describe changes in stroke width or rotation of a digit, changes in viewpoint or lighting in a three-dimensional scene but also the *arrow of time* [Pic+14; Wei+18] in time-series, describing how a video changes from one frame to the next, see Fig. 7.1.

We tackle this problem by introducing a novel formalism based on *geometric measure theory* [Fed69], which we find to be interesting in itself. To motivate our application in generative modeling, recall the manifold hypothesis which states that the distribution of real-world data tends to concentrate nearby a low-dimensional manifold, see [FMN16] and the references therein. Under that hypothesis, a possible unifying view on prominent methods in unsupervised and representation learning such as generative adversarial networks (GANs) [Goo+14] and variational auto-encoders (VAEs) [KW14; RMW14] is the following: both approaches aim to approximate the true distribution concentrating near the manifold with a distribution on some low-dimensional latent space $\mathcal{Z} \subset \mathbf{R}^l$ that is pushed through a decoder or generator $g : \mathcal{Z} \to \mathcal{X}$ mapping to the (high-dimensional) data space $\mathcal{X} \subset \mathbf{R}^d$ [GPC17; Bot+17].

We argue that treating data as a distribution potentially ignores useful available geometric information such as orientation and tangent vectors to the data manifold. Such tangent vectors describe the aforementioned local variations or pertubations. Therefore we postulate that *data should not be viewed as a distribution but rather as a k-current.*

We postpone the definition of k-currents [Rha55] to Sec. 7.3, and informally think of them as distributions over k-dimensional oriented planes. For the limiting case $k = 0$, currents simply reduce to distributions in the sense of [Sch51] and positive 0-currents with unit mass are probability measures. A seminal work in the theory of currents was written by [FF60], which established compactness theorems for subsets of currents (*normal* and *integral currents*). In this paper, we will work in the space of normal k-currents with compact support in $\mathcal{X} \subset \mathbf{R}^d$, denoted by $\boldsymbol{N}_{k,\mathcal{X}}(\mathbf{R}^d)$.

Similarly as probabilistic models build upon f-divergences [CS+04], integral probability metrics [Sri+12] or more general optimal transportation related divergences [PC18; Fey+18], we require a sensible notion to measure "distance" between k-currents.

In this work, we will focus on the flat norm[1] due to [Whi57]. To be precise, we consider a scaled variant introduced and studied by [MV07; Vix+10]. This choice

[1] The terminology "flat" carries no geometrical significance and refers to Whitney's use of musical notation flat $|\cdot|^\flat$ and sharp $|\cdot|^\sharp$.

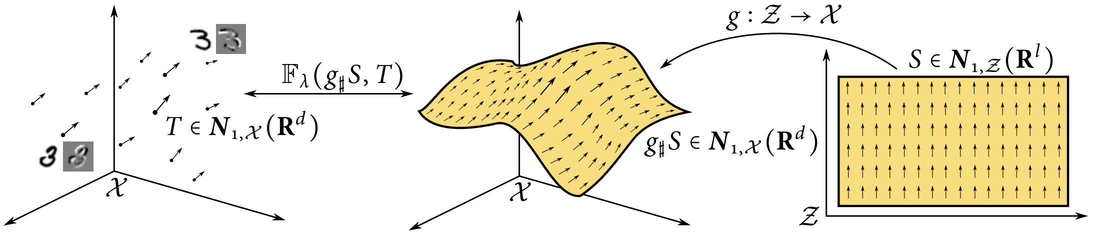

$\mathbb{F}_\lambda(g_\sharp S, T)$
$T \in \mathbf{N}_{1,\mathcal{X}}(\mathbf{R}^d)$
$\mathcal{X}$
$\mathcal{X}$
$g : \mathcal{Z} \to \mathcal{X}$
$g_\sharp S \in \mathbf{N}_{1,\mathcal{X}}(\mathbf{R}^d)$
$S \in \mathbf{N}_{1,\mathcal{Z}}(\mathbf{R}^l)$
$\mathcal{Z}$

Figure 7.2: **Illustration of the proposed idea.** We suggest the novel perspective to view observed data (here the MNIST dataset) as a k-current T, shown as the dots with attached arrows on the left. The arrows indicate the oriented tangent space, and we selected $k = 1$ to be rotational deformation. We propose to minimize the flat distance of T to the *pushforward* $g_\sharp S$ (shown in the middle) of a current S on a low-dimensional latent space $\mathcal{Z}$ (right) with respect to a "generator" map $g : \mathcal{Z} \to \mathcal{X}$. For 0-currents (no selected tangent vectors) and sufficiently large λ, the proposed "FlatGAN" formulation specializes to the Wasserstein GAN [ACB17].

is motivated in Sec. 7.4, where we show that the flat norm enjoys certain attractive properties similar to the celebrated Wasserstein distances. For example, it metrizes the weak*-convergence for normal currents.

A potential alternative to the flat norm are kernel metrics on spaces of currents [VG05; Gla+08]. These have been proposed for diffeomorphic registration, but kernel distances on distributions have also been sucessfully employed for generative modeling, see [Li+17]. Constructions similar to the Kantorovich relaxation in optimal transport but generalized to k-currents recently appeared in the context of convexifications for certain variational problems [MC19a].

7.2 Related Work

Our main idea is illustrated in Fig. 7.2, which was inspired from the optimal transportation point of view on GANs given by [GPC17].

Tangent vectors of the data manifold, either prespecified [Sim+92; Sim+98; Fra+03] or learned with a contractive autoencoder [Rif+11], have been used to train classifiers that aim to be *invariant* to changes relative to the data manifold. In contrast to these works, we use tangent vectors to learn interpretable representations and a generative model that aims to be *equivariant*. The principled introduction of tangent k-vectors into probabilistic generative models is one of our main contributions.

Various approaches to learning informative or disentangled latent representations in a completely unsupervised fashion exist [Sch92; Hig+16; Che+16; KM18]. Our approach is orthogonal to these works, as specifying tangent vectors further encourages informative representations to be learned. For example, our GAN formulation could be combined with a mutual information term as in InfoGAN [Che+16].

Our work is more closely related to semi-supervised approaches on learning disentangled latent representations, which similarly also require some form of knowledge of the underlying factors [HKW11; Den+17; Mat+16; Nar+17] and also to conditional GANs [MO14; OOS17]. However, the difference is the connection to geometric measure theory which we believe to be completely novel, and our specific FlatGAN formulation that seamlessly extends the Wasserstein GAN [ACB17], cf. Fig. 7.2.

Since the concepts we need from geometric measure theory are not commonly used in machine learning, we briefly review them in the following section.

7.3 Geometric Measure Theory

The book by [Fed69] is still the formidable, definitive reference on the subject. As a more accessible introduction we recommend [KP08] or [Mor16]. While our aim is to keep the manuscript self-contained, we invite the interested reader to consult Chapter 4 in [Mor16], which in turn refers to the corresponding chapters in the book of [Fed69] for more details.

7.3.1 Grassmann algebra

Notation. Denote $\{e_1, \cdots, e_d\}$ a basis of $\mathbf{R}^d$ with dual basis $\{\mathrm{d}x_1, \cdots, \mathrm{d}x_d\}$ such that $\mathrm{d}x_i : \mathbf{R}^d \to \mathbf{R}$ is the linear functional that maps every $x = (x_1, \cdots, x_d)$ to the i-th component x_i. For $k \leq d$, denote $I(d,k)$ as the ordered multi-indices $\mathbf{i} = (i_1, \cdots, i_k)$ with $1 \leq i_1 < \cdots < i_k \leq d$.

One can multiply vectors in $\mathbf{R}^d$ to obtain a new object:

$$\xi = v_1 \wedge \cdots \wedge v_k, \tag{7.1}$$

called a k-vector ξ in $\mathbf{R}^d$. The wedge (or exterior) product $\wedge$ is characterized by multilinearity

$$\begin{aligned} c v_1 \wedge v_2 = v_1 \wedge c v_2 = c(v_1 \wedge v_2), \quad \text{for } c \in \mathbf{R}, \\ (u_1 + v_1) \wedge (u_2 + v_2) = \\ u_1 \wedge u_2 + u_1 \wedge v_2 + v_1 \wedge u_2 + v_1 \wedge v_2, \end{aligned} \tag{7.2}$$

and it is alternating

$$u \wedge v = -v \wedge u, \qquad u \wedge u = 0. \tag{7.3}$$

In general, any k-vector can be written as

$$\xi = \sum_{\mathbf{i} \in I(d,k)} a_{\mathbf{i}} \cdot e_{i_1} \wedge \cdots \wedge e_{i_k} = \sum_{\mathbf{i} \in I(d,k)} a_{\mathbf{i}} \cdot e_{\mathbf{i}}, \tag{7.4}$$

for coefficients $a_\mathbf{i} \in \mathbf{R}$. The vector space of k-vectors is denoted by $\Lambda_k \mathbf{R}^d$ and has dimension $\binom{d}{k}$. We define for two k-vectors $v = \sum_\mathbf{i} a_\mathbf{i} e_\mathbf{i}$, $w = \sum_\mathbf{i} b_\mathbf{i} e_\mathbf{i}$ an inner product $\langle v, w \rangle = \sum_\mathbf{i} a_\mathbf{i} b_\mathbf{i}$ and the Euclidean norm $|v| = \sqrt{\langle v, v \rangle}$.

A simple (or decomposable) k-vector is any $\xi \in \Lambda_k \mathbf{R}^d$ that can be written using products of 1-vectors. Simple k-vectors such as (7.1) are uniquely determined by the k-dimensional space spanned by the $\{v_i\}$, their orientation and the norm $|v|$ corresponding to the area of the parallelotope spanned by the $\{v_i\}$. Simple k-vectors with unit norm can therefore be thought of as oriented k-dimensional subspaces and the rules (7.2)-(7.3) can be thought of as equivalence relations.

It turns out that the inner product of two simple k-vectors can be computed by the $k \times k$-determinant

$$\langle w_1 \wedge \cdots \wedge w_k, v_1 \wedge \cdots \wedge v_k \rangle = \det \left(W^\top V \right), \tag{7.5}$$

where the columns of $W \in \mathbf{R}^{d \times k}$, $V \in \mathbf{R}^{d \times k}$ contain the individual 1-vectors. This will be useful later for our practical implementation.

Not all k-vectors are simple. An illustrative example is $e_1 \wedge e_2 + e_3 \wedge e_4 \in \Lambda_2 \mathbf{R}^4$, which describes two 2-dimensional subspaces in $\mathbf{R}^4$ intersecting only at zero.

The dual space of $\Lambda_k \mathbf{R}^d$ is denoted as $\Lambda^k \mathbf{R}^d$, and its elements are called k-covectors. They are similarly represented as (7.4) but with dual basis $\mathrm{d}x_\mathbf{i}$. Analogously to the previous page, we can define an inner product between k-vectors and k-covectors. Next to the Euclidean norm $|\cdot|$, we define two additional norms due to [Whi57].

Definition 8 (Mass and comass)**.** *The comass norm defined for k-covectors* $w \in \Lambda^k \mathbf{R}^n$ *is given by*

$$\|w\|^* = \sup \left\{ \langle w, v \rangle : v \textit{ is simple }, |v| = 1 \right\}, \tag{7.6}$$

and the mass norm for $v \in \Lambda_k \mathbf{R}^n$ *is given by*

$$\begin{aligned} \|v\| &= \sup \left\{ \langle v, w \rangle : \|w\|^* \leq 1 \right\} \\ &= \inf \left\{ \sum_i |\xi_i| : \xi_i \textit{ are simple}, v = \sum_i \xi_i \right\}. \end{aligned} \tag{7.7}$$

The mass norm is by construction the largest norm that agrees with the Euclidean norm on simple k-vectors. For the non-simple 2-vector from before, we compute

$$\|e_1 \wedge e_2 + e_3 \wedge e_4\| = 2, \; |e_1 \wedge e_2 + e_3 \wedge e_4| = \sqrt{2}. \tag{7.8}$$

Interpreting the non-simple vector as two tangent planes, we see that the mass norm gives the correct area, while the Euclidean norm underestimates it. The comass $\|\cdot\|^*$ will be used later to define the mass of currents and the flat norm.

7.3.2 Differential forms

In order to define currents, we first need to introduce differential forms. A differential k-form is a k-covectorfield $\omega : \mathbf{R}^d \to \Lambda^k \mathbf{R}^d$. The support $\operatorname{spt} \omega$ is defined as the closure of the set $\{x \in \mathbf{R}^d : \omega(x) \neq 0\}$.

Differential forms allow one to perform coordinate-free integration over oriented manifolds. Given some manifold $\mathcal{M} \subset \mathbf{R}^d$, possibly with boundary, an *orientation* is a continuous map $\tau_{\mathcal{M}} : \mathcal{M} \to \Lambda_k \mathbf{R}^d$ which assigns to each point a simple k-vector with unit norm that spans the tangent space at that point. Integration of a differential form over an oriented manifold $\mathcal{M}$ is then defined by:

$$\int_{\mathcal{M}} \omega = \int_{\mathcal{M}} \langle \omega(x), \tau_{\mathcal{M}}(x) \rangle \, \mathrm{d}\mathcal{H}^k(x), \tag{7.9}$$

where the second integral is the standard Lebesgue integral with respect to the k-dimensional Hausdorff measure $\mathcal{H}^k$ restricted to $\mathcal{M}$, i.e., $(\mathcal{H}^k \llcorner \mathcal{M})(A) = \mathcal{H}^k(A \cap \mathcal{M})$. The k-dimensional Hausdorff measure assigns to sets in $\mathbf{R}^d$ their k-dimensional volume, see Chapter 2 in [Mor16] for a nice illustration. For $k = d$ the Hausdorff measure coincides with the Lebesgue measure.

The exterior derivative of a differential k-form is the $(k+1)$-form $d\omega : \mathbf{R}^d \to \Lambda^{k+1}\mathbf{R}^d$ defined by

$$\langle d\omega(x), v_1 \wedge \cdots \wedge v_{k+1} \rangle = \lim_{h \to 0} \frac{1}{h^{k+1}} \int_{\partial P} \omega, \tag{7.10}$$

where ∂P is the oriented boundary of the parallelotope spanned by the $\{hv_i\}$ at point x. The above definition is for example used in the textbook of [HH15]. To get an intuition, note that for $k = 0$ this reduces to the familiar directional derivative $\langle d\omega(x), v_1 \rangle = \lim_{h \to 0} \frac{1}{h} \left(\omega(x + hv_1) - \omega(x) \right)$. In case $\omega : \mathbf{R}^d \to \Lambda^k \mathbf{R}^d$ is sufficiently smooth, the limit in (7.10) is given by

$$\begin{aligned} &\langle d\omega(x), v_1 \wedge \cdots \wedge v_{k+1} \rangle = \\ &\sum_{i=1}^{k+1} (-1)^{i-1} \nabla_x \langle \omega(x), v_1 \wedge \cdots \wedge \hat{v}_i \wedge \cdots \wedge v_k \rangle \cdot v_i, \end{aligned} \tag{7.11}$$

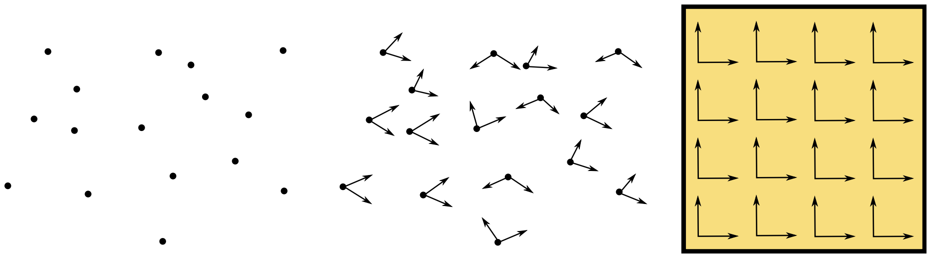
where $\hat{v}_i$ means that the vector v_i is omitted. The formulation (7.11) will be used in the practical implementation. Interestingly, with (7.9) and (7.10) in mind, Stokes' theorem

$$\int_{\mathcal{M}} d\omega = \int_{\partial \mathcal{M}} \omega, \tag{7.12}$$

becomes almost obvious, as (informally speaking) integrating (7.10) one obtains (7.12) since the oppositely oriented boundaries of neighbouring parallelotopes cancel each other out in the interior of $\mathcal{M}$.

To define the pushforward of currents which is central to our formulation, we require the pullback of differential forms. The pullback $g^\sharp \omega : \mathbf{R}^l \to \Lambda^k \mathbf{R}^l$ by a map $g : \mathbf{R}^l \to \mathbf{R}^d$ of the k-form $\omega : \mathbf{R}^d \to \Lambda^k \mathbf{R}^d$ is given by

$$\langle g^\sharp \omega, v_1 \wedge .. \wedge v_k \rangle = \langle \omega \circ g, D_{v_1} g \wedge .. \wedge D_{v_k} g \rangle, \tag{7.13}$$



(a) $\sum_i \delta_{x_i}$ **(b)** $\sum_i \delta_{x_i} \wedge T_i$ **(c)** $\mathcal{H}^2 \llcorner [0,1]^2 \wedge e_{12}$

Figure 7.3: Example of a 0-current **(a)**, and 2-currents **(b)**, **(c)**.

where $D_{v_i} g := \nabla g \cdot v_i$ and $\nabla g \in \mathbf{R}^{d \times l}$ is the Jacobian. We will also require (7.13) for the practical implementation.

7.3.3 Currents

We have now the necessary tools to define currents and the required operations on them, which will be defined through duality with differential forms. Consider the space of compactly supported and smooth k-forms in $\mathbf{R}^d$ which we denote by $\mathcal{D}^k(\mathbf{R}^d)$. When furnished with an appropriate topology (cf. §4.1 in [Fed69] for the details) this is a locally convex topological vector space. k-currents are continuous linear functionals on smooth, compactly supported differential forms, i.e., elements from the topological dual space $\mathcal{D}_k(\mathbf{R}^d) = \mathcal{D}^k(\mathbf{R}^d)'$. Some examples for currents are given in Fig. 7.3. The 0-current in **(a)** could be an empirical data distribution, and the 2-current in **(b)** represents the data distribution with a two dimensional oriented tangent space at each data point. The 2-current in **(c)** simply represents the set $[0,1]^2$ as an oriented manifold, its action on a differential form is given as in (7.9).

A natural notion of convergence for currents is given by the weak* topology:

$$T_i \overset{*}{\rightharpoonup} T \text{ iff } T_i(\omega) \to T(\omega), \text{ for all } \omega \in \mathcal{D}^k(\mathbf{R}^d). \tag{7.14}$$

The support of a current $T \in \mathcal{D}_k(\mathbf{R}^d)$, $\operatorname{spt} T$, is the complement of the largest open set, so that when testing T with compactly supported forms on that open set the answer is zero. Currents with compact support are denoted by $\mathcal{E}_k(\mathbf{R}^d)$. The boundary operator $\partial : \mathcal{D}_k(\mathbf{R}^d) \to \mathcal{D}_{k-1}(\mathbf{R}^d)$ is defined using exterior derivative

$$\partial T(\omega) = T(d\omega), \tag{7.15}$$

and Stokes' theorem (7.12) ensures that this coincides with the intuitive notion of boundary for currents which are represented by integration over manifolds in the sense of (7.9).

The pushforward of a current is defined using the pullback

$$g_\sharp T(\omega) = T(g^\sharp \omega), \tag{7.16}$$

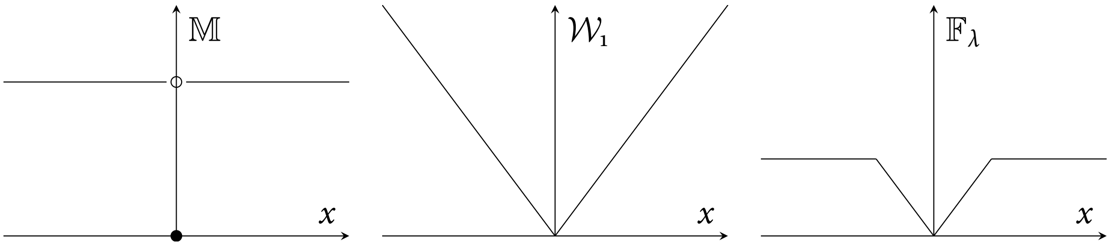


Figure 7.4: Illustration of distances between 0-currents on the example of two Dirac measures δ_x, δ_0. The flat metric $\mathbb{F}_\lambda$ has the following advantages: unlike the mass $\mathbb{M}$ it is continuous, and unlike Wasserstein-1 it easily generalizes to k-currents (see Fig. 7.5).

where the intuition is that the pushforward transforms the current with the map g, see the illustration in Fig. 7.2.

The mass of a current $T \in \mathcal{D}_k(\mathbf{R}^d)$ is given by

$$\mathbb{M}(T) = \sup\left\{T(\omega) : \|\omega(x)\|^* \le 1\right\}. \tag{7.17}$$

If the current T is an oriented manifold then the mass $\mathbb{M}(T)$ is the *volume* of that manifold. One convenient way to construct k-currents, is by combining a smooth k-vectorfield $\xi : \mathbf{R}^d \to \Lambda_k \mathbf{R}^d$ with a Radon measure μ:

$$(\mu \wedge \xi)(\psi) = \int \langle \xi, \psi \rangle \, \mathrm{d}\mu, \text{ for all } \psi \in \mathcal{D}^k(\mathbf{R}^d). \tag{7.18}$$

A concrete example is illustrated in Fig. 7.3 **(b)**, where given samples $\{x_1, \cdots, x_N\} \subset \mathbf{R}^d$ and tangent 2-vectors $\{T_1, \cdots, T_N\} \subset \Lambda_2 \mathbf{R}^d$ a 2-current is constructed.

For currents with finite mass there is a measure $\|T\|$ and a map $\vec{T} : \mathbf{R}^d \to \Lambda_k \mathbf{R}^d$ with $\|\vec{T}(\cdot)\| = 1$ almost everywhere so that we can represent it by integration as follows:

$$T(\omega) = \int \langle \omega(x), \vec{T}(x) \rangle \, \mathrm{d}\|T\|(x) = \|T\| \wedge \vec{T}(\omega). \tag{7.19}$$

Another perspective is that finite mass currents are simply k-vector valued Radon measures. Currents with finite mass and finite boundary mass are called *normal currents* [FF60]. The space of normal currents with support in a compact set $\mathcal{X}$ is denoted by $\boldsymbol{N}_{k,\mathcal{X}}(\mathbf{R}^d)$.

7.4 The Flat Metric

As indicated in Fig. 7.2, we wish to fit a current $g_\sharp S$ that is the pushforward of a low-dimensional latent current S to the current T given by the data. A more meaningful norm on currents than the mass $\mathbb{M}$ turns out to be the flat norm.

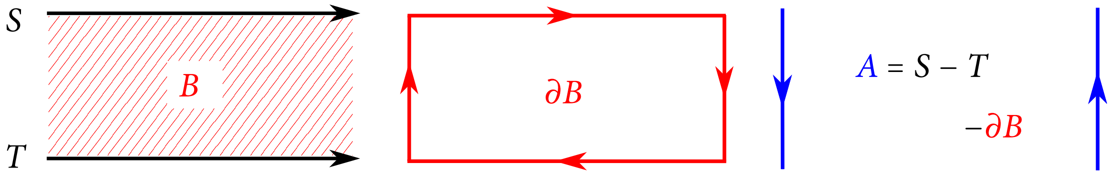


Figure 7.5: The flat metric $\mathbb{F}_\lambda(S, T)$ is given an optimal decomposition $S - T = A + \partial B$ into a k-current A and the boundary of a $(k+1)$-current B with minimal weighted mass $\lambda \mathbb{M}(A) + \mathbb{M}(B)$. An intuition is that $\lambda \mathbb{M}(A)$ is a penalty that controls how closely ∂B should approximate $S - T$, while $\mathbb{M}(B)$ is the $(k+1)$-dimensional volume of B.

Definition 9 (Flat norm and flat metric). *The flat norm with scale*[2] $\lambda > 0$ *is defined for any k-current $T \in \mathcal{D}_k(\mathbf{R}^d)$ as*

$$\mathbb{F}_\lambda(T) = \sup\Big\{T(\omega) \mid \omega \in \mathcal{D}^k(\mathbf{R}^d),\ \textit{with} \\ \|\omega(x)\|^* \leq \lambda, \|d\omega(x)\|^* \leq 1,\ \textit{for all } x\Big\}. \tag{7.20}$$

For $\lambda = 1$ we simply write $\mathbb{F}(\cdot) \equiv \mathbb{F}_1(\cdot)$ and $\mathbb{F}_\lambda(S, T) = \mathbb{F}_\lambda(S - T)$ will be denoted as the flat metric.

The flat norm also has a primal formulation

$$\mathbb{F}_\lambda(T) = \min_{B \in \mathcal{E}_{k+1}(\mathbf{R}^d)} \lambda \mathbb{M}(T - \partial B) + \mathbb{M}(B) \tag{7.21}$$

$$= \min_{T = A + \partial B} \lambda \mathbb{M}(A) + \mathbb{M}(B), \tag{7.22}$$

where the minimum in (7.21)–(7.22) can be shown to exist, see §4.1.12 in [Fed69]. The flat norm is finite if T is a normal current and it can be verified that it is indeed a norm.

To get an intuition, we compare the flat norm to the mass (7.17) and the Wasserstein-1 distance in Fig. 7.4 on the example of Dirac measures δ_x, δ_0. The mass $x \mapsto \mathbb{M}(\delta_x - \delta_0)$ is discontinuous and has zero gradient and is therefore unsuitable as a distance between currents. While the Wasserstein-1 metric $x \mapsto \mathcal{W}(\delta_x, \delta_0)$ is continuous in x, it does not easily generalize from probability measures to k-currents. In contrast, the flat metric $x \mapsto \mathbb{F}_\lambda(\delta_x, \delta_0)$ has a meaningful geometric interpretation also for arbitrary k-currents. In Fig. 7.5 we illustrate the flat norm for two 1-currents. In that figure, if S and T are of length one and are ε apart, then $\mathbb{F}_\lambda(S, T) \leq (1 + 2\lambda)\varepsilon$ which converges to zero for $\varepsilon \to 0$.

Note that for 0-currents, the flat norm (7.20) is strongly related to the Wasserstein-1 distance except for the additional constraint on the dual variable $\|\omega(x)\|^* \leq \lambda$,

[2] We picked a different convention for λ as in [MV07], where it bounds the other constraint, to emphasize the connection to the Wasserstein-1 distance.

which in the example of Fig. 7.4 controls the truncation cutoff. Notice also the similarity of (7.21) to the Beckmann formulation of the Wasserstein-1 distance [Bec52; San15], with the difference being the implementation of the "divergence constraint" with a soft penalty $\lambda\mathbb{M}(T - \partial B)$. Considering the case $\lambda = \infty$ as in the Wasserstein distance is problematic in case we have $k > 0$, since not every current $T \in \mathcal{D}_k(\mathbf{R}^n)$ is the boundary of a $(k+1)$-current, see the example above in Fig. 7.5.

The following proposition studies the effect of the scale parameter $\lambda > 0$ on the flat norm.

Proposition 18. *For any $\lambda > 0$, the following relation holds*

$$\min\{1, \lambda\} \cdot \mathbb{F}(T) \leq \mathbb{F}_\lambda(T) \leq \max\{1, \lambda\} \cdot \mathbb{F}(T), \tag{7.23}$$

meaning that $\mathbb{F}$ and $\mathbb{F}_\lambda$ are equivalent norms.

Proof. By a result of [MV07] we have the interesting relation

$$\mathbb{F}_\lambda(T) = \lambda^k \mathbb{F}(d_{\lambda^{-1}\sharp}T), \tag{7.24}$$

where d_λ is the λ-dilation. Using the bound $\mathbb{F}(f_\sharp T) \leq \sup\{\mathrm{Lip}(f)^k, \mathrm{Lip}(f)^{k+1}\}\mathbb{F}(T)$, §4.1.14 in [Fed69], and the fact that $\mathrm{Lip}(d_{\lambda^{-1}}) = \lambda^{-1}$, one inequality directly follows. For the other side, notice that

$$\begin{aligned} \mathbb{F}(T) &= \mathbb{F}(d_{\lambda\sharp}d_{\lambda^{-1}\sharp}T) = \mathbb{F}_{\lambda^{-1}}(d_{\lambda^{-1}\sharp}T)\lambda^k \\ &\leq \max\{1, \lambda^{-1}\}\mathbb{F}(d_{\lambda^{-1}\sharp}T)\lambda^k \\ &= \max\{1, \lambda^{-1}\}\mathbb{F}_\lambda(T). \end{aligned} \tag{7.25}$$

and dividing by $\max\{1, \lambda^{-1}\}$ yields the result. □

The importance of the flat norm is due to the fact that it metrizes the weak*-convergence (7.14) on compactly supported normal currents with uniformly bounded mass and boundary mass.

Proposition 19. *Let $\mathcal{X} \subset \mathbf{R}^d$ be a compact set and $c > 0$ some fixed constant. For a sequence $\{T_j\} \subset \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ with $\mathbb{M}(T_j) + \mathbb{M}(\partial T_j) < c$ we have that:*

$$\mathbb{F}_\lambda(T, T_j) \to 0 \quad \textit{if and only if} \quad T_j \overset{*}{\rightharpoonup} T. \tag{7.26}$$

Proof. Due to Prop. 18 it is enough to consider the case $\lambda = 1$, which is given by Corollary 7.3 in the paper of [FF60]. □

7.5 Flat Metric Minimization

Motivated by the theoretical properties of the flat metric shown in the previous section, we consider the following optimization problem:

$$\min_{\theta \in \Theta} \mathbb{F}_\lambda(g_{\theta\sharp}S, T), \tag{7.27}$$

where $S \in \boldsymbol{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ and $T \in \boldsymbol{N}_{k,\mathcal{X}}(\mathbf{R}^d)$. We will assume that $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ is parametrized with parameters in a compact set $\Theta \subset \mathbf{R}^n$ and write $g_\theta : \mathcal{Z} \to \mathcal{X}$ to abbreviate $g(\cdot, \theta)$ for some $\theta \in \Theta$. We need the following assumption to be able to prove the existence of minimizers for the problem (7.27).

Assumption 1. *The map $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ is smooth in z with uniformly bounded derivative. Furthermore, we assume that $g(z, \cdot)$ is locally Lipschitz continuous and that the parameter set $\Theta \subset \mathbf{R}^n$ is compact.*

Under this assumption, we will show that the objective in (7.27) is Lipschitz continuous. This will in turn guarantee existence of minimizers, as the domain is assumed to be compact.

Proposition 20. *Let $S \in \boldsymbol{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$, $T \in \boldsymbol{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ be normal currents with compact support. If the pushforward map $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ fulfills Assumption 1, then the function $\theta \mapsto \mathbb{F}_\lambda(g_{\theta\,\sharp} S, T)$ is Lipschitz continuous and hence differentiable almost everywhere.*

Proof. In Appendix A.4. □

7.5.1 Application to generative modeling

We now turn towards our considered application illustrated in Fig. 7.2. There, we denote by $k \geq 0$ the number of tangent vectors we specify at each sample point. The latent current $S \in \boldsymbol{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ is constructed by combining a probability distribution $\mu \in \boldsymbol{N}_{0,\mathcal{Z}}(\mathbf{R}^l)$, which could for example be the uniform distribution, with the unit k-vectorfield as follows:

$$S = \mu \wedge (e_1 \wedge \cdots \wedge e_k). \tag{7.28}$$

For an illustration, see the right side of Fig. 7.2 and Fig. 7.3. The data current $T \in \boldsymbol{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ is constructed from the samples $\{x_i\}_{i=1}^N$ and tangent vectorfields $T_i : \mathcal{X} \to \Lambda_k \mathbf{R}^d$.

$$T = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i} \wedge T_i, \tag{7.29}$$

The tangent k-vectorfields $T_i(x) = T_{i,1} \wedge \cdots \wedge T_{i,k}$ are given by individual tangent vectors to the data manifold $T_{i,j} \in \mathbf{R}^d$. For an illustration, see the left side of Fig. 7.2 or Fig. 7.3. After solving (7.27), the map $g_\theta : \mathcal{Z} \to \mathcal{X}$ will be our generative model, where changes in the latent space $\mathcal{Z}$ along the unit directions $e_1, \cdots, e_k$ are expected to behave equivariantly to the specified tangent directions $T_{i,1}, \cdots, T_{i,k}$ near $g(z)$.

7.6 Experiments

The specific hyperparameters, architectures and tangent vector setups used in practice are detailed in Appendix B.2.

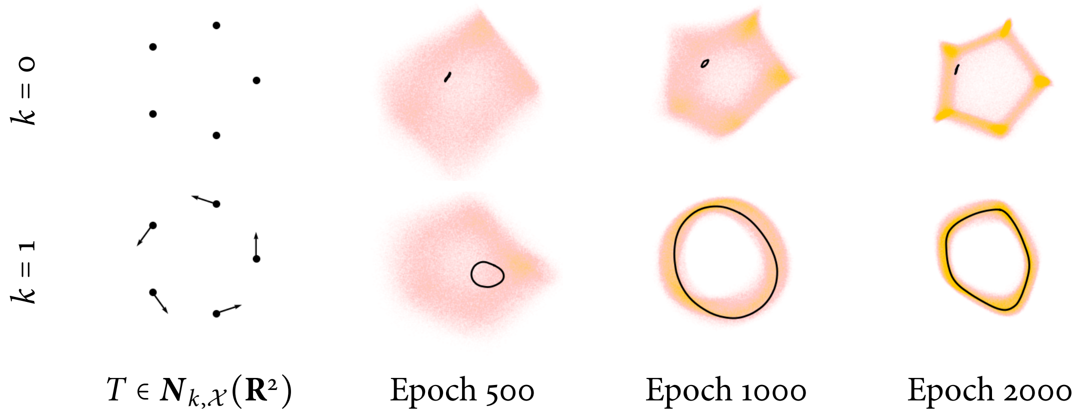


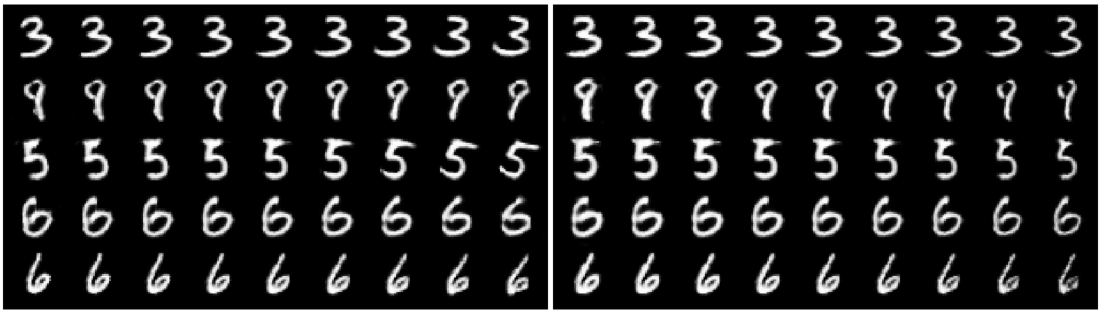
Figure 7.6: We illustrate the effect of moving from $k = 0$ to $k = 1$ and plot the measure $\|g_\sharp S\|$ of the pushforward of a k-current $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^5)$ (shown in orange) for different epochs. The black curve illustrates a walk along the first latent dimension z_1. For $k = 0$, which is similar to WGAN-GP [Gul+17], the latent walk is not meaningful. The proposed approach ($k = 1$) allows to specify tangent vectors at the samples to which the first latent dimension behaves equivariantly, yielding an interpretable representation.

7.6.1 Illustrative 2D example

As a first proof of concept, we illustrate the effect of moving from $k = 0$ to $k = 1$ on a very simple dataset consisting of five points on a circle. As shown in Fig. 7.6, for $k = 0$ (corresponding to a WGAN-GP formulation) varying the first latent variable has no clear meaning. In contrast, with the proposed FlatGAN formulation, we can specify vectors tangent to the circle from which the data is sampled. This yields an interpretable latent representation that corresponds to an angular movement along the circle. As the number of epochs is increasing, both formulations tend to concentrate most of the probability mass on the five data points. However, since $g_\theta : \mathcal{Z} \to \mathcal{X}$ is continuous by construction an interpretable path remains.

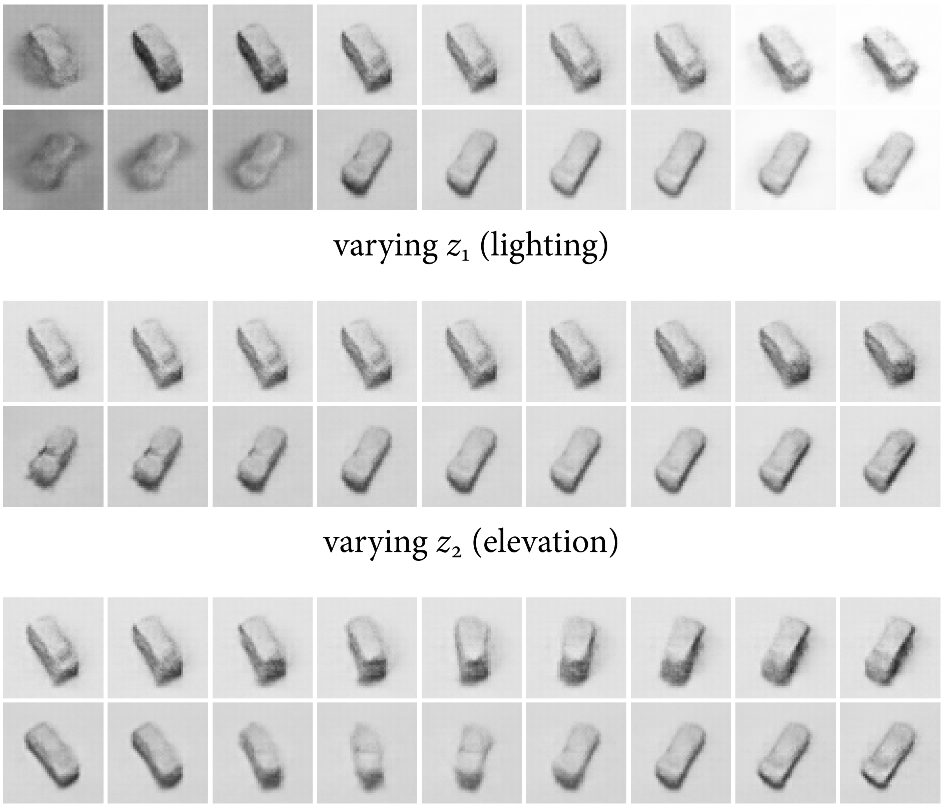
7.6.2 Equivariant representation learning

In Fig. 7.7 and Fig. 7.8 we show examples for $k = 2$ and $k = 3$ on MNIST respectively the smallNORB dataset of [LHB04]. For MNIST, we compute the tangent vectors manually by rotation and dilation of the digits, similar as done by [Sim+92; Sim+98]. For the smallNORB example, the tangent vectors are given as differences between the corresponding images. As observed in the figures, the proposed formulation leads to interpretable latent codes which behave equivariantly with the generated images. We remark that the goal was not to achieve state-of-the-art image quality but rather to demonstrate that specifying tangent vectors yields disentangled rep-



varying z_1 (rotation) varying z_2 (thickness)

Figure 7.7: We show the effect of varying the first two components in 128-dimensional latent space, corresponding to the two selected tangent vectors which are rotation and thickness. As seen in the figure, varying the corresponding latent representation yields an interpretable effect on the output, corresponding to the specified tangent direction.



varying z_1 (lighting)

varying z_2 (elevation)

varying z_3 (azimuth)

Figure 7.8: From left to right we vary the latent codes in $[-1, 1]$ after training on the smallNORB dataset [LHB04].

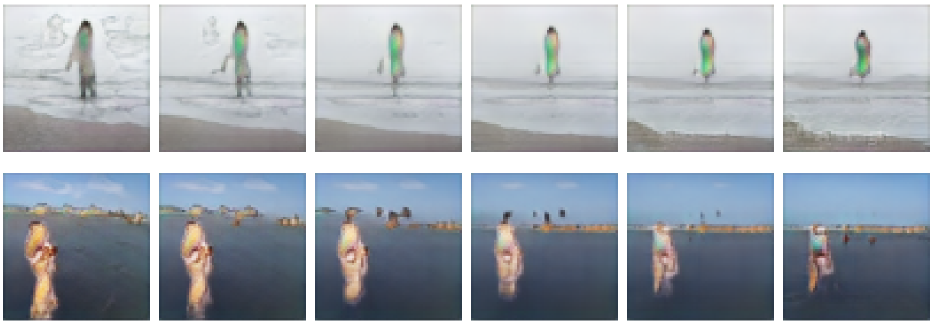


Figure 7.9: Varying the learned latent representation of time. The model captures behaviours such as people walking on the beach, see also the results shown in Fig. 7.1.

resentations. As remarked by [Jad+15], representing a 3D scene with a sequence of 2D convolutions is challenging and a specialized architecture based on a voxel representation would be more appropriate for the smallNORB example.

7.6.3 Discovering the arrow of time

In our last experiment, we set $k = 1$ and specify the tangent vector as the difference of two neighbouring frames in video data. We train on the tinyvideo beach dataset [VPT16], which consists of more than 36 million frames. After training for about half an epoch, we can already observe a learned latent representation of time, see Fig. 7.1 and Fig. 7.9. We generate individual frames by varying the latent coordinate z_1 from -12.5 to 12.5.

Even though the model is trained on individual frames in random order, a somewhat coherent representation of time is discovered which captures phenomena such as ocean waves or people walking on the beach.

7.7 Discussion and Conclusion

In this work, we demonstrated that k-currents can be used introduce a notion of orientation into probabilistic models. Furthermore, in experiments we have shown that specifying partial tangent information to the data manifold leads to interpretable and equivariant latent representations such as the camera position and lighting in a 3D scene or the arrow of time in time series data.

The difference to purely unsupervised approaches such as InfoGAN or β-VAE is, that we can encourage potentially very complex latent representations to be learned. Nevertheless, an additional mutual information term as in [Che+16] can be directly added to the formulation so that some representations could be encouraged through tangent vectors and the remaining ones are hoped to be discovered in an unsupervised fashion. Generally speaking, we believe that geometric measure theory is a rather underexploited field with many possible application areas in probabilistic machine learning. We see this work as a step towards leveraging this potential.

Part III

Conclusion

Chapter 8

Thesis Summary

In this thesis, we revisited convex relaxation approaches for continuous variational problems. The proposed formulations lead to more accurate results under a coarser discretization, boosting the practicality and applicability of previous relaxations. The formulations were analyzed from the perspective of a sublabel-accurate multilabeling method, as well as from the perspective of dual finite element discretizations. We also presented a novel relaxation for vectorial variational problems with polyconvex regularization, based on currents and differential forms. Finally, we also demonstrated that the formalisms of geometric measure theory find further applications beyond the relaxation of variational problems.

In the following, we briefly summarize the contributions made in each chapter.

Chapter 3. Starting from a discrete multilabeling perspective, we introduced a framework that assigns meaningful costs to fractional labelings. Such fractional labelings are understood as a sublabel-accurate solution to the original multilabeling problem. The resulting convex optimization problem is implemented using a first-order primal-dual method and epigraphical projections. We presented efficient projections onto the epigraph for piecewise linear and quadratic functions. When implemented on GPUs, the sublabel-accurate formulation entails only a little overhead over standard multilabeling approaches while requiring far fewer labels. We demonstrated the effectiveness of the approach in various experiments. To summarize, this chapter proposes a modification to existing continuous multilabeling methods, which significantly boosts their practicality and solution accuracy.

Chapter 4. In this chapter, we demonstrated that the previous sublabel-accurate multilabeling approaches can be generalized from scalar to vectorial label spaces. The partitioning of the label space into a triangular mesh lead to a convex relaxation of the label cost on each triangle. We showed that the aforementioned epigraphical projections can still be carried out in this vectorial setting. In the case of piecewise linear costs, an efficient active set method is employed. In experiments on large-displacement optical flow, the method was shown to outperform baselines for vectorial multilabeling.

Chapter 5. In this chapter, we demonstrated that the sublabel-accurate multilabeling methods can be understood as a finite-element discretization of a dual representation of certain label-continuous relaxations. Specifically, the previous sublabel-accurate methods were obtained by choosing a piecewise linear representation of the dual variable. This crucial insight allowed us to extend previous sublabel-accurate multilabeling approaches to more general regularizations. Moreover, it suggests a principled way to discretize continuous variational problems by considering a subspace discretization in the dual. The generality and effectiveness of this principle was demonstrated by discretizing a recent relaxation of the vectorial Mumford-Shah functional. This viewpoint further suggests the use of higher-order finite elements to obtain even more accurate formulations. Furthermore, the proposed discretization based on continuous piecewise-linear dual variables is connected to recent methods for discrete-continuous MRFs. In contrast, these have been shown to correspond to a discontinuous piecewise-linear approximation of the dual problem, which requires more variables.

Chapter 6. Given that sublabel-accurate multilabeling methods are discretizations of specific continuous formulations, the aim of this chapter was to extend the applicability of continuous convex relaxations. In particular, we proposed a novel lifting for vectorial variational problems with general polyconvex regularizations. The lifting is obtained by interpreting the original variational problem as an anisotropic minimal surface problem in higher dimension. This insight revealed the deep geometric structure underlying variational problems. The main challenge is the fact that the codimension of the minimal surface is generally larger than one. Hence it can not be represented as the boundary or derivative of an indicator function anymore. In order to deal with surfaces in codimension larger than one, the necessary tools from geometric measure theory, such as currents and differential forms were introduced. Finally, a convex formulation was obtained by performing a relaxation from oriented surfaces to more general sets of currents.

A dual representation over a constrained set of differential forms allowed for a principled sublabel-accurate discretization using the insights from the previous chapter. In experiments, we demonstrated that the proposed formulation recovers existing sublabel-accurate solutions in codimension one. As an example of a problem in codimension larger than one, we tackled the problem of global nonrigid registration of two-dimensional shapes. The proposed formulation was used to find bijective and smooth correspondences by solving a single convex optimization problem.

The proposed convexification of minimal surface energies in general codimension also provides an answer to the open question posed in [GSC12; ZB14]. There is indeed a convex (dual) formulation for the nonconvex Beltrami minimal surface regularizations [KMSoo]. The convexification and its dual formulation correspond to the mass and comass norm from geometric measure theory. Due to the sublabel-accurate formulation, it can be efficiently implemented, especially for convex data terms, where a discretization with the minimal number of labels is sufficient.

Chapter 7. As an outlook and to put things into a broader perspective and context, in this chapter, we demonstrated that concepts from geometric measure theory are useful in applications beyond the convexification and relaxation of variational problems. A driving factor behind the development of geometric measure theory has been the study of minimal surface problems in arbitrary dimension and codimension. The well-known *manifold hypothesis* in machine learning states that the distribution of real-world data (such as natural images) concentrates near a manifold of much lower dimension than the ambient vector space. Therefore, from a geometric point of view, real-world data can be understood as a somewhat irregular and diffuse surface of high codimension and dimension. It seems natural to represent it using the generalized notions of surfaces from geometric measure theory, which fulfill all the required criteria. Based on these ideas, we proposed to represent data as a k-current. This representation seamlessly generalizes the usual way of representing data as probability distributions, which are special 0-currents. More importantly, it allows one to introduce a partial orientation to the data by attaching oriented k-dimensional tangent planes to each data point. This data current is best interpreted as a first-order approximation to the actual underlying data manifold. Using the flat norm, we extended the recently proposed Wasserstein generative adversarial networks to work with general currents rather than just probability distributions. This extension enabled us to learn a generative model whose first k latent variables behave equivariantly to the specified tangent vectors. We see the results presented in this chapter as a first step towards leveraging the potential of geometric measure theory in machine learning and high-dimensional data analysis.

Chapter 9

Future Research

In this chapter, we will discuss several directions for future work. Some directions are rather straightforward extensions of the presented material, while others are open research challenges. We will also discuss recent work that has been carried out in parallel or is based on the results presented in this thesis.

As a remark, an overview of new perspectives in the convexification of variational problems has also been given in [BP18, Section 5].

9.1 Extending the Applicability of Lifting Methods

The class of energies we considered in this thesis was restricted to local integral functionals of the form

$$E(f) = \int_{\mathcal{X}} c(x, f(x), \nabla f(x)) \, \mathrm{d}x, \tag{9.1}$$

with convexity or polyconvexity assumptions in the last argument. For discontinuous f, more general Mumford-Shah type energies which additionally penalize the jump-set were also considered.

While many practical problems fit into this template, it is nevertheless desirable to extend the scope of convex relaxations methods to a larger class of energies.

As lifting strategies for more complicated functionals come with even higher complexity, efficient representations are crucial. In that sense, the sublabel-accurate formulations presented in this thesis pave the way for the next generation of functional lifting methods.

More general regularizations, higher-order derivatives. One desirable generalization includes a dependency of the cost also on higher-order derivatives of the function:

$$E(f) = \int_{\mathcal{X}} c(x, f(x), \nabla f(x), \nabla^2 f(x), \ldots) \, \mathrm{d}x. \tag{9.2}$$

In imaging applications, the use of higher order or mixed derivatives has shown to yield superior results, see [BKP10].

One possibility is to reduce the higher-order variational problem to a sequence of simpler problems by pursuing an alternating minimization strategy. This idea leads to a sequential approach, where each problem is either already convex or can be solved with the standard functional lifting approach [RPB13]. This strategy has been shown to work well, but due to the iterative approach, it is not easy to verify whether the computed solution is globally optimal. It seems natural to have a lifting approach, which requires only the solution of a single problem. Recent advances in this direction include direct liftings for Laplacian regularization [LL18; VL19], the total generalized variation [SG18] or even more general classes of regularizers [Vog20].

Another apparent limitation is the convexity assumption of the cost in the gradient argument. While Mumford-Shah type energies allow for certain nonconvex penalizations, the class of energies is still somewhat restricted.

Relaxations for variational problems that depend in a general nonconvex way on the gradient or Jacobian have been studied in the context of Young measures, see [Ped99; CR00]. In contrast, these works usually assume convexity of the cost in the first two arguments. For curvature regularization, convex relaxations have been studied in the discrete [SKC09] and continuous setting [BPW13; BPW15; CP19].

In the case of scalar-valued problems, a different direction to introduce higher-order regularizations and a nonconvex dependence on the gradient is to apply the vectorial lifting procedure from Chapter 6 to the gradient field of the scalar function. An additional curl-free constraint could ensure that the vector-field is a gradient of some function. Possibly, the current associated with the lifting of the gradient field might be connected to the curvature tensor of the graph of the original scalar function.

Nonlocal data terms. In many applications such as inverse problems, a certain nonlocality is present in the functional. In deblurring or deconvolution problems, the observed data z is assumed to be given by a convolution with a kernel k, i.e., $z \approx k * f + \eta$. It remains an open challenge to derive relaxations, which can naturally handle such nonlocal forward models. One possibility would be again to resort to heuristic strategies based on alternating minimization or variable splitting, where the nonlocal operator is split out into a convex subproblem. A different recent approach is based on majorization-minimization, in which a nonconvex majorizer is optimized using lifting methods [GM18].

Generalization to Manifolds. The relaxations presented in this thesis assume that both the domain and codomain have Euclidean structure. Generalizations to the case in which the codomain is a manifold have been proposed; see [Lel+13a]. Recently, based on the finite-element discretization view of Chapter 5, an extension of [Lel+13a] to the sublabel-accurate setting has been proposed in [Vog+19] using finite-element methods for manifolds [DE13].

For surfaces in codimension one, global optimality results as the ones presented in [Poc+10] could perhaps carry over to the setting where domain and codomain are orientable manifolds. For example, it was shown in [DHK11] that the top-dimensional boundary operator for a triangulation of an oriented compact manifold is totally unimodular, guaranteeing the exactness of linear programming relaxations. A related optimality result for area minimizing currents in homology classes was proven in a continuous setting in [Fed74, §5.10]. It would be interesting to see whether such results can be used to obtain global optimality guarantees, for example, in image processing problems involving functions with values in one-dimensional manifolds as considered in [CS13].

9.2 Rounding and Optimality Guarantees

Even in scenarios in which the convex relaxation is tight, one might still get a superposition of integral solutions. This superposition appears since convex combinations of solutions are also feasible solutions. As remarked in [Lel+13b], a simple binary thresholding strategy to obtain an integral solution would destroy the sublabel-accurate representation: a superposition of discrete solutions to approximate continuous integral solutions beyond the mesh accuracy is still desirable. Properly disentangling such desirable superpositions from non-desirable ones remains an open challenge.

For liftings in higher dimension or codimension or with more general regularizations, there are so far no known guarantees or sufficient conditions under which the relaxation is tight. The current tightness results in literature so far are based on decompositions or foliations of relaxed solutions into a superposition of integral solutions. In codimension one, such a decomposition relies on the celebrated coarea formula [Poc+10]. In dimension one, such a decomposition can also be guaranteed [Smi93; PS13]. The situation in dimension and codimension larger than one is more complicated. Decompositions of normal currents into a foliation of integral currents in this general setting have recently been studied [AM17; AMS19]. Imposing the required involutivity condition would amount to a nonconvex constraint. Therefore, it remains unclear under which conditions the relaxation proposed in Chapter 6 is tight in general dimension and codimension. The efficient global optimization of vectorial variational problems with provable a priori optimality guarantees remains a major open challenge. It is expected to be difficult, as related discrete optimization problems are known to be quite challenging as well.

9.3 Discretization Aspects

We have seen in Chapter 5 that the sublabel-accurate multilabeling methods presented in this thesis can be interpreted as a specific discretization to a continuous model. This illustrates the strength and flexibility of the function space approach: A single formulation can lead to a variety of implementations. Under this viewpoint, it might be interesting to explore different discretization approaches further.

Concurrently to the work presented in Chapter 5, the paper [ZH17] also considers ”sublabel-accurate” discretizations of relaxations for the piecewise smooth Mumford-Shah functional. In particular, the paper provides an in-depth study of the resulting finite-dimensional optimization problem with the aim to understand fractional labelings from a discrete viewpoint.

Product-spaces. From a practical viewpoint, the product space formulation [GSC13; SCC14] enables the efficient optimization of vectorial problems with modest memory requirements. The idea behind these methods is an efficient representation of the vectorial function via a collection of graph surfaces in codimension one. This representation is in contrast to Chapter 6, where a single surface in higher codimension is considered. It would be interesting to investigate the precise relationship of [GSC13; SCC14] to the formalism presented in Chapter 6. In particular, it seems possible to recover the product space formulation by restricting the dual variable in Chapter 6 to a certain simpler form. Since a restriction of the dual variable leads to a lower bound, this would be a formal argument supporting the intuition that the relaxation based on the full space $\mathcal{X} \times \mathcal{Y}$ is tighter than the ones considered in [GSC13; SCC14]. While straightforward to derive, a sublabel-accurate discretization of the works [GSC13; SCC14] would be desirable for practical applications. The results in [Bac19] might provide insights into situations under which the factored approximation is exact: Specifically, the paper analyzes a formulation that is quite similar to the data term proposed in [GSC13; SCC14] from the perspective of continuous submodularity.

Convergence guarantees. An aspect that has not been considered in this thesis is an analysis of the convergence rate of the discrete approximation to the exact continuous solution. Furthermore, it would be interesting to analyze the rate at which the discrete approximation converges, which could give some theoretical justification for the considered approach. The dual discretization approach might be particularly well-suited to such considerations. By taking a subspace on the dual, one gets a lower bound to the continuous energy, and subdividing the discretization mesh will monotonically increase this lower bound. If the existence and specific regularity properties (such as Lipschitz continuity) of the optimal dual variable in the continuous setting are known, it might be possible to derive approximation rates. This strategy was applied to derive convergence rates for discrete-continuous MRFs in [FA14]. Recently, dual discretizations of continuous variational problems with Raviart-Thomas finite elements have been analyzed [CP20; Bar20a; Her+19; Bar20b; CC20]. In codimension one, these discretizations are quite related to the one pursued in Chapter 5 and Chapter 6. The work [CC20] shows that the total variation based on such Raviart-Thomas elements achieves an improved

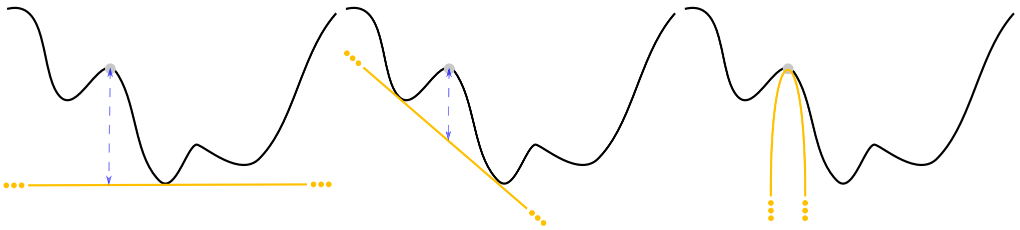


Figure 9.1: The approximation space for the dual variable dictates the shape of the "tool" with which we try to caress the function from below. The gray point illustrates the primal variable, which here is a Dirac measure. Considering linear instead of constant functions reduces the duality gap, indicated by the blue arrow. For quadratic dual variables, the gap can be further reduced.

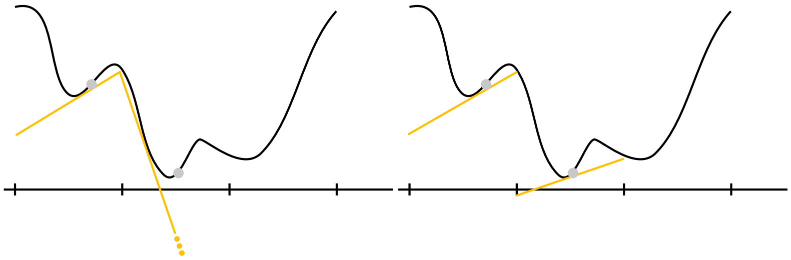


Figure 9.2: Comparison of an approximation with continuous (left) and discontinuous (right) piecewise linear dual variables on a fixed mesh. In case of a non-integer solution (two gray dots), the discontinuous approximation leads to a better lower bound for the function shown here.

convergence rate over the standard ”isotropic” total variation. Such a convergence analysis could possibly be adapted to the lifting methods. In codimension one, the lifted problem is essentially an anisotropic total variation problem in one dimension higher, so it is expected that results can carry over.

Higher-order finite elements and other discretizations. In this thesis, we have considered only piecewise constant and piecewise linear discretizations of the dual variable. Furthermore, the piecewise linear dual variable was chosen to be continuous. Clearly, there are other possible choices, which are illustrated in Fig. 9.1 and Fig. 9.2.

In these figures, we illustrate the effect of dual approximations on the following energy:

$$\boldsymbol{\rho}(\mu) = \sup_{\varphi \leq \rho} \int \varphi(x)\, \mathrm{d}\mu(x). \tag{9.3}$$

This energy is a simplified setting of the lifting approach, where we consider only the ”data term” ρ at a single point. Clearly it holds that $\boldsymbol{\rho}(\delta_x) = \rho(x)$. However, by restricting the dual variable φ we get certain lower bounds. As seen in Fig. 9.1, an affine dual variable cannot reach into the nonconvex dents of the function, which leads to a convexification of the cost ρ which we have seen in Chapter 3 and Chapter 4.

By considering higher-order dual variables (such as polynomials), the cost can be represented more accurately. Indeed, as seen in Fig. 9.1, a quadratic dual variable can reach into the nonconvexities of the function. Therefore, one can more accurately approximate the data term. It remains an open challenge whether such higher-order discretizations can be implemented efficiently.

Instead of increasing the degree of the dual variable, we considered a piecewise linear approximation. There, one has the choice of considering continuous or discontinuous piecewise linear dual variables. As shown in Fig. 9.2, the discontinuous dual variable can lead to a strictly better relaxation in case the measure μ (shown as the two gray dots) is not a Dirac measure. It was shown in [FA14] that the discrete-continuous MRF [ZK12] is based on such a discontinuous dual approximation. While being tighter, this representation requires storing two variables per interval as opposed to the continuous representation, which requires only one variable per node. The empirical results in Chapter 3 demonstrated that the representation with continuous dual variables lead to comparable results as the discontinuous one proposed in [ZK12], while requiring only half the amount of variables. A theoretical study on the differences could be an interesting avenue for future work.

A discretization of the dual variable using continuous piecewise linear functions can also be motivated from a functional-analytic viewpoint. For the full model, the subgraph indicator function in the primal is a function of bounded variation. The predual space of BV consists of the square-integrable vector fields whose divergence is also a square-integrable function. A conforming discretization of this $H(\text{div}, \Omega)$-space is given by Raviart-Thomas vector fields. The continuous piecewise linear approximation (corresponding to the ”2-sparse” sublabel-accurate representation) is obtained by lowest-order Raviart-Thomas fields. These considerations also motivated the use of Whitney forms in Chapter 6, which are a generalization of Raviart-Thomas elements. In retrospect, it is a happy coincidence that our representation, which was at first derived from a somewhat ”heuristic” viewpoint in Chapter 3, turned out to correspond to a conforming finite-element discretization of the dual problem. A rigorous analysis from the perspective of finite-element methods is another direction for future work.

Another interesting direction is to consider adaptive discretizations, for example, considered in [BER17]. Adaptive finite element approaches for a similar model in the context of branched transportation were recently considered in [DW20]. The discretization is based on piecewise linear dual variables and hinges on similar constraint set reductions, as presented in Chapter 5.

Given the flexibility and representational capacity of deep neural networks as powerful function approximators, it could be promising to employ them to discretize the continuous dual problem. The strategy of approximating the dual variable with a neural network was considered with some success in the context of optimal transport, specifically for the Wasserstein–1 distance [Seg+18; ACB17] and for the flat norm in Chapter 7. It could be interesting to explore whether such approximations can be beneficial in the context of the presented lifting methods.

Prolonging the discretization and Moreau-Yosida smoothing. Instead of discretizing the continuous energy with some approximating family of functions as discussed in the previous paragraph, it might be beneficial to stay in the infinite-dimensional setting and derive a system of optimality conditions. For example, this system might come from the primal-dual optimality relations given by the general Fenchel-Rockafellar duality theorems [Roc74].

The possible advantage is that an accurate discretization of the continuous optimality system could be designed to respect certain key quantities of the continuous model at optimality. For such considerations, a discrete theory that mimics and preserves geometric relations of the continuous setting, such as discrete exterior calculus, seems vital. Note that there might be no corresponding "equivalent" discretization on the energy level, i.e., solving the optimality system (variational inequality) does not correspond to the minimization of any discrete energy. Therefore, this approach could be more general than the one pursued in this thesis and potentially allows for additional freedom regarding the discretization.

For the presented energies in this thesis, the continuous optimality system is nonlinear and nonsmooth. A possible approach is to employ a continuation-based Moreau-Yosida regularization together with semi-smooth Newton methods [Ulb02; HIK02; IK03; Ulb11] to regularize and solve the system of equations. An advantage over the first-order methods we considered in this thesis is that the convergence of the numerical scheme is independent of the mesh spacing [Ulb11]. While each iteration requires one to solve a linear system which still scales with the problem dimension, a matrix-free implementation with a problem-specific preconditioner could lead to excellent performance. While promising in theory, it has yet to be confirmed that such a strategy can handle the huge scale problems of this thesis, which we were able to tackle rather directly with the first-order primal-dual algorithm implemented on GPUs.

An encouraging example of the above approach for a slightly related optimization problem posed in the space of measures is the recent work [SSC18].

Independent of the above considerations, smoothing and regularization strategies could significantly improve the performance of the lifting methods. Due to the dual discretization, the primal solution is not unique: Restricting the set of test functions hinders the ability to distinguish two distributions. Formally, the dual discretization introduces a quotient space into equivalence classes in the primal. A strictly convex regularization term such as the entropy or a quadratic norm will pick a certain candidate from the equivalence class, which might aid the convergence.

In the setting of optimal transport, such regularization strategies have enabled efficient optimization using Sinkhorn algorithms [Cut13]. Furthermore, regularizations might also make the formulation more amenable to *stochastic* optimization algorithms, as was recently demonstrated in the setting of optimal transport in [Gen+16].

Part IV

Appendix

Appendix A

Proofs

A.1 Chapter 3

Proof of Proposition 1. The proof follows from a direct calculation. We start with the definition of the biconjugate:

$$\begin{aligned}\boldsymbol{\rho}^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{v}\in\mathbf{R}^k} \langle \boldsymbol{u}, \boldsymbol{v}\rangle - \left(\min_{1\leq i\leq k} \boldsymbol{\rho}_i(\boldsymbol{u})\right)^* \\ &= \sup_{\boldsymbol{v}\in\mathbf{R}^k} \langle \boldsymbol{u}, \boldsymbol{v}\rangle - \max_{1\leq i\leq k} \boldsymbol{\rho}_i^*(\boldsymbol{u}).\end{aligned} \tag{A.1}$$

This shows the first equation inside the proposition. For the individual $\boldsymbol{\rho}_i^*$ we again start with the definition of the convex conjugate:

$$\begin{aligned}\boldsymbol{\rho}_i^*(\boldsymbol{v}) &= \sup_{\alpha\in[0,1]} \langle \alpha\mathbf{1}_i + (1-\alpha)\mathbf{1}_{i-1}, \boldsymbol{v}\rangle - \rho(\alpha\gamma_{i+1} + (1-\alpha)\gamma_i) \\ &= \sup_{\alpha\in[0,1]} \langle \mathbf{1}_{i-1}, \boldsymbol{v}\rangle + \alpha\boldsymbol{v}_i - \rho(\gamma_i^\alpha).\end{aligned} \tag{A.2}$$

Applying the substitution $\gamma_i^\alpha = \alpha\gamma_{i+1} + (1-\alpha)\gamma_i$ and consequently $\alpha = \frac{\gamma_i^\alpha - \gamma_i}{\gamma_{i+1}-\gamma_i}$ yields:

$$\begin{aligned}\boldsymbol{\rho}_i^*(\boldsymbol{v}) &= \sup_{\gamma_i^\alpha\in\Gamma_i} \langle \mathbf{1}_{i-1}, \boldsymbol{v}\rangle + \frac{\gamma_i^\alpha - \gamma_i}{\gamma_{i+1}-\gamma_i}\boldsymbol{v}_i - \rho(\gamma_i^\alpha) \\ &= \langle \mathbf{1}_{i-1}, \boldsymbol{v}\rangle - \frac{\gamma_i}{\gamma_{i+1}-\gamma_i}\boldsymbol{v}_i + \sup_{\gamma_i^\alpha\in\Gamma_i} \gamma_i^\alpha \frac{\boldsymbol{v}_i}{\gamma_{i+1}-\gamma_i} - \rho(\gamma_i^\alpha) \\ &= \langle \mathbf{1}_{i-1}, \boldsymbol{v}\rangle - \frac{\gamma_i}{\gamma_{i+1}-\gamma_i}\boldsymbol{v}_i + (\rho + \delta_{\Gamma_i})^*\left(\frac{\boldsymbol{v}_i}{\gamma_{i+1}-\gamma_i}\right) \\ &=: c_i(\boldsymbol{v}) + \rho_i^*\left(\frac{\boldsymbol{v}_i}{\gamma_{i+1}-\gamma_i}\right).\end{aligned} \tag{A.3}$$

□

Proof of Proposition 2. It is easy to see that

$$\boldsymbol{\sigma}^*(\boldsymbol{v}) = \max_{i\in\{1,\dots,L\}} \left(\sum_{l=1}^{i-1} \boldsymbol{v}_l - \rho(\gamma_i) \right).$$

To compute the biconjugate, we write any input argument $\boldsymbol{u} = \sum_{i=1}^{k} \mu_i \mathbf{1}_{i+1}$, and use $\boldsymbol{\sigma}^{**} = \boldsymbol{\rho}^{**}$ to obtain

$$\begin{aligned}\boldsymbol{\rho}^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{v}} \langle \boldsymbol{u}, \boldsymbol{v}\rangle - \max_{i\in\{1,\dots,L\}} \left(\sum_{l=1}^{i-1} \boldsymbol{v}_l - \rho(\gamma_i) \right) \\ &= \sup_{\boldsymbol{v}} \sum_{i=1}^{k} \mu_i \sum_{l=1}^{i} \boldsymbol{v}_l - \max_{i\in\{1,\dots,L\}} \left(\sum_{l=1}^{i-1} \boldsymbol{v}_l - \rho(\gamma_i) \right).\end{aligned}$$

Instead of taking the supremum of all $\boldsymbol{v}$, we might as well take the supremum over all vectors $\mathbf{p}$ with $\mathbf{p}_i = \sum_{l=1}^{i} \boldsymbol{v}_l$. Care has to be taken of the first summand in the second term of the above formulation. We obtain

$$\begin{aligned}&\sup_{\boldsymbol{v}} \sum_{i=1}^{k} \mu_i \sum_{l=1}^{i} \boldsymbol{v}_l - \max_{i\in\{1,\dots,L\}} \left(\sum_{l=1}^{i-1} \boldsymbol{v}_l - \rho(\gamma_i) \right), \\ &= \sup_{\mathbf{p}} \sum_{i=1}^{k} \mu_i \mathbf{p}_i - \max_{i\in\{2,\dots,L\}} \max(\mathbf{p}_{i-1} - \rho(\gamma_i), -\rho(\gamma_1)), \\ &= \sup_{\mathbf{p}} \sum_{i=1}^{k} \mu_i \mathbf{p}_i - \max_{i\in\{1,\dots,k\}} \max(\mathbf{p}_i - \rho(\gamma_{i+1}), -\rho(\gamma_1)), \\ &= \sum_{i=1}^{k} \mu_i \, \rho(\gamma_{i+1}) + \sup_{\mathbf{p}} \sum_{i=1}^{k} \mu_i \mathbf{p}_i - \max_{i\in\{1,\dots,k\}} \max(\mathbf{p}_i, -\rho(\gamma_1)),\end{aligned}$$

Note that for any μ_i being negative, the supremum immediately yields infinity by taking $\mathbf{p}_i \to -\infty$. Similarly, if $\sum_{i=1}^{k} \mu_i > 1$ yields infinity by taking all $\mathbf{p}_i \to \infty$. For $\mu_i \geq 0$ for all i, and $\sum_{i=1}^{k} \mu_i \leq 1$, we know that $\sum_{i=1}^{k} \mu_i \mathbf{p}_i \leq (\max_i \mathbf{p}_i) \sum_{i=1}^{k} \mu_i$. Since equality can be obtained by choosing $\mathbf{p}_l = \max_i \mathbf{p}_i$ for all l, we can reduce the above supremum to

$$\sup_{z} \left(z \sum_{i=1}^{k} \mu_i - \max(z, -\rho(\gamma_1)) \right) = \left(1 - \sum_{i=1}^{k} \mu_i \right) \rho(\gamma_1),$$

where we used that the supremum over z is attained at $z = -\rho(\gamma_1)$ (still assuming that $\sum_{i=1}^{k} \mu_i \leq 1$). Let us now undo our change of variable. It is easy to see that $\mu_k = \boldsymbol{u}_k$, and $\mu_i = \boldsymbol{u}_i - \boldsymbol{u}_{i+1}$ for $i = 1, \dots, k-1$. The latter leads to

$$\begin{aligned}&\sum_{i=1}^{k} \mu_i \, \rho(\gamma_{i+1}) + \left(1 - \sum_{i=1}^{k} \mu_i \right) \rho(\gamma_1) \\ &= \rho(\gamma_{k+1}) \boldsymbol{u}_k + \sum_{i=1}^{k-1} (\boldsymbol{u}_i - \boldsymbol{u}_{i+1}) \, \rho(\gamma_{i+1}) + (1 - \boldsymbol{u}_1) \rho(\gamma_1) \\ &= \rho(\gamma_1) + \langle \boldsymbol{u}, \mathbf{r} \rangle,\end{aligned}$$

for $\mathbf{r}_i = \rho(\gamma_{i+1}) - \rho(\gamma_i)$. Considering the aforementioned constraints of $\mu_i \geq 0$, and $\sum_{i=1}^{k} \mu_i \leq 1$, we finally find

$$\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \begin{cases} \rho(\gamma_1) + \langle \boldsymbol{u}, \mathbf{r} \rangle & \text{if } 1 \geq \boldsymbol{u}_1 \geq \ldots \geq \boldsymbol{u}_k \geq 0, \\ \infty, & \text{else.} \end{cases}$$

□

Proof of Proposition 3. For the special case $k = 1$ the biconjugate from (A.1) is just:

$$\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \sup_{\boldsymbol{v} \in \mathbf{R}} \boldsymbol{uv} - \boldsymbol{\rho}_1^{*}(\boldsymbol{v}) = \boldsymbol{\rho}_1^{**}(\boldsymbol{u}). \tag{A.4}$$

Now using the first line in (A.3), $\boldsymbol{\rho}_1^{**}$ becomes:

$$\begin{aligned} \boldsymbol{\rho}_1^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{v} \in \mathbf{R}} \boldsymbol{uv} - \sup_{\gamma \in \Gamma} \frac{\gamma - \gamma_1}{\gamma_2 - \gamma_1} \boldsymbol{v} - \rho(\gamma) \\ &= \sup_{\boldsymbol{v} \in \mathbf{R}} \boldsymbol{v} \left(\boldsymbol{u} + \frac{\gamma_1}{\gamma_2 - \gamma_1} \right) - \sup_{\gamma \in \Gamma} \gamma \frac{\boldsymbol{v}}{\gamma_2 - \gamma_1} - \rho(\gamma) \\ &= \sup_{\boldsymbol{v} \in \mathbf{R}} \boldsymbol{v} \left(\boldsymbol{u} + \frac{\gamma_1}{\gamma_2 - \gamma_1} \right) - \rho^{*} \left(\frac{\boldsymbol{v}}{\gamma_2 - \gamma_1} \right) \\ &= \sup_{\tilde{\boldsymbol{v}} \in \mathbf{R}} \tilde{\boldsymbol{v}}(\gamma_1 + \boldsymbol{u}(\gamma_2 - \gamma_1)) - \rho^{*}(\tilde{\boldsymbol{v}}) \\ &= \rho^{**}(\gamma_1 + \boldsymbol{u}(\gamma_2 - \gamma_1)), \end{aligned} \tag{A.5}$$

where we used $\operatorname{dom}(\rho) = \Gamma$ as well as the substitution $\boldsymbol{v} = (\gamma_2 - \gamma_1)\tilde{\boldsymbol{v}}$. □

Proof of Proposition 4. We compute the individual conjugate as:

$$\begin{aligned} \boldsymbol{\Phi}_{i,j}^{*}(\boldsymbol{q}) &= \sup_{\boldsymbol{g} \in \mathbf{R}^{d \times k}} \langle g, \boldsymbol{q} \rangle - \boldsymbol{\Phi}_{i,j}(\boldsymbol{q}) \\ &= \sup_{\alpha, \beta \in [0,1]} \sup_{v \in \mathbf{R}^d} \langle \boldsymbol{q}, (\mathbf{1}_i^{\alpha} - \mathbf{1}_j^{\beta}) v^{\top} \rangle - \left| \gamma_i^{\alpha} - \gamma_j^{\beta} \right| |v|_2 \\ &= \sup_{\alpha, \beta \in [0,1]} \sup_{v \in \mathbf{R}^d} \langle \boldsymbol{q}^{\top} (\mathbf{1}_i^{\alpha} - \mathbf{1}_j^{\beta}), v \rangle - \left| \gamma_i^{\alpha} - \gamma_j^{\beta} \right| |v|_2 \\ &= \sup_{\alpha, \beta \in [0,1]} \sup_{v \in \mathbf{R}^d} \langle \boldsymbol{q}^{\top} (\mathbf{1}_i^{\alpha} - \mathbf{1}_j^{\beta}), v \rangle - \left| \gamma_i^{\alpha} - \gamma_j^{\beta} \right| |v|_2. \end{aligned} \tag{A.6}$$

The inner maximum over v is the conjugate of the ℓ_2-norm scaled by $\left| \gamma_i^{\alpha} - \gamma_j^{\beta} \right|$ evaluated at $\boldsymbol{q}^{\top} \left(\mathbf{1}_i^{\alpha} - \mathbf{1}_j^{\beta} \right)$. This yields:

$$\boldsymbol{\Phi}_{i,j}^{*}(\boldsymbol{q}) = \begin{cases} 0, & \text{if } \left| \boldsymbol{q}^{\top} \left(\mathbf{1}_i^{\alpha} - \mathbf{1}_j^{\beta} \right) \right|_2 \leq \left| \gamma_i^{\alpha} - \gamma_j^{\beta} \right|, \\ & \qquad\qquad\qquad\qquad \forall \alpha, \beta \in [0,1], \\ \infty, & \text{else.} \end{cases} \tag{A.7}$$

For the overall biconjugate we have:

$$\begin{aligned}\boldsymbol{\Phi}^{**}(\boldsymbol{g}) &= \sup_{\boldsymbol{q}\in\mathbf{R}^{k\times d}} \langle \boldsymbol{q}, \boldsymbol{g}\rangle - \max_{1\leq i,j\leq k} \boldsymbol{\Phi}_{i,j}^{*}(\boldsymbol{q}) \\ &= \sup_{\boldsymbol{q}\in\mathcal{K}} \langle \boldsymbol{q}, \boldsymbol{g}\rangle.\end{aligned} \tag{A.8}$$

Since we have the max over all $1 \leq i, j \leq k$ conjugates, the set $\mathcal{K}$ is given as the intersection of the sets described by the individual indicator functions $\boldsymbol{\Phi}_{i,j}$:

$$\begin{aligned}\mathcal{K} = \Big\{\boldsymbol{q} \in \mathbf{R}^{k\times d} \ : \ &\left|\boldsymbol{q}^{\top}(\mathbf{1}_i^{\alpha} - \mathbf{1}_j^{\beta})\right|_2 \leq \left|\gamma_i^{\alpha} - \gamma_j^{\beta}\right|, \\ &\forall\ 1 \leq i \leq j \leq k,\ \forall \alpha, \beta \in [0,1]\Big\}.\end{aligned} \tag{A.9}$$

□

Proof of Proposition 5. First we rewrite (A.9) by expanding the matrix-vector product into sums:

$$\left|\sum_{l=j}^{i-1} \boldsymbol{q}_l + \alpha\boldsymbol{q}_i - \beta\boldsymbol{q}_j\right|_2 \leq \left|\gamma_i^{\alpha} - \gamma_j^{\beta}\right|, \forall\ 1 \leq j \leq i \leq k,\ \forall\alpha, \beta \in [0,1]. \tag{A.10}$$

Since the other cases for $1 \leq i \leq j \leq k$ in (A.9) are equivalent to (A.10), it is enough to consider (A.10) instead of (A.9).

Let $\gamma_1 < \gamma_2 < \cdots < \gamma_L$. In the following, we will show the equivalences:

$$\text{(A.10)}$$

$$\Leftrightarrow$$

$$\left|\sum_{l=j}^{i} \boldsymbol{q}_l\right|_2 \leq \gamma_{i+1} - \gamma_j,\ \forall\ 1 \leq j \leq i \leq k. \tag{A.11}$$

$$\Leftrightarrow$$

$$|\boldsymbol{q}_i|_2 \leq \gamma_{i+1} - \gamma_i,\ \forall\ 1 \leq i \leq k. \tag{A.12}$$

The direction "(A.10) ⇒ (A.11)" follows by setting $\alpha = 1$ and $\beta = 0$ in (A.10), and "(A.11) ⇒ (A.12)" follows by setting $i = j$ in (A.11).

The direction "(A.12) ⇒ (A.11)" can be proven by a quick calculation:

$$\left|\sum_{l=j}^{i} \boldsymbol{q}_l\right|_2 \leq \sum_{l=j}^{i} |\boldsymbol{q}_l|_2 \leq \sum_{l=j}^{i} \gamma_{l+1} - \gamma_l = \gamma_{i+1} - \gamma_j. \tag{A.13}$$

It remains to show "(A.11) ⇒ (A.10)". We start with the case $j = i$:

$$\begin{aligned}|\alpha\boldsymbol{q}_i - \beta\boldsymbol{q}_i|_2 &= |\alpha - \beta||\boldsymbol{q}_i|_2 \\ &\leq |\alpha - \beta|(\gamma_{i+1} - \gamma_i) \\ &= |(\gamma_{i+1} - \gamma_i)\alpha - (\gamma_{i+1} - \gamma_i)\beta| \\ &= |(\alpha - \beta)(\gamma_{i+1} - \gamma_i)| = |\gamma_i^{\alpha} - \gamma_i^{\beta}|.\end{aligned} \tag{A.14}$$

Now let $j < i$. Since $\gamma_j < \gamma_i$ it also holds that $\gamma_j^\beta \leq \gamma_i^\alpha$, thus it is equivalent to show (A.10) without the absolute value on the right hand side.

First we show that "(A.11) ⇒ (A.10)" for $\beta \in \{0, 1\}$ and $\alpha \in [0, 1]$:

$$\begin{aligned} \left| \sum_{l=j+1}^{i-1} \boldsymbol{q}_l + \alpha \boldsymbol{q}_i + (1-\beta)\boldsymbol{q}_j \right|_2 &\leq \left| \sum_{l=j+1}^{i-1} \boldsymbol{q}_l + (1-\beta)\boldsymbol{q}_j \right|_2 + \alpha |\boldsymbol{q}_i|_2 \\ &\overset{\text{for } \beta=0 \text{ or } \beta=1}{\leq} \gamma_i - \gamma_j^\beta + \alpha(\gamma_{i+1} - \gamma_i) \\ &= \gamma_i^\alpha - \gamma_j^\beta. \end{aligned} \tag{A.15}$$

Using a similar argument we show that, using the above, "(A.11) ⇒ (A.10)" for all $\alpha, \beta \in [0, 1]$.

$$\begin{aligned} \left| \sum_{l=j+1}^{i-1} \boldsymbol{q}_l + \alpha \boldsymbol{q}_i + (1-\beta)\boldsymbol{q}_j \right|_2 &\leq \left| \sum_{l=j+1}^{i-1} \boldsymbol{q}_l + \alpha\boldsymbol{q}_i \right|_2 + (1-\beta) |\boldsymbol{q}_j|_2 \\ &\overset{\text{using (A.15)}, \beta=1}{\leq} \gamma_i^\alpha - \gamma_{j+1} + (1-\beta)(\gamma_{j+1} - \gamma_j) \\ &= \gamma_i^\alpha - \gamma_j^\beta. \end{aligned} \tag{A.16}$$

□

A.2 Chapter 4

Proof of Proposition 6. By definition the biconjugate of $\boldsymbol{\rho}$ is given as

$$\begin{aligned} \boldsymbol{\rho}^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{v} \in \mathbf{R}^{|\mathcal{V}|}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \left(\min_{1 \leq i \leq |\mathcal{T}|} \boldsymbol{\rho}_i(\boldsymbol{v}) \right)^* \\ &= \sup_{\boldsymbol{v} \in \mathbf{R}^{|\mathcal{V}|}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \max_{1 \leq i \leq |\mathcal{T}|} \boldsymbol{\rho}_i^*(\boldsymbol{v}). \end{aligned} \tag{A.17}$$

We proceed computing the conjugate of $\boldsymbol{\rho}_i$:

$$\begin{aligned} \boldsymbol{\rho}_i^*(\boldsymbol{v}) &= \sup_{\boldsymbol{u} \in \mathbf{R}^{|\mathcal{V}|}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \boldsymbol{\rho}_i(\boldsymbol{u}) \\ &= \sup_{\boldsymbol{\alpha} \in \Delta_{n+1}^U} \langle E_i \alpha, \boldsymbol{v} \rangle - \rho\left(T_i \alpha\right), \end{aligned} \tag{A.18}$$

We introduce the substitution $r := T_i \alpha \in \Delta_i$ and obtain

$$\alpha = K_i^{-1} \begin{pmatrix} r \\ 1 \end{pmatrix}, \quad K_i := \begin{pmatrix} T_i \\ \mathbf{1}^\top \end{pmatrix} \in \mathbf{R}^{n+1 \times n+1}, \tag{A.19}$$

since K_i is invertible for $(\mathcal{V},\mathcal{T})$ being a non-degenerate triangulation and $\sum_{j=1}^{n+1}\alpha_j = 1$. With this we can further rewrite the conjugate as

$$\begin{aligned} \cdots &= \sup_{\boldsymbol{r}\in\Delta_i}\langle A_i r + b_i, E_i^\top \boldsymbol{v}\rangle - \rho(r) \\ &= \langle E_i b_i, \boldsymbol{v}\rangle + \sup_{r\in\mathbf{R}^n}\langle r, A_i^\top E_i^\top \boldsymbol{v}\rangle - \rho(r) - \delta_{\Delta_i}(r) \\ &= \langle E_i b_i, \boldsymbol{v}\rangle + \rho_i^*(A_i^\top E_i^\top \boldsymbol{v}). \end{aligned} \tag{A.20}$$

□

Proof of Proposition 7. Define $\boldsymbol{\Psi}_{i,j}$ as

$$\boldsymbol{\Psi}_{i,j}(\boldsymbol{p}) := \begin{cases} \|T_i\alpha - T_j\beta\| \cdot \|v\| & \text{if } \boldsymbol{p} = (E_i\alpha - E_j\beta)v^\top,\ \alpha,\beta\in\Delta_{n+1}^U,\ v\in\mathbf{R}^d, \\ \infty & \text{otherwise.} \end{cases} \tag{A.21}$$

Then, $\boldsymbol{\Psi}$ can be rewritten as a pointwise minimum over the individual $\boldsymbol{\Psi}_{i,j}$

$$\boldsymbol{\Psi}(\boldsymbol{p}) = \min_{1\le i,j\le|\mathcal{T}|}\boldsymbol{\Psi}_{i,j}(\boldsymbol{p}). \tag{A.22}$$

We begin computing the conjugate of $\boldsymbol{\Psi}_{i,j}$

$$\begin{aligned} \boldsymbol{\Psi}_{i,j}^*(\boldsymbol{q}) &= \sup_{\boldsymbol{p}\in\mathbf{R}^{d\times|\mathcal{V}|}}\langle \boldsymbol{p},\boldsymbol{q}\rangle - \boldsymbol{\Psi}_{i,j}(\boldsymbol{p}) \\ &= \sup_{\alpha,\beta\in\Delta_{n+1}^U}\sup_{v\in\mathbf{R}^d}\langle Q_i\alpha - Q_j\beta, v\rangle - \|T_i\alpha - T_j\beta\|\cdot\|v\| \\ &= \sup_{\alpha,\beta\in\Delta_{n+1}^U}\left(\|T_i\alpha - T_j\beta\|\cdot\|\cdot\|\right)^*(Q_i\alpha - Q_j\beta) \\ &= \delta_{\mathcal{K}_{i,j}}(\boldsymbol{q}), \end{aligned} \tag{A.23}$$

with the set $K_{i,j}$ being defined as

$$\mathcal{K}_{i,j} := \left\{\boldsymbol{q}\in\mathbf{R}^{d\times|\mathcal{V}|} \;\middle|\; \|Q_i\alpha - Q_j\beta\| \le \|T_i\alpha - T_j\beta\|,\ \alpha,\beta\in\Delta_{n+1}^U\right\}. \tag{A.24}$$

Since the maximum over indicator functions of sets is equal to the indicator function of the intersection of the sets we obtain for $\boldsymbol{\Psi}^*$

$$\boldsymbol{\Psi}^*(\boldsymbol{q}) = \max_{1\le i,j\le|\mathcal{T}|}\boldsymbol{\Psi}_{i,j}^*(\boldsymbol{q}) = \delta_{\mathcal{K}}(\boldsymbol{q}). \tag{A.25}$$

□

Proof of Proposition 8. Let $\boldsymbol{q}\in\mathbf{R}^{d\times|\mathcal{V}|}$ s.t. $\|Q_i\alpha - Q_j\beta\| \le \|T_i\alpha - T_j\beta\|$ for all $\alpha,\beta\in\Delta_{n+1}^U$ and $1\le i,j\le|\mathcal{T}|$. For any $1\le i\le|\mathcal{T}|$ define

$$\begin{aligned} &f_i : \mathbb{R}^n \to \mathbb{R}^n, \\ &(\alpha_1,\ldots,\alpha_n) \mapsto \sum_{l=1}^n \alpha_l t^{i_l} + (1 - \sum_{l=1}^n \alpha_l)t^{i_{n+1}} = T_i\alpha, \end{aligned} \tag{A.26}$$

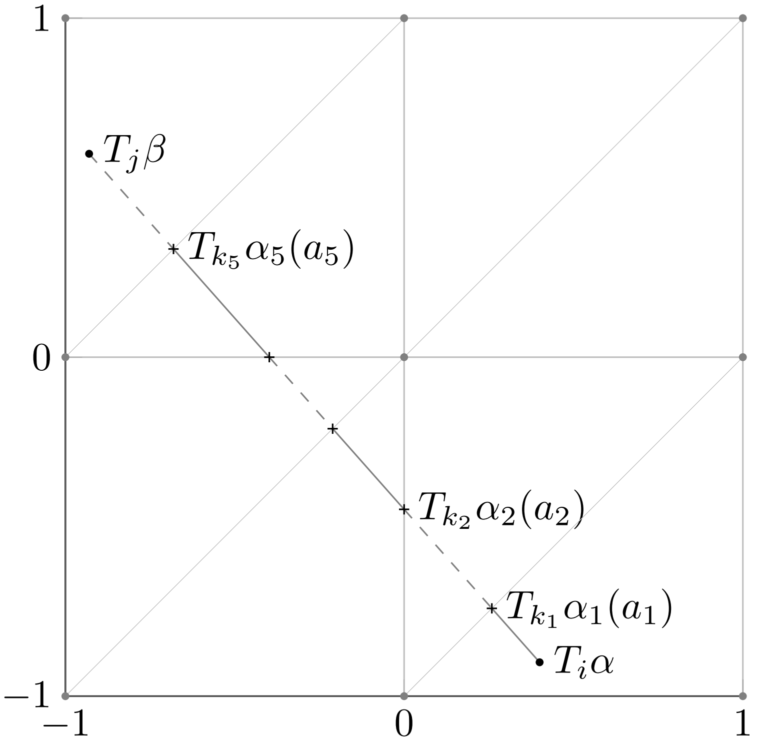


Figure A.1: Figure illustrating the second direction of the proof. The gray dots and lines visualize the triangulation $(\mathcal{V},\mathcal{T})$. The line segment between $T_i\alpha$ and $T_j\beta$ is composed of shorter line segments which are fully contained in one of the triangles. On each of the triangles the inequality (A.31) holds, which allows to conclude that it holds for the whole line segment.

and analogously

$$\begin{aligned} g_i : \mathbb{R}^n &\to \mathbb{R}^{|\mathcal{V}|} \\ (\alpha_1,\dots,\alpha_n) &\mapsto \sum_{l=1}^{n} \alpha_l \boldsymbol{q}^{i_l} + (1-\sum_{l=1}^{n}\alpha_l)\boldsymbol{q}^{i_{n+1}} = Q_i\alpha. \end{aligned} \tag{A.27}$$

Let us choose an $\alpha \in \mathbb{R}^n$ such that $\alpha_i > 0$, $\sum_l \alpha_l < 1$. Then $\|Q_i\alpha - Q_j\beta\| \le \|T_i\alpha - T_j\beta\|$ for all $\alpha,\beta \in \Delta^U_{n+1}$ and $1 \le i,j \le |\mathcal{T}|$ implies that

$$\|g_i(\alpha) - g_i(\alpha - h)\| \le \|f_i(\alpha) - f_i(\alpha - h)\|, \tag{A.28}$$

holds for all vectors h with sufficiently small entries. Inserting the definitions of g_i and f_i we find that

$$\|Q_iDh\| \le \|T_iDh\| \tag{A.29}$$

holds for all h with sufficiently small entries. For a non-degenerate triangle, T_iD is invertible and a simple substitution yields that

$$\|Q_iD(T_iD)^{-1}\tilde{h}\|_2 \le \|\tilde{h}\|, \tag{A.30}$$

holds for all $\tilde{h}$ with sufficiently small entries. This means that the operator norm of $D^i_{\boldsymbol{q}}$ induced by the ℓ^2 norm, i.e. the S^∞ norm, is bounded by one.

Let us now show the other direction. For $\boldsymbol{q} \in \mathbf{R}^{d\times|\mathcal{V}|}$ s.t. $\|D^i_{\boldsymbol{q}}\|_{S^\infty} \leq 1$, $1 \leq i \leq |\mathcal{T}|$, note that inverting the above computation immediately yields that

$$\|Q_k\alpha - Q_k\beta\| \leq \|T_k\alpha - T_k\beta\| \tag{A.31}$$

holds for all $1 \leq k \leq |\mathcal{T}|$, $\alpha, \beta \in \Delta^U_{n+1}$. Our goal is to show that having this inequality on each simplex is sufficient to extend it to arbitrary pairs of simplices. The overall idea of this part of the proof is illustrated in Fig. A.1.

Let $1 \leq i, j \leq |\mathcal{T}|$ and $\alpha, \beta \in \mathbb{R}^n$ with $\alpha_l, \beta_l \geq 0$, $\sum_l \alpha_l \leq \sum_l \beta_l \leq 1$ be given. Consider the line segment

$$\begin{aligned} c(\gamma) : [0,1] &\to \mathbb{R}^d \\ \gamma &\mapsto \gamma\, T_j\beta + (1-\gamma)\, T_i\alpha. \end{aligned} \tag{A.32}$$

Since the triangulated domain is convex, there exist $0 = a_0 < a_1 < \cdots < a_r = 1$ and functions $\alpha_l(\gamma)$ such that for $\gamma \in [a_l, a_{l+1}]$, $0 \leq l \leq r-1$ one can write $c(\gamma) = \gamma\, T_j\beta + (1-\gamma)\, T_i\alpha = T_{k_l}\alpha_l(\gamma)$ for some $1 \leq k_l \leq T$. The continuity of $c(\gamma)$ implies that $T_{k_l}\alpha_l(a_{l+1}) = T_{k_{l+1}}\alpha_{l+1}(a_{l+1})$, i.e. these points correspond to both simplices, k_l and k_{l+1}. Note that this also means that $Q_{k_l}\alpha_l(a_{l+1}) = Q_{k_{l+1}}\alpha_{l+1}(a_{l+1})$. The intuition of this construction is that the $c(a_{l+1})$ are located on the boundaries of adjacent simplices on the line segment. We find

$$\begin{aligned} \|T_i\alpha - T_j\beta\| &= \sum_{l=0}^{r-1}(a_{l+1} - a_l)\|T_i\alpha - T_j\beta\| \\ &= \sum_{l=0}^{r-1}\|(a_{l+1} - a_l)(T_i\alpha - T_j\beta)\| \\ &= \sum_{l=0}^{r-1}\|a_{l+1}T_i\alpha - a_lT_i\alpha - a_{l+1}T_j\beta + a_lT_j\beta\| \\ &= \sum_{l=0}^{r-1}\left\|a_lT_j\beta + (1-a_l)T_i\alpha - \big(a_{l+1}T_j\beta + (1-a_{l+1})T_i\alpha\big)\right\| \\ &= \sum_{l=0}^{r-1}\|T_{k_l}\alpha_l(a_l) - T_{k_l}\alpha_l(a_{l+1})\| \\ &\overset{(A.31)}{\geq} \sum_{l=0}^{r-1}\|Q_{k_l}\alpha_l(a_l) - Q_{k_l}\alpha_l(a_{l+1})\| \\ &\geq \left\|\sum_{l=0}^{r-1}(Q_{k_l}\alpha_l(a_l) - Q_{k_l}\alpha_l(a_{l+1}))\right\| \\ &= \left\|\sum_{l=0}^{r-1}(Q_{k_l}\alpha_l(a_l) - Q_{k_{l+1}}\alpha_{l+1}(a_{l+1}))\right\| \\ &= \|Q_{k_0}\alpha_0(a_0) - Q_{k_r}\alpha_r(a_r)\| \\ &= \left\|Q_i\alpha - Q_j\beta\right\|, \end{aligned} \tag{A.33}$$

which yields the assertion.

□

Proof of Proposition 9. Let $\Delta = \operatorname{conv}\{t^1, \cdots, t^{n+1}\}$ be given by affinely independent vertices $t^i \in \mathbf{R}^n$. We show that our lifting approach applied to the label space Δ solves the convexified unlifted problem, where the dataterm was replaced by its convex hull on Δ. Let the matrices $T \in \mathbf{R}^{n\times(n+1)}$ and $D \in \mathbf{R}^{(n+1)\times n}$ be defined through

$$T = \begin{pmatrix} t^1, & \cdots, & t^{n+1} \end{pmatrix}, \; D = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ -1 & \cdots & -1 \end{pmatrix}, \; TD = \begin{pmatrix} t^1 - t^{n+1}, & \cdots, & t^n - t^{n+1} \end{pmatrix}, \tag{A.34}$$

The transformation $x \mapsto t^{n+1} + TDx$ maps $\Delta_e = \operatorname{conv}\{0, e^1, \cdots, e^n\} \subset \mathbf{R}^n$ to Δ. Now consider the following lifted function $\boldsymbol{u} : \Omega \to \mathbf{R}^{n+1}$ parametrized through $\tilde{u} : \Omega \to \Delta_e$:

$$\boldsymbol{u}(x) = \begin{pmatrix} \tilde{u}_1(x), & \cdots, & \tilde{u}_n(x), & 1 - \sum_{j=1}^n \tilde{u}_j(x) \end{pmatrix}. \tag{A.35}$$

Consider a fixed $x \in \Omega$. Plugging this lifted representation into the biconjugate of the lifted dataterm $\boldsymbol{\rho}$ yields:

$$\begin{aligned}
\boldsymbol{\rho}^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{v}\in\mathbf{R}^{n+1}} \langle \boldsymbol{u}, \boldsymbol{v}\rangle - \sup_{\alpha\in\Delta_{n+1}^U} \langle \alpha, \boldsymbol{v}\rangle - \rho(T\alpha) \\
&= \sup_{\boldsymbol{v}\in\mathbf{R}^{n+1}} \left\langle \left(\tilde{u}_1(x), \cdots, \tilde{u}_n(x), 1 - \sum_{j=1}^n \tilde{u}_j(x) \right), \boldsymbol{v} \right\rangle - \\
&\qquad \sup_{\alpha\in\Delta_{n+1}^U} \langle \alpha, \boldsymbol{v}\rangle - \rho(T\alpha) \\
&= \sup_{\boldsymbol{v}\in\mathbf{R}^{n+1}} \langle \tilde{u}, D^\top \boldsymbol{v}\rangle + \boldsymbol{v}_{n+1} - \\
&\qquad \sup_{\alpha\in\Delta_{n+1}^U} \left\langle \left(\alpha_1, \cdots, \alpha_n, 1 - \sum_{j=1}^n \alpha_j \right), \boldsymbol{v} \right\rangle - \\
&\qquad\qquad \rho\left(\sum_{j=1}^n \alpha_j t^j + \left(1 - \sum_{j=1}^n \alpha_j \right) t^{n+1} \right) \\
&= \sup_{\boldsymbol{v}\in\mathbf{R}^{n+1}} \langle \tilde{u}, D^\top \boldsymbol{v}\rangle + \boldsymbol{v}_{n+1} - \sup_{\alpha\in\Delta_{n+1}^U} \boldsymbol{v}_{n+1} + \langle \alpha, D^\top \boldsymbol{v}\rangle - \rho(t^{n+1} + TD\alpha)
\end{aligned} \tag{A.36}$$

Since $D^\top$ is surjective, we can apply the substitution $\tilde{v} = D^\top \boldsymbol{v}$:

$$\begin{aligned}
\cdots &= \sup_{\tilde{v}\in\mathbf{R}^n} \langle \tilde{u}, \tilde{v}\rangle - \sup_{\alpha\in\Delta_{n+1}^U} \langle \alpha, \tilde{v}\rangle - \rho(t^{n+1} + TD\alpha) \\
&= \sup_{\tilde{v}\in\mathbf{R}^n} \langle \tilde{u}, \tilde{v}\rangle - \sup_{w\in\Delta} \langle (TD)^{-1}(w - t^{n+1}), \tilde{v}\rangle - \rho(w).
\end{aligned} \tag{A.37}$$

In the last step the substitution $w = t^{n+1} + TD\alpha \Leftrightarrow \alpha = (TD)^{-1}(w - t^{n+1})$ was performed. This can be further simplified to

$$\begin{aligned}
\cdots &= \sup_{\tilde{v}\in\mathbf{R}^n} \langle \tilde{u}, \tilde{v}\rangle + \langle (TD)^{-1}t^{n+1}, \tilde{v}\rangle - (\rho + \delta_\Delta)^*((TD)^{-T}\tilde{v}) \\
&= \sup_{\tilde{v}\in\mathbf{R}^n} \langle \tilde{u} + (TD)^{-1}t^{n+1}, \tilde{v}\rangle - (\rho + \delta_\Delta)^*((TD)^{-T}\tilde{v}) \\
&= \sup_{\tilde{v}\in\mathbf{R}^n} \langle TD\tilde{u} + t^{n+1}, (TD)^{-T}\tilde{v}\rangle - (\rho + \delta_\Delta)^*((TD)^{-T}\tilde{v}).
\end{aligned} \tag{A.38}$$

Since TD is invertible we can perform another substitution $v' = (TD)^{-T}\tilde{v}$.

$$\begin{aligned}
\cdots &= \sup_{v'\in\mathbf{R}^n} \langle TD\tilde{u} + t^{n+1}, v'\rangle - (\rho + \delta_\Delta)^*(v') \\
&= (\rho + \delta_\Delta)^{**}(t^{n+1} + TD\tilde{u}).
\end{aligned} \tag{A.39}$$

The lifted regularizer is given as:

$$\boldsymbol{R}(\boldsymbol{u}) = \sup_{\boldsymbol{q}:\Omega\to\mathbf{R}^{d\times n+1}} \int_\Omega \langle \boldsymbol{u}, \operatorname{div} \boldsymbol{q}\rangle - \Psi^*(\boldsymbol{q})\, \mathrm{d}x \tag{A.40}$$

Using the parametrization by $\tilde{u}$, this can be equivalently written as

$$\sup_{\boldsymbol{q}(x)\in\mathcal{K}} \int_\Omega \sum_{j=1}^{n} \tilde{u}_j \operatorname{div}(\boldsymbol{q}_j - \boldsymbol{q}_{n+1}) + \operatorname{div} \boldsymbol{q}_{n+1}\, \mathrm{d}x, \tag{A.41}$$

where the set $\mathcal{K} \subset \mathbf{R}^{d\times n+1}$ can be written as

$$\mathcal{K} = \{\boldsymbol{q} \in \mathbf{R}^{d\times n+1} \mid \|D^\top \boldsymbol{q}^\top (TD)^{-1}\|_{S^\infty} \le 1\}. \tag{A.42}$$

Note that since $\boldsymbol{q}_{n+1} \in C_c^\infty(\Omega, \mathbf{R}^d)$, the last term $\operatorname{div} \boldsymbol{q}_{n+1}$ in (A.41) vanishes by partial integration. With the substituion $\tilde{q}(x) = D^\top \boldsymbol{q}(x)^\top$ we have

$$\sup_{\tilde{q}\in\tilde{\mathcal{K}}} \int_\Omega \langle \tilde{u}, \operatorname{div} \tilde{q}\rangle\, \mathrm{d}x, \tag{A.43}$$

with set $\tilde{\mathcal{K}} \subset \mathbf{R}^{d\times n}$:

$$\tilde{\mathcal{K}} = \{q \in \mathbf{R}^{d\times n} \mid \|q(TD)^{-1}\|_{S^\infty} \le 1\}. \tag{A.44}$$

Note that since $\boldsymbol{q}_i \in C_c^\infty(\Omega, \mathbf{R}^d)$, the same holds for the linearly transformed $\tilde{q}$. With another substituion $q'(x) = \tilde{q}(x)(TD)^{-1}$ we have

$$\begin{aligned}
\cdots &= \sup_{q'\in\mathcal{K}'} \int_\Omega \langle \tilde{u}, \operatorname{div} q' TD\rangle\, \mathrm{d}x \\
&= \sup_{q'\in\mathcal{K}'} \int_\Omega \langle TD\tilde{u}, \operatorname{div} q'\rangle\, \mathrm{d}x
\end{aligned} \tag{A.45}$$

where the set $\mathcal{K}' \subset \mathbf{R}^{d\times n+1}$ is given as

$$\mathcal{K}' = \{q \in \mathbf{R}^{d\times n} \mid \|q\|_{S^\infty} \le 1\}, \tag{A.46}$$

which is the usual unlifted definition of the total variation $TV(t^{n+1} + TD\tilde{u})$.

This shows that the lifting method solves

$$\min_{\tilde{u}:\Omega\to\Delta_e} \int_\Omega (\rho(x,\cdot) + \delta_\Delta)^{**}(t^{n+1} + TD\tilde{u}(x))\mathrm{d}x + \lambda TV(t^{n+1} + TD\tilde{u}), \tag{A.47}$$

which is equivalent to the original problem but with a convexified data term. □

A.3 Chapter 5

Proposition 21. *For concave $\kappa : \mathbf{R}_0^+ \to \mathbf{R}$ with $\kappa(a) = 0 \Leftrightarrow a = 0$, the constraints*

$$\begin{aligned} &\Big\|(1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l) + \beta\hat{\varphi}_x(j)\Big\| \\ &\quad\le \frac{\kappa(\gamma_j^\beta - \gamma_i^\alpha)}{h}, \ \forall 1 \le i \le j \le k, \alpha, \beta \in [0,1], \end{aligned} \tag{A.48}$$

are equivalent to

$$\Big\|\sum_{l=i}^{j} \hat{\varphi}_x(l)\Big\| \le \frac{\kappa(\gamma_{j+1} - \gamma_i)}{h}, \forall 1 \le i \le j \le k. \tag{A.49}$$

Proof. The implication (A.48) ⇒ (A.49) clearly holds. Let us now assume the constraints (A.49) are fulfilled. First we show that the constraints (A.48) also hold for $\alpha \in [0,1]$ and $\beta \in \{0,1\}$. First, we start with $\beta = 0$:

$$\begin{aligned} &\Big\|(1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l)\Big\| \\ &= \Big\|(1-\alpha)\sum_{l=i}^{j-1} \hat{\varphi}_x(l) + \alpha \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l)\Big\| \\ &\le (1-\alpha)\Big\|\sum_{l=i}^{j-1} \hat{\varphi}_x(l)\Big\| + \alpha\Big\|\sum_{l=i+1}^{j-1} \hat{\varphi}_x(l)\Big\| \\ &\overset{\text{by (A.49)}}{\le} (1-\alpha)\frac{1}{h}\kappa(\gamma_j - \gamma_i) + \alpha\frac{1}{h}\kappa(\gamma_j - \gamma_{i+1}) \\ &\overset{\text{concavity}}{\le} \frac{1}{h}\big(\kappa((1-\alpha)(\gamma_j - \gamma_i) + \alpha(\gamma_j - \gamma_{i+1}))\big) = \frac{1}{h}\kappa(\gamma_j^0 - \gamma_i^\alpha). \end{aligned} \tag{A.50}$$

In the same way, it can be shown that for $\beta = 1$ we have:

$$\Big\|(1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l) + 1\cdot\hat{\varphi}_x(j)\Big\| \le \frac{1}{h}\kappa(\gamma_j^1 - \gamma_i^\alpha). \tag{A.51}$$

We have shown the constraints to hold for $\alpha \in [0,1]$ and $\beta \in \{0,1\}$. Finally we show they also hold for $\beta \in [0,1]$:

$$\begin{aligned}
&\|(1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l) + \beta\hat{\varphi}_x(j)\| \\
&= \|(1-\alpha)\hat{\varphi}_x(i) + (1-\beta)\sum_{l=i+1}^{j-1} \hat{\varphi}_x(l) + \beta \sum_{l=i+1}^{j} \hat{\varphi}_x(l)\| \\
&= \|(1-\beta)\left((1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l)\right) + \beta\left((1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j} \hat{\varphi}_x(l)\right)\| \\
&\leq (1-\beta)\|(1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j-1} \hat{\varphi}_x(l)\| + \beta\|(1-\alpha)\hat{\varphi}_x(i) + \sum_{l=i+1}^{j} \hat{\varphi}_x(l)\| \\
&\overset{(A.50),(A.51)}{\leq} \frac{1}{h}(1-\beta)\kappa(\gamma_j^0 - \gamma_i^\alpha) + \beta\kappa(\gamma_j^1 - \gamma_i^\alpha) \\
&\overset{\text{concavity}}{\leq} \frac{1}{h}\kappa((1-\beta)(\gamma_j^0 - \gamma_i^\alpha) + \beta(\gamma_j^1 - \gamma_i^\alpha)) = \frac{1}{h}\kappa(\gamma_j^\beta - \gamma_i^\alpha)
\end{aligned} \tag{A.52}$$

Noticing that (A.49) is precisely (A.48) for $\alpha, \beta \in \{0,1\}$ (as $\kappa(a) = 0 \Leftrightarrow a = 0$) completes the proof. □

Proposition 22. *For convex one-homogeneous η the discretization with piecewise constant φ_t and φ_x leads to the traditional discretization as proposed in [Poc+09a], except with min-pooled instead of sampled unaries.*

Proof. The constraints in [Poc+09a, Eq. 18] have the form

$$\hat{\varphi}_t(i) \geq \eta^*(\hat{\varphi}_x(i)) - \rho(\gamma_i), \tag{A.53}$$

$$\left\|\sum_{l=i}^{j} \hat{\varphi}_x(l)\right\| \leq \kappa(\gamma_{j+1} - \gamma_i), \tag{A.54}$$

with $\rho(u) = \lambda(u-f)^2$, $\eta(g) = \|g\|^2$ and $\kappa(a) = \nu[\![a > 0]\!]$. The constraints (A.54) are equivalent to (A.49) up to a rescaling of $\hat{\varphi}_x$ with h. For the constraints (A.53) (cf. [Poc+09a, Eq. 18]), the unaries are sampled at the labels γ_i. The discretization with piecewise constant duals leads to a similar form, except for a min-pooling on dual intervals, $\forall 1 \leq i \leq k$:

$$\begin{aligned}
\hat{\varphi}_t(i) &\geq \eta^*(\hat{\varphi}_x(i)) - \inf_{t \in [\gamma_i, \gamma_i^*]} \rho(t), \\
\hat{\varphi}_t(i+1) &\geq \eta^*(\hat{\varphi}_x(i)) - \inf_{t \in [\gamma_i^*, \gamma_{i+1}]} \rho(t).
\end{aligned} \tag{A.55}$$

The similarity between (A.55) and (A.53) becomes more evident by assuming convex one-homogeneous η. Then (A.55) reduces to the following:

$$\begin{aligned} -\hat{\varphi}_t(1) &\leq \inf_{t\in[\gamma_1,\gamma_1^*]} \rho(t), \\ -\hat{\varphi}_t(i) &\leq \inf_{t\in\Gamma_i^*} \rho(t), \ \forall i \in \{2,\cdots,\ell-1\}, \\ -\hat{\varphi}_t(\ell) &\leq \inf_{t\in[\gamma_{\ell-1}^*,\gamma_\ell]} \rho(t), \end{aligned} \tag{A.56}$$

as well as

$$\hat{\varphi}_x(i) \in \operatorname{dom}(\eta^*), \forall i \in \{1,\cdots,k\}. \tag{A.57}$$

□

Proposition 23. *The constraints*

$$\inf_{t\in\Gamma_i} \hat{\varphi}_t(i)\frac{\gamma_{i+1}-t}{h} + \hat{\varphi}_t(i+1)\frac{t-\gamma_i}{h} + \rho(t) \geq \eta^*(\hat{\varphi}_x(i)). \tag{A.58}$$

can be equivalently reformulated by introducing additional variables $a \in \mathbf{R}^k$, $b \in \mathbf{R}^k$, where $\forall i \in \{1,\cdots,k\}$:

$$\begin{aligned} &r(i) = (\hat{\varphi}_t(i) - \hat{\varphi}_t(i+1))/h, \\ &a(i) + b(i) - (\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(x,i+1)\gamma_i)/h = 0, \\ &r(i) \geq \rho_i^*\left(a(i)\right), \hat{\varphi}_x(i) \geq \eta^*\left(b(i)\right), \end{aligned} \tag{A.59}$$

with $\rho_i(x,t) = \rho(x,t) + \delta\{t \in \Gamma_i\}$.

Proof. Rewriting the infimum in (A.58) as minus the convex conjugate of ρ_i, and multiplying the inequality with -1 the constraints become:

$$\begin{aligned} &\rho_i^*(r(i)) + \eta^*(\hat{\varphi}_x(i)) - \frac{\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(i+1)\gamma_i}{h} \leq 0, \\ &r(i) = (\hat{\varphi}(i) - \hat{\varphi}(i+1))/h. \end{aligned} \tag{A.60}$$

To show that (A.60) and (A.59) are equivalent, we prove that they imply each other. Assume (A.60) holds. Then without loss of generality set $a(i) = \rho_i^*(r(i)) + \xi_1$, $b(i) = \eta_i^*(\varphi_x(i)) + \xi_2$ for some $\xi_1, \xi_2 \geq 0$. Clearly, this choice fulfills (A.60). Since for $\xi_1 = \xi_2 = 0$ we have by assumption that

$$a(i) + b(i) - (\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(x,i+1)\gamma_i)/h \leq 0, \tag{A.61}$$

there exists some $\xi_1, \xi_2 \geq 0$ such that (A.59) holds.

Now conversely assume (A.59) holds. Since $a(i) \geq \rho_i^*\left(r(i)\right)$, $b(i) \geq \eta^*\left(\hat{\varphi}_x(i)\right)$, and

$$a(i) + b(i) - (\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(x,i+1)\gamma_i)/h = 0, \tag{A.62}$$

this directly implies

$$\rho_i^*(r(i)) + \eta^*(\hat{\varphi}_x(i)) - \frac{\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(i+1)\gamma_i}{h} \leq 0, \tag{A.63}$$

since the left-hand side becomes smaller by plugging in the lower bound. □

Proposition 24. *The discretization with piecewise linear φ_t and piecewise constant φ_x together with the choice $\eta(g) = \|g\|$ and $\kappa(a) = a$ is equivalent to the relaxation [Möl+16].*

Proof. Since $\eta(g) = \|g\|$, the constraints (A.58) become

$$\begin{aligned}&\inf_{t \in \Gamma_i} \hat{\varphi}_t(i)\frac{\gamma_{i+1} - t}{h} + \hat{\varphi}_t(i+1)\frac{t - \gamma_i}{h} + \rho(t) \geq 0.\\ &\varphi_x \in \operatorname{dom}(\eta^*).\end{aligned} \tag{A.64}$$

This decouples the constraints into data term and regularizer. The data term constraints can be written using the convex conjugate of $\rho_i = \rho + \delta\{\cdot \in \Gamma_i\}$ as the following:

$$\frac{\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(i+1)\gamma_i}{h} - \rho_i^*\left(\frac{\hat{\varphi}_t(i) - \hat{\varphi}_t(i+1)}{h}\right) \geq 0. \tag{A.65}$$

Let $\boldsymbol{v}_i = \hat{\varphi}_t(i) - \hat{\varphi}_t(i+1)$ and $q = \hat{\varphi}_t(1)$. Then we can write (A.65) as a telescope sum over the $\boldsymbol{v}_i$

$$q - \sum_{j=1}^{i-1} \boldsymbol{v}_j + \frac{\gamma_i}{h}\boldsymbol{v}_i - \rho_i^*\left(\frac{\boldsymbol{v}_i}{h}\right) \geq 0, \tag{A.66}$$

which is the same as the constraints in [Möl+16, Eq. 9, Eq. 22]. The cost function is given as

$$-\hat{\varphi}_t(1) - \sum_{i=1}^{k} \hat{v}(i)\left[\hat{\varphi}_t(i+1) - \hat{\varphi}_t(i)\right] = \langle \hat{v}, \boldsymbol{v} \rangle - q, \tag{A.67}$$

which is exactly the first part of [Möl+16, Eq. 21]. Finally, for the regularizer we get

$$\Big\|\sum_{l=i}^{j} \hat{\varphi}_x(l)\Big\| \leq \frac{|\gamma_{j+1} - \gamma_i|}{h}, \; \|\hat{\varphi}_x(i)\| \leq 1, \tag{A.68}$$

which clearly reduces to the same set as in [Möl+16, Proposition 5], by applying that proposition (and with the rescaling/substitution $p = h \cdot \varphi_x$).

□

Proposition 25. *The data term from [Möl+16] (which is in turn a special case of the discretization with piecewise linear φ_t) can be (pointwise) brought into the primal form*

$$\mathcal{D}(\widehat{v}) = \inf_{\substack{x_i \geq 0, \sum_i x_i = 1\\ \widehat{v} = y/h + I^\top x}} \sum_{i=1}^{k} x_i \rho_i^{**}\left(\frac{y_i}{x_i}\right), \tag{A.69}$$

where $I \in \mathbf{R}^{k \times k}$ is a discretized integration operator.

Proof. The equivalence of the sublabel accurate data term proposed in [Möl+16] to the discretization with piecewise linear φ_t is established in Proposition 24 (cf. (A.66) and (A.67)). It is given pointwise as

$$\mathcal{D}(\widehat{v}) = \max_{\boldsymbol{v},q} \langle \boldsymbol{v}, \widehat{v} \rangle - q - \sum_{i=1}^{k} \delta\left\{\left(\frac{\boldsymbol{v}_i}{h}, [q\mathbf{1}_k - I\boldsymbol{v}]_i\right) \in \operatorname{epi}(\rho_i^*)\right\}, \tag{A.70}$$

where $\widehat{v} \in \mathbf{R}^k$, $\boldsymbol{v} \in \mathbf{R}^k$, $q \in \mathbf{R}$, and k is the number of pieces and $\mathbf{1}_k \in \mathbf{R}^k$ is the vector consisting only of ones. Furthermore, $\rho_i(t) = \rho(t) + \delta\{t \in \Gamma_i\}, \operatorname{dom}(\rho_i) = \Gamma_i = [\gamma_i, \gamma_{i+1}]$. The integration operator $I \in \mathbf{R}^{k \times k}$ is defined as

$$I = \begin{bmatrix} -\frac{\gamma_1}{h} & & & \\ 1 & -\frac{\gamma_2}{h} & & \\ & & \ddots & \\ 1 & \cdots & 1 & -\frac{\gamma_k}{h} \end{bmatrix}. \tag{A.71}$$

Using convex duality, and the substitution $h\tilde{v} = \boldsymbol{v}$ we can rewrite (A.70) as

$$\min_{x} \max_{\tilde{v},q,z} \langle \tilde{v}, h \cdot \widehat{v} \rangle - q - \langle x, z - (q\mathbf{1}_k - hI\tilde{v}) \rangle - \sum_{i=1}^{k} \delta\left\{(\tilde{v}_i, z_i) \in \operatorname{epi}(\rho_i^*)\right\}. \tag{A.72}$$

The convex conjugate of $F_i(z, v) = \delta\{(v, -z) \in \operatorname{epi}(\rho_i^*)\}$ is the lower-semicontinuous envelope of the perspective [Roc96, Section 15], and since $\rho_i : \Gamma_i \to \mathbf{R}$ has bounded domain, is given as the following (cf. also [ZK12, Appendix 3])

$$F_i^*(x, y) = \begin{cases} x\rho_i^{**}(y/x), & \text{if } x > 0, \\ 0, & \text{if } x = 0 \wedge y = 0, \\ \infty, & \text{if } x < 0 \vee (x = 0 \wedge y \neq 0). \end{cases} \tag{A.73}$$

Thus with the convention that $0/0 = 0$ equation (A.72) can be rewritten as convex conjugates:

$$\begin{aligned} &\min_{x} \left(\max_{q} q(\mathbf{1}_k^\top x) - q\right) + \\ &\left(\max_{\tilde{v},z} \langle \tilde{v}, h \cdot (\widehat{v} - I^\top x) \rangle + \langle -z, x \rangle - \sum_{i=1}^{k} F_i(-z_i, \tilde{v}_i)\right) \\ &= \min_{x} \delta\left\{\sum_i x_i = 1\right\} + \sum_i F_i^*\left(x_i, [h(\widehat{v} - I^\top x)]_i\right). \end{aligned} \tag{A.74}$$

Hence we have that

$$\mathcal{D}(\widehat{v}) = \min_{\substack{x,y \\ y = h(\widehat{v} - I^\top x) \\ x_i \geq 0 \\ \sum_i x_i = 1 \\ y_i/x_i \in \operatorname{dom}(\rho_i^{**})}} \sum_i x_i \rho_i^{**}\left(\frac{y_i}{x_i}\right), \tag{A.75}$$

which can be rewritten in the form (A.70). □

Proposition 26. *Let $\gamma = \kappa(\gamma_2 - \gamma_1)$ and $\ell = 2$. The approximation with piecewise linear φ_t and piecewise constant φ_x of the continuous optimization problem*

$$\inf_{v \in \mathcal{C}} \sup_{\varphi \in \mathcal{K}} \int_{\Omega \times \mathbf{R}} \langle \varphi, Dv \rangle. \tag{A.76}$$

is equivalent to

$$\inf_{u : \Omega \to \Gamma} \int_{\Omega} \rho^{**}(x, u(x)) + (\eta^{**} \,\square\, \gamma \| \cdot \|)(\nabla u(x)) \, \mathrm{d}x, \tag{A.77}$$

where $(\eta \,\square\, \gamma \| \cdot \|)(x) = \inf_y \, \eta(x - y) + \gamma \|y\|$ denotes the infimal convolution (cf. [Roc96, Section 5]).

Proof. Plugging in the representations for piecewise linear φ_t and piecewise constant φ_x we have the coefficient functions $\hat{v} : \Omega \to [0, 1]$, $\hat{\varphi}_t : \Omega \times \{1, 2\} \to \mathbf{R}$, $\hat{\varphi}_x : \Omega \to \mathbf{R}^n$ and the following optimization problem:

$$\begin{aligned} &\inf_{\hat{v}} \sup_{\hat{\varphi}_x, \hat{\varphi}_t} \int_{\Omega} -\hat{\varphi}_t(x, 1) - \hat{v}(x) \left[\hat{\varphi}_t(x, 2) - \hat{\varphi}_t(x, 1) \right] - h \cdot \hat{v}(x) \cdot \operatorname{div}_x \hat{\varphi}_x(x) \, \mathrm{d}x \\ &\text{subject to } \inf_{t \in \Gamma} \hat{\varphi}_t(x, 1) \frac{\gamma_2 - t}{h} + \hat{\varphi}_t(x, 2) \frac{t - \gamma_1}{h} + \rho(x, t) \geq \eta^*(x, \hat{\varphi}_x(x)) \\ &\qquad\qquad \| \hat{\varphi}_x(x) \| \leq \kappa(\gamma_2 - \gamma_1) =: \gamma. \end{aligned} \tag{A.78}$$

Using the convex conjugate of $\rho : \Omega \times \Gamma \to \mathbf{R}$ (in its second argument), we rewrite the first constraint as

$$\frac{\hat{\varphi}_t(x, 1)\gamma_2 - \hat{\varphi}_t(x, 2)\gamma_1}{h} \geq \rho^* \left(x, \frac{\hat{\varphi}_t(x, 1) - \hat{\varphi}_t(x, 2)}{h} \right) + \eta^*(x, \hat{\varphi}_x(x)). \tag{A.79}$$

Using the substitution $\tilde{\varphi}(x) = \frac{\hat{\varphi}_t(x,1) - \hat{\varphi}_t(x,2)}{h}$ we can reformulate the constraints as

$$\hat{\varphi}_t(x, 1) \geq \rho^*(x, \tilde{\varphi}(x)) + \eta^*(x, \hat{\varphi}_x(x)) - \gamma_1 \tilde{\varphi}(x), \tag{A.80}$$

and the cost function as

$$\sup_{\tilde{\varphi}, \hat{\varphi}_t, \hat{\varphi}_x} \int_{\Omega} -\hat{\varphi}_t(x, 1) + h\hat{v}(x)\tilde{\varphi}(x) - h\hat{v}(x) \operatorname{div}_x \hat{\varphi}_x(x) \mathrm{d}x. \tag{A.81}$$

The pointwise supremum over $-\hat{\varphi}_t(x, 1)$ is attained where the constraint (A.80) is sharp, which means we can pull it into the cost function to arrive at

$$\begin{aligned} \sup_{\tilde{\varphi}, \hat{\varphi}_x} \int_{\Omega} &-\rho^*(x, \tilde{\varphi}(x)) - \eta^*(x, \hat{\varphi}_x(x)) - \delta\{ \| \hat{\varphi}_x(x) \leq \gamma \| \} + \\ &\gamma_1 \tilde{\varphi}(x) + h\hat{v}(x)\tilde{\varphi}(x) - h\hat{v}(x) \operatorname{div}_x \hat{\varphi}_x(x) \mathrm{d}x, \end{aligned} \tag{A.82}$$

where we wrote the second constraint in (A.78) as an indicator function. As the supremum decouples in $\tilde{\varphi}$ and $\hat{\varphi}_x$, we can rewrite it using convex (bi-)conjugates, by interchanging integration and supremum (cf. [RWW98, Theorem 14.60]):

$$\begin{aligned}\sup_{\tilde{\varphi}} \int_\Omega \gamma_1\tilde{\varphi}(x) + h\hat{v}(x)\tilde{\varphi}(x) - \rho^*(x,\tilde{\varphi}(x))\mathrm{d}x \\ = \int_\Omega \sup_{\tilde{\varphi}}\ \gamma_1\tilde{\varphi} + h\hat{v}(x)\tilde{\varphi} - \rho^*(x,\tilde{\varphi})\mathrm{d}x \\ = \int_\Omega \rho^{**}(x,\gamma_1 + h\hat{v}(x))\,\mathrm{d}x.\end{aligned} \tag{A.83}$$

For the part in $\hat{\varphi}_x$ we assume that $\hat{v}$ is sufficiently smooth and apply partial integration ($\hat{\varphi}_x$ vanishes on the boundary), and then perform a similar calculation to the previous one:

$$\begin{aligned}&\sup_{\hat{\varphi}_x} \int_\Omega -(\eta^* + \delta\{\|\cdot\| \le \gamma\})(x,\hat{\varphi}_x(x)) - h\hat{v}(x)\,\mathrm{div}_x\,\hat{\varphi}_x(x)\mathrm{d}x \\ &= \sup_{\hat{\varphi}_x} \int_\Omega -(\eta^* + \delta\{\|\cdot\| \le \gamma\})(x,\hat{\varphi}_x(x)) + h\langle\nabla_x\hat{v}(x),\hat{\varphi}_x(x)\rangle\mathrm{d}x \\ &= \int_\Omega \sup_{\hat{\varphi}_x} -(\eta^* + \delta\{\|\cdot\| \le \gamma\})(x,\hat{\varphi}_x) + h\langle\nabla_x\hat{v}(x),\hat{\varphi}_x\rangle\mathrm{d}x \\ &= \int_\Omega (\eta^* + \delta\{\|\cdot\| \le \gamma\})^*(x,h\nabla_x\hat{v}(x))\mathrm{d}x \\ &= \int_\Omega (\eta^{**} \ \Box\ \gamma\|\cdot\|)(x,h\nabla_x\hat{v}(x))\mathrm{d}x \\ &= \int_\Omega (\eta \ \Box\ \gamma\|\cdot\|)(x,h\nabla_x\hat{v}(x))\mathrm{d}x.\end{aligned} \tag{A.84}$$

Here we used also the result that $(f^* + g)^* = f^{**} \ \Box\ g^*$ [RWW98, Theorem 11.23]. Combining (A.83) and (A.84) and using the substitution $u = \gamma_1 + h\hat{v}$, we finally arrive at:

$$\int_\Omega \rho^{**}(x,u(x)) + (\eta^{**} \ \Box\ \gamma\|\cdot\|)(x,\nabla u(x))\,\mathrm{d}x, \tag{A.85}$$

which is the same as (A.77). □

A.4 Chapter 7

Proof of Proposition 20. Since $g_{\theta\,\sharp}S$ and T are normal currents we know that $\mathbb{F}_\lambda(g_{\theta\,\sharp}S, T) < \infty$ for all $\theta \in \Theta$. We now directly show Lipschitz continuity. First notice that

$$\mathbb{F}_\lambda(g_{\theta\,\sharp}S - T) = \mathbb{F}_\lambda(g_{\theta\,\sharp}S + g_{\theta'\,\sharp}S - g_{\theta'\,\sharp}S - T) \tag{A.86}$$

$$\le \mathbb{F}_\lambda(g_{\theta\,\sharp}S - g_{\theta'\,\sharp}S) + \mathbb{F}_\lambda(g_{\theta'\,\sharp}S - T), \tag{A.87}$$

yields the following bound:

$$|\mathbb{F}_\lambda(g_{\theta\,\sharp}S - T) - \mathbb{F}_\lambda(g_{\theta'\,\sharp}S - T)| \le \mathbb{F}_\lambda(g_{\theta\,\sharp}S - g_{\theta'\,\sharp}S). \tag{A.88}$$

Due to Prop. 18 we have that

$$\mathbb{F}_\lambda(g_{\theta\,\sharp}S - g_{\theta'\,\sharp}S) \le \max\{1, \lambda\} \cdot \mathbb{F}(g_{\theta\,\sharp}S - g_{\theta'\,\sharp}S). \tag{A.89}$$

Now define the compact set $C \subset \mathbf{R}^d$ as

$$C = \left\{(1-t)g_\theta(z) + tg_{\theta'}(z) : z \in \operatorname{spt} S, 0 \le t \le 1\right\}, \tag{A.90}$$

and as in §4.1.12 in [Fed69] for compact $K \subset \mathbf{R}^d$ the "stronger" flat norm

$$\begin{aligned}\mathbb{F}_K(T) = \sup\{&T(\omega) \mid \omega \in \mathcal{D}^k(\mathbf{R}^d), \text{ with} \\ &\|\omega(x)\|^* \le 1, \|d\omega(x)\|^* \le 1 \text{ for all } x \in K\}.\end{aligned} \tag{A.91}$$

Since the constraint in the supremum in (A.91) is less restrictive than in the definition of the flat norm (7.20), we have

$$\mathbb{F}(g_{\theta\,\sharp}S - g_{\theta'\,\sharp}S) \le \mathbb{F}_C(g_{\theta\,\sharp}S - g_{\theta'\,\sharp}S). \tag{A.92}$$

Then, the inequality after §4.1.13 in [Fed69] bounds the right side of (A.92) for $k > 0$ by

$$\begin{aligned}&\mathbb{F}_C(g_{\theta\,\sharp}S - g_{\theta'\,\sharp}S) \le \\ &\quad \|S\|(|g_\theta - g_{\theta'}|\rho^k) + \|\partial S\|(|g_\theta - g_{\theta'}|\rho^{k-1}),\end{aligned} \tag{A.93}$$

where $\rho(z) = \max\{\|\nabla_Z g(z,\theta)\|, \|\nabla_Z g(z,\theta')\|\} < \infty$ due to Assumption 1 and we write $\|S\|(f) = \int f(z)\,\mathrm{d}\|S\|(z)$, where $\|S\|$ is defined in the sense of (7.19). For $k = 0$, a similar bound can be derived without the term $\|\partial S\|$.

For $k > 0$, by setting $\mu_S = \|\partial S\| + \|S\|$ we can further bound the term in (A.93) by

$$\begin{aligned}&\|S\|(|g_\theta - g_{\theta'}|\rho^k) + \|\partial S\|(|g_\theta - g_{\theta'}|\rho^{k-1}) \le \\ &\quad c_1 \cdot \int \|g_\theta(z) - g_{\theta'}(z)\|\mathrm{d}\mu_S(z),\end{aligned} \tag{A.94}$$

where $c_1 = \sup_z \max\{\rho^k(z), \rho^{k-1}(z)\}$. For $k = 0$, the bound is derived analogously.

Now since $g(z,\cdot)$ is locally Lipschitz and $\Theta \subset \mathbf{R}^n$ is compact, $g(z,\cdot)$ is Lipschitz and we denote the constant as $\operatorname{Lip}(g)$, leading to the bound

$$\int \|g_\theta(z) - g_{\theta'}(z)\|\mathrm{d}\mu_S(z) \le \mu_S(\mathcal{Z}) \operatorname{Lip}(g) \cdot \|\theta - \theta'\|. \tag{A.95}$$

Since $S \in \boldsymbol{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ is a normal current, $\mu_S(\mathcal{Z}) < \infty$. Thus by combining (A.88), (A.89), (A.92), (A.93), (A.94) and (A.95) there is a finite $c_2 = \max\{1,\lambda\} \cdot c_1 \cdot \mu_S(\mathcal{Z}) \cdot \operatorname{Lip}(g) < \infty$ such that

$$|\mathbb{F}_\lambda(g_{\theta\,\sharp}S - T) - \mathbb{F}_\lambda(g_{\theta'\,\sharp}S - T)| \le c_2\|\theta - \theta'\|. \tag{A.96}$$

Therefore, the cost $\mathbb{F}_\lambda(g_{\theta\,\sharp}S, T)$ in (7.27) is Lipschitz in θ and by Rademacher's theorem, §3.1.6 in [Fed69], also differentiable almost everywhere. □

Appendix B

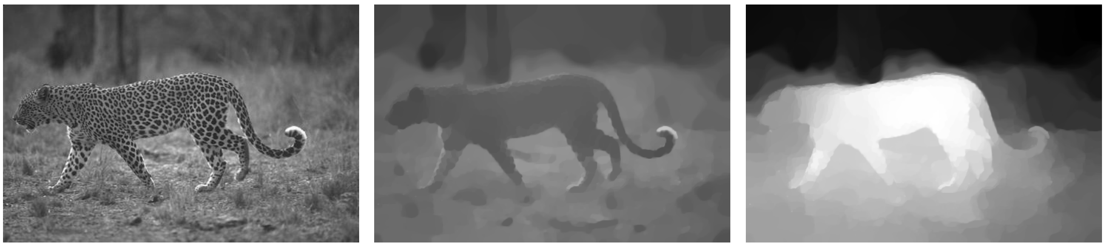
Supplementary Materials

B.1 Chapter 4

In this experiment we jointly estimate the mean μ and variance σ of an image $I : \Omega \to \mathbf{R}$ according to a Gaussian model. The label space is chosen as $\Gamma = [0, 255] \times [1, 10]$ and the dataterm as proposed in [GSC13]:

$$\rho(x, \mu(x), \sigma(x)) = \frac{(\mu(x) - I(x))^2}{2\sigma(x)^2} + \frac{1}{2}\log(2\pi\sigma(x)^2). \tag{B.1}$$

As the projection onto the epigraph of $(\rho + \delta_\Delta)^*$ seems difficult to compute, we approximate ρ by a piecewise linear function using 29 × 29 sublabels and convexify it using the quickhull algorithm [BDH96]. In Fig. B.1 we show the result of minimizing (B.1) with total variation regularization.



Input image

Mean μ

Variance σ

Figure B.1: Joint estimation of mean and variance. Our formulation can optimize difficult nonconvex joint optimization problems with continuous label spaces.

B.2 Chapter 6

In Fig. B.2 we show an additional experiment, in which we compute a minimal current under the shortest path energy $c(x, y, \xi) = \sqrt{1 + \|\xi\|^2}$. The center of mass coincides with the analytical (straight line) solution, despite the coarse mesh.

In Fig. B.3 we show that the bijective correspondences computed by solving a linear assignment problem can be quite noisy. The linear assignment problem (corresponding to the Kantorovich relaxation in optimal transport) globally optimizes a cost $c(x, y) = \|I_1(x) - I_2(y)\|$, which does not contain any spatial regularization.

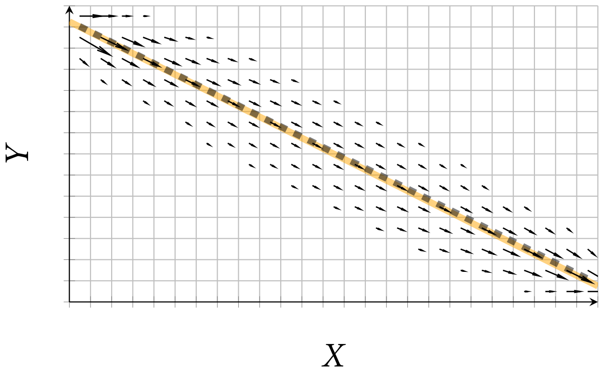
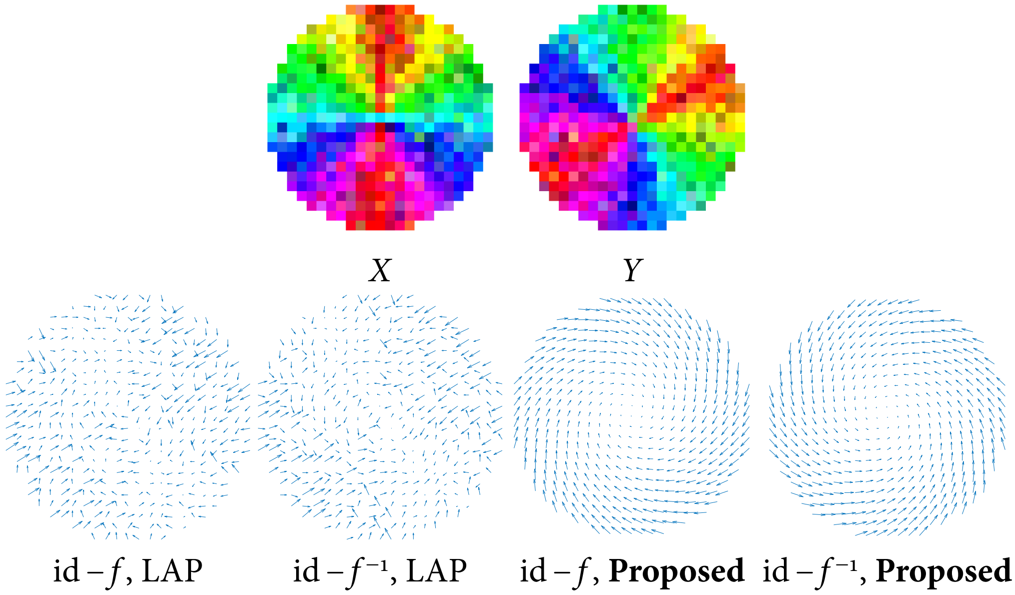


Figure B.2: Minimal current on a cubical complex of size 25 × 14, here under a minimal surface (shortest path) energy with Dirichlet boundary conditions.



$\mathrm{id} - f$, LAP $\quad$ $\mathrm{id} - f^{-1}$, LAP $\quad$ $\mathrm{id} - f$, **Proposed** $\quad$ $\mathrm{id} - f^{-1}$, **Proposed**

Figure B.3: Solving a linear assignment problem (denoted by LAP) yields a quite noisy solution when compared to the spatially regularized correspondence field obtained by the proposed relaxation.

B.3 Chapter 7

For all experiments we use Adam optimizer [KB14], with step size 10^{-4} and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.9$. The batch size is set to 50 in all experiments except the first one (which runs full batch with batch size 5). We always set $\lambda = 1$.

Illustrative 2D Example. We pick the same parameters for $k \in \{0, 1\}$. We set the penalty to $\rho = 10$ and use 5 discriminator updates per generator update as in [Gul+17]. The generator is a 5 – 6 – 250 – 250 – 250 – 2 fully connected network with leaky ReLU activations. The first layer ensures that the latent coordinate z_1 has the topology of a circle, i.e., it is implemented as $(\cos(z_1), \sin(z_1), z_2, z_3, z_4, z_5)$. The discriminators ω^0 and $\omega^{1,1}$ are 2 – 100 – 100 – 100 – 1 respectively 2 – 100 – 100 – 2 nets with leaky ReLUs. The distribution on the latent is a uniform $z_1 \sim U([-\pi, \pi])$ and $z_i \sim \mathcal{N}(0, 1)$ for the remaining 4 latent codes.

MNIST. For the remaining experiments, we use only 1 discriminator update per iteration. The digits are resized to 32 × 32. For generator we use DCGAN architecture [RMC15] without batch norm and with ELU activations, see Table B.1. The discriminators are given by the architectures in Table B.2, with leaky ReLUs between the layers.

layer name	output size	filters
Reshape	$128 \times 1 \times 1$	–
Conv2DTranspose	$32F \times 4 \times 4$	$128 \to 32F$
Conv2DTranspose	$16F \times 8 \times 8$	$32F \to 16F$
Conv2DTranspose	$4F \times 16 \times 16$	$16F \to 4F$
Conv2DTranspose	$1 \times 32 \times 32$	$4F \to 1$

Table B.1: Generator architecture for MNIST experiment, $F = 32$.

Before computing $\langle \omega^{1,1}(x) \wedge \omega^{1,2}(x), v_1 \wedge v_2 \rangle$, the tangent images $v_1, v_2 \in \mathbf{R}^{32 \cdot 32}$ are convolved with a Gaussian with a standard deviation of 2 and downsampled to 8 × 8 using average pooling. The distributions on the latent space are given by $z_1 \sim U([-7.5, 7.5])$, $z_2 \sim U([-0.5, 0.5])$ and $z_i \sim \mathcal{N}(0, 1)$ for the remaining 126 latent variables. The tangent vectors at each sample are computed by a 2 degree rotation and a dilation with radius one.

SmallNORB. We downsample the smallNORB images to 48 × 48. The architectures and parameters are chosen similar to the previous MNIST example, see Table B.3 and Table B.4.

layer name	output size	filters
Reshape	$1 \times 32 \times 32$	–
Conv2D	$2F \times 16 \times 16$	$1 \to 2F$
Conv2D	$4F \times 8 \times 8$	$2F \to 4F$
Conv2D	$32F \times 4 \times 4$	$4F \to 32F$
Conv2D	$1 \times 1 \times 1$	$32F \to 1$
Conv2DTranspose	$1 \times 8 \times 8$	$32F \to 1$

Table B.2: The discriminator ω^0 has $F = 32$ and red last layer. The discriminators $\omega^{1,1}$, $\omega^{1,2}$ have $F = 8$ and last layer in blue.

layer name	output size	filters
Reshape	$128 \times 1 \times 1$	–
Conv2DTranspose	$32F \times 4 \times 4$	$128 \to 32F$
Conv2DTranspose	$16F \times 8 \times 8$	$32F \to 16F$
Conv2DTranspose	$16F \times 12 \times 12$	$16F \to 16F$
Conv2DTranspose	$4F \times 24 \times 24$	$16F \to 4F$
Conv2DTranspose	$1 \times 48 \times 48$	$4F \to 1$

Table B.3: Generator for smallNORB experiment, $F = 24$.

layer name	output size	filters
Reshape	$1 \times 48 \times 48$	–
Conv2D	$2F \times 24 \times 24$	$1 \to 2F$
Conv2D	$4F \times 12 \times 12$	$2F \to 4F$
Conv2D	$32F \times 6 \times 6$	$4F \to 32F$
Conv2D	$1 \times 1 \times 1$	$32F \to 1$
Conv2DTranspose	$1 \times 12 \times 12$	$32F \to 1$

Table B.4: SmallNORB discriminator ω^0, $F = 32$, last layer in shown in red, and tangent discriminators $\omega^{1,1}$, $\omega^{1,2}$, $\omega^{1,3}$ where $F = 8$ and last layer is highlighted in blue.

Tinyvideos. The architectures for the tinyvideo experiment are borrowed from the recent work [MGN18], see Table B.5 and Table B.6.

layer name	output size	filters
Fully Connected	8192	–
Reshape	512 × 4 × 4	–
ResNet-Block	512 × 4 × 4	512 → 512 → 512
NN-Upsampling	512 × 8 × 8	–
ResNet-Block	256 × 8 × 8	512 → 256 → 256
NN-Upsampling	256 × 16 × 16	–
ResNet-Block	128 × 16 × 16	256 → 128 → 128
NN-Upsampling	128 × 32 × 32	–
ResNet-Block	64 × 32 × 32	128 → 64 → 64
NN-Upsampling	64 × 64 × 64	–
ResNet-Block	64 × 64 × 64	64 → 64 → 64
Conv2D	3 × 64 × 64	64 → 3

Table B.5: Generator architecture for tinyvideos experiment.

layer name	output size	filters
Conv2D	64 × 64 × 64	3 → 64
ResNet-Block	64 × 64 × 64	64 → 64 → 64
AvgPool2D	64 × 32 × 32	–
ResNet-Block	128 × 32 × 32	64 → 64 → 128
AvgPool2D	128 × 16 × 16	–
ResNet-Block	256 × 16 × 16	128 → 128 → 256
AvgPool2D	256 × 8 × 8	–
ResNet-Block	512 × 8 × 8	256 → 256 → 512
AvgPool2D	512 × 4 × 4	–
ResNet-Block	1024 × 4 × 4	512 → 512 → 1024
Conv2D	1 × 1 × 1	1024 → 1
ResNet-Block	256 × 16 × 16	128 → 256 → 256
Conv2D	3 × 16 × 16	256 → 3

Table B.6: Discriminator architectures for tinyvideos experiment. Last layers of ω^0 are highlighted in red, and the last layers of the temporal discriminator $\omega^{1,1}$ are highlighted in blue.

Appendix C

Open Source Codes

The prost Framework

`https://github.com/tum-vision/prost`: A C++/CUDA implementation of the first-order primal dual algorithm described in Chapter 2. Was used for all experiments in Chapter 3 – Chapter 6.

Sublabel-Accurate Multilabeling

`https://github.com/tum-vision/sublabel_relax`: Codes to reproduce the results from Chapter 3 and Chapter 4.

FlatGAN

`https://github.com/moellenh/flatgan`: Codes to reproduce the generative modeling results from Chapter 7.

Appendix D

Original Publications

In the following, we include reprints of the accepted versions of the original publications this cumulative thesis is based on. In particular, Chapters 3–7 are minor layout and content adaptations of the publications included in this appendix.

Sublabel-Accurate Relaxation of Nonconvex Energies

Authors	Thomas Möllenhoff[1,♣]	thomas.moellenhoff@tum.de
	Emanuel Laude[1,♣]	emanuel.laude@tum.de
	Michael Moeller[2]	michael.moeller@uni-siegen.de
	Jan Lellmann[3]	jan.lellmann@mic.uni-luebeck.de
	Daniel Cremers[1]	cremers@tum.de

[1] Technische Universität München, Germany
[2] Universität Siegen, Germany
[3] Universität Lübeck, Germany
♣ equal contribution

Publication

T. Möllenhoff, E. Laude, M. Moeller, J. Lellmann, D. Cremers. Sublabel-accurate relaxation of nonconvex energies. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. Received the best paper honorable mention award at CVPR 2016. DOI: http://dx.doi.org/10.1109/CVPR.2016.428. eprint: https://arxiv.org/abs/1512.01383. Copyright © 2016 IEEE. Reprinted with permission from IEEE and the authors.

Contribution	Problem definition	*significantly contributed*
	Literature survey	*significantly contributed*
	Mathematical derivations	*significantly contributed*
	Numerical implementation	*significantly contributed*
	Experimental evaluation	*significantly contributed*
	Preparation of manuscript	*significantly contributed*

Notice

Following the *IEEE Thesis / Dissertation Reuse Permissions*, we include here the *accepted version* of the original publication [Möl+16].

Sublabel–Accurate Relaxation of Nonconvex Energies

Thomas Möllenhoff*
TU München
moellenh@in.tum.de

Emanuel Laude*
TU München
laudee@in.tum.de

Michael Moeller
TU München
moellerm@in.tum.de

Jan Lellmann
University of Lübeck
lellmann@mic.uni-luebeck.de

Daniel Cremers
TU München
cremers@tum.de

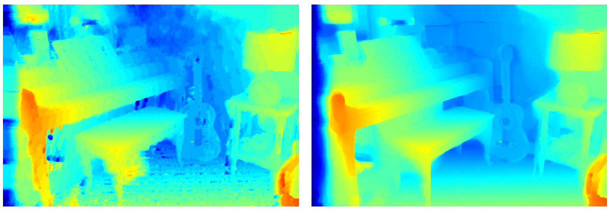
Abstract

We propose a novel spatially continuous framework for convex relaxations based on functional lifting. Our method can be interpreted as a sublabel–accurate solution to multilabel problems. We show that previously proposed functional lifting methods optimize an energy which is linear between two labels and hence require (often infinitely) many labels for a faithful approximation. In contrast, the proposed formulation is based on a piecewise convex approximation and therefore needs far fewer labels – see Fig. 1. In comparison to recent MRF-based approaches, our method is formulated in a spatially continuous setting and shows less grid bias. Moreover, in a local sense, our formulation is the tightest possible convex relaxation. It is easy to implement and allows an efficient primal-dual optimization on GPUs. We show the effectiveness of our approach on several computer vision problems.

1. Introduction

Energy minimization methods have become the central paradigm for solving practical problems in computer vision. The energy functional can often be written as the sum of a data fidelity and a regularization term. One of the most popular regularizers is the total variation (TV) due to its many favorable properties [4]. Hence, an important class of optimization problems is given as

$$\min_{u:\Omega\to\Gamma} \int_{\Omega} \rho(x,u(x))\,\mathrm{d}x + \lambda\,TV(u), \tag{1}$$



Pock *et al.* [17], 48 labels, 1.49 GB, $52s$. Proposed, 8 labels, 0.49 GB, $30s$.

Figure 1. We propose a convex relaxation for the variational model (1), which opposed to existing functional lifting methods [17, 18] allows continuous label spaces *even after* discretization. Our method (here applied to stereo matching) avoids label space discretization artifacts, while saving on memory and runtime.

defined for functions u with finite total variation, arbitrary, possibly nonconvex dataterms $\rho : \Omega \times \Gamma \to \mathbb{R}$, label spaces Γ which are closed intervals in $\mathbb{R}$, $\Omega \subset \mathbb{R}^d$, and $\lambda \in \mathbb{R}^+$. The multilabel interpretation of the dataterm is that $\rho(x,u(x))$ represents the costs of assigning label $u(x)$ to point x. For (weakly) differentiable functions $TV(u)$ equals the integral over the norm of the derivative, and therefore favors a spatially coherent label configuration. The difficultly of minimizing the nonconvex energy (1) has motivated researchers to develop convex reformulations.

Convex representations of (1) and more general related energies have been studied in the context of the calibration method for the Mumford-Shah functional [1]. Based on these works, relaxations for the piecewise constant [15] and piecewise smooth Mumford-Shah functional [16] have been proposed. Inspired by Ishikawa's graph-theoretic globally

[*] Those authors contributed equally.

optimal solution to discrete variants of (1), continuous analogues have been considered by Pock *et al.* in [17, 18]. Continuous relaxations for multilabeling problems with finite label spaces Γ have also been studied in [11].

Interestingly, the discretization of the aforementioned continuous relaxations is very similar to the linear programming relaxations proposed for MAP inference in the Markov Random Field (MRF) community [10, 22, 24, 26]. Both approaches ultimately discretize the range Γ into a finite set of labels. A closer analysis of these relaxations reveals, however, that they are not well-suited to represent the continuous valued range that we face in most computer vision problems such as stereo matching or optical flow. More specifically, the above relaxations are not designed to assign meaningful cost values to non-integral configurations. As a result, a large number of labels is required to achieve a faithful approximation. Solving real-world vision problems therefore entails large optimization problems with high memory and runtime requirement. To address this problem, Zach and Kohli [27], Zach [25] and Fix and Agarwal [7] introduced MRF-based approaches which retain continuous label spaces after discretization. For manifold-valued labels, this issue was addressed by Lellmann *et al.* [12], however with the sole focus on the regularizer.

1.1. Contributions

We propose the first sublabel–accurate convex relaxation of nonconvex problems in a spatially continuous setting. It exhibits several favorable properties:

- In contrast to existing spatially continuous lifting approaches [17, 18], the proposed method provides substantially better solutions with far fewer labels – see Fig. 1. This provides savings in runtime and memory.
- In Sec. 3 we show that the functional lifting methods [17, 18] are a special case of the proposed framework.
- In Sec. 3 we show that, in a local sense, our formulation is the tightest convex relaxation which takes dataterm and regularizer into account separately. It is unknown whether this "local convex envelope" property also holds for the discrete approach [27].
- Our formulation is compact and requires only half the amount of variables for the dataterm than the formulation in [27]. We prove that the sublabel–accurate total variation can be represented in a very simple way, introducing no overhead compared to [17, 18]. In contrast, the regularizer in [27] is much more involved.
- Since our method is derived in a spatially continuous setting, the proposed approach easily allows different gradient discretizations. In contrast to [25, 27] the regularizer is isotropic leading to noticeably less grid bias.

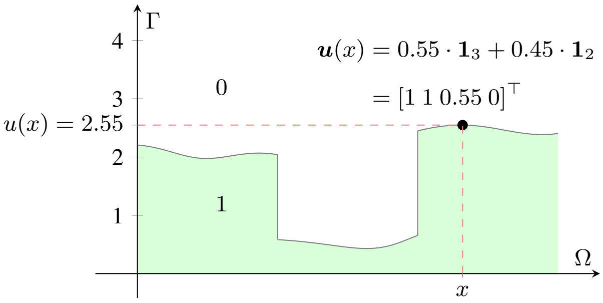

Γ
4
3
2
1
$\boldsymbol{u}(x) = 0.55 \cdot \mathbf{1}_3 + 0.45 \cdot \mathbf{1}_2$
$= [1\ 1\ 0.55\ 0]^\top$
0
1
$u(x) = 2.55$
Ω
x

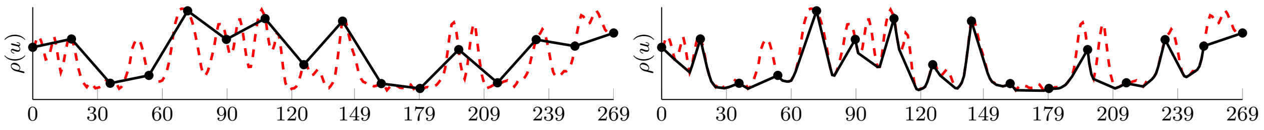
Figure 2. Lifted representation. Instead of optimizing over the function $u : \Omega \to \Gamma$, we optimize over all possible graph functions (here shaded in green) on $\Omega \times \Gamma$. The main idea behind our approach is the finite dimensional representation of the graph at every $x \in \Omega$ by means of $\boldsymbol{u} : \Omega \to \mathbb{R}^k$ (here $k = 4$).

2. Notation and Mathematical Preliminaries

We make heavy use of the convex conjugate, which is given as $f^*(y) = \sup_{x \in \mathbb{R}^n} \langle y, x \rangle - f(x)$ for functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. The biconjugate f^{**} denotes its *convex envelope*, i.e. the largest lower-semicontinuous convex under-approximation of f. For a set C we denote by δ_C the function which maps any element from C to 0 and is ∞ otherwise. For a comprehensive introduction to convex analysis, we refer the reader to [19]. Vector valued functions $\boldsymbol{u} : \Omega \to \mathbb{R}^k$ are written in bold symbols. If it is clear from the context, we will drop the $x \in \Omega$ inside the functions, e.g., we write $\rho(u)$ for $\rho(x, u(x))$, or α for $\alpha(x)$.

3. Functional Lifting

To derive a convex representation of (1), we rely on the framework of functional lifting. The idea is to reformulate the optimization problem in a higher dimensional space. We numerically show in Sec. 5 that considering the convex envelope of the dataterm and regularizer in this higher dimensional space leads to a better approximation of the original nonconvex energy. We start by sampling the range



(a) (b)

Figure 3. We show the nonconvex energy $\rho(u)$ at a fixed point $x \in \Omega$ (red dashed line in both plots) from the stereo matching experiment in Fig. 9 over the full range of 270 disparities. The black dots indicate the positions of the labels and the black curves show the approximations used by the respective methods. Fig. 3a: The baseline lifting method [17] uses a piecewise linear approximation with labels as nodes. Fig. 3b: The proposed method uses an optimal piecewise convex approximation. As we can see, the piecewise convex approximation is closer to the original nonconvex energy and therefore more accurate.

Γ at $L = k + 1$ labels $\gamma_1 < \ldots < \gamma_L \in \Gamma$. This partitions the range into k intervals $\Gamma_i = [\gamma_i, \gamma_{i+1}]$ so that $\Gamma = \Gamma_1 \cup \ldots \cup \Gamma_k$. For any value in the range of $u : \Omega \to \Gamma$ there exist a label index $1 \leq i \leq k$ and $\alpha \in [0, 1]$ such that

$$u(x) = \gamma_i^\alpha := \gamma_i + \alpha(\gamma_{i+1} - \gamma_i). \tag{2}$$

We represent a value in the range Γ by a vector in $\mathbb{R}^k$

$$\boldsymbol{u}(x) = \mathbf{1}_i^\alpha := \alpha \mathbf{1}_i + (1 - \alpha)\mathbf{1}_{i-1}, \tag{3}$$

where $\mathbf{1}_i$ denotes a vector starting with i ones followed by $k - i$ zeros. We call $\boldsymbol{u} : \Omega \to \mathbb{R}^k$ the *lifted* representation of u, representing the graph of u. This notation is depicted in Fig. 2 for $k = 4$. Back-projecting the lifted $\boldsymbol{u}(x)$ to the range of u using the layer cake formula yields a one-to-one correspondence between $u(x) = \gamma_i^\alpha$ and $\boldsymbol{u}(x) = \mathbf{1}_i^\alpha$ via

$$u(x) = \gamma_1 + \sum_{i=1}^{k} \boldsymbol{u}_i(x)(\gamma_{i+1} - \gamma_i). \tag{4}$$

We write problem (1) in terms of such graph functions, a technique that is used in the theory of Cartesian currents [8].

3.1. Convexification of the Dataterm

For now, we consider a fixed $x \in \Omega$. Then the dataterm from (1) is a possibly nonconvex real-valued function (cf. Fig. 3) that we seek to minimize over a compact interval Γ:

$$\min_{u \in \Gamma} \rho(u). \tag{5}$$

Due to the one-to-one correspondence between γ_i^α and $\mathbf{1}_i^\alpha$ it is clear that solving problem (5) is equivalent to finding a minimizer of the lifted energy:

$$\boldsymbol{\rho}(\boldsymbol{u}) = \min_{1 \leq i \leq k} \boldsymbol{\rho}_i(\boldsymbol{u}), \tag{6}$$

$$\boldsymbol{\rho}_i(\boldsymbol{u}) = \begin{cases} \rho(\gamma_i^\alpha), & \text{if } \boldsymbol{u} = \mathbf{1}_i^\alpha,\ \alpha \in [0, 1], \\ \infty, & \text{else.} \end{cases} \tag{7}$$

Note that the constraint in (7) is essentially the nonconvex special ordered set of type 2 (SOS2) constraint [3]. More precisely, we demand that the "derivative" in label direction $(\partial_\gamma \boldsymbol{u})_i := \boldsymbol{u}_{i+1} - \boldsymbol{u}_i$ is zero, except for two neighboring elements, which add up to one. In the following proposition, we derive the tightest convex relaxation of $\boldsymbol{\rho}$.

Proposition 1. *The convex envelope of (6) is given as:*

$$\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \sup_{\boldsymbol{v} \in \mathbb{R}^k} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \max_{1 \leq i \leq k} \boldsymbol{\rho}_i^*(\boldsymbol{v}), \tag{8}$$

where the conjugate of the individual $\boldsymbol{\rho}_i$ is

$$\boldsymbol{\rho}_i^*(\boldsymbol{v}) = c_i(\boldsymbol{v}) + \rho_i^*\left(\frac{\boldsymbol{v}_i}{\gamma_{i+1} - \gamma_i}\right), \tag{9}$$

with $c_i(\boldsymbol{v}) = \langle \mathbf{1}_{i-1}, \boldsymbol{v} \rangle - \frac{\gamma_i}{\gamma_{i+1} - \gamma_i} \boldsymbol{v}_i$ and $\rho_i = \rho + \delta_{\Gamma_i}$.

Proof. See supplementary material. □

The above proposition reveals that the convex relaxation implicitly convexifies the dataterm ρ on each interval Γ_i. The equality $\rho_i^* = \rho_i^{***}$ implies that starting with ρ_i yields exactly the same convex relaxation as starting with ρ_i^{**}.

Corollary 1. *If ρ is linear on each Γ_i, then the convex envelopes of $\boldsymbol{\rho}(\boldsymbol{u})$ and $\boldsymbol{\sigma}(\boldsymbol{u})$ coincide, where the latter is:*

$$\boldsymbol{\sigma}(\boldsymbol{u}) = \begin{cases} \rho(\gamma_i^\alpha), & \textit{if } \exists i :\ \boldsymbol{u} = \mathbf{1}_i^\alpha,\ \alpha \in \{0, 1\}, \\ \infty, & \textit{else.} \end{cases} \tag{10}$$

Proof. Consider an additional constraint $\delta_{\{\gamma_i, \gamma_{i+1}\}}$ for each ρ_i, which corresponds to selecting $\alpha \in \{0, 1\}$ in (7). The fact that our relaxation is independent of whether we choose ρ_i or ρ_i^{**}, along with the fact that the convex hull of two points is a line, yields the assertion. □

For the piecewise linear case, it is possible to find an explicit form of the biconjugate.

Proposition 2. *Let us denote by $\boldsymbol{r} \in \mathbb{R}^k$ the vector with*

$$\boldsymbol{r}_i = \rho(\gamma_{i+1}) - \rho(\gamma_i), \ \ 1 \le i \le k. \tag{11}$$

Under the assumptions of Prop. 1, one obtains:

$$\boldsymbol{\sigma}^{**}(\boldsymbol{u}) = \begin{cases} \rho(\gamma_1) + \langle \boldsymbol{u}, \boldsymbol{r} \rangle, & \text{if } \boldsymbol{u}_i \ge \boldsymbol{u}_{i+1}, \boldsymbol{u}_i \in [0,1], \\ \infty, & \text{else.} \end{cases} \tag{12}$$

Proof. See supplementary material. □

Up to an offset (which is irrelevant for the optimization), one can see that (12) coincides with the dataterm of [15], the discretizations of [17, 18], and – after a change of variable – with [11]. This not only proves that the latter is optimizing a convex envelope, but also shows that our method naturally generalizes the work from piecewise linear to arbitrary piecewise convex energies. Fig. 3a and Fig. 3b illustrate the difference of $\boldsymbol{\sigma}^{**}$ and $\boldsymbol{\rho}^{**}$ on the example of a nonconvex stereo matching cost.

Because our method allows arbitrary convex functions on each Γ_i, we can prove that, for the two label case, our approach optimizes the convex envelope of the dataterm.

Proposition 3. *In the case of binary labeling, i.e., $L = 2$, the convex envelope of (6) reduces to*

$$\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \rho^{**}\left(\gamma_1 + \boldsymbol{u}(\gamma_2 - \gamma_1)\right), \ \textit{with } \boldsymbol{u} \in [0,1]. \tag{13}$$

Proof. See supplementary material. □

3.2. A Lifted Representation of the Total Variation

We now want to find a lifted convex formulation that emulates the total variation regularization in (1). We follow [5] and define an appropriate integrand of the functional

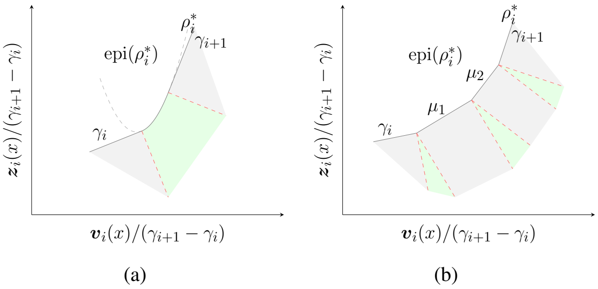
$$TV(\boldsymbol{u}) = \int_\Omega \boldsymbol{\Phi}(x, D\boldsymbol{u}), \tag{14}$$

where the distributional derivative $D\boldsymbol{u}$ is a finite $\mathbb{R}^{k \times d}$-valued Radon measure [2]. We define

$$\boldsymbol{\Phi}(\boldsymbol{g}) = \min_{1 \le i \le j \le k} \boldsymbol{\Phi}_{i,j}(\boldsymbol{g}). \tag{15}$$

The individual $\boldsymbol{\Phi}_{i,j} : \mathbb{R}^{k \times d} \to \mathbb{R} \cup \{\infty\}$ are given by:

$$\boldsymbol{\Phi}_{i,j}(\boldsymbol{g}) = \begin{cases} \left|\gamma_i^\alpha - \gamma_j^\beta\right| \cdot |\nu|_2, & \text{if } \boldsymbol{g} = \left(\mathbf{1}_i^\alpha - \mathbf{1}_j^\beta\right) \nu^\top, \\ \infty, & \text{else,} \end{cases} \tag{16}$$


epi(ρ_i^*)
ρ_i^*
γ_{i+1}
γ_i
$\boldsymbol{z}_i(x)/(\gamma_{i+1} - \gamma_i)$
$\boldsymbol{v}_i(x)/(\gamma_{i+1} - \gamma_i)$
μ_1
μ_2

(a)

(b)

Figure 4. Illustration of the epigraph projection. In the left subfigure the projection onto the epigraph of the conjugate of a convex quadratic ρ_i is shown. In the right subfigure the piecewise linear case is illustrated. In the both cases all points that lie in the gray sets are orthogonally projected onto the respective linear parts whereas the points that lie in the green sets are projected onto the parabolic part (in the quadratic case) respectively the kinks (in the piecewise linear case). In the piecewise linear case the green sets are normal cones. The red dashed lines correspond to the boundary cases. γ_i, γ_{i+1}, μ_1, μ_2 are the slopes of the segments of ρ_i^* respectively the (sub-)label positions of ρ_i.

for some $\alpha, \beta \in [0,1]$ and $\nu \in \mathbb{R}^d$. The intuition is that $\Phi_{i,j}$ penalizes a jump from γ_i^α to γ_j^β in the direction of ν. Since $\boldsymbol{\Phi}$ is nonconvex we compute the convex envelope.

Proposition 4. *The convex envelope of (15) is*

$$\boldsymbol{\Phi}^{**}(\boldsymbol{g}) = \sup_{\boldsymbol{p} \in \mathcal{K}} \langle \boldsymbol{p}, \boldsymbol{g} \rangle, \tag{17}$$

where $\mathcal{K} \subset \mathbb{R}^{k \times d}$ is given as:

$$\mathcal{K} = \Big\{ \boldsymbol{p} \in \mathbb{R}^{k \times d} \ \Big| \ \left|\boldsymbol{p}^\top (\mathbf{1}_i^\alpha - \mathbf{1}_j^\beta)\right|_2 \le \left|\gamma_i^\alpha - \gamma_j^\beta\right|, \\ \forall\, 1 \le i \le j \le k, \ \forall \alpha, \beta \in [0,1] \Big\}. \tag{18}$$

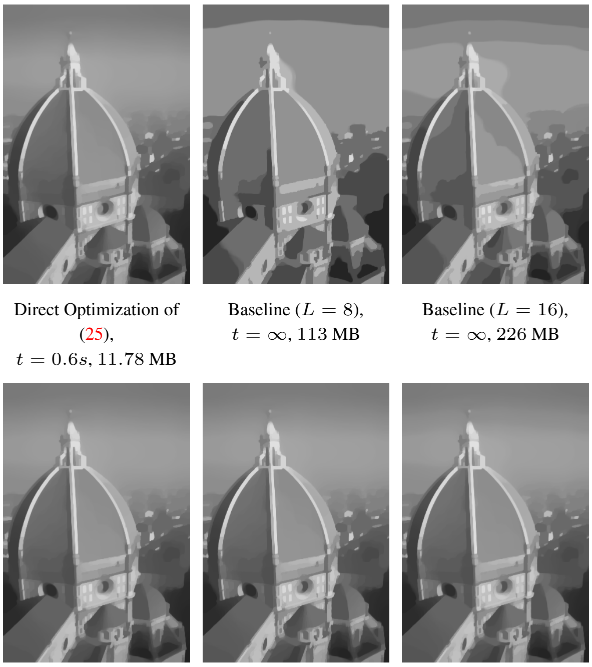
Proof. See supplementary material. □

The set $\mathcal{K}$ from Eq. (18) involves infinitely many constraints which makes numerical optimization difficult. As the following proposition reveals, the infinite number of constraints can be reduced to only linearly many, allowing to enforce the constraint $\boldsymbol{p} \in \mathcal{K}$ exactly.

Proposition 5. *If the labels are ordered ($\gamma_1 < \gamma_2 < \ldots < \gamma_L$) then the constraint set $\mathcal{K}$ from Eq. (18) is equal to*

$$\mathcal{K} = \{ \boldsymbol{p} \in \mathbb{R}^{k \times d} \ | \ |\boldsymbol{p}_i|_2 \le \gamma_{i+1} - \gamma_i, \ \forall i \}. \tag{19}$$

Proof. See supplementary material. □



Direct Optimization of (25), $t = 0.6s$, 11.78 MB

Baseline ($L = 8$), $t = \infty$, 113 MB

Baseline ($L = 16$), $t = \infty$, 226 MB

Baseline ($L = 256$), $t = \infty$, 3619 MB

Proposed ($L = 2$) $t = 1s$, 27 MB

Proposed ($L = 10$) $t = 15s$, 211 MB

Figure 5. Denoising comparison. We compare the proposed method to the baseline method [17] on the convex ROF problem. We show the time in seconds required for each method to produce a solution within a certain energy gap to the optimal solution. As the baseline method optimizes a piecewise linear approximation of the quadratic dataterm, it fails to reach that optimality gap even for $L = 256$ (indicated by $t = \infty$). In contrast, while the proposed lifting method can solve a large class of non-convex problems, it is almost as efficient as direct methods on convex problems.

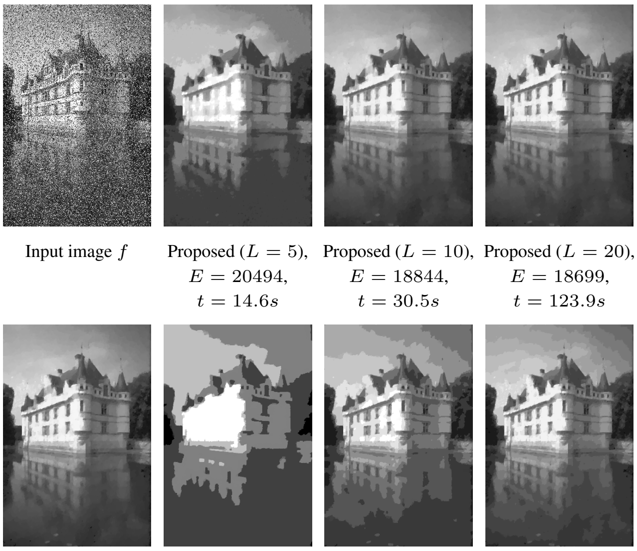
This shows that the proposed regularizer coincides with the total variation from [5], where it has been derived based on (16) for α and β restricted to $\{0, 1\}$. Prop. 5 together with Prop. 3 show that for $k = 1$ our formulation amounts to unlifted TV optimization with a convexified dataterm.

4. Numerical Optimization

Discretizing $\Omega \subset \mathbb{R}^d$ as a d-dimensional Cartesian grid, the relaxed energy minimization problem becomes

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^k} \sum_{x\in\Omega} \boldsymbol{\rho}^{**}(x, \boldsymbol{u}(x)) + \boldsymbol{\Phi}^{**}(x, \nabla\boldsymbol{u}(x)), \tag{20}$$

where ∇ denotes a forward-difference operator with $\nabla\boldsymbol{u} : \Omega \to \mathbb{R}^{k\times d}$. We rewrite the dataterm given in equation (8) by replacing the pointwise maximum over the conjugates $\boldsymbol{\rho}_i^*$ with a maximum over a real number $q \in \mathbb{R}$ and obtain the following saddle point formulation of problem (20):



Input image f

Proposed ($L = 5$), $E = 20494$, $t = 14.6s$

Proposed ($L = 10$), $E = 18844$, $t = 30.5s$

Proposed ($L = 20$), $E = 18699$, $t = 123.9s$

Baseline ($L = 256$), $E = 18660$, $t = 1001s$

Baseline ($L = 5$), $E = 23864$, $t = 4.7s$

Baseline ($L = 10$), $E = 19802$, $t = 6.3s$

Baseline ($L = 20$), $E = 18876$, $t = 12.8s$

Figure 6. Denoising using a robust truncated quadratic dataterm. The top row shows the input image along with the result obtained by our approach for a varying number of labels L. The bottom row illustrates the results obtained by the baseline method [17]. The energy of the final solution as well as the total runtime are given below each image.

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^k} \max_{\substack{(\boldsymbol{v},q)\in\mathcal{C} \\ \boldsymbol{p}:\Omega\to\mathcal{K}}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \sum_{x\in\Omega} q(x) + \langle \boldsymbol{p}, \nabla\boldsymbol{u} \rangle, \tag{21}$$

$$\mathcal{C} = \{(\boldsymbol{v}, q) : \Omega \to \mathbb{R}^k \times \mathbb{R} \mid q(x) \geq \boldsymbol{\rho}_i^*(\boldsymbol{v}(x)),\ \forall x, \forall i\}. \tag{22}$$

We numerically compute a minimizer of problem (21) using a first-order primal-dual method [6, 16] with diagonal preconditioning [14] and adaptive steps [9]. It alternates between a gradient descent step in the primal variable and a gradient ascent step in the dual variable. Subsequently the dual variables are orthogonally projected onto the sets $\mathcal{C}$ respectively $\mathcal{K}$. In the following we give some hints on the implementation of the individual steps. For a detailed discussion we refer to [9]. The projection onto the set $\mathcal{K}$ is a simple ℓ_2-ball projection. To simplify the projection onto $\mathcal{C}$, we transform the k-dimensional epigraph constraints in (22) into 1-dimensional scaled epigraph constraints by introducing an additional variable $z : \Omega \to \mathbb{R}^k$ with:

$$\boldsymbol{z}_i(x) = [q(x) - c_i\,(\boldsymbol{v}(x))]\,(\gamma_{i+1} - \gamma_i)\,. \tag{23}$$

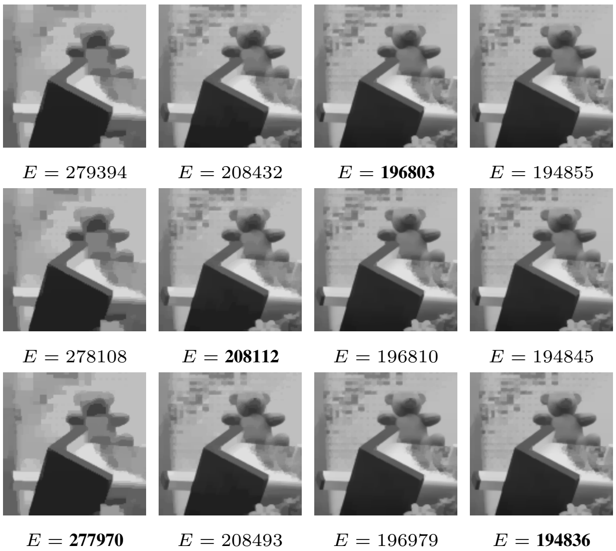


Figure 7. Comparison to the MRF approach presented in [27]. The first row shows DC-Linear, second row DC-MRF and third row our results for 4, 8, 16 and 32 convex pieces on the truncated quadratic energy (26). Below the figures we show the final nonconvex energy. We achieve competitive results while using a more compact representation and generalizing to isotropic regularizers.

Using equation (9) we can write the constraints in (22) as

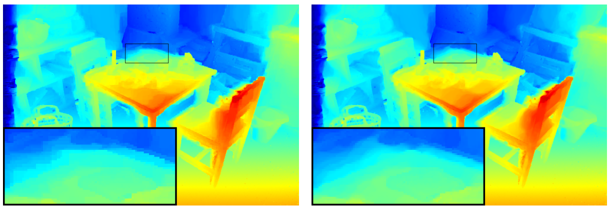
$$\frac{\boldsymbol{z}_i(x)}{\gamma_{i+1} - \gamma_i} \geq \rho_i^* \left(\frac{\boldsymbol{v}_i(x)}{\gamma_{i+1} - \gamma_i} \right). \tag{24}$$

We implement the newly introduced equality constraints (23) introducing a Lagrange multiplier $\boldsymbol{s} : \Omega \to \mathbb{R}^k$. It remains to discuss the orthogonal projections onto the epigraphs of the conjugates ρ_i^*. Currently we support quadratic and piecewise linear convex pieces ρ_i. For the piecewise linear case, the conjugate ρ_i^* is a piecewise linear function with domain $\mathbb{R}$. The slopes correspond to the x-positions of the sublabels and the intercepts correspond to the function values at the sublabel positions.

The conjugates as well as the epigraph projections of both, a quadratic and a piecewise linear piece are depicted in Fig. 4. For the quadratic case, the projection onto the epigraph of a parabola is computed using [23, Appendix B.2].

5. Experiments

We implemented the primal-dual algorithm in CUDA to run on GPUs. [1] For $d = 2$, our implementation of the functional lifting framework [17], which will serve as a baseline method, requires $4N(L-1)$ optimization variables, while the proposed method requires $6N(L-1) + N$ variables, where N is the number of points used to discretize the domain $\Omega \subset \mathbb{R}^d$. As we will show, our method requires much fewer labels to yield comparable results, thus, leading to an improvement in accuracy, memory usage, and speed.



(a) Anisotropic Regularization (b) Isotropic Regularization

Figure 8. We compare the proposed relaxation with anistropic regularizer to isotropic regularization on the stereo matching example. Using an anisotropic formulation as in [27] leads to grid bias.

5.1. Rudin-Osher-Fatemi Model

As a proof of concept, we first evaluate the novel relaxation on the well-known Rudin-Osher-Fatemi (ROF) model [20]. It corresponds to (1) with the following dataterm:

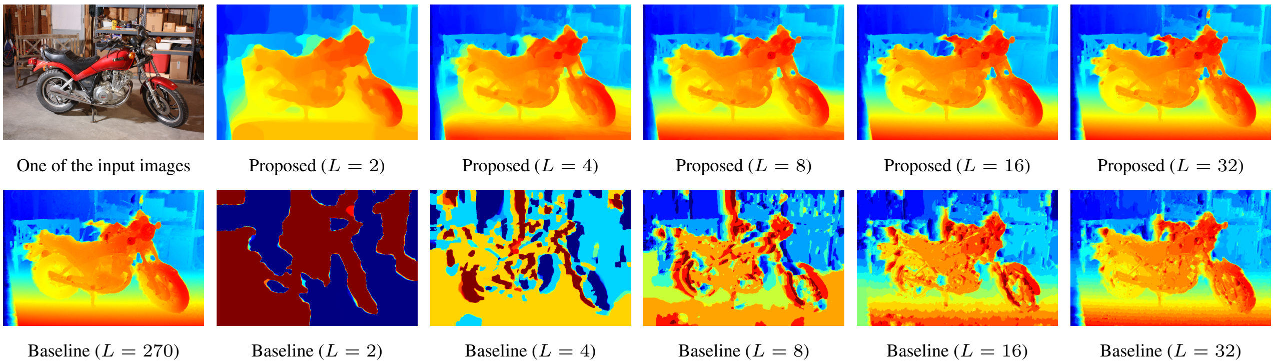
$$\rho(x, u(x)) = (u(x) - f(x))^2, \tag{25}$$

where $f : \Omega \to \mathbb{R}$ denotes the input data. While there is no practical use in applying convex relaxation methods to an already convex problem such as the ROF model, the purpose of this is two-fold. Firstly, it allows us to measure the overhead introduced by our method by comparing it to standard convex optimization methods which do not rely on functional lifting. Secondly, we can experimentally verify that the relaxation is tight for a convex dataterm.

In Fig. 5 we solve (25) directly using the primal-dual algorithm [9], using the baseline functional lifting method [17] and using our proposed algorithm. First, the globally optimal energy was computed using the direct method with a very high number of iterations. Then we measure how long each method took to reach this global optimum to a fixed tolerance.

The baseline method fails to reach the global optimum even for 256 labels. While the lifting framework introduces a certain overhead, the proposed method finds the same globally optimal energy as the direct unlifted optimization approach and generalizes to nonconvex energies.

[1] https://github.com/tum-vision/sublabel_relax



One of the input images

Proposed ($L = 2$)

Proposed ($L = 4$)

Proposed ($L = 8$)

Proposed ($L = 16$)

Proposed ($L = 32$)

Baseline ($L = 270$)

Baseline ($L = 2$)

Baseline ($L = 4$)

Baseline ($L = 8$)

Baseline ($L = 16$)

Baseline ($L = 32$)

Figure 9. Stereo comparison. We compare the proposed method to the baseline method on the example of stereo matching. The first column shows one of the two input images and below the baseline method with the full number of labels. The proposed relaxation requires much fewer labels to reach a smooth depth map. Even for $L = 32$, the label space discretization of the baseline method is strongly visible, while the proposed method yields a smooth result already for $L = 8$.

5.2. Robust Truncated Quadratic Dataterm

The quadratic dataterm in (25) is often not well suited for real-world data as it comes from a pure Gaussian noise assumption and does not model outliers. We now consider a robust truncated quadratic dataterm:

$$\rho(x, u(x)) = \frac{\alpha}{2} \min \left\{ (u(x) - f(x))^2, \nu \right\}. \tag{26}$$

To implement (26), we use a piecewise polynomial approximation of the dataterm. In Fig. 6 we degraded the input image with additive Gaussian and salt and pepper noise. The parameters in (26) were chosen as $\alpha = 25$, $\nu = 0.025$ and $\lambda = 1$. It can be seen that the proposed method requires fewer labels to find lower energies than the baseline.

5.3. Comparison to the Method of Zach and Kohli

We remark that Prop. 4 and Prop. 5 hold for arbitrary convex one-homogeneous functionals $\phi(\nu)$ instead of $|\nu|_2$ in equation (16). In particular, they hold for the anisotropic total variation $\phi(\nu) = |\nu|_1$. This generalization allows us to directly compare our convex relaxation to the MRF approach of Zach and Kohli [27].

In Fig. 7 we show the results of optimizing the two models entitled “DC-Linear” and “DC-MRF” proposed in [27], and of our proposed method with anisotropic regularization on the robust truncated denoising energy (26). We picked the parameters as $\alpha = 0.2$, $\nu = 500$, and $\lambda = 1$. The label space is also chosen as $\Gamma = [0, 256]$ as described in [27]. Note that overall, all the energies are better than the ones reported in [27]. It can be seen from Fig. 7 that the proposed relaxation is competitive to the one proposed by Zach and Kohli. In addition, the proposed relaxation uses a more compact representation and extends to isotropic and convex one-homogeneous regularizers. To illustrate the advantages of isotropic regularizations, Fig. 8a and Fig. 8b show a comparison of our proposed method for isotropic and anisotropic regularization for the example of stereo matching discussed in the next section.

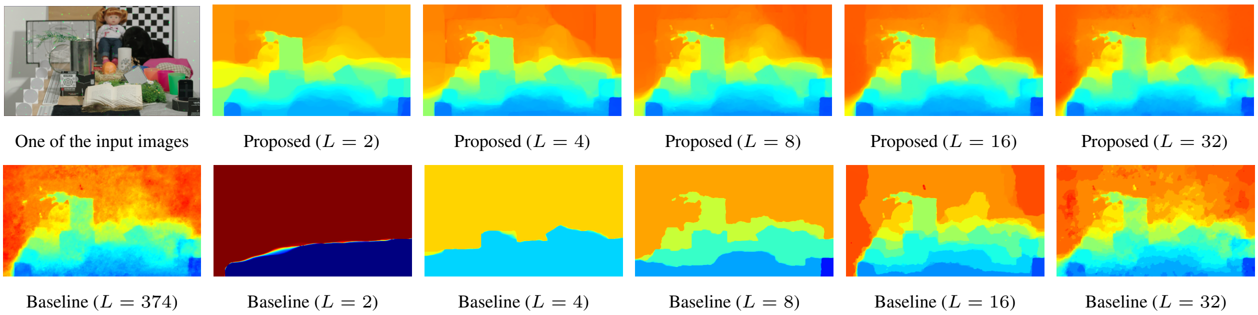
5.4. Stereo Matching

Given a pair of rectified images, the task of finding a correspondence between the two images can be formulated as an optimization problem over a scalar field $u : \Omega \to \Gamma$ where each point $u(x) \in \Gamma$ denotes the displacement along the epipolar line associated with each $x \in \Omega$. The overall cost functional fits Eq. (1). In our experiments, we computed $\rho(x, u(x))$ for 270 disparities on the Middlebury stereo benchmark [21] in a 4×4 patch using a truncated sum of absolute gradient differences. We convexify the matching cost ρ in each range Γ_i by numerically computing the convex envelope using the gift wrapping algorithm.

The first row in Fig. 9 shows the result of the proposed relaxation using the convexified energy between two labels. The second row shows the baseline approach using the same amount of labels. Even for $L = 2$, the proposed method produces a reasonable depth map while the baseline approach basically corresponds to a two region segmentation.

5.5. Phase Unwrapping

Many sensors such as time-of-flight cameras or interferometric synthetic aperture radar (SAR) yield cyclic data ly-



One of the input images | Proposed ($L = 2$) | Proposed ($L = 4$) | Proposed ($L = 8$) | Proposed ($L = 16$) | Proposed ($L = 32$)

Baseline ($L = 374$) | Baseline ($L = 2$) | Baseline ($L = 4$) | Baseline ($L = 8$) | Baseline ($L = 16$) | Baseline ($L = 32$)

Figure 10. Depth from focus comparison. We compare our method to the baseline approach on the problem of depth from focus. First column: one of the 374 differently focused input images and the baseline method for full number of labels. Following columns: proposed relaxation (top row) vs. baseline (bottom row) for 2, 4, 8, 16 and 32 labels each.

ing on the circle $\mathcal{S}^1$. Here we consider the task of total variation regularized unwrapping. As is shown on the left in Fig. 11, the dataterm is a nonconvex function where each minimum corresponds to a phase shift by 2π:

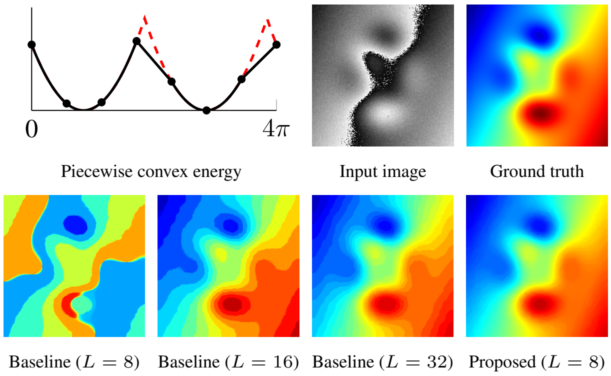
$$\rho\left(x, u(x)\right) = d_{\mathcal{S}^1}\left(u(x), f(x)\right)^2. \tag{27}$$

For the experiments, we approximated the nonconvex energy by quadratic pieces as depicted in Fig. 11. The label space is chosen as $\Gamma = [0, 4\pi]$ and the regularization parameter was set to $\lambda = 0.005$. Again, it is visible in Fig. 11 that the baseline method shows label space discretization and fails to unwrap the depth map correctly if the number of labels is chosen too low. The proposed method yields a smooth unwrapped result using only 8 labels.

5.6. Depth From Focus

In depth from focus the task is to recover the depth of a scene, given a stack of images each taken from a constant position but in a different focal setting, so that in each image only the objects of a certain depth are sharp. images. We compute the dataterm cost ρ by using the modified Laplacian function [13] as a contrast measure.

Similar to the stereo experiments, we convexify the cost on each label range by computing the convex hull. The results are shown in Fig. 10. While the baseline method clearly shows the label space discretization, the proposed approach yields a smooth depth map. Since the proposed method uses a convex lower bound of the lifted energy, the regularizer has slightly more influence on the final result. This explains why the resulting depth maps in Fig. 10 and Fig. 9 look overall less noisy.



Piecewise convex energy | Input image | Ground truth

Baseline ($L = 8$) | Baseline ($L = 16$) | Baseline ($L = 32$) | Proposed ($L = 8$)

Figure 11. We show the piecewise convex approximation of the phase unwrapping energy, followed by the cyclic input image and the unwrapped ground truth. With only 8 labels, the proposed method already yields a smooth reconstruction. The baseline method fails to unwrap the heightmap correctly using 8 labels, and for 16 and 32 labels, the discretization is still noticable.

6. Conclusion

In this work we proposed a tight convex relaxation that can be interpreted as a sublabel–accurate formulation of classical multilabel problems. The final formulation is a simple saddle-point problem that admits fast primal-dual optimization. Our method maintains sublabel accuracy even after discretization and for that reason outperforms existing spatially continuous methods. Interesting directions for future work include higher dimensional label spaces, manifold valued data and more general regularizers.

Acknowledgement This work was supported by the ERC Starting Grant “ConvexVision”. We greatly acknowledge NVIDIA for providing us a Titan X GPU.

References

[1] G. Alberti, G. Bouchitté, and G. D. Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Calc. Var. Partial Dif.*, 3(16):299–333, 2003. 1

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford University Press, 2000. 4

[3] E. Beale and J. Tomlin. Special facilities in a general mathematical programming system for nonconvex problems using ordered sets of variables. *Proceedings of the fifth international conference on operational research*, pages 447–454, 1970. 3

[4] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9:263–340, 2010. 1

[5] A. Chambolle, D. Cremers, and T. Pock. A convex approach to minimal partitions. *SIAM Journal on Imaging Sciences*, 5(4):1113–1158, 2012. 4, 5

[6] E. Esser, X. Zhang, and T. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010. 5

[7] A. Fix and S. Agarwal. Duality and the continuous graphical model. In *Computer Vision ECCV 2014*, volume 8691 of *Lecture Notes in Computer Science*, pages 266–281. Springer International Publishing, 2014. 2

[8] M. Giaquinta, G. Modica, and J. Souček. *Cartesian currents in the calculus of variations I, II.*, volume 37-38 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3.* Springer-Verlag, Berlin, 1998. 3

[9] T. Goldstein, E. Esser, and R. Baraniuk. Adaptive Primal-Dual Hybrid Gradient Methods for Saddle-Point Problems. *arXiv Preprint*, 2013. 5, 6

[10] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(10):1333–1336, 2003. 2

[11] J. Lellmann and C. Schnörr. Continuous multiclass labeling approaches and algorithms. *SIAM J. Imaging Sciences*, 4(4):1049–1096, 2011. 2, 4

[12] J. Lellmann, E. Strekalovskiy, S. Koetter, and D. Cremers. Total variation regularization for functions with values in a manifold. In *ICCV*, December 2013. 2

[13] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(8):824–831, Aug. 1994. 8

[14] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *ICCV*, pages 1762–1769, 2011. 5

[15] T. Pock, A. Chambolle, H. Bischof, and D. Cremers. A convex relaxation approach for computing minimal partitions. In *CVPR*, pages 810–817, 2009. 1, 4

[16] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the piecewise smooth Mumford-Shah functional. In *ICCV*, 2009. 1, 5

[17] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM J. Imaging Sci.*, 3(4):1122–1145, 2010. 1, 2, 3, 4, 5, 6

[18] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. A convex formulation of continuous multi-label problems. In *European Conference on Computer Vision (ECCV)*, Marseille, France, October 2008. 1, 2, 4

[19] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996. 2

[20] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 6

[21] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nei, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, volume 8753, pages 31–42. Springer, 2014. 7

[22] M. Schlesinger. Sintaksicheskiy analiz dvumernykh zritelnikh signalov v usloviyakh pomekh (Syntactic analysis of two-dimensional visual signals in noisy conditions). *Kibernetika*, 4:113–130, 1976. 2

[23] E. Strekalovskiy, A. Chambolle, and D. Cremers. Convex relaxation of vectorial problems with coupled regularization. *SIAM Journal on Imaging Sciences*, 7(1):294–336, 2014. 6

[24] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(7):1165–1179, July 2007. 2

[25] C. Zach. Dual decomposition for joint discrete-continuous optimization. In *AISTATS*, pages 632–640, 2013. 2

[26] C. Zach, C. Hane, and M. Pollefeys. What is optimized in tight convex relaxations for multi-label problems? In *CVPR*, pages 1664–1671, 2012. 2

[27] C. Zach and P. Kohli. A convex discrete-continuous approach for markov random fields. In *ECCV*, volume 7577, pages 386–399. Springer Berlin Heidelberg, 2012. 2, 6, 7

Sublabel-Accurate Convex Relaxation of Vectorial Multilabel Energies

Authors	Emanuel Laude[1, ♣]	emanuel.laude@tum.de
	Thomas Möllenhoff[1, ♣]	thomas.moellenhoff@tum.de
	Michael Moeller[2]	michael.moeller@uni-siegen.de
	Jan Lellmann[3]	jan.lellmann@mic.uni-luebeck.de
	Daniel Cremers[1]	cremers@tum.de
	[1] Technische Universität München, Germany [2] Universität Siegen, Germany [3] Universität Lübeck, Germany [♣] equal contribution	
Publication	E. Laude, T. Möllenhoff, M. Moeller, J. Lellmann, D. Cremers. Sublabel-Accurate Convex Relaxation of Vectorial Multilabel Energies. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. DOI: http://dx.doi.org/10.1007/978-3-319-46448-0_37. eprint: https://arxiv.org/abs/1604.01980. Copyright © 2016 Springer. Reprinted with permission from the Springer Nature Customer Service Centre GmbH.	
Contribution	Problem definition	*significantly contributed*
	Literature survey	*significantly contributed*
	Mathematical derivations	*significantly contributed*
	Numerical implementation	*significantly contributed*
	Experimental evaluation	*significantly contributed*
	Preparation of manuscript	*significantly contributed*
Notice	Following the Springer Nature license for *reuse in a dissertation/thesis*, we include here the *accepted version* of the original publication [Lau+16].	

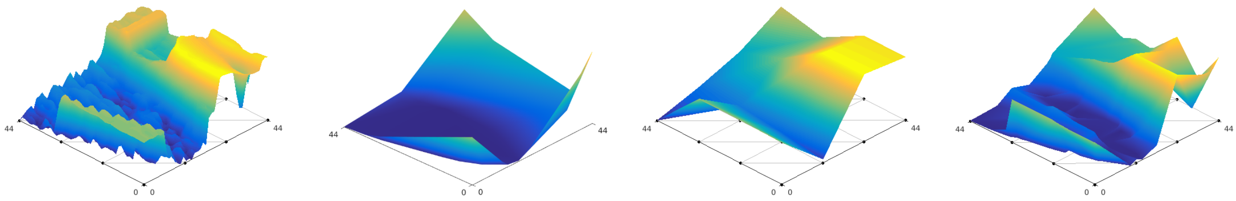
Sublabel-Accurate Convex Relaxation of Vectorial Multilabel Energies

Emanuel Laude[⋆1], Thomas Möllenhoff[⋆1], Michael Moeller[1],
Jan Lellmann[2], and Daniel Cremers[1]

[1]Technical University of Munich[⋆⋆] [2]University of Lübeck

Abstract. Convex relaxations of multilabel problems have been demonstrated to produce provably optimal or near-optimal solutions to a variety of computer vision problems. Yet, they are of limited practical use as they require a fine discretization of the label space, entailing a huge demand in memory and runtime. In this work, we propose the first sublabel accurate convex relaxation for vectorial multilabel problems. Our key idea is to approximate the dataterm in a piecewise convex (rather than piecewise linear) manner. As a result we have a more faithful approximation of the original cost function that provides a meaningful interpretation for fractional solutions of the relaxed convex problem.

Keywords: Convex Relaxation, Optimization, Variational Methods



(a) Original dataterm (b) Without lifting (c) Classical lifting (d) Proposed lifting

Fig. 1: In (a) we show a nonconvex dataterm. Convexification without lifting would result in the energy (b). Classical lifting methods [11] (c), approximate the energy piecewise linearly between the labels, whereas the proposed method results in an approximation that is convex on each triangle (d). Therefore, we are able to capture the structure of the nonconvex energy much more accurately.

[⋆] These authors contributed equally.

[⋆⋆] This work was supported by the ERC Starting Grant “Convex Vision”.

1 Introduction

1.1 Nonconvex Vectorial Problems

In this paper, we derive a sublabel-accurate convex relaxation for vectorial optimization problems of the form

$$\min_{u:\Omega\to\Gamma} \int_\Omega \rho\big(x, u(x)\big)\, \mathrm{d}x \, + \, \lambda TV(u), \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $\Gamma \subset \mathbb{R}^n$ and $\rho : \Omega \times \Gamma \to \mathbb{R}$ denotes a generally nonconvex pointwise dataterm. As regularization we focus on the *total variation* defined as:

$$TV(u) = \sup_{q \in C_c^\infty(\Omega, \mathbb{R}^{n\times d}), \|q(x)\|_{S^\infty} \leq 1} \int_\Omega \langle u, \operatorname{Div} q\rangle \, \mathrm{d}x, \tag{2}$$

where $\|\cdot\|_{S^\infty}$ is the Schatten-∞ norm on $\mathbb{R}^{n\times d}$, i.e., the largest singular value. For differentiable functions u we can integrate (2) by parts to find

$$TV(u) = \int_\Omega \|\nabla u(x)\|_{S^1} \, \mathrm{d}x, \tag{3}$$

where the dual norm $\|\cdot\|_{S^1}$ penalizes the sum of the singular values of the Jacobian, which encourages the individual components of u to jump in the same direction. This type of regularization is part of the framework of Sapiro and Ringach [19].

1.2 Related Work

Due to its nonconvexity the optimization of (1) is challenging. For the scalar case ($n = 1$), Ishikawa [9] proposed a pioneering technique to obtain globally optimal solutions in a spatially discrete setting, given by the minimum s-t-cut of a graph representing the space $\Omega \times \Gamma$. A continuous formulation was introduced by Pock et al. [15] exhibiting several advantages such as less grid bias and parallelizability.

In a series of papers [16,14], connections of the above approaches were made to the mathematical theory of *cartesian currents* [6] and the calibration method for the Mumford-Shah functional [1], leading to a generalization of the convex relaxation framework [15] to more general (in particular nonconvex) regularizers.

In the following, researchers have strived to generalize the concept of functional lifting and convex relaxation to the vectorial setting ($n > 1$). If the dataterm and the regularizer are both separable in the label dimension, one can simply apply the above convex relaxation approach in a channel-wise manner to each component separately. But when either the dataterm or the regularizer couple the label components, the situation becomes more complex [8,20].

The approach which is most closely related to our work, and which we consider as a baseline method, is the one by Lellmann et al. [11]. They consider coupled dataterms with coupled total variation regularization of the form (2).

A drawback shared by all mentioned papers is that ultimately one has to discretize the label space. While Lellmann et al. [11] propose a sublabel-accurate regularizer, we show that their dataterm leads to solutions which still have a strong bias towards the label grid. For the scalar-valued setting, continuous label spaces have been considered in the MRF community by Zach et al. [22] and Fix et al. [5]. The paper [21] proposes a method for mixed continuous and discrete vectorial label spaces, where everything is derived in the spatially discrete MRF setting. Möllenhoff et al. [12] recently proposed a novel formulation of the scalar-valued case which retains fully continuous label spaces even after discretization. The contribution of this work is to extend [12] to vectorial label spaces, thereby complementing [11] with a sublabel-accurate dataterm.

1.3 Contribution

In this work we propose the first sublabel-accurate convex formulation of vectorial labeling problems. It generalizes the formulation for scalar-valued labeling problems [12] and thus includes important applications such as optical flow estimation or color image denoising. We show that our method, derived in a spatially continuous setting, has a variety of interesting theoretical properties as well as practical advantages over the existing labeling approaches:

- We generalize existing functional lifting approaches (see Sec. 2.2).
- We show that our method is the best convex under-approximation (in a local sense), see Prop. 1 and Prop. 2.
- Due to its sublabel-accuracy our method requires only a small amount of labels to produce good results which leads to a drastic reduction in memory. We believe that this is a vital step towards the real-time capability of lifting and convex relaxation methods. Moreover, our method eliminates the label bias, that previous lifting methods suffer from, even for many labels.
- In Sec. 2.3 we propose a regularizer that couples the different label components by enforcing a joint jump normal. This is in contrast to [8], where the components are regularized separately.
- For convex dataterms, our method is equivalent to the unlifted problem – see Prop. 4. Therefore, it allows a seamless transition between direct optimization and convex relaxation approaches.

1.4 Notation

We write $\langle x, y\rangle = \sum_i x_i y_i$ for the standard inner product on $\mathbb{R}^n$ or the Frobenius product if x, y are matrices. Similarly $\|\cdot\|$ without any subscript denotes the usual Euclidean norm, respectively the Frobenius norm for matrices.

We denote the convex conjugate of a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ by $f^*(y) = \sup_{x\in\mathbb{R}^n} \langle y, x\rangle - f(x)$. It is an important tool for devising convex relaxations, as the biconjugate f^{**} is the largest lower-semicontinuous (lsc.) convex function below f. For the indicator function of a set C we write δ_C, i.e., $\delta_C(x) = 0$ if $x \in C$ and ∞ otherwise. $\Delta_n^U \subset \mathbb{R}^n$ stands for the unit n-simplex.

2 Convex Formulation

2.1 Lifted Representation

Motivated by Fig. 1, we construct an equivalent representation of (1) in a higher dimensional space, before taking the convex envelope.

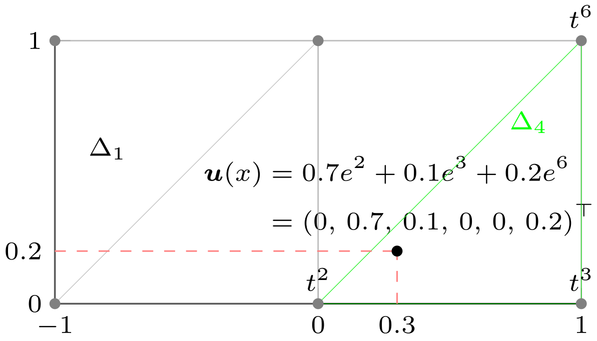
Let $\Gamma \subset \mathbb{R}^n$ be a compact and convex set. We partition Γ into a set $\mathcal{T}$ of n-simplices Δ_i so that Γ is a disjoint union of Δ_i up to a set of measure zero. Let t^{i_j} be the j-th vertex of Δ_i and denote by $\mathcal{V} = \{t^1, \dots, t^{|\mathcal{V}|}\}$ the union of all vertices, referred to as labels, with $1 \le i \le |\mathcal{T}|$, $1 \le j \le n+1$ and $1 \le i_j \le |\mathcal{V}|$. For $u : \Omega \to \Gamma$, we refer to $u(x)$ as a *sublabel*. Any sublabel can be written as a convex combination of the vertices of a simplex Δ_i with $1 \le i \le |\mathcal{T}|$ for appropriate barycentric coordinates $\alpha \in \Delta_n^U$:

$$u(x) = T_i\alpha := \sum_{j=1}^{n+1} \alpha_j t^{i_j},\ T_i := (t^{i_1},\ t^{i_2},\ \dots,\ t^{i_{n+1}}) \in \mathbb{R}^{n\times n+1}. \tag{4}$$

By encoding the vertices $t^k \in \mathcal{V}$ using a one-of-$|\mathcal{V}|$ representation e^k we can identify any $u(x) \in \Gamma$ with a sparse vector $\boldsymbol{u}(x)$ containing at least $|\mathcal{V}| - n$ many zeros and vice versa:

$$\begin{aligned} \boldsymbol{u}(x) &= E_i\alpha := \sum_{j=1}^{n+1} \alpha_j e^{i_j},\ E_i := (e^{i_1},\ e^{i_2}, \dots,\ e^{i_{n+1}}) \in \mathbb{R}^{|\mathcal{V}|\times n+1}, \\ u(x) &= \sum_{k=1}^{|\mathcal{V}|} t^k \boldsymbol{u}_k(x),\ \alpha \in \Delta_n^U,\ 1 \le i \le |\mathcal{T}|\,. \end{aligned} \tag{5}$$

The entries of the vector e^{i_j} are zero except for the (i_j)-th entry, which is equal to one. We refer to $\boldsymbol{u} : \Omega \to \mathbb{R}^{|\mathcal{V}|}$ as the *lifted* representation of u. This one-to-one-correspondence between $u(x) = T_i\alpha$ and $\boldsymbol{u}(x) = E_i\alpha$ is shown in Fig. 2. Note that both, α and i depend on x. However, for notational convenience we drop the dependence on x whenever we consider a fixed point $x \in \Omega$.



t^6
1
Δ_1
Δ_4
$\boldsymbol{u}(x) = 0.7e^2 + 0.1e^3 + 0.2e^6$
$= (0,\ 0.7,\ 0.1,\ 0,\ 0,\ 0.2)^\top$
0.2
t^2
t^3
0
-1
0
0.3
1

Fig. 2: This figure illustrates our notation and the one-to-one correspondence between $u(x) = (0.3, 0.2)^\top$ and the lifted $\boldsymbol{u}(x)$ containing the barycentric coordinates $\alpha = (0.7, 0.1, 0.2)^\top$ of the sublabel $u(x) \in \Delta_4 = \operatorname{conv}\{t^2, t^3, t^6\}$. The triangulation $(\mathcal{V}, \mathcal{T})$ of $\Gamma = [-1; 1] \times [0; 1]$ is visualized via the gray lines, corresponding to the triangles and the gray dots, corresponding to the vertices $\mathcal{V} = \{(-1, 0)^\top, (0, 0)^\top, \ldots, (1, 1)^\top\}$, that we refer to as the labels.

2.2 Convexifying the Dataterm

Let for now the weight of the regularizer in (1) be zero. Then, at each point $x \in \Omega$ we minimize a generally nonconvex energy over a compact set $\Gamma \subset \mathbb{R}^n$:

$$\min_{u \in \Gamma} \rho(u). \tag{6}$$

We set up the lifted energy so that it attains finite values if and only if the argument $\boldsymbol{u}$ is a sparse representation $\boldsymbol{u} = E_i \alpha$ of a sublabel $u \in \Gamma$:

$$\boldsymbol{\rho}(\boldsymbol{u}) = \min_{1 \leq i \leq |\mathcal{T}|} \boldsymbol{\rho}_i(\boldsymbol{u}), \qquad \boldsymbol{\rho}_i(\boldsymbol{u}) = \begin{cases} \rho(T_i \alpha), & \text{if } \ \boldsymbol{u} = E_i \alpha, \ \alpha \in \Delta_n^U, \\ \infty, & \text{otherwise.} \end{cases} \tag{7}$$

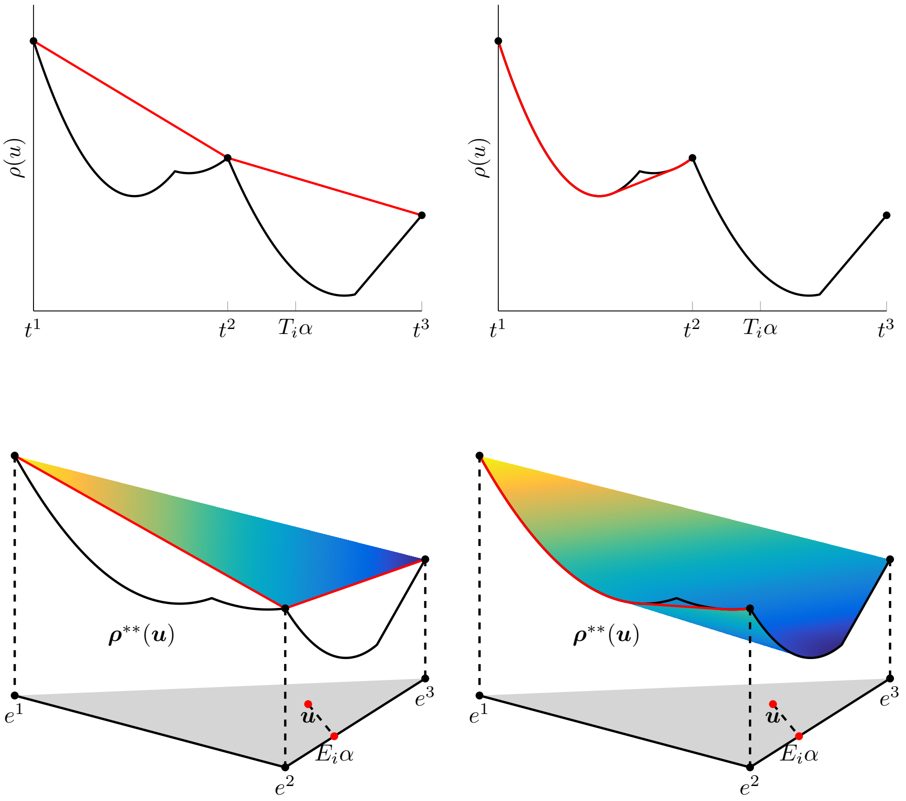
Problems (6) and (7) are equivalent due to the one-to-one correspondence of $u = T_i \alpha$ and $\boldsymbol{u} = E_i \alpha$. However, energy (7) is finite on a nonconvex set only. In order to make optimization tractable, we minimize its convex envelope.

Proposition 1 *The convex envelope of* (7) *is given as:*

$$\begin{aligned} \boldsymbol{\rho}^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{v} \in \mathbb{R}^{|\mathcal{V}|}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \max_{1 \leq i \leq |\mathcal{T}|} \boldsymbol{\rho}_i^*(\boldsymbol{v}), \\ \boldsymbol{\rho}_i^*(\boldsymbol{v}) &= \langle E_i b_i, \boldsymbol{v} \rangle + \rho_i^*(A_i^\top E_i^\top \boldsymbol{v}), \quad \rho_i := \rho + \delta_{\Delta_i}. \end{aligned} \tag{8}$$

b_i and A_i are given as $b_i := M_i^{n+1}$, $A_i := \left(M_i^1,\ M_i^2,\ \ldots,\ M_i^n\right)$, where M_i^j are the columns of the matrix $M_i := (T_i^\top, \mathbf{1})^{-\top} \in \mathbb{R}^{n+1 \times n+1}$.

Proof. Follows from a calculation starting at the definition of $\boldsymbol{\rho}^{**}$. See supplementary material for a detailed derivation.



Standard lifting [11]

Proposed lifting

Fig. 3: Geometrical intuition for the proposed lifting and standard lifting [11] for the special case of 1-dimensional range $\mathit{\Gamma} = [a, b]$ and 3 labels $\{t^1, t^2, t^3\}$. The standard lifting correponds to a linear interpolation of the original cost in between the locations t^1, t^2, t^3, which are associated with the vertices e^1, e^2, e^3 in the lifted energy (lower left). The proposed method extends the cost to the relaxed set in a more precise way: The original cost is preserved on the connecting lines between adjacent e^i (black lines on the bottom right) up to concave parts (red graphs and lower surface on the right). This information, which may influence the exact location of the minimizer, is lost in the standard formulation. If the solution of the lifted formulation $\boldsymbol{u}$ is in the interior (gray area) an approximate solution to the original problem can still be obtained via Eq. (5).

The geometric intuition of this construction is depicted in Fig. 3. Note that if one prescribes the value of $\boldsymbol{\rho}_i$ in (7) only on the *vertices* of the unit simplices Δ_n^U, i.e., $\boldsymbol{\rho}(\boldsymbol{u}) = \rho(t^k)$ if $\boldsymbol{u} = e^k$ and $+\infty$ otherwise, one obtains the linear biconjugate $\boldsymbol{\rho}^{**}(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{s} \rangle$, $\boldsymbol{s} = (\rho(t^i), \ldots, \rho(t^L))$ on the feasible set. This coincides with the standard relaxation of the dataterm used in [16,10,4,11]. In that sense, our approach can be seen as a relaxing the dataterm in a more precise way, by incorporating the true value of ρ not only on the finite set of labels $\mathcal{V}$, but also everywhere in between, i.e., on every *sublabel*.

2.3 Lifting the Vectorial Total Variation

We define the lifted vectorial total variation as

$$\boldsymbol{TV}(\boldsymbol{u}) = \int_{\Omega} \boldsymbol{\Psi}(D\boldsymbol{u}), \tag{9}$$

where $D\boldsymbol{u}$ denotes the distributional derivative of $\boldsymbol{u}$ and $\boldsymbol{\Psi}$ is positively one-homogeneous, i.e., $\boldsymbol{\Psi}(c\boldsymbol{u}) = c\boldsymbol{\Psi}(\boldsymbol{u}), c \geqslant 0$. For such functions, the meaning of (9) can be made fully precise using the polar decomposition of the Radon measure $D\boldsymbol{u}$ [2, Cor. 1.29, Thm. 2.38]. However, in the following we restrict ourselves to an intuitive motivation for the derivation of $\boldsymbol{\Psi}$ for smooth functions.

Our goal is to find $\boldsymbol{\Psi}$ so that $\boldsymbol{TV}(\boldsymbol{u}) = TV(u)$ whenever $\boldsymbol{u} : \Omega \to \mathbb{R}^{|\mathcal{V}|}$ corresponds to some $u : \Omega \to \Gamma$, in the sense that $\boldsymbol{u}(x) = E_i\alpha$ whenever $u(x) = T_i\alpha$. In order for the equality to hold, it must in particular hold for all u that are classically differentiable, i.e., $Du = \nabla u$, and whose Jacobian $\nabla u(x)$ is of rank 1, i.e., $\nabla u(x) = (T_i\alpha - T_j\beta) \otimes \nu(x)$ for some $\nu(x) \in \mathbb{R}^d$. This rank 1 constraint enforces the different components of u to have the same jump normal, which is desirable in many applications. In that case, we observe

$$TV(u) = \int_{\Omega} \|T_i\alpha - T_j\beta\| \cdot \|\nu(x)\| \, \mathrm{d}x. \tag{10}$$

For the corresponding lifted representation $\boldsymbol{u}$, we have $\nabla\boldsymbol{u}(x) = (E_i\alpha - E_j\beta) \otimes \nu(x)$. Therefore it is natural to require $\boldsymbol{\Psi}(\nabla\boldsymbol{u}(x)) = \boldsymbol{\Psi}\left((E_i\alpha - E_j\beta) \otimes \nu(x)\right) := \|T_i\alpha - T_j\beta\| \cdot \|\nu(x)\|$ in order to achieve the goal $\boldsymbol{TV}(\boldsymbol{u}) = TV(u)$. Motivated by these observations, we define

$$\boldsymbol{\Psi}(\boldsymbol{p}) := \begin{cases} \|T_i\alpha - T_j\beta\| \cdot \|\nu\| & \text{if } \boldsymbol{p} = (E_i\alpha - E_j\beta) \otimes \nu, \\ \infty & \text{otherwise,} \end{cases} \tag{11}$$

where $\alpha, \beta \in \Delta_{n+1}^U$, $\nu \in \mathbb{R}^d$ and $1 \leq i,j \leq |\mathcal{T}|$. Since the convex envelope of (9) is intractable, we derive a "locally" tight convex underapproximation:

$$\boldsymbol{R}(\boldsymbol{u}) = \sup_{\boldsymbol{q}:\Omega \to \mathbb{R}^{d\times|\mathcal{V}|}} \int_{\Omega} \langle \boldsymbol{u}, \operatorname{Div} \boldsymbol{q} \rangle - \boldsymbol{\Psi}^*(\boldsymbol{q}) \, \mathrm{d}x. \tag{12}$$

Proposition 2 *The convex conjugate of $\boldsymbol{\Psi}$ is*

$$\boldsymbol{\Psi}^*(\boldsymbol{q}) = \delta_{\mathcal{K}}(\boldsymbol{q}) \tag{13}$$

with convex set

$$\mathcal{K} = \bigcap_{1 \leq i,j \leq |\mathcal{T}|} \left\{ \boldsymbol{q} \in \mathbb{R}^{d\times|\mathcal{V}|} \mid \|Q_i\alpha - Q_j\beta\| \leq \|T_i\alpha - T_j\beta\|, \ \alpha, \beta \in \Delta_{n+1}^U \right\}, \tag{14}$$

and $Q_i = (\boldsymbol{q}^{i_1}, \ \boldsymbol{q}^{i_2}, \ \ldots, \ \boldsymbol{q}^{i_{n+1}}) \in \mathbb{R}^{d\times n+1}$. $\boldsymbol{q}^j \in \mathbb{R}^d$ are the columns of $\boldsymbol{q}$.

Proof. Follows from a calculation starting at the definition of the convex conjugate $\boldsymbol{\Psi}^*$. See supplementary material.

Interestingly, although in its original formulation (14) the set $\mathcal{K}$ has infinitely many constraints, one can equivalently represent $\mathcal{K}$ by finitely many.

Proposition 3 *The set* $\mathcal{K}$ *in equation* (14) *is the same as*

$$\mathcal{K} = \left\{ \boldsymbol{q} \in \mathbb{R}^{d \times |\mathcal{V}|} \mid \left\| D^i_{\boldsymbol{q}} \right\|_{S^\infty} \leq 1, \ 1 \leq i \leq |\mathcal{T}| \right\}, \ D^i_{\boldsymbol{q}} = Q_i D \left(T_i D \right)^{-1}, \tag{15}$$

where the matrices $Q_i D \in \mathbb{R}^{d \times n}$ *and* $T_i D \in \mathbb{R}^{n \times n}$ *are given as*

$$Q_i D := \left(\boldsymbol{q}^{i_1} - \boldsymbol{q}^{i_{n+1}}, \ \ldots, \ \boldsymbol{q}^{i_n} - \boldsymbol{q}^{i_{n+1}} \right), \ T_i D := \left(t^{i_1} - t^{i_{n+1}}, \ \ldots, \ t^{i_n} - t^{i_{n+1}} \right).$$

Proof. Similar to the analysis in [11], equation (14) basically states the Lipschitz continuity of a piecewise linear function defined by the matrices $\boldsymbol{q} \in \mathbb{R}^{d \times |\mathcal{V}|}$. Therefore, one can expect that the Lipschitz constraint is equivalent to a bound on the derivative. For the complete proof, see supplementary material.

2.4 Lifting the Overall Optimization Problem

Combining dataterm and regularizer, the overall optimization problem is given

$$\min_{\boldsymbol{u}:\Omega \to \mathbb{R}^{|\mathcal{V}|}} \ \sup_{\boldsymbol{q}:\Omega \to \mathcal{K}} \int_\Omega \boldsymbol{\rho}^{**}(\boldsymbol{u}) + \langle \boldsymbol{u}, \operatorname{Div} \boldsymbol{q} \rangle \ \mathrm{d}x. \tag{16}$$

A highly desirable property is that, opposed to any other vectorial lifting approach from the literature, our method with just one simplex applied to a convex problem yields the same solution as the unlifted problem.

Proposition 4 *If the triangulation contains only 1 simplex,* $\mathcal{T} = \{\Delta\}$, *i.e.,* $|\mathcal{V}| = n + 1$, *then the proposed optimization problem* (16) *is equivalent to*

$$\min_{u:\Omega \to \Delta} \int_\Omega (\rho + \delta_\Delta)^{**}(x, u(x)) \ \mathrm{d}x + \lambda TV(u), \tag{17}$$

which is (1) *with a globally convexified dataterm on* Δ.

Proof. For $u = t^{n+1} + TD\tilde{u}$ the substitution $\boldsymbol{u} = \left(\tilde{u}_1, \ldots, \tilde{u}_n, 1 - \sum_{j=1}^n \tilde{u}_j \right)$ into $\boldsymbol{\rho}^{**}$ and $\boldsymbol{R}$ yields the result. For a complete proof, see supplementary material.

3 Numerical Optimization

3.1 Discretization

For now assume that $\Omega \subset \mathbb{R}^d$ is a d-dimensional Cartesian grid and let Div denote a finite-difference divergence operator with $\operatorname{Div} \boldsymbol{q} : \Omega \to \mathbb{R}^{|\mathcal{V}|}$. Then the relaxed energy minimization problem becomes

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^{|\mathcal{V}|}} \max_{\boldsymbol{q}:\Omega\to\mathcal{K}} \sum_{x\in\Omega} \boldsymbol{\rho}^{**}(x, \boldsymbol{u}(x)) + \langle \operatorname{Div} \boldsymbol{q}, \boldsymbol{u} \rangle. \tag{18}$$

In order to get rid of the pointwise maximum over $\boldsymbol{\rho}_i^*(\boldsymbol{v})$ in Eq. (8), we introduce additional variables $w(x) \in \mathbb{R}$ and additional constraints $(\boldsymbol{v}(x), w(x)) \in \mathcal{C}$, $x \in \Omega$ so that $w(x)$ attains the value of the pointwise maximum:

$$\min_{\boldsymbol{u}:\Omega\to\mathbb{R}^{|\mathcal{V}|}} \max_{\substack{(\boldsymbol{v},w):\Omega\to\mathcal{C} \\ \boldsymbol{q}:\Omega\to\mathcal{K}}} \sum_{x\in\Omega} \langle \boldsymbol{u}(x), \boldsymbol{v}(x) \rangle - w(x) + \langle \operatorname{Div} \boldsymbol{q}, \boldsymbol{u} \rangle, \tag{19}$$

where the set $\mathcal{C}$ is given as

$$\mathcal{C} = \bigcap_{1\leq i\leq|\mathcal{T}|} \mathcal{C}_i, \quad \mathcal{C}_i := \left\{ (x, y) \in \mathbb{R}^{|\mathcal{V}|+1} \mid \boldsymbol{\rho}_i^*(x) \leq y \right\}. \tag{20}$$

For numerical optimization we use a GPU-based implementation[1] of a first-order primal-dual method [14]. The algorithm requires the orthogonal projections of the dual variables onto the sets $\mathcal{C}$ respectively $\mathcal{K}$ in every iteration. However, the projection onto an epigraph of dimension $|\mathcal{V}| + 1$ is difficult for large values of $|\mathcal{V}|$. We rewrite the constraints $(\boldsymbol{v}(x), w(x)) \in \mathcal{C}_i$, $1 \leq i \leq |\mathcal{T}|$, $x \in \Omega$ as $(n+1)$-dimensional epigraph constraints introducing variables $r^i(x) \in \mathbb{R}^n$, $s_i(x) \in \mathbb{R}$:

$$\rho_i^* \left(r^i(x) \right) \leq s_i(x), \quad r^i(x) = A_i^\top E_i^\top \boldsymbol{v}(x), \quad s_i(x) = w(x) - \langle E_i b_i, \boldsymbol{v}(x) \rangle. \tag{21}$$

These equality constraints can be implemented using Lagrange multipliers. For the projection onto the set $\mathcal{K}$ we use an approach similar to [7, Figure 7].

3.2 Epigraphical Projections

Computing the Euclidean projection onto the epigraph of ρ_i^* is a central part of the numerical implementation of the presented method. However, for $n > 1$ this is nontrivial. Therefore we provide a detailed explanation of the projection methods used for different classes of ρ_i. We will consider quadratic, truncated quadratic and piecewise linear ρ.

[1] `https://github.com/tum-vision/sublabel_relax`

Quadratic case: Let ρ be of the form $\rho(u) = \frac{a}{2}\, u^\top u + b^\top u + c$. A direct projection onto the epigraph of $\rho_i^* = (\rho + \delta_{\Delta_i})^*$ for $n > 1$ is difficult. However, the epigraph can be decomposed into separate epigraphs for which it is easier to project onto: For proper, convex, lsc. functions f, g the epigraph of $(f+g)^*$ is the Minkowski sum of the epigraphs of f^* and g^* (cf. [17, Exercise 1.28, Theorem 11.23a]). This means that it suffices to compute the projections onto the epigraphs of a quadratic function $f^* = \rho^*$ and a convex, piecewise linear function $g^*(v) = \max_{1 \le j \le n+1} \langle t^{i_j}, v \rangle$ by rewriting constraint (21) as

$$\rho^*(r_f) \le s_f, \;\; {\delta_{\Delta_i}}^*(c_g) \le d_g \;\; \text{s.t. } (r, s) = (r_f, s_f) + (c_g, d_g). \tag{22}$$

For the projection onto the epigraph of a n-dimensional quadratic function we use the method described in [20, Appendix B.2]. The projection onto a piecewise linear function is described in the last paragraph of this section.

Truncated quadratic case: Let ρ be of the form $\rho(u) = \min\{\, \nu, \; \frac{a}{2}\, u^\top u + b^\top u + c\,\}$ as it is the case for the nonconvex robust ROF with a truncated quadratic dataterm in Sec. 4.2. Again, a direct projection onto the epigraph of ρ_i^* is difficult. However, a decomposition of the epigraph into simpler epigraphs is possible as the epigraph of $\min\{f, g\}^*$ is the intersection of the epigraphs of f^* and g^*. Hence, one can separately project onto the epigraphs of $(\nu + \delta_{\Delta_i})^*$ and $(\frac{a}{2}\, u^\top u + b^\top u + c + \delta_{\Delta_i})^*$. Both of these projections can be handled using the methods from the other paragraphs.

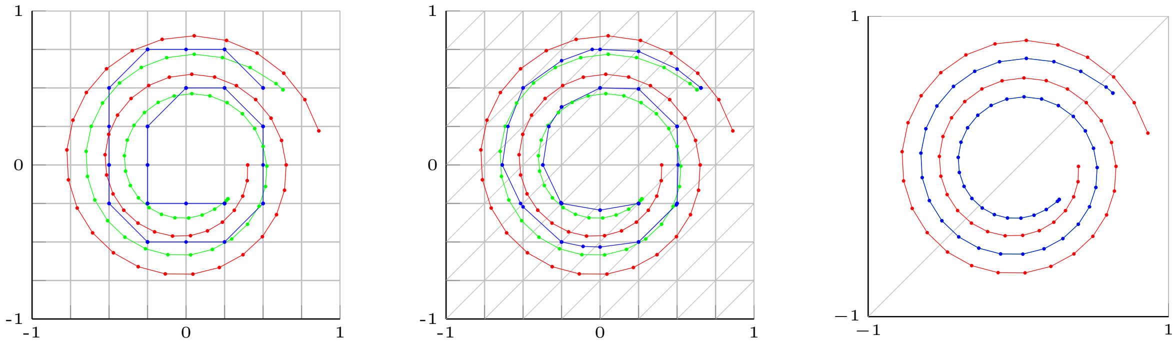
Piecewise linear case: In case ρ is piecewise linear on each Δ_i, i.e., ρ attains finite values at a discrete set of sampled sublabels $\mathcal{V}_i \subset \Delta_i$ and interpolates linearly between them, we have that

$$(\rho + \delta_{\Delta_i})^*(v) = \max_{\tau \in \mathcal{V}_i} \; \langle \tau, v \rangle - \rho(\tau). \tag{23}$$

Again this is a convex, piecewise linear function. For the projection onto the epigraph of such a function, a quadratic program of the form

$$\min_{(x,y) \in \mathbb{R}^{n+1}} \; \frac{1}{2}\|x - c\|^2 + \frac{1}{2}\|y - d\|^2 \;\; \text{s.t. } \langle \tau, x \rangle - \rho(\tau) \le y, \forall \tau \in \mathcal{V}_i \tag{24}$$

needs to be solved. We implemented the primal active-set method described in [13, Algorithm 16.3], and found it solves the program in a few (usually $2-10$) iterations for a moderate number of constraints.



Naive, 81 labels. [11], 81 labels. Ours, **4 labels**.

Fig. 4: ROF denoising of a vector-valued signal $f : [0,1] \to [-1,1]^2$, discretized on 50 points (shown in red). We compare the proposed approach (right) with two alternative techniques introduced in [11] (left and middle). The labels are visualized by the gray grid. While the naive (standard) multilabel approach from [11] (left) provides solutions that are constrained to the chosen set of labels, the sublabel accurate regularizer from [11] (middle) does allow sublabel solutions, yet – due to the dataterm bias – these still exhibit a strong preference for the grid points. In contrast, the proposed approach does not exhibit any visible grid bias providing fully sublabel-accurate solutions: With only 4 labels, the computed solutions (shown in blue) coincide with the "unlifted" problem (green).

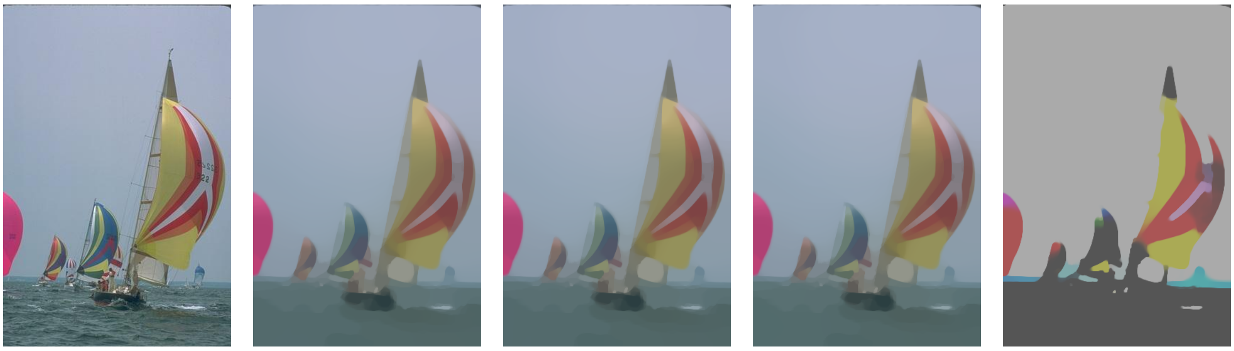
4 Experiments

4.1 Vectorial ROF Denoising

In order to validate experimentally, that our model is exact for convex dataterms, we evaluate it on the Rudin-Osher-Fatemi [18] (ROF) model with vectorial TV (2). In our model this corresponds to defining $\rho(x, u(x)) = \frac{1}{2}\|u(x) - I(x)\|^2$. As expected based on Prop. 4 the energy of the solution of the unlifted problem is equal to the energy of the projected solution of our method for $|\mathcal{V}| = 4$ up to machine precision, as can be seen in Fig. 4 and Fig. 5. We point out, that the sole purpose of this experiment is a proof of concept as our method introduces an overhead and convex problems can be solved via direct optimization. It can be seen in Fig. 4 and Fig. 5, that the baseline method [11] has a strong label bias.

4.2 Denoising with Truncated Quadratic Dataterm

For images degraded with both, Gaussian and salt-and-pepper noise we define the dataterm as $\rho(x, u(x)) = \min \left\{ \frac{1}{2}\|u(x) - I(x)\|^2, \nu \right\}$. We solve the problem



Input image

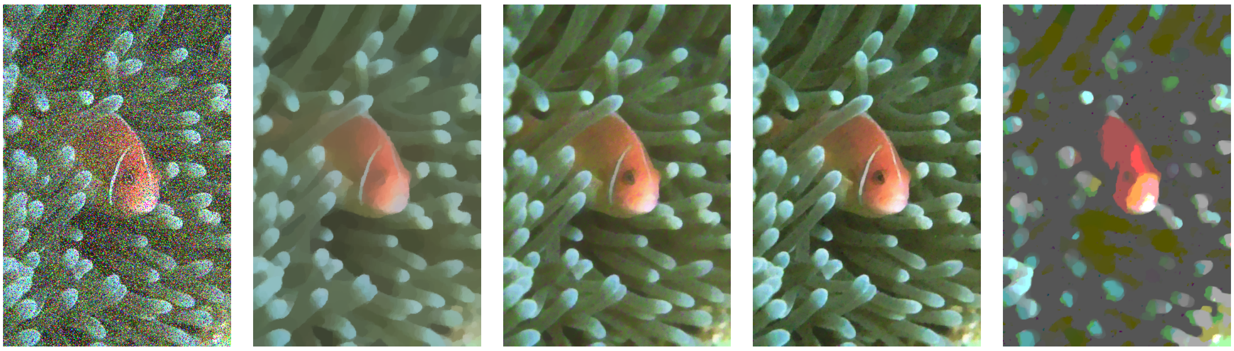
Unlifted Problem, $E = 992.50$

Ours, $|\mathcal{T}| = 1$, $|\mathcal{V}| = 4$, $E = 992.51$

Ours, $|\mathcal{T}| = 6$ $|\mathcal{V}| = 2 \times 2 \times 2$ $E = 993.52$

Baseline, $|\mathcal{V}| = 4 \times 4 \times 4$, $E = 2255.81$

Fig. 5: Convex ROF with vectorial TV. Direct optimization and proposed method yield the same result. In contrast to the baseline method [11] the proposed approach has no discretization artefacts and yields a lower energy. The regularization parameter is chosen as $\lambda = 0.3$.



Noisy input

Ours, $|\mathcal{T}| = 1$, $|\mathcal{V}| = 4$, $E = 2849.52$

Ours, $|\mathcal{T}| = 6$, $|\mathcal{V}| = 2 \times 2 \times 2$, $E = 2806.18$

Ours, $|\mathcal{T}| = 48$, $|\mathcal{V}| = 3 \times 3 \times 3$, $E = 2633.83$

Baseline, $|\mathcal{V}| = 4 \times 4 \times 4$, $E = 3151.80$

Fig. 6: ROF with a truncated quadratic dataterm ($\lambda = 0.03$ and $\nu = 0.025$). Compared to the baseline method [11] the proposed approach yields much better results, already with a very small number of 4 labels.

using the epigraph decomposition described in the second paragraph of Sec. 3.2. It can be seen, that increasing the number of labels $|\mathcal{V}|$ leads to lower energies and at the same time to a reduced effect of the TV. This occurs as we always compute a piecewise convex underapproximation of the original nonconvex dataterm, that gets tighter with a growing number of labels. The baseline method [11] again produces strong discretization artefacts even for a large number of labels $|\mathcal{V}| = 4 \times 4 \times 4 = 64$.

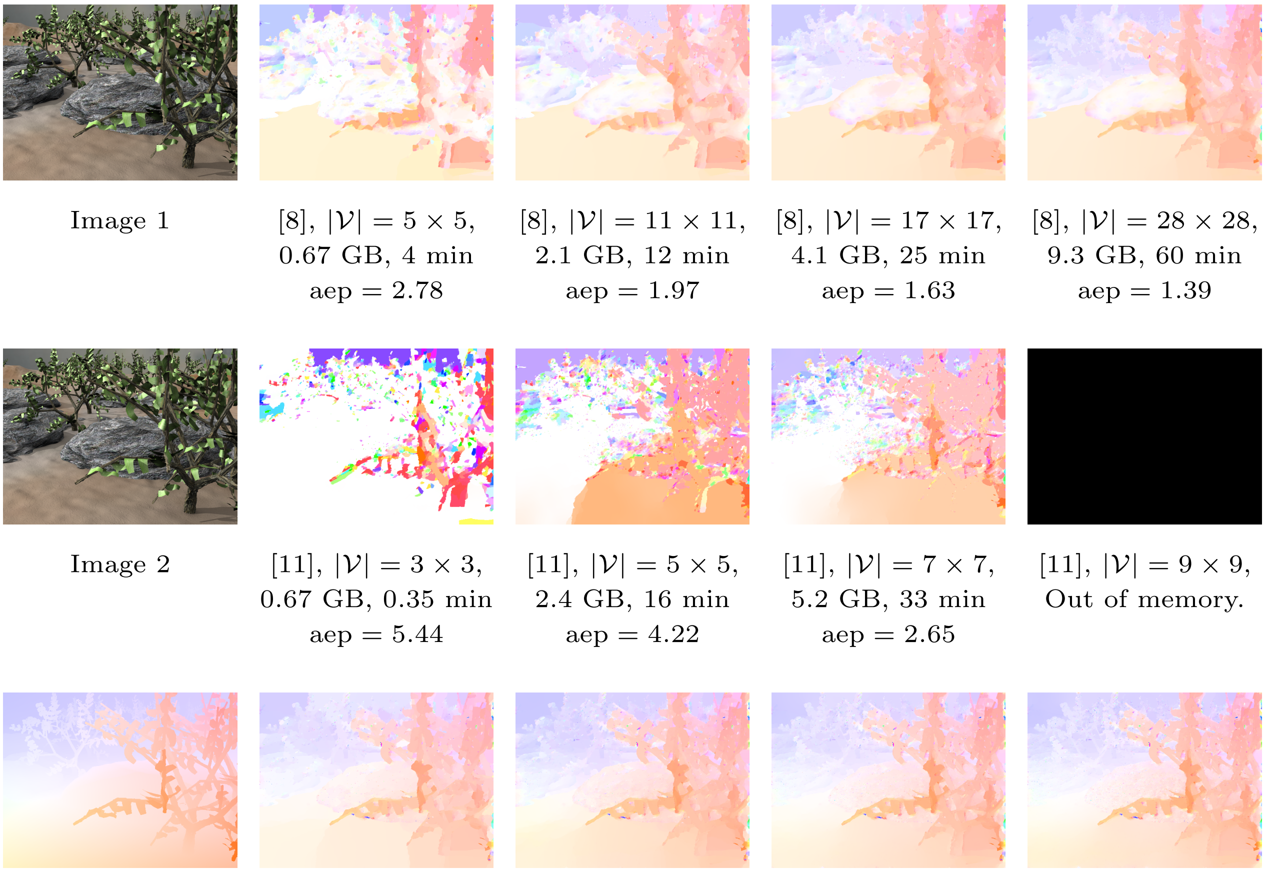


Image 1

[8], $|\mathcal{V}| = 5 \times 5$, 0.67 GB, 4 min aep = 2.78

[8], $|\mathcal{V}| = 11 \times 11$, 2.1 GB, 12 min aep = 1.97

[8], $|\mathcal{V}| = 17 \times 17$, 4.1 GB, 25 min aep = 1.63

[8], $|\mathcal{V}| = 28 \times 28$, 9.3 GB, 60 min aep = 1.39

Image 2

[11], $|\mathcal{V}| = 3 \times 3$, 0.67 GB, 0.35 min aep = 5.44

[11], $|\mathcal{V}| = 5 \times 5$, 2.4 GB, 16 min aep = 4.22

[11], $|\mathcal{V}| = 7 \times 7$, 5.2 GB, 33 min aep = 2.65

[11], $|\mathcal{V}| = 9 \times 9$, Out of memory.

Ground truth

Ours, $|\mathcal{V}| = 2 \times 2$, 0.63 GB, 17 min aep = 1.28

Ours, $|\mathcal{V}| = 3 \times 3$, 1.9 GB, 34 min aep = 1.07

Ours, $|\mathcal{V}| = 4 \times 4$, 4.1 GB, 41 min aep = 0.97

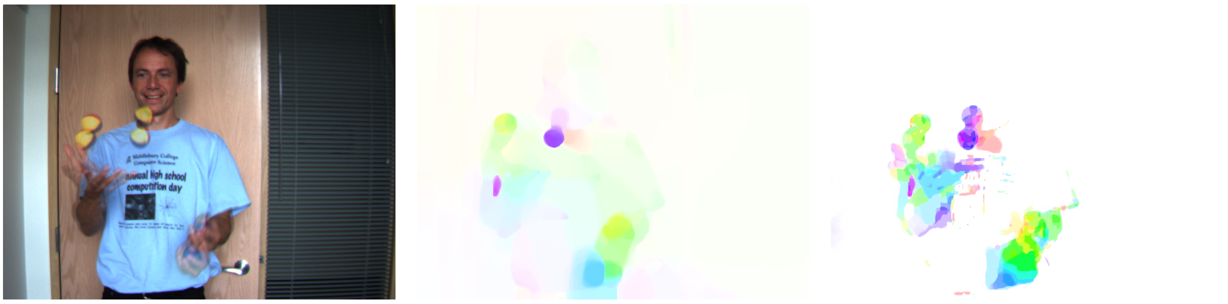
Ours, $|\mathcal{V}| = 6 \times 6$, 10.1 GB, 56 min aep = 0.9

Fig. 7: We compute the optical flow using our method, the product space approach [8] and the baseline method [11] for a varying amount of labels and compare the average endpoint error (aep). The product space method clearly outperforms the baseline, but our approach finds the overall best result already with 2 × 2 labels. To achieve a similarly precise result as the product space method, we require 150 times fewer labels, 10 times less memory and 3 times less time. For the same number of labels, the proposed approach requires more memory as it has to store a convex approximation of the energy instead of a linear one.

4.3 Optical Flow

We compute the optical flow $v : \Omega \to \mathbb{R}^2$ between two input images I_1, I_2. The label space $\Gamma = [-d, d]^2$ is chosen according to the estimated maximum displacement $d \in \mathbb{R}$ between the images. The dataterm is $\rho(x, v(x)) = \|I_2(x) - I_1(x + v(x))\|$, and $\lambda(x)$ is based on the norm of the image gradient $\nabla I_1(x)$.

In Fig. 7 we compare the proposed method to the product space approach [8]. Note that we implemented the product space dataterm using Lagrange mul-



(a) Image 1 and 2 (b) Proposed, $|\mathcal{V}| = 2 \times 2$ (c) Baseline, $|\mathcal{V}| = 7 \times 7$

Fig. 8: Large displacement flow between two 640×480 images (a) using a 81×81 search window. The result of our method with 4 labels is shown in (b), the baseline [11] in (c). Our method can correctly identify the large motion.

tipliers, also referred to as the *global* approach in [8]. While this increases the memory consumption, it comes with lower computation time and guaranteed convergence. For our method, we sample the label space $\Gamma = [-15, 15]^2$ on 150×150 sublabels and subsequently convexify the energy on each triangle using the quickhull algorithm [3]. For the product space approach we sample the label space at equidistant labels, from 5×5 to 27×27. As the regularizer from the product space approach is different from the proposed one, we chose μ differently for each method. For the proposed method, we set $\mu = 0.5$ and for the product space and baseline approach $\mu = 3$. We can see in Fig. 7, our method outperforms the product space approach w.r.t. the average end-point error. Our method outperforms previous lifting approaches: In Fig. 8 we compare our method on large displacement optical flow to the baseline [11]. To obtain competitive results on the Middlebury benchmark, one would need to engineer a better dataterm.

5 Conclusions

We proposed the first sublabel-accurate convex relaxation of vectorial multilabel problems. To this end, we approximate the generally nonconvex dataterm in a piecewise convex manner as opposed to the piecewise linear approximation done in the traditional functional lifting approaches. This assures a more faithful approximation of the original cost function and provides a meaningful interpretation for the non-integral solutions of the relaxed convex problem. In experimental validations on large-displacement optical flow estimation and color image denoising, we show that the computed solutions have superior quality to the traditional convex relaxation methods while requiring substantially less memory and runtime.

References

1. Alberti, G., Bouchitté, G., Maso, G.D.: The calibration method for the Mumford-Shah functional and free-discontinuity problems. Calc. Var. Partial Dif. 3(16), 299–333 (2003)
2. Ambrosio, L., Fusco, N., Pallara, D.: Functions of bounded variation and free discontinuity problems. Oxford Mathematical Monographs, The Clarendon Press Oxford University Press, New York (2000)
3. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS) 22(4), 469–483 (1996)
4. Chambolle, A., Cremers, D., Pock, T.: A convex approach to minimal partitions. SIAM Journal on Imaging Sciences 5(4), 1113–1158 (2012)
5. Fix, A., Agarwal, S.: Duality and the continuous graphical model. In: Computer Vision ECCV 2014, Lecture Notes in Computer Science, vol. 8691, pp. 266–281. Springer International Publishing (2014), http://dx.doi.org/10.1007/978-3-319-10578-9_18
6. Giaquinta, M., Modica, G., Souček, J.: Cartesian currents in the calculus of variations I, II., Ergebnisse der Mathematik und ihrer Grenzgebiete. 3., vol. 37-38. Springer-Verlag, Berlin (1998)
7. Goldluecke, B., Strekalovskiy, E., Cremers, D.: The natural total variation which arises from geometric measure theory. SIAM Journal on Imaging Sciences 5(2), 537–563 (2012)
8. Goldluecke, B., Strekalovskiy, E., Cremers, D.: Tight convex relaxations for vector-valued labeling. SIAM Journal on Imaging Sciences 6(3), 1626–1664 (2013)
9. Ishikawa, H.: Exact optimization for Markov random fields with convex priors. IEEE Trans. Pattern Analysis and Machine Intelligence 25(10), 1333–1336 (2003)
10. Lellmann, J., Schnörr, C.: Continuous multiclass labeling approaches and algorithms. SIAM Journal on Imaging Sciences 4(4), 1049–1096 (2011)
11. Lellmann, J., Strekalovskiy, E., Koetter, S., Cremers, D.: Total variation regularization for functions with values in a manifold. In: ICCV (December 2013)
12. Möllenhoff, T., Laude, E., Moeller, M., Lellmann, J., Cremers, D.: Sublabel-accurate relaxation of nonconvex energies. In: CVPR (2016)
13. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York, 2nd edn. (2006)
14. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: An algorithm for minimizing the piecewise smooth Mumford-Shah functional. In: ICCV (2009)
15. Pock, T., Schoenemann, T., Graber, G., Bischof, H., Cremers, D.: A convex formulation of continuous multi-label problems. In: European Conference on Computer Vision (ECCV). Marseille, France (October 2008)
16. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: Global solutions of variational models with convex regularization. SIAM J. Imaging Sci. 3(4), 1122–1145 (2010)
17. Rockafellar, R., Wets, R.B.: Variational Analysis. Springer (1998)

18. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1), 259–268 (1992)
19. Sapiro, G., Ringach, D.: Anisotropic diffusion of multivalued images with applications to color filtering. IEEE Trans. Img. Proc. 5(11), 1582–1586 (1996)
20. Strekalovskiy, E., Chambolle, A., Cremers, D.: Convex relaxation of vectorial problems with coupled regularization. SIAM Journal on Imaging Sciences 7(1), 294–336 (2014)
21. Zach, C.: Dual decomposition for joint discrete-continuous optimization. In: AISTATS. pp. 632–640 (2013)
22. Zach, C., Kohli, P.: A convex discrete-continuous approach for markov random fields. In: ECCV, vol. 7577, pp. 386–399. Springer Berlin Heidelberg (2012)

Sublabel-Accurate Discretization of Nonconvex Free Discontinuity Problems

Authors	Thomas Möllenhoff [1]	thomas.moellenhoff@tum.de
	Daniel Cremers [1]	cremers@tum.de
	[1] Technische Universität München, Germany	
Publication	T. Möllenhoff, D. Cremers. Sublabel-Accurate Discretization of Nonconvex Free-Discontinuity Problems. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. DOI: http://dx.doi.org/10.1109/ICCV.2017.134. eprint: https://arxiv.org/abs/1611.06987. Copyright © 2017 IEEE. Reprinted with permission from IEEE and the authors.	
Contribution	Problem definition	*significantly contributed*
	Literature survey	*significantly contributed*
	Mathematical derivations	*significantly contributed*
	Numerical implementation	*significantly contributed*
	Experimental evaluation	*significantly contributed*
	Preparation of manuscript	*significantly contributed*
Notice	Following the *IEEE Thesis / Dissertation Reuse Permissions*, we include here the *accepted version* of the original publication [MC17].	

Sublabel-Accurate Discretization of Nonconvex Free-Discontinuity Problems

Thomas Möllenhoff   Daniel Cremers
Technical University of Munich
{thomas.moellenhoff,cremers}@tum.de

Abstract

In this work we show how sublabel-accurate multilabeling approaches [15, 18] can be derived by approximating a classical label-continuous convex relaxation of nonconvex free-discontinuity problems. This insight allows to extend these sublabel-accurate approaches from total variation to general convex and nonconvex regularizations. Furthermore, it leads to a systematic approach to the discretization of continuous convex relaxations. We study the relationship to existing discretizations and to discrete-continuous MRFs. Finally, we apply the proposed approach to obtain a sublabel-accurate and convex solution to the vectorial Mumford-Shah functional and show in several experiments that it leads to more precise solutions using fewer labels.

1. Introduction

1.1. A class of continuous optimization problems

Many tasks particularly in low-level computer vision can be formulated as optimization problems over mappings $u : \Omega \to \Gamma$ between sets Ω and Γ. The energy functional is usually designed in such a way that the minimizing argument corresponds to a mapping with the desired solution properties. In classical discrete Markov random field (MRF) approaches, which we refer to as *fully discrete optimization*, Ω is typically a set of nodes (e.g., pixels or superpixels) and Γ a set of labels $\{1, \dots, \ell\}$.

However, in many problems such as image denoising, stereo matching or optical flow where $\Gamma \subset \mathbb{R}^d$ is naturally modeled as a continuum, this discretization into *labels* can entail unreasonably high demands in memory when using a fine sampling, or it leads to a strong label bias when using a coarser sampling, see Figure 1. Furthermore, as jump discontinuities are ubiquitous in low-level vision (e.g., caused by object edges, occlusion boundaries, changes in albedo, shadows, etc.), it is important to model them in a meaningful manner. By restricting either Ω or Γ to a discrete set, one loses the ability to mathematically distinguish between continuous and discontinuous mappings.

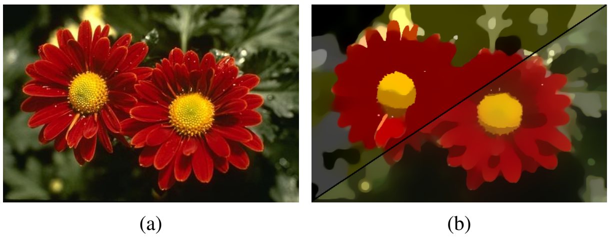


Figure 1: The classical way to discretize continuous convex relaxations such as the vectorial Mumford-Shah functional [26] leads to solutions (**b**), top-left) with a strong bias towards the chosen labels (here an equidistant $5 \times 5 \times 5$ sampling of the RGB space). This can be seen in the bottom left part of the image, where the green color is truncated to the nearest label which is gray. The proposed sublabel-accurate approximation of the continuous relaxation leads to bias-free solutions (**b**), bottom-right).

Motivated by these two points we consider *fully-continuous* optimization approaches, where the idea is to postpone the discretization of $\Omega \subset \mathbb{R}^n$ and $\Gamma \subset \mathbb{R}$ as long as possible. The prototypical class of continuous optimization problems which we consider in this work are nonconvex free-discontinuity problems, inspired by the celebrated Mumford-Shah functional [4, 19]:

$$E(u) = \int_{\Omega \setminus J_u} f\left(x, u(x), \nabla u(x)\right) \mathrm{d}x \\ + \int_{J_u} d\left(x, u^-(x), u^+(x), \nu_u(x)\right) \mathrm{d}\mathcal{H}^{n-1}(x). \tag{1}$$

The first integral is defined on the region $\Omega \setminus J_u$ where u is continuous. The integrand $f : \Omega \times \Gamma \times \mathbb{R}^n \to [0, \infty]$ can be thought of as a combined data term and regularizer, where the regularizer can penalize variations in terms of the (weak) gradient ∇u. The second integral is defined on the $(n-1)$-dimensional discontinuity set $J_u \subset \Omega$ and $d : \Omega \times \Gamma \times \Gamma \times \mathcal{S}^{n-1} \to [0, \infty]$ penalizes jumps from u^- to u^+ in unit direction ν_u. The appropriate function space for (1) are the *special functions of bounded variation*. These are functions of bounded variation (cf. Section 2 for a defintion) whose distributional derivative Du can be decomposed into a continuous part and a jump part in the spirit of (1):

$$Du = \nabla u \cdot \mathcal{L}^n + \left(u^+ - u^-\right) \nu_u \cdot \mathcal{H}^{n-1} \llcorner J_u, \qquad (2)$$

where $\mathcal{L}^n$ denotes the n-dimensional Lebesgue measure and $\mathcal{H}^{n-1} \llcorner J_u$ the $(n-1)$-dimensional Hausdorff measure restricted to the jump set J_u. For an introduction to functions of bounded variation and the study of existence of minimizers to (1) we refer the interested reader to [2].

Note that due to the possible nonconvexity of f in the first two variables a surprisingly large class of low-level vision problems fits the general framework of (1). While (1) is a difficult nonconvex optimization problem, the state-of-the-art are convex relaxations [1, 6, 9]. We give an overview of the idea behind the convex reformulation in Section 3.

Extensions to the vectorial setting, i.e., $\dim(\Gamma) > 1$, have been studied by Strekalovskiy *et al.* in various works [12, 26, 27] and recently using the theory of currents by Windheuser and Cremers [29]. The case when Γ is a manifold has been considered by Lellmann *et al.* [17]. These advances have allowed for a wide range of difficult vectorial and joint optimization problems to be solved within a convex framework.

1.2. Related work

The first practical implementation of (1) was proposed by Pock *et al.* [20], using a simple finite differencing scheme in both Ω and Γ which has remained the standard way to discretize convex relaxations. This leads to a strong label bias (see Figure 1b), top-left) *despite* the initially label-continuous formulation.

In the MRF community, a related approach to overcome this label-bias are *discrete-continuous* models (discrete Ω and continuous Γ), pioneered by Zach *et al.* [30, 31]. Most similar to the present work is the approach of Fix and Agarwal [11]. They derive the discrete-continuous approaches as a discretization of an infinite dimensional dual linear program. Their approach differs from ours, as we start from a different (nonlinear) infinite-dimensional optimization problem and consider a representation of the dual variables which enforces continuity. The recent work of Bach [3] extends the concept of submodularity from discrete to continuous Γ along with complexity estimates.

There are also *continuous-discrete* models, i.e. the range Γ is discretized into labels but Ω is kept continuous [10, 16]. Recently, these spatially continuous multilabeling models have been extended to allow for so-called *sublabel accurate* solutions [15, 18], i.e., solutions which lie between two labels. These are, however, limited to total variation regularization, due to the separate convexification of data term and regularizer. We show in this work that for general regularizers a joint convex relaxation is crucial.

Finally, while not focus of this work, there are of course also *fully-discrete* approaches, among many [14, 25, 28], which inspired some of the continuous formulations.

1.3. Contribution

In this work, we propose an approximation strategy for *fully-continuous* relaxations which retains continuous Γ even after discretization (see Figure 1b), bottom-right). We summarize our contributions as:

- We generalize the work [18] from total variation to general convex and nonconvex regularization.
- We prove (see Prop. 2 and Prop. 4) that different approximations to a convex relaxation of (1) give rise to existing relaxations [20] and [18]. We investigate the relationship to discrete-continuous MRFs in Prop. 5.
- On the example of the vectorial Mumford-Shah functional [26] we show that our framework yields also sublabel-accurate formulations of extensions to (1).

2. Notation and preliminaries

We denote the Iverson bracket as $[\![\cdot]\!]$. Indicator functions from convex analysis which take on values 0 and ∞ are denoted by $\delta\{\cdot\}$. We denote by f^* the convex conjugate of $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Let $\Omega \subset \mathbb{R}^n$ be a bounded open set. For a function $u \in L^1(\Omega; \mathbb{R})$ its total variation is defined by

$$TV(u) = \sup \left\{ \int_\Omega u \operatorname{Div} \varphi \, \mathrm{d}x : \varphi \in C_c^1(\Omega; \mathbb{R}^n) \right\}. \qquad (3)$$

The space of functions of bounded variation, i.e., for which $TV(u) < \infty$ (or equivalently for which the distributional derivative Du is a finite Radon measure) is denoted by $\mathrm{BV}(\Omega; \mathbb{R})$ [2]. We write $u \in \mathrm{SBV}(\Omega; \mathbb{R})$ for functions $u \in \mathrm{BV}(\Omega; \mathbb{R})$ whose distributional derivative admits the decomposition (2). For the rest of this work, we will make the following simplifying assumptions:

- The Lagrangian f in (1) is separable, i.e.,

  $$f(x, t, g) = \rho(x, t) + \eta(x, g), \qquad (4)$$

  for possibly nonconvex $\rho : \Omega \times \Gamma \to \mathbb{R}$ and regularizers $\eta : \Omega \times \mathbb{R}^n \to \mathbb{R}$ which are convex in g.
- The jump regularizer in (1) is isotropic and induced by a concave function $\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}$:

  $$d(x, u^-, u^+, \nu_u) = \kappa(|u^- - u^+|) \|\nu_u\|_2, \qquad (5)$$

  with $\kappa(a) = 0 \Leftrightarrow a = 0$.
- The range $\Gamma = [\gamma_1, \gamma_\ell] \subset \mathbb{R}$ is a compact interval.

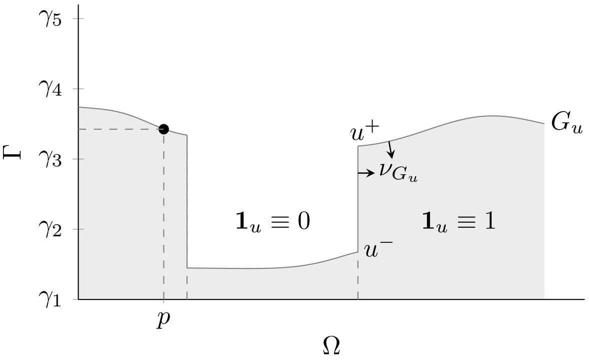


Figure 2: The central idea behind the convex relaxation for problem (1) is to reformulate the functional in terms of the complete graph $G_u \subset \Omega \times \Gamma$ of $u : \Omega \to \Gamma$ in the product space. This procedure is often referred to as "lifting", as one lifts the dimensionality of the problem.

3. The convex relaxation

In [1, 6, 9] the authors propose a convex relaxation for the problem (1). Their basic idea is to reformulate the energy (1) in terms of the *complete graph* of u, i.e. lifting the problem to one dimension higher as illustrated in Figure 2. The complete graph $G_u \subset \Omega \times \Gamma$ is defined as the (measure-theoretic) boundary of the characteristic function of the subgraph $\mathbf{1}_u : \Omega \times \mathbb{R} \to \{0, 1\}$ given by:

$$\mathbf{1}_u(x,t) = [\![t < u(x)]\!]. \tag{6}$$

Furthermore we denote the inner unit normal to $\mathbf{1}_u$ with ν_{G_u}. It is shown in [1] that for $u \in \mathrm{SBV}(\Omega; \mathbb{R})$ one has

$$E(u) = F(\mathbf{1}_u) = \sup_{\varphi \in \mathcal{K}} \int_{G_u} \langle \varphi, \nu_{G_u} \rangle \, \mathrm{d}\mathcal{H}^n, \tag{7}$$

with constraints on the dual variables $\varphi \in \mathcal{K}$ given by

$$\begin{aligned} \mathcal{K} = \Big\{ & (\varphi_x, \varphi_t) \in C_c^1(\Omega \times \mathbb{R}; \mathbb{R}^n \times \mathbb{R}) : \\ & \varphi_t(x,t) + \rho(x,t) \geq \eta^*(x, \varphi_x(x,t)), && (8) \\ & \Big\| \int_t^{t'} \varphi_x(x,t) \mathrm{d}t \Big\|_2 \leq \kappa(|t - t'|), \forall t, t', \forall x \Big\}. && (9) \end{aligned}$$

The functional (7) can be interpreted as the maximum flux of admissible vector fields $\varphi \in \mathcal{K}$ through the cut given by the complete graph G_u. The set $\mathcal{K}$ can be seen as capacity constraints on the flux field φ. This is reminiscent to constructions from the discrete optimization community [14]. The constraints (8) correspond to the first integral in (1) and the non-local constraints (9) to the jump penalization.

Using the fact that the distributional derivative of the subgraph indicator function $\mathbf{1}_u$ can be written as

$$D\mathbf{1}_u = \nu_{G_u} \cdot \mathcal{H}^m \llcorner G_u, \tag{10}$$

one can rewrite the energy (7) as

$$F(\mathbf{1}_u) = \sup_{\varphi \in \mathcal{K}} \int_{\Omega \times \Gamma} \langle \varphi, D\mathbf{1}_u \rangle. \tag{11}$$

A convex formulation is then obtained by relaxing the set of admissible primal variables to a convex set:

$$\begin{aligned} \mathcal{C} = \Big\{ & v \in \mathrm{BV}_{\mathrm{loc}}(\Omega \times \mathbb{R}; [0,1]) : \\ & v(x,t) = 1 \;\; \forall t \leq \gamma_1, v(x,t) = 0 \;\; \forall t > \gamma_\ell, && (12) \\ & v(x, \cdot) \text{ non-increasing} \Big\}. \end{aligned}$$

This set can be thought of as the convex hull of the subgraph functions $\mathbf{1}_u$. The final optimization problem is then a convex-concave saddle point problem given by:

$$\inf_{v \in \mathcal{C}} \sup_{\varphi \in \mathcal{K}} \int_{\Omega \times \mathbb{R}} \langle \varphi, Dv \rangle. \tag{13}$$

In dimension one ($n = 1$), this convex relaxation is tight [8, 9]. For $n > 1$ global optimality can be guaranteed by means of a thresholding theorem in case $\kappa \equiv \infty$ [7, 21]. If the primal solution $\widehat{v} \in \mathcal{C}$ to (13) is binary, the global optimum u^* of (1) can be recovered simply by pointwise thresholding $\widehat{u}(x) = \sup\{t : \widehat{v}(x,t) > \frac{1}{2}\}$. If $\widehat{v}$ is not binary, in the general setting it is not clear how to obtain the global optimal solution from the relaxed solution. An a posteriori optimality bound to the global optimum $E(u^*)$ of (1) for the thresholded solution $\widehat{u}$ can be computed by:

$$|E(\widehat{u}) - E(u^*)| \leq |F(\mathbf{1}_{\widehat{u}}) - F(\widehat{v})|. \tag{14}$$

Using that bound, it has been observed that solutions are usually near globally optimal [26]. In the following section, we show how different discretizations of the continuous problem (13) lead to various existing lifting approaches and to generalizations of the recent sublabel-accurate continuous multilabeling approach [18].

4. Sublabel-accurate discretization

4.1. Choice of primal and dual mesh

In order to discretize the relaxation (13), we partition the range $\Gamma = [\gamma_1, \gamma_\ell]$ into $k = \ell - 1$ intervals. The individual intervals $\Gamma_i = [\gamma_i, \gamma_{i+1}]$ form a one dimensional *simplicial complex* (see e.g., [13]), and we have $\Gamma = \Gamma_1 \cup \ldots \cup \Gamma_k$. The points $\gamma_i \in \Gamma$ are also referred to as *labels*. We assume that the labels are equidistantly spaced with label distance $h = \gamma_{i+1} - \gamma_i$. The theory generalizes also to non-uniformly spaced labels, as long as the spacing is homogeneous in Ω. Furthermore, we define $\gamma_0 = \gamma_1 - h$ and $\gamma_{\ell+1} = \gamma_\ell + h$.

The mesh for dual variables is given by *dual complex*, which is formed by the intervals $\Gamma_i^* = [\gamma_{i-1}^*, \gamma_i^*]$ with nodes $\gamma_i^* = \frac{\gamma_i + \gamma_{i+1}}{2}$. An overview of the notation and the considered finite dimensional approximations is given in Figure 3.

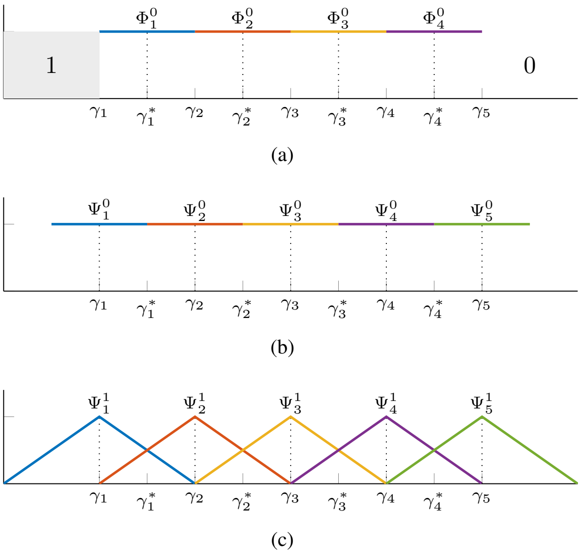


Figure 3: Overview of the notation and proposed finite dimensional approximation spaces.

4.2. Representation of the primal variable

As $\mathbf{1}_u$ is a discontinuous jump function, we consider a piecewise constant approximation for $v \in \mathcal{C}$,

$$\Phi_i^0(t) = [\![t \in \Gamma_i]\!],\ 1 \le i \le k, \tag{15}$$

see Figure 3a). Due to the boundary conditions in Eq. (12), we set v outside of Γ to 1 on the left and 0 on the right. Note that the non-decreasing constraint in $\mathcal{C}$ is implicitly realized as $\varphi_t \in \mathcal{K}$ can be arbitrarily large.

For coefficients $\hat{v} : \Omega \times \{1, \dots, k\} \to \mathbb{R}$ we have

$$v(x,t) = \sum_{i=1}^{k} \hat{v}(x,i)\Phi_i^0(t). \tag{16}$$

As an example of this representation, consider the approximation of $\mathbf{1}_u$ at point p shown in Figure 2:

$$\begin{aligned} \widehat{v}(p,\cdot) &= \sum_{i=1}^{k} e_i \int_\Gamma \Phi_i^0(t)\mathbf{1}_u(p,t)\mathrm{d}t \\ &= h \cdot \begin{bmatrix} 1 & 1 & 0.4 & 0 \end{bmatrix}^\top . \end{aligned} \tag{17}$$

This leads to the sublabel-accurate representation also considered in [18]. In that work, the representation from the above example (17) encodes a convex combination between the labels γ_3 and γ_4 with interpolation factor 0.4. Here it is motivated from a different perspective: we take a finite dimensional subspace approximation of the infinite dimensional optimization problem (13).

4.3. Representation of the dual variables

4.3.1 Piecewise constant φ_t

The simplest discretization of the dual variable φ_t is to pick a piecewise constant approximation on the dual intervals Γ_i^* as shown in Figure 3b): The functions are given by

$$\Psi_i^0(t) = [\![t \in \Gamma_i^*]\!],\ 1 \le i \le \ell, \tag{18}$$

As φ is a vector field in C_c^1, the functions Ψ vanish outside of Γ. For coefficient functions $\hat{\varphi}_t : \Omega \times \{1, \dots, \ell\} \to \mathbb{R}$ and $\hat{\varphi}_x : \Omega \times \{1, \dots, k\} \to \mathbb{R}^n$ we have:

$$\varphi_t(t) = \sum_{i=1}^{\ell} \hat{\varphi}_t(i)\Psi_i^0(t),\ \varphi_x(t) = \sum_{i=1}^{k} \hat{\varphi}_x(i)\Phi_i^0(t). \tag{19}$$

To avoid notational clutter, we dropped $x \in \Omega$ in (19) and will do so also in the following derivations. Note that for φ_x we chose the same piecewise constant approximation as for v, as we keep the model continuous in Ω, and ultimately discretize it using finite differences in x.

Discretization of the constraints In the following, we will plug in the finite dimensional approximations into the constraints from the set $\mathcal{K}$. We start by reformulating the constraints in (8). Taking the infimum over $t \in \Gamma_i$ they can be equivalently written as:

$$\inf_{t \in \Gamma_i} \varphi_t(t) + \rho(t) - \eta^*\left(\varphi_x(t)\right) \ge 0,\ 1 \le i \le \ell. \tag{20}$$

Plugging in the approximation (19) into the above leads to the following constraints for $1 \le i \le k$:

$$\begin{aligned} \hat{\varphi}_t(i) + &\inf_{t \in [\gamma_i, \gamma_i^*]} \rho(t) \ge \eta^*(\hat{\varphi}_x(i)), \\ \hat{\varphi}_t(i+1) + &\underbrace{\inf_{t \in [\gamma_i^*, \gamma_{i+1}]}}_{\text{min-pooling}} \rho(t) \ge \eta^*(\hat{\varphi}_x(i)). \end{aligned} \tag{21}$$

These constraints can be seen as min-pooling of the continuous unary potentials in a symmetric region centered on the label γ_i. To see that more easily, assume one-homogeneous regularization so that $\eta^* \equiv 0$ on its domain. Then two consecutive constraints from (21) can be combined into one where the infimum of ρ is taken over $\Gamma_i^* = [\gamma_i^*, \gamma_{i+1}^*]$ centered the label γ_i. This leads to capacity constraints for the flow in vertical direction $-\hat{\varphi}_t(i)$ of the form

$$-\hat{\varphi}_t(i) \le \inf_{t \in \Gamma_i^*} \rho(t),\ 2 \le i \le \ell - 1, \tag{22}$$

as well as similar constraints on $\hat{\varphi}_t(1)$ and $\hat{\varphi}_t(\ell)$. The effect of this on a nonconvex energy is shown in Figure 4 on the left. The constraints (21) are convex inequality constraints,

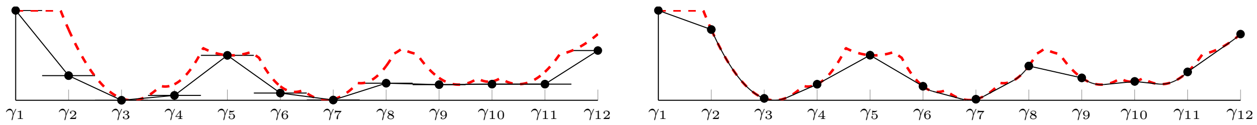


Figure 4: **Left:** piecewise constant dual variables φ_t lead to a linear approximation (shown in black) to the original cost function (shown in red). The unaries are determined through min-pooling of the continuous cost in the Voronoi cells around the labels. **Right:** continuous piecewise linear dual variables φ_t convexify the costs on each interval.

which can be implemented using standard proximal optimization methods and orthogonal projections onto the epigraph epi(η^*) as described in [21, Section 5.3].

For the second part of the constraint set (9) we insert again the finite-dimensional representation (19) to arrive at:

$$\begin{aligned}&\Big\|(1-\alpha)\hat\varphi_x(i)+\sum_{l=i+1}^{j-1}\hat\varphi_x(l)+\beta\hat\varphi_x(j)\Big\| \\ &\quad\le\frac{\kappa(\gamma_j^\beta-\gamma_i^\alpha)}{h},\ \forall\, 1\le i\le j\le k, \alpha,\beta\in[0,1],\end{aligned}\tag{23}$$

where $\gamma_i^\alpha := (1-\alpha)\gamma_i + \alpha\gamma_{i+1}$. These are infinitely many constraints, but similar to [18] these can be implemented with finitely many constraints.

Proposition 1. *For concave $\kappa : \mathbb{R}_0^+ \to \mathbb{R}$ with $\kappa(a) = 0 \Leftrightarrow a = 0$, the constraints (23) are equivalent to*

$$\Big\|\sum_{l=i}^{j}\hat\varphi_x(l)\Big\|\le\frac{\kappa(\gamma_{j+1}-\gamma_i)}{h},\ \forall 1\le i\le j\le k.\tag{24}$$

Proof. Proofs are given in the supplementary material. □

This proposition reveals that only information from the labels γ_i enters into the jump regularizer κ. For $\ell = 2$ we expect all regularizers to behave like the total variation.

Discretization of the energy For the discretization of the saddle point energy (13) we apply the divergence theorem

$$\int_{\Omega\times\mathbb{R}}\langle\varphi, Dv\rangle=\int_{\Omega\times\mathbb{R}}-\operatorname{Div}\varphi\cdot v\ \mathrm{d}t\ \mathrm{d}x,\tag{25}$$

and then discretize the divergence by inserting the piecewise constant representations of φ_t and v:

$$\begin{aligned}&\int_{\mathbb{R}}-\partial_t\varphi_t(t)v(t)\ \mathrm{d}t= \\ &-\hat\varphi_t(1)-\sum_{i=1}^{k}\hat v(i)\left[\hat\varphi_t(i+1)-\hat\varphi_t(i)\right].\end{aligned}\tag{26}$$

The discretization of the other parts of the divergence are given as the following:

$$\int_{\mathbb{R}}-\partial_{x_j}\varphi_x(t)v(t)\ \mathrm{d}t=-h\sum_{i=1}^{k}\partial_{x_j}\hat\varphi_x(i)\hat v(i),\tag{27}$$

where the spatial derivatives ∂_{x_j} are ultimately discretized using standard finite differences. It turns out that the above discretization can be related to the one from [20]:

Proposition 2. *For convex one-homogeneous η the discretization with piecewise constant φ_t and φ_x leads to the traditional discretization as proposed in [20], except with min-pooled instead of sampled unaries.*

4.3.2 Piecewise linear φ_t

As the dual variables in $\mathcal{K}$ are continuous vector fields, a more faithful approximation is given by a continuous piecewise linear approximation, given for $1 \le i \le \ell$ as:

$$\Psi_i^1(t)=\begin{cases}\frac{t-\gamma_{i-1}}{h}, & \text{if } t\in[\gamma_{i-1},\gamma_i],\\ \frac{\gamma_{i+1}-t}{h}, & \text{if } t\in[\gamma_i,\gamma_{i+1}],\\ 0 & \text{otherwise.}\end{cases}\tag{28}$$

They are shown in Figure 3c), and we set:

$$\varphi_t(t)=\sum_{i=1}^{\ell}\hat\varphi_t(i)\Psi_i^1(t).\tag{29}$$

Note that the piecewise linear dual representation considered by Fix *et al.* in [11] differs in this point, as they do not ensure a continuous representation. Unlike the proposed approach their approximation does not take a true subspace of the original infinite dimensional function space.

Discretization of the constraints We start from the reformulation (20) of the original constraints (8). With (29) for φ_t and (19) for φ_x, we have for $1 \le i \le k$:

$$\begin{aligned}\inf_{t\in\Gamma_i}\ &\hat\varphi_t(i)\frac{\gamma_{i+1}-t}{h}+\hat\varphi_t(i+1)\frac{t-\gamma_i}{h}\\ &+\rho(t)\ge\eta^*(\hat\varphi_x(i)).\end{aligned}\tag{30}$$

While the constraints (30) seem difficult to implement, they can be reformulated in a simpler way involving ρ^*.

Proposition 3. *The constraints* (30) *can be equivalently reformulated by introducing additional variables $a \in \mathbb{R}^k$, $b \in \mathbb{R}^k$, where $\forall i \in \{1, \dots, k\}$:*

$$\begin{aligned} &r(i) = (\hat{\varphi}_t(i) - \hat{\varphi}_t(i+1))/h, \\ &a(i) + b(i) - (\hat{\varphi}_t(i)\gamma_{i+1} - \hat{\varphi}_t(x, i+1)\gamma_i)/h = 0, \\ &r(i) \geq \rho_i^* \left(a(i)\right), \hat{\varphi}_x(i) \geq \eta^* \left(b(i)\right), \end{aligned} \tag{31}$$

with $\rho_i(x,t) = \rho(x,t) + \delta\{t \in \Gamma_i\}$.

The constraints (31) are implemented by projections onto the epigraphs of η^* and ρ_i^*, as they can be written as:

$$(r(i), a(i)) \in \mathrm{epi}(\rho_i^*), \ (\hat{\varphi}_x(i), b(i)) \in \mathrm{epi}(\eta^*). \tag{32}$$

Epigraphical projections for quadratic and piecewise linear ρ_i are described in [18]. In Section 5.1 we describe how to implement piecewise quadratic ρ_i. As the convex conjugate of ρ_i enters into the constraints, it becomes clear that this discretization only sees the *convexified* unaries on each interval, see also the right part of Figure 4.

Discretization of the energy It turns out that the piecewise linear representation of φ_t leads to the same discrete bilinear saddle point term as (26). The other term remains unchanged, as we pick the same representation of φ_x.

Relation to existing approaches In the following we point out the relationship of the approximation with piecewise linear φ_t to the sublabel-accurate multilabeling approaches [18] and the discrete-continuous MRFs [31].

Proposition 4. *The discretization with piecewise linear φ_t and piecewise constant φ_x, together with the choice $\eta(g) = \|g\|$ and $\kappa(a) = a$ is equivalent to the relaxation [18].*

Thus we extend the relaxation proposed in [18] to more general regularizations. The relaxation [18] was derived starting from a discrete label space and involved a separate relaxation of data term and regularizer. To see this, first note that the convex conjugate of a convex one-homogeneous function is the indicator function of a convex set [23, Corollary 13.2.1]. Then the constraints (8) can be written as

$$-\varphi_t(x,t) \leq \rho(x,t), \tag{33}$$

$$\varphi_x(x,t) \in \mathrm{dom}\{\eta^*\}, \tag{34}$$

where (33) is the data term and (34) the regularizer. This provides an intuition why the separate convex relaxation of data term and regularizer in [18] worked well. However, for general choices of η a joint relaxation of data term and regularizer as in (30) is crucial. The next proposition establishes the relationship between the data term from [31] and the one from [18].

Proposition 5. *The data term from [18] (which is in turn a special case of the discretization with piecewise linear φ_t) can be (pointwise) brought into the primal form*

$$\mathcal{D}(\widehat{v}) = \inf_{\substack{x_i \geq 0, \sum_i x_i = 1 \\ \widehat{v} = y/h + I^\top x}} \sum_{i=1}^k x_i \rho_i^{**}\left(\frac{y_i}{x_i}\right), \tag{35}$$

where $I \in \mathbb{R}^{k \times k}$ is a discretized integration operator.

The data term of Zach and Kohli [31] is precisely given by (35) except that the optimization is directly performed on $x, y \in \mathbb{R}^k$. The variable x can be interpreted as 1-sparse indicator of the interval Γ_i and $y \in \mathbb{R}^k$ as a sublabel offset. The constraint $\widehat{v} = y/h + I^\top x$ connects this representation to the subgraph representation $\widehat{v}$ via the operator $I \in \mathbb{R}^{k \times k}$ (see supplementary material for the definition). For general regularizers η, the discretization with piecewise linear φ_t differs from [18] as we perform a *joint convexification* of data term and regularizer and from [31] as we consider the spatially continuous setting. Another important question to ask is which primal formulation is actually optimized after discretization with piecewise linear φ_t. In particular the distinction between jump and smooth regularization only makes sense for continuous label spaces, so it is interesting to see what is optimized after discretizing the label space.

Proposition 6. *Let $\gamma = \kappa(\gamma_2 - \gamma_1)$ and $\ell = 2$. The approximation with piecewise linear φ_t and piecewise constant φ_x of the continuous optimization problem* (13) *is equivalent to*

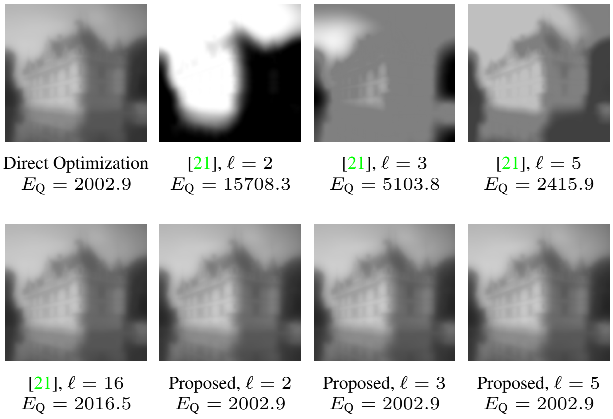
$$\inf_{u:\Omega \to \Gamma} \int_\Omega \rho^{**}(x, u(x)) + (\eta^{**} \,\square\, \gamma\|\cdot\|)(\nabla u(x)) \, \mathrm{d}x, \tag{36}$$

where $(\eta \,\square\, \gamma\|\cdot\|)(x) = \inf_y \eta(x-y) + \gamma\|y\|$ denotes the infimal convolution (cf. [23, Section 5]).

From Proposition 6 we see that the minimal discretization with $\ell = 2$ amounts to approximating problem (1) by globally convexifying the data term. Furthermore, we can see that Mumford-Shah (truncated quadratic) regularization ($\eta(g) = \alpha\|g\|^2$, $\kappa(a) \equiv \lambda[\![a > 0]\!]$) is approximated by a convex Huber regularizer in case $\ell = 2$. This is because the infimal convolution between x^2 and $|x|$ corresponds to the Huber function. While even for $\ell = 2$ this is a reasonable approximation to the original model (1), we can gradually increase the number of labels to get an increasingly faithful approximation of the original nonconvex problem.

4.3.3 Piecewise quadratic φ_t

For piecewise quadratic φ_t the main difficulty are the constraints in (20). For piecewise linear φ_t the infimum over a linear function plus ρ_i lead to (minus) the convex conjugate of ρ_i. Quadratic dual variables lead to so called generalized Φ-conjugates [24, Chapter 11L*, Example 11.66].



Direct Optimization $E_{\mathrm{Q}} = 2002.9$

[21], $\ell = 2$ $E_{\mathrm{Q}} = 15708.3$

[21], $\ell = 3$ $E_{\mathrm{Q}} = 5103.8$

[21], $\ell = 5$ $E_{\mathrm{Q}} = 2415.9$

[21], $\ell = 16$ $E_{\mathrm{Q}} = 2016.5$

Proposed, $\ell = 2$ $E_{\mathrm{Q}} = 2002.9$

Proposed, $\ell = 3$ $E_{\mathrm{Q}} = 2002.9$

Proposed, $\ell = 5$ $E_{\mathrm{Q}} = 2002.9$

Figure 5: To verify the tightness of the approximation, we optimize a convex problem (quadratic data term with quadratic regularization). The discretization with piecewise linear φ_t recovers the exact solution with 2 labels and remains tight (numerically) for all $\ell > 2$, while the traditional discretization from [21] leads to a strong label bias.

Such conjugates were also theoretically considered in the recent work [11] for discrete-continuous MRFs, however an efficient implementation seems challenging. The advantage of this representation would be that one can avoid convexification of the unaries on each interval Γ_i and thus obtain a tighter approximation. While in principle the resulting constraints could be implemented using techniques from convex algebraic geometry and semi-definite programming [5] we leave this direction open to future work.

5. Implementation and extensions

5.1. Piecewise quadratic unaries ρ_i

In some applications such as robust fusion of depth maps, the data term ρ has a piecewise quadratic form:

$$\rho(u) = \sum_{m=1}^{M} \min \left\{ \nu_m, \alpha_m \left(u - f_m\right)^2 \right\}. \tag{37}$$

The intervals on which the above function is a quadratic are formed by the breakpoints $f_m \pm \sqrt{\nu_m / \alpha_m}$. In order to optimize this within our framework, we need to compute the convex conjugate of ρ on the intervals Γ_i, see Eq. (31). We can write the data term (37) on each Γ_i as

$$\min_{1 \leq j \leq n_i} \underbrace{a_{i,j} u^2 + b_{i,j} u + c_{i,j} + \delta\{u \in I_{i,j}\}}_{=:\rho_{i,j}(u)}, \tag{38}$$

where n_i denotes the number of pieces and the intervals $I_{i,j}$ are given by the breakpoints and Γ_i. The convex conjugate is then given by $\rho_i^*(v) = \max_{1 \leq j \leq n_i} \rho_{i,j}^*(v)$. As the epigraph of the maximum is the intersection of the epigraphs, $\mathsf{epi}(\rho_i^*) = \bigcap_{j=1}^{n_j} \mathsf{epi}\left(\rho_{i,j}^*\right)$, the constraints for the data term $(r^i, a^i) \in \mathsf{epi}(\rho_i^*)$, can be broken down:

$$\left(r^{i,j}, a^{i,j}\right) \in \mathsf{epi}\left(\rho_{i,j}^*\right), r^i = r^{i,j}, a^i = a^{i,j}, \forall j. \tag{39}$$

The projection onto the epigraphs of the $\rho_{i,j}^*$ are carried out as described in [18]. Such a convexified piecewise quadratic function is shown on the right in Figure 4.

5.2. The vectorial Mumford-Shah functional

Recently, the free-discontinuity problem (1) has been generalized to vectorial functions $u : \Omega \to \mathbb{R}^{n_c}$ by Strekalovskiy *et al.* [26]. The model they propose is

$$\sum_{c=1}^{n_c} \int_{\Omega \setminus J_u} f_c(x, u_c(x), \nabla_x u_c(x)) \, \mathrm{d}x + \lambda \mathcal{H}^{n-1}(J_u), \tag{40}$$

which consists of a separable data term and separable regularization on the continuous part. The individual channels are coupled through the jump part regularizer $\mathcal{H}^{n-1}(J_u)$ of the joint jump set across all channels. Using the same strategy as in Section 4, applied to the relaxation described in [26, Section 3], a sublabel-accurate representation of the vectorial Mumford-Shah functional can be obtained.

5.3. Numerical solution

We solve the final finite dimensional optimization problem after finite-difference discretization in spatial direction using the primal-dual algorithm [20] implemented in the convex optimization framework `prost` [1].

6. Experiments

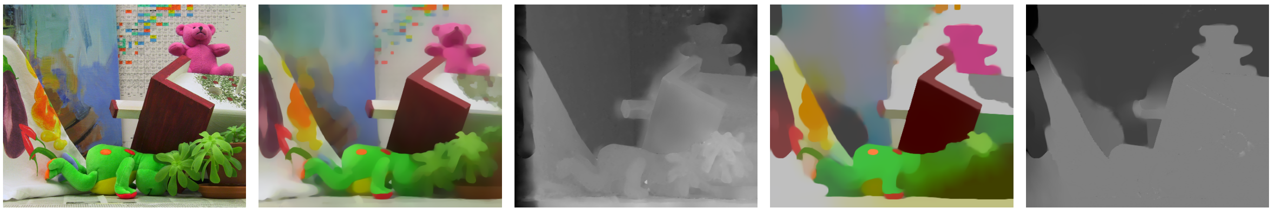
6.1. Exactness in the convex case

We validate our discretization in Figure 5 on the convex problem $\rho(u) = (u - f)^2$, $\eta(\nabla u) = \lambda |\nabla u|^2$. The global minimizer of the problem is obtained by solving $(I - \lambda \Delta) u = f$. For piecewise linear φ_t we recover the exact solution using only 2 labels, and remain (experimentally) exact as we increase the number of labels. The discretization from [21] shows a strong label bias due to the piecewise constant dual variable φ_t. Even with 16 labels their solution is different from the ground truth energy.

6.2. The vectorial Mumford-Shah functional

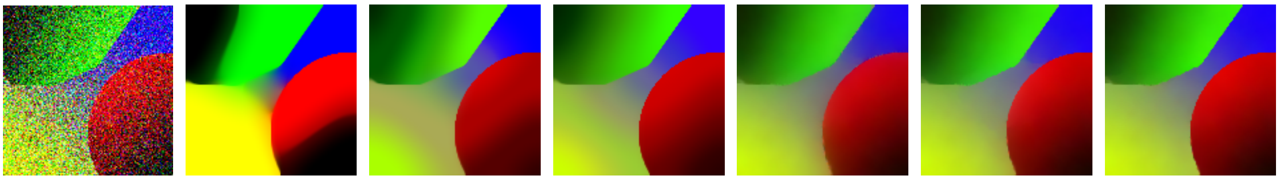
Joint depth fusion and segmentation We consider the problem of joint image segmentation and robust depth fusion from [22] using the vectorial Mumford-Shah functional from Section 5.2. The data term for the depth channel is given by (37), where f_m are the input depth hypotheses, α_m is a depth confidence and ν_m is a truncation parameter to be robust towards outliers. For the segmentation, we use

[1] https://github.com/tum-vision/prost



(a) Left input image (b) Proposed, (Segmentation) (c) Proposed, (Depth map) (d) [26], (Segmentation) (e) [26], (Depth map)

Figure 6: Joint segmentation and stereo matching. **b), c)** Using the proposed discretization we can arrive at smooth solutions using a moderate $(5 \times 5 \times 5 \times 5)$ discretization of the 4-dimensional RGB-D label space. **d), e)** When using such a coarse sampling of the label space, the classical discretization used in [26] leads to a strong label bias. Note that with the proposed approach, a piecewise constant segmentation as in **d)** could also be obtained by increasing the smoothness parameter.



Noisy Input, (PSNR=10.4) [26], $\ell = 2 \times 2 \times 2$ (PSNR=14.7) [26], $\ell = 4 \times 4 \times 4$ (PSNR=25.0) [26], $\ell = 6 \times 6 \times 6$ (PSNR=29.3) Ours, $\ell = 2 \times 2 \times 2$, (PSNR=24.8) Ours, $\ell = 4 \times 4 \times 4$, (PSNR=28.0) Ours, $\ell = 6 \times 6 \times 6$, (PSNR=$\mathbf{30.0}$)

Figure 7: Denoising of a synthetic piecewise smooth image degraded with 30% Gaussian noise. The standard discretization of the vectorial Mumford-Shah functional shows a strong bias towards the chosen labels (see also Figure 8), while the proposed discretization has no bias and leads to the highest overall peak signal to noise ratio (PSNR).

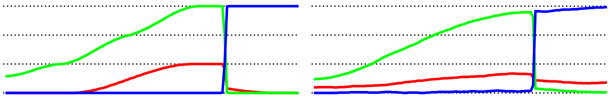


Figure 8: We show a 1D-slice through the resulting image in Figure 7 (with $\ell = 4 \times 4 \times 4$). The discretization [26] (left) shows a strong bias towards the labels, while the proposed discretization (right) yields a sublabel-accurate solution.

a quadratic difference dataterm in RGB space. For Figure 6 we computed multiple depth hypotheses f_m on a stereo pair using different matching costs (sum of absolute (gradient) differences, and normalized cross correlation) with varying patch radii (0 to 2). Even for a moderate label space of $5 \times 5 \times 5 \times 5$ we have no label discretization artifacts.

The piecewise linear approximation of the unaries in [26] leads to an almost piecewise constant segmentation of the image. To highlight the sublabel-accuracy of the proposed approach we chose a small smoothness parameter which leads to a piecewise smooth segmentation, but with a higher smoothness term or different choice of unaries a piecewise constant segmentation could also be obtained.

Piecewise-smooth approximations In Figure 7 we compare the discretizations for the vectorial Mumford-Shah functional. We see that the approach [26] shows strong label bias (see also Figure 8 and 1) while the discretiziation with piecewise linear duals leads to a sublabel-accurate result.

7. Conclusion

We proposed a framework to numerically solve *fully-continuous* convex relaxations in a sublabel-accurate fashion. The key idea is to implement the dual variables using a piecewise linear approximation. We prove that different choices of approximations for the dual variables give rise to various existing relaxations: in particular piecewise constant duals lead to the traditional lifting [20] (with min-pooling of the unary costs), whereas piecewise linear duals lead to the sublabel lifting that was recently proposed for total variation regularized problems [18]. While the latter method is not easily generalized to other regularizers due to the separate convexification of data term and regularizer, the proposed representation generalizes to arbitrary convex and non-convex regularizers such as the scalar and the vectorial Mumford-Shah problem. The proposed approach provides a systematic technique to derive sublabel-accurate discretizations for continuous convex relaxation approaches, thereby boosting their memory and runtime efficiency for challenging large-scale applications.

References

[1] G. Alberti, G. Bouchitté, and G. Dal Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Calc. Var. Partial Differential Equations*, 16(3):299–333, 2003. 2, 3

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford University Press, USA, 2000. 2

[3] F. Bach. Submodular functions: from discrete to continous domains. *arXiv:1511.00394*, 2015. 2

[4] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, 1987. 1

[5] G. Blekherman, P. A. Parrilo, and R. R. Thomas. *Semidefinite Optimization and Convex Algebraic Geometry*. SIAM, 2012. 6

[6] G. Bouchitté. Recent convexity arguments in the calculus of variations. *Lecture notes from the 3rd Int. Summer School on the Calculus of Variations, Pisa*, 1998. 2, 3

[7] G. Bouchitté and I. Fragalà. Duality for non-convex variational problems. *Comptes Rendus Mathematique*, 353(4):375–379, 2015. 3

[8] M. Carioni. A discrete coarea-type formula for the Mumford-Shah functional in dimension one. *arXiv preprint arXiv:1610.01846*, 2016. 3

[9] A. Chambolle. Convex representation for lower semicontinuous envelopes of functionals in L^1. *J. Convex Anal.*, 8(1):149–170, 2001. 2, 3

[10] A. Chambolle, D. Cremers, and T. Pock. A convex approach to minimal partitions. *SIAM J. Imaging Sciences*, 5(4):1113–1158, 2012. 2

[11] A. Fix and S. Agarwal. Duality and the continuous graphical model. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2014. 2, 5, 6

[12] B. Goldluecke, E. Strekalovskiy, and D. Cremers. Tight convex relaxations for vector-valued labeling. *SIAM J. Imaging Sciences*, 6(3):1626–1664, 2013. 2

[13] A. N. Hirani. *Discrete exterior calculus*. PhD thesis, California Institute of Technology, 2003. 4

[14] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(10):1333–1336, 2003. 2, 3

[15] E. Laude, T. Möllenhoff, M. Moeller, J. Lellmann, and D. Cremers. Sublabel-accurate convex relaxation of vectorial multilabel energies. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2016. 1, 2

[16] J. Lellmann and C. Schnörr. Continuous multiclass labeling approaches and algorithms. *SIAM J. Imaging Sciences*, 4(4):1049–1096, 2011. 2

[17] J. Lellmann, E. Strekalovskiy, S. Koetter, and D. Cremers. Total variation regularization for functions with values in a manifold. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, 2013. 2

[18] T. Möllenhoff, E. Laude, M. Moeller, J. Lellmann, and D. Cremers. Sublabel-accurate relaxation of nonconvex energies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016. 1, 2, 3, 4, 5, 6, 7

[19] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42(5):577–685, 1989. 1

[20] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the piecewise smooth Mumford-Shah functional. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, 2009. 2, 5, 7

[21] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM J. Imaging Sci.*, 3(4):1122–1145, 2010. 3, 4, 7, 10

[22] T. Pock, C. Zach, and H. Bischof. Mumford-Shah meets stereo: Integration of weak depth hypotheses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2007. 7

[23] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996. 6

[24] R. T. Rockafellar, R. J.-B. Wets, and M. Wets. *Variational analysis*. Springer, 1998. 6

[25] M. Schlesinger. Sintaksicheskiy analiz dvumernykh zritelnikh signalov v usloviyakh pomekh (Syntactic analysis of two-dimensional visual signals in noisy conditions). *Kibernetika*, 4:113–130, 1976. 2

[26] E. Strekalovskiy, A. Chambolle, and D. Cremers. A convex representation for the vectorial Mumford-Shah functional. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2012. 1, 2, 3, 7, 11

[27] E. Strekalovskiy, A. Chambolle, and D. Cremers. Convex relaxation of vectorial problems with coupled regularization. *SIAM J. Imaging Sciences*, 7(1):294–336, 2014. 2

[28] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(7):1165–1179, 2007. 2

[29] T. Windheuser and D. Cremers. A convex solution to spatially-regularized correspondence problems. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2016. 2

[30] C. Zach. Dual decomposition for joint discrete-continuous optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, AISTATS*, 2013. 2

[31] C. Zach and P. Kohli. A convex discrete-continuous approach for Markov random fields. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2014. 2, 5, 6

Lifting Vectorial Variational Problems: A Natural Formulation based on Geometric Measure Theory and Discrete Exterior Calculus

Authors	Thomas Möllenhoff[1]	thomas.moellenhoff@tum.de
	Daniel Cremers[1]	cremers@tum.de
	[1] Technische Universität München, Germany	

Publication T. Möllenhoff, D. Cremers. Lifting Vectorial Variational Problems: A Natural Formulation based on Geometric Measure Theory and Discrete Exterior Calculus. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. DOI: http://dx.doi.org/10.1109/CVPR.2019.01137. eprint: https://arxiv.org/abs/1905.00851. Copyright © 2019 IEEE. Reprinted with permission from IEEE and the authors.

Contribution	Problem definition	*significantly contributed*
	Literature survey	*significantly contributed*
	Mathematical derivations	*significantly contributed*
	Numerical implementation	*significantly contributed*
	Experimental evaluation	*significantly contributed*
	Preparation of manuscript	*significantly contributed*

Notice Following the *IEEE Thesis / Dissertation Reuse Permissions*, we include here the *accepted version* of the original publication [MC19a].

Lifting Vectorial Variational Problems: A Natural Formulation based on Geometric Measure Theory and Discrete Exterior Calculus

Thomas Möllenhoff and Daniel Cremers
Technical University of Munich
{thomas.moellenhoff,cremers}@tum.de

Abstract

Numerous tasks in imaging and vision can be formulated as variational problems over vector-valued maps. We approach the relaxation and convexification of such vectorial variational problems via a lifting to the space of currents. To that end, we recall that functionals with polyconvex Lagrangians can be reparametrized as convex one-homogeneous functionals on the graph of the function. This leads to an equivalent shape optimization problem over oriented surfaces in the product space of domain and codomain. A convex formulation is then obtained by relaxing the search space from oriented surfaces to more general currents. We propose a discretization of the resulting infinite-dimensional optimization problem using Whitney forms, which also generalizes recent "sublabel-accurate" multilabeling approaches.

1. Introduction

We consider functionals of C^1-mappings $f : \mathcal{X} \to \mathcal{Y}$

$$E(f) = \int_{\mathcal{X}} c\,(x, f(x), \nabla f(x))\,\mathrm{d}x, \tag{1}$$

where $\mathcal{X} \subset \mathbf{R}^n$, $\mathcal{Y} \subset \mathbf{R}^N$ are bounded and open. The cost function $c \equiv c(x, y, \xi)$ is assumed to be a nonnegative (possibly nonconvex) continuous function on $\mathcal{X} \times \mathcal{Y} \times \mathbf{R}^{N \times n}$ that is *polyconvex* (see Def. 2) in the Jacobian matrix ξ.

This work is concerned with relaxation and global optimization of (1) when, both, dimension and codimension are possibly larger than one ($n > 1$, $N > 1$). This is expected to be difficult: In the discrete setting problems with $n = 1$ or $N = 1$ typically correspond to polynomial-time solvable shortest path ($n = 1$) or graph cut ($N = 1$) problems [11, 60, 24, 53], whereas for $n, N > 1$, the arising multilabel problems with unordered label spaces are known to be NP-hard - see [35]. Nevertheless, heuristic strategies have been shown to yield excellent results in tasks such as optical flow [10] or shape matching [55, 9]. In contrast to such well-established Markov random field (MRF) works [30, 31, 29, 55, 39, 9, 10, 14] we consider the way less explored continuous (infinite-dimensional) setting.

Our motivation partly stems from the fact that formulations in function space are very general and admit a variety of discretizations. Finite difference discretizations of continuous relaxations often lead to models that are reminiscent of MRFs [70], while piecewise-linear approximations are related to discrete-continuous MRFs [71], see [17, 40]. More recently, for the Kantorovich relaxation in optimal transport, approximations with deep neural networks were considered and achieved promising performance, for example in generative modeling [2, 54].

We further argue that fractional (non-integer) solutions to a careful discretization of the continuous model can implicitly approximate an "integer" continuous solution. Therefore one can achieve accuracies that go substantially beyond the mesh size. The resulting models would be difficult to interpret and derive from a finite-dimensional viewpoint such that the continuous considerations are required for the final implementation. Also, formulations arising from continuous relaxations allow one to introduce isotropic smoothness potentials without reverting to higher-order terms in the cost, and, as we show in this work, one can impose general polyconvex regularizations using only local constraints. An example of a polyconvex function (which is in general nonconvex) is the surface area of the graph, sometimes referred to as "Beltrami regularization" in the image processing community, see e.g., [28].

In contrast to the discrete multi-labeling setting, an important question is whether variational problems involving the energy (1) admit a minimizer. A fruitful approach to address this question is to suitably relax the notion of solution, thereby enlarging the search space of admissible candidates ("lifting the problem to a larger space"). The origins of this idea can be traced back[1] to the turn of the century, see Hilbert's twentieth problem [21]. An example of that principle is the celebrated Kantorovich relaxation [26] of Monge's transportation problem [42]. There,

[1] We refer the interested reader to the historical remarks in L. C. Young's book on the calculus of variations [69, pp. 122–123].

the search over maps $f : \mathcal{X} \to \mathcal{Y}$ is relaxed to one over probability measures on the product space $\mathcal{X} \times \mathcal{Y}$. Each map can be identified in that extended space with a measure concentrated on its graph. Existence of optimal transportation plans follows directly due to good compactness properties of the larger space. Furthermore, the nonlinearly constrained and nonconvex optimization problem is transformed into one of linear programming, leading to rich duality theories and fast numerical algorithms [47].

One may ask whether the relaxed solution in the extended space has certain regularity properties, for example whether it is the graph of a (sufficiently regular) map and thus can be considered a solution to the original ("unlifted") problem. In the case of optimal transport, such regularity theory can be guaranteed under some assumptions [63, 52]. Establishing existence and regularity for problems in which the cost additionally depends on the Jacobian (for example minimal surface problems) has been a driving factor in the development of geometric measure theory, see [44] for an introduction. In this work, we will use ideas from geometric measure theory to pursue the above relaxation and lifting principle for the energy (1). The main idea is to reformulate the original variational problem as a shape optimization problem over oriented manifolds representing the graph of the map $f : \mathcal{X} \to \mathcal{Y}$ in the product space $\mathcal{X} \times \mathcal{Y}$. To obtain a convex formulation we enlarge the search space from oriented manifolds to currents.

1.1. Related Work

A common strategy to solve problems involving (1) is to revert to local gradient descent minimization based on the Euler-Lagrange equations. But for nonconvex problems solutions might depend on the initialization and the computed stationary points may be quite suboptimal. Therefore, we pursue the aforementioned lifting of the energy (1) to currents. This lifting has been previously considered in geometric measure theory to establish the aforementioned existence and regularity theory for vectorial variational problems in a very broad setting, see e.g., [15, 16, 5]. In contrast to such impressive theoretical achievements, this paper is concerned with a discretization and implementation.

There is also a variety of related applied works. The paper [68] tackles the problem of bijective and smooth shape matching using linear programming. Similar to the present work, the authors also look for graph surfaces in $\mathcal{X} \times \mathcal{Y}$ but they consider the discrete setting and use a different notion of boundary operator. We study the continuous setting, but also our discrete formulation is quite different.

For $N = 1$, the proposed continuous formulation specializes to [1, 48]. To tackle the setting of $N > 1$ in a memory efficient manner, Strekalovskiy *et al.* [58, 19, 59] keep a collection of N surfaces with codimension one under the factorization assumption that $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_N$. In contrast, we consider only one surface of codimension N, we do not require an assumption on $\mathcal{Y}$, our approach is applicable to a larger class of functionals and we expect it to yield a tighter relaxation. The lifting approaches [34, 20] also tackle vectorial problems by considering the full product space, but are limited to total variation regularization (with the former allowing $\mathcal{Y}$ to be a manifold). The recent work [67] is most related to the present one, however their relaxation considers a specific instance of (1). Moreover, the above works are based on finite difference discretizations of the continuous model. In contrast, the proposed discretization using discrete exterior calculus yields solutions beyond the mesh accuracy as in recent sublabel-accurate approaches. The latter are restricted to $N = 1$ [41, 40] or total variation regularization [33]. Recent works also include extensions to total generalized variation or Laplacian regularization [57, 64, 36].

Recent approaches in shape analysis [56, 62, 61] also operate in the product space $\mathcal{X} \times \mathcal{Y}$. However, these are based on local minimizations of the Gromov-Wasserstein distance [37] and spectral variants thereof [38] which leads to (nonconvex) quadratic assignment problems. While the goal to find a smooth (possibly bijective) map is similar, the formulations appear to be quite different. To alleviate the increased cost of the product space formulation, computationally efficient representations of densities in $\mathcal{X} \times \mathcal{Y}$ have been studied in the context of functional maps [46, 51].

2. Notation and Preliminaries

Throughout this paper we will introduce notions from geometric measure theory, as they are not commonly used in the vision community. While the subject is rather technical, our aim is to keep the presentation light and to focus on the geometric intuition and aspects which are important for a practical implementation. We invite the reader to consult chapter 4 in the book [44] and the chapter on exterior calculus in [13], which both contain many illuminating illustrations. For a more technical treatment we refer to [15, 32].

In the following, we denote a basis in $\mathbf{R}^d$ as $\{e_1, \ldots, e_d\}$ with dual basis $\{\mathrm{d}x_1, \ldots, \mathrm{d}x_d\}$ where $\mathrm{d}x_i : \mathbf{R}^d \to \mathbf{R}$ is the linear functional that maps every $x = (x_1, \ldots, x_d)$ to the i-th component x_i. Given an integer $k \leq d$, $I(d, k)$ are the multi-indices $\mathbf{i} = (i_1, \ldots, i_k)$ with $1 \leq i_1 < \ldots < i_k \leq d$.

As we will consider n-surfaces in $\mathcal{X} \times \mathcal{Y} \subset \mathbf{R}^{n+N}$, most of the time we set $d = n+N$ and $k = n$. To further simplify notation, we denote the basis vectors $\{e_{n+1}, \ldots, e_{n+N}\}$ by $\{\varepsilon_1, \ldots, \varepsilon_N\}$ and similarly refer to the dual basis as $\{\mathrm{d}x_1, \ldots \mathrm{d}x_n, \mathrm{d}y_1, \ldots, \mathrm{d}y_N\}$. When it is clear from the context, we treat vectors $e_i \in \mathbf{R}^n$ and $\varepsilon_i \in \mathbf{R}^N$ in the sense that $e_i \simeq (e_i, \mathbf{0}_N) \in \mathbf{R}^{n+N}$, $\varepsilon_i \simeq (\mathbf{0}_n, \varepsilon_i) \in \mathbf{R}^{n+N}$. As an example, for $\nabla f(x) \in \mathbf{R}^{N \times n}$ we can define the expression $e_i + \nabla f(x) e_i$ and read it as $(e_i, \nabla f(x) e_i) \in \mathbf{R}^{n+N}$.

2.1. Convex Analysis

The extended reals are denoted by $\overline{\mathbf{R}} = \mathbf{R} \cup \{+\infty\}$. For a finite-dimensional real vector space V and $\Psi : V \to \overline{\mathbf{R}}$ we denote the convex conjugate as $\Psi^* : V^* \to \overline{\mathbf{R}}$ and the biconjugate as $\Psi^{**} : V \to \overline{\mathbf{R}}$. Ψ^{**} is the largest lower-semicontinuous convex function below Ψ. In our notation, for functions with several arguments, the conjugate is always taken only in the last argument. As a general reference to convex analysis, we refer the reader to the books [23, 50].

2.2. Multilinear Algebra

The formalism of multi-vectors we introduce in this section is central to this work, as the idea of the relaxation is to represent the oriented graph of f by a k-vectorfield (more precisely: a k-current) in the product space $\mathcal{X} \times \mathcal{Y}$. Basically, one can multiply $v_i \in \mathbf{R}^d$ to obtain an object

$$v = v_1 \wedge \ldots \wedge v_k, \tag{2}$$

called a *simple* k-vector in $\mathbf{R}^d$. The geometric intuition of simple k-vectors is, that they describe the k-dimensional space spanned by the $\{v_i\}$, together with an orientation and the area of the parallelotope given by the $\{v_i\}$. Thus, simple k-vectors can be thought of oriented parallelotopes as shown in orange in Fig. 1. In general, k-vectors are defined to be formal sums

$$v = \sum_{\mathbf{i} \in I(d,k)} v^{\mathbf{i}} \cdot e_{i_1} \wedge \ldots \wedge e_{i_k} = \sum_{\mathbf{i} \in I(d,k)} v^{\mathbf{i}} \cdot e_{\mathbf{i}}, \tag{3}$$

for coefficients $v^{\mathbf{i}} \in \mathbf{R}$. They form the vector space $\mathbf{\Lambda}_k \mathbf{R}^d$, which has dimension $\binom{d}{k}$.

The dual space $\mathbf{\Lambda}^k \mathbf{R}^d$ of k-covectors is defined analogously, with $\langle \mathrm{d}x_{\mathbf{i}}, e_{\mathbf{j}} \rangle = \delta_{\mathbf{ij}}$. We define for two k-vectors (and also for k-covectors) $v = \sum_{\mathbf{i}} v_{\mathbf{i}} e_{\mathbf{i}}$, $w = \sum_{\mathbf{i}} w_{\mathbf{i}} e_{\mathbf{i}}$ an inner product $\langle v, w \rangle = \sum_{\mathbf{i}} v_{\mathbf{i}} w_{\mathbf{i}}$ and norm $|v| = \sqrt{\langle v, v \rangle}$.

k-vectors (elements of $\mathbf{\Lambda}_k \mathbf{R}^d$) are called *simple*, if they can be written as the *wedge product* of 1-vectors as in (2). Unfortunately, for $1 < k < d - 1$, not all k-vectors are simple and the set of simple k-vectors is a nonconvex cone in $\mathbf{\Lambda}_k \mathbf{R}^d$, called the Grassmann cone [7]. This is one aspect why the setting of $n > 1$ and $N > 1$ is more challenging.

Later on, we will consider a relaxation from the nonconvex set of simple k-vectors to general k-vectors. Naturally, for the relaxation to be good, we want the convex energy to be *as large as possible* on non-simple k-vectors. For the Euclidean norm, a good convex extension is the *mass* norm

$$\|v\| = \inf \left\{ \sum_i |\xi_i| : \xi_i \text{ are simple}, v = \sum_i \xi_i \right\}. \tag{4}$$

The dual norm is the *comass* norm given by:

$$\|w\|^* = \sup \left\{ \langle w, v \rangle : v \text{ is simple }, |v| \leq 1 \right\}. \tag{5}$$

The mass norm can be understood as the largest norm that agrees with the Euclidean norm on simple k-vectors.

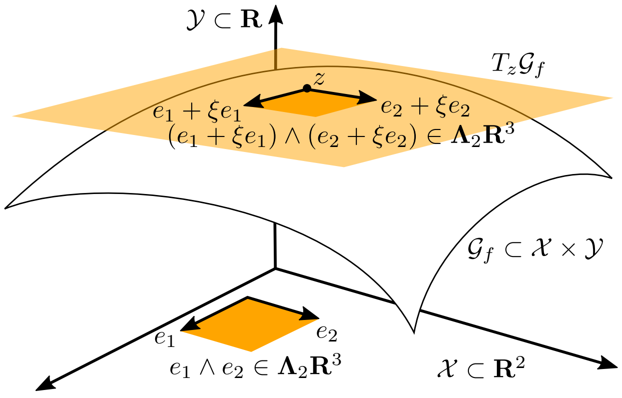

$\mathcal{Y} \subset \mathbf{R}$
$T_z\mathcal{G}_f$
z
$e_1 + \xi e_1$
$e_2 + \xi e_2$
$(e_1 + \xi e_1) \wedge (e_2 + \xi e_2) \in \mathbf{\Lambda}_2 \mathbf{R}^3$
$\mathcal{G}_f \subset \mathcal{X} \times \mathcal{Y}$
e_1
e_2
$e_1 \wedge e_2 \in \mathbf{\Lambda}_2 \mathbf{R}^3$
$\mathcal{X} \subset \mathbf{R}^2$

Figure 1: Illustration for the setting of $n = 2$, $N = 1$. The graph $\mathcal{G}_f$ of the C^1-map $f : \mathcal{X} \to \mathcal{Y}$ is a smooth oriented manifold embedded in the product space $\mathcal{X} \times \mathcal{Y}$. The tangent space at $z = (x, f(x))$ is spanned by the simple n-vector $(e_1 + \xi e_1) \wedge \ldots \wedge (e_n + \xi e_n) \in \mathbf{\Lambda}_n \mathbf{R}^{n+N}$, where $\xi = \nabla f(x) \in \mathbf{R}^{N \times n}$ is the Jacobian.

3. Lifting to Graphs in the Product Space

With the necessary preliminaries in mind, our goal is now to reparametrize the original energy (1) to the graph $\mathcal{G}_f \subset \mathcal{X} \times \mathcal{Y}$. As shown in Fig. 1, the graph is an oriented n-dimensional manifold in the product space with global parametrization $u(x) = (x, f(x))$.

Definition 1 (Orientation). *If $\mathcal{M} \subset \mathbf{R}^d$ is a k-dimensional smooth manifold in $\mathbf{R}^d$ (possibly with boundary), an **orientation** of $\mathcal{M}$ is a continuous map $\tau_{\mathcal{M}} : \mathcal{M} \to \mathbf{\Lambda}_k \mathbf{R}^d$ such that $\tau_{\mathcal{M}}(z)$ is a simple k-vector with unit norm that spans the tangent space $T_z\mathcal{M}$ at every point $z \in \mathcal{M}$.*

From differential geometry we know that the tangent space $T_z\mathcal{G}_f$ at $z = (x, f(x))$ is spanned by $\partial_i u(u^{-1}(z)) = e_i + \nabla f(x) e_i$. Therefore, an orientation of $\mathcal{G}_f$ is given by

$$\tau_{\mathcal{G}_f}(z) = \frac{M(\nabla f(\pi_1 z))}{|M(\nabla f(\pi_1 z))|}, \tag{6}$$

where the map $M : \mathbf{R}^{N \times n} \to \mathbf{\Lambda}_n \mathbf{R}^{n+N}$ is given by

$$M(\xi) = (e_1 + \xi e_1) \wedge \ldots \wedge (e_n + \xi e_n), \tag{7}$$

and $\pi_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ is the canonical projection onto the first argument. In order to derive the reparametrization, we have to connect a simple n-vector (representing an oriented tangent plane of the graph) with the Jacobian of the original energy. For that, we need an inverse of the map given in (7).

To derive such an inverse, we first introduce further helpful notations. For $\mathbf{i} \in I(m, l)$ we denote by $\bar{\mathbf{i}} \in I(m, m - l)$ the element which complements $\mathbf{i}$ in $\{1, 2, \ldots, m\}$ in increasing order, denote $\bar{0} = \{1, \ldots, m\}$ and 0 as the empty multi-index. Every $v \in \mathbf{\Lambda}_n \mathbf{R}^{n+N}$ can be written as

$$v = \sum_{|\mathbf{i}|+|\mathbf{j}|=n} v^{\mathbf{i},\mathbf{j}} e_{\mathbf{i}} \wedge \varepsilon_{\mathbf{j}}, \tag{8}$$

where $\mathbf{i} \in I(n,l)$, $\mathbf{j} \in I(N,l')$, $l + l' = n$. To give an example, the $\binom{5}{2} = 10$ coefficients of a 2-vector $v \in \mathbf{\Lambda}_2 \mathbf{R}^5$ according to the notation (8) are:

$$\begin{array}{llll} v^{\bar{0},0} & & & \\ v^{1,1} & v^{2,1} & & \\ v^{1,2} & v^{2,2} & v^{0,(1,2)} & \\ v^{1,3} & v^{2,3} & v^{0,(1,3)} & v^{0,(2,3)}, \end{array} \tag{9}$$

where we highlighted the $N \times n$ coefficients with $|\mathbf{j}| = 1$. Now note that the vector $v = M(\xi)$ is by construction a simple n-vector with first component $v^{\bar{0},0} = 1$. To any $v \in \mathbf{\Lambda}_n \mathbf{R}^{n+N}$ with $v^{\bar{0},0} = 1$ we associate $\xi(v) \in \mathbf{R}^{N \times n}$ given by

$$[\xi(v)]_{j,i} = (-1)^{n-i} v^{\bar{i},j}. \tag{10}$$

If and only if $v \in \mathbf{\Lambda}_n \mathbf{R}^{n+N}$ is simple with first component $v^{\bar{0},0} = 1$ then $v = M(\xi(v))$. A proof is given in [18, Vol. I, Ch. 2.1, Prop. 1]. Thus, on the set of simple n-vectors with first component $v^{\bar{0},0} = 1$,

$$\Sigma_1 = \{v \in \mathbf{\Lambda}_n \mathbf{R}^{n+N} : v = M(\xi) \text{ for } \xi \in \mathbf{R}^{N \times n}\}, \tag{11}$$

the inverse of the map (7) is given by (10).

Using the above notations, we can define a generalized notion of convexity, which essentially states that there is a convex reformulation on k-vectors.

Definition 2 (Polyconvexity)**.** *A map $c : \mathbf{R}^{N \times n} \to \overline{\mathbf{R}}$ is **polyconvex** if there is a convex function $\bar{c} : \mathbf{\Lambda}_n \mathbf{R}^{n+N} \to \overline{\mathbf{R}}$ such that we have*

$$c(\xi) = \bar{c}(M(\xi)) \quad \textit{for all} \quad \xi \in \mathbf{R}^{N \times n}. \tag{12}$$

*Equivalently one has that $c(\xi(v)) = \bar{c}(v)$ for all $v \in \Sigma_1$. We also refer to the convex function $\bar{c}$ as a **polyconvex extension**.*

In general, the polyconvex extension is not unique. Any convex function has an obvious polyconvex extension by (10), but as discussed in the previous section we would like the convex extension to be as large as possible for $v \notin \Sigma_1$. The largest polyconvex extension which agrees with the original function on Σ_1 can be formally defined using the convex biconjugate, but is often hard to explicitly compute. The mass norm (4) corresponds to such a construction.

Nevertheless, given any polyconvex extension, we can now reparametrize the original energy (1) on the oriented graph $\mathcal{G}_f$, as we show in the following central proposition.

Proposition 1. *Let $\bar{c} : \mathcal{X} \times \mathcal{Y} \times \mathbf{\Lambda}_n \mathbf{R}^{n+N} \to \overline{\mathbf{R}}$ be a polyconvex extension of the original cost c in the last argument. Define the function $\Psi : \mathcal{X} \times \mathcal{Y} \times \mathbf{\Lambda}_n \mathbf{R}^{n+N} \to \overline{\mathbf{R}}$,*

$$\Psi(z, v) = \begin{cases} v^{\bar{0},0} \bar{c}(\pi_1 z, \pi_2 z, v / v^{\bar{0},0}), & \textit{if } v^{\bar{0},0} > 0, \\ +\infty, & \textit{otherwise,} \end{cases} \tag{13}$$

where $\pi_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and $\pi_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ are the canonical projections onto the first and second argument. Then we can reparametrize (1) as follows:

$$\begin{aligned} &\int_{\mathcal{X}} c(x, f(x), \nabla f(x)) \, \mathrm{d}\mathcal{L}^n(x) \\ &\qquad = \int_{\mathcal{G}_f} \Psi(z, \tau_{\mathcal{G}_f}(z)) \, \mathrm{d}\mathcal{H}^n(z), \end{aligned} \tag{14}$$

where the second integral is the standard Lebesgue integral with respect to the n-dimensional Hausdorff measure on $\mathbf{R}^{n+N}$ restricted to the graph $\mathcal{G}_f$.

Proof. We directly calculate:

$$\int_{\mathcal{X}} c\,(x, f(x), \nabla f(x)) \, \mathrm{d}\mathcal{L}^n(x) \tag{15}$$

$$= \int_{\mathcal{X}} \Psi\,(x, f(x), M(\nabla f(x))) \, \mathrm{d}\mathcal{L}^n(x) \tag{16}$$

$$= \int_{\mathcal{G}_f} \Psi\,(z, M(\nabla f(\pi_1 z))) \, \frac{1}{|M(\nabla f(\pi_1 z))|} \mathrm{d}\mathcal{H}^n(z) \tag{17}$$

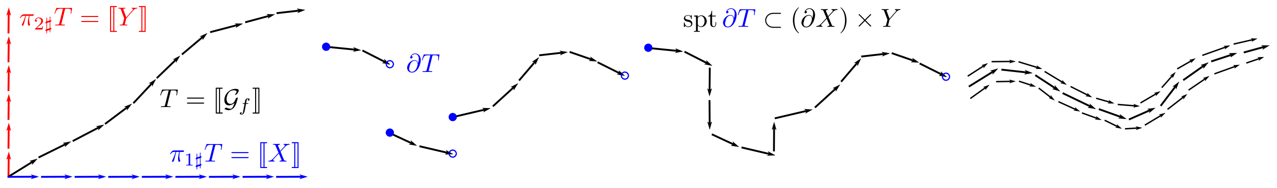
$$= \int_{\mathcal{G}_f} \Psi\,\left(z, \tau_{\mathcal{G}_f}(z)\right) \mathrm{d}\mathcal{H}^n(z). \tag{18}$$

The step from (15) to (16) uses that $\bar{c}$ is a polyconvex extension (so that we can apply (12)) and the fact that for $v = M(\nabla f(x))$ we have $v^{\bar{0},0} = 1$. To arrive at (17), an application of the area formula [32, Corollary 5.1.13] suffices and for (18) we used positive one-homogenity of Ψ and the definition of $\tau_{\mathcal{G}_f}$ in (6). □

Interestingly, the function (13) is convex and one-homogeneous in the last argument, as it is the *perspective* of a convex function. However, the search space of oriented graphs of C^1 mappings is nonconvex. Therefore we relax from oriented graphs to the larger set of currents, which we will introduce in the following section. Since currents form a vector space, we therefore obtain a convex functional over a convex domain.

4. From Oriented Graphs to Currents

Throughout this section, let $U \subset \mathbf{R}^d$ be an open set, which will later be a neighbourhood of $X \times Y \subset \mathbf{R}^{n+N}$, where $X = \mathrm{cl}(\mathcal{X})$, $Y = \mathrm{cl}(\mathcal{Y})$ are the closures of $\mathcal{X}, \mathcal{Y}$. The main idea of our relaxation and the geometric intuitions of *pushforward* and *boundary* operator we introduce in this section are summarized in the following Fig. 2. Currents are defined in duality with differential forms, which we will briefly introduce in the following section.


$\pi_{2\sharp}T = [\![Y]\!]$
$T = [\![\mathcal{G}_f]\!]$
$\pi_{1\sharp}T = [\![X]\!]$
∂T
$\operatorname{spt} \partial T \subset (\partial X) \times Y$

(a) Graph of diffeomorphism f (b) Graph of function with jumps (c) "Stitched" graph (d) Current which is not a graph

Figure 2: The idea of our relaxation is to move from oriented graphs in the product space to the larger set of currents. These include oriented graphs as special cases, as shown in (a). For a diffeomorphism, the pushforwards $\pi_{1\sharp}T$, $\pi_{2\sharp}T$ yield currents induced by domain and codomain, which will be a linear constraint in the relaxed problem. In (b) we show the current given by the graph of a discontinuous function. Since it has holes, the boundary operator ∂T has support inside the domain. We will constrain the support of the boundary to exclude such cases. (c) Stitching jumps yields a current with vertical parts at the jump points, which corresponds to the limiting case in the perspective function (13). To obtain an overall convex formulation, we will also allow currents (d) which don't necessarily concentrate on the graph of a function.

4.1. Differential Forms

A differential form of order k (short: k-form) is a map $\omega : U \to \mathbf{\Lambda}^k \mathbf{R}^d$. The *support* of a differential form $\operatorname{spt} \omega$ is defined as the closure of $\{z \in U : \omega(z) \neq 0\}$. Integration of a k-form over an oriented k-dimensional manifold is defined by

$$\int_{\mathcal{M}} \omega := \int_{\mathcal{M}} \langle \omega(z), \tau_{\mathcal{M}}(z) \rangle \, \mathrm{d}\mathcal{H}^k(z). \tag{19}$$

A notion of derivative for k-forms is the exterior derivative $d\omega$, which is a $(k+1)$-form given by:

$$\langle d\omega(z), v_1 \wedge \ldots \wedge v_{k+1} \rangle = \lim_{h \to 0} \frac{1}{h^{k+1}} \int_{\partial P} \omega, \tag{20}$$

where ∂P is the oriented boundary of the parallelotope spanned by the $\{hv_i\}$ at point z. To get an intuition, note that for $k = 0$ this reduces to the familiar directional derivative $\langle d\omega(x), v_1 \rangle = \lim_{h \to 0} \frac{1}{h} (\omega(x + hv_1) - \omega(x))$. With (19) and (20) in mind, one sees why Stokes' theorem

$$\int_{\mathcal{M}} d\omega = \int_{\partial \mathcal{M}} \omega. \tag{21}$$

should hold intuitively. Given a map $\pi : \mathbf{R}^d \to \mathbf{R}^q$, the *pullback* $\pi^\sharp \omega$ of the k-form ω is determined by

$$\langle \pi^\sharp \omega, v_1 \wedge .. \wedge v_k \rangle = \langle \omega \circ \pi, D_{v_1} \pi \wedge .. \wedge D_{v_k} \pi \rangle, \tag{22}$$

where $D_{v_i} \pi = \nabla \pi \cdot v_i$ and $\nabla \pi \in \mathbf{R}^{q \times d}$ is the Jacobian.

4.2. Currents

Denote the space of smooth k-forms with compact support on U as $\mathcal{D}^k(U)$. *Currents* are elements of the dual space $\mathcal{D}_k(U) = \mathcal{D}^k(U)'$, i.e., linear functionals acting on differential forms. As shown in Fig. 2a, an oriented k-dimensional manifold $\mathcal{M} \subset U$ induces a current by

$$[\![\mathcal{M}]\!](\omega) = \int_{\mathcal{M}} \omega. \tag{23}$$

However, since $\mathcal{D}_k(U)$ is a vector space, not all elements look like k-dimensional manifolds, see Fig. 2d. The *boundary* of the k-current $T \in \mathcal{D}_k(U)$ is the $(k-1)$-current $\partial T \in \mathcal{D}_{k-1}(U)$ defined via the exterior derivative:

$$\partial T(\omega) = T(d\omega), \text{ for all } \omega \in \mathcal{D}^{k-1}(U). \tag{24}$$

Stokes' theorem (21) ensures that for currents which are given by k-dimensional oriented manifolds, the boundary of the current agrees with the usual notion, see Fig. 2b.

The *support* of a current, denoted by $\operatorname{spt} T$, is the complement of the biggest open set V such that

$$T(\omega) = 0 \text{ whenever } \operatorname{spt}(\omega) \subset V. \tag{25}$$

Given a map $\pi : \mathbf{R}^d \to \mathbf{R}^q$ the *pushforward* $\pi_\sharp T$ of the k-current $T \in \mathcal{D}_k(U)$ is given by

$$\pi_\sharp T(\omega) = T(\pi^\sharp \omega), \text{ for all } \omega \in \mathcal{D}^k(\mathbf{R}^q). \tag{26}$$

Intuitively, it transforms the current using the map π, as illustrated in Fig. 2a. The *mass* of a current $T \in \mathcal{D}_k(U)$ is

$$\mathbb{M}(T) = \sup \left\{ T(\omega) : \omega \in \mathcal{D}^k(U), \|\omega(z)\|^* \leq 1 \right\}, \tag{27}$$

and as expected $\mathbb{M}([\![\mathcal{M}]\!]) = \mathcal{H}^k(\mathcal{M})$. We denote the space of k-currents with finite mass and compact support by $\mathbf{M}_k(U)$. These are *representable by integration*, meaning there is a measure $\|T\|$ on U and a map $\vec{T} : U \to \mathbf{\Lambda}_k \mathbf{R}^d$ such that $\|\vec{T}(z)\| = 1$ for $\|T\|$-almost all z such that

$$T(\omega) = \int \langle \omega(z), \vec{T}(z) \rangle \, \mathrm{d}\|T\|(z). \tag{28}$$

The decomposition (28) is crucial, and we will use it to define the relaxation in the next section.

4.3. The Relaxed Energy

We lift the original energy (1) to the space of finite mass currents $T \in \mathbf{M}_n(U)$ with $\operatorname{spt} T \subset X \times Y$ as follows:

$$\mathbf{E}(T) = \int \Psi^{**}\left(\pi_1 z, \pi_2 z, \vec{T}(z)\right) \, \mathrm{d}\|T\|(z). \tag{29}$$

Since for $T = [\![\mathcal{G}_f]\!]$ we have $\vec{T} = \tau_{\mathcal{G}_f}$, $\|T\| = \mathcal{H}^n \llcorner \mathcal{G}_f$ the desirable property $\mathbf{E}([\![\mathcal{G}_f]\!]) = E(f)$ holds due to Prop. 1.

Note that in (29) we use the lower-semicontinuous regularization Ψ^{**} which extends (13) at $v^{\bar{0},0} = 0$ with the correct value. Interestingly, this point corresponds to the situation when the graph has vertical parts, which cannot occur for C^1 functions but can happen for general currents, see Fig. 2c. In [43] it was shown that one can penalize such jumps in a way depending on the jump distance and direction. We will not consider such additional regularization due to space limitations, but remark that they could be integrated by adding further constraints to the following dual representation, which is a consequence of [18, Vol. II, Sec. 1.3.1, Thm. 2].

Proposition 2. *For $T \in \mathbf{M}_n(U)$ with $\operatorname{spt} T \subset X \times Y$, we have the dual representation*

$$\mathbf{E}(T) = \sup_{\omega \in \mathcal{K}} T(\omega), \tag{30}$$

where the constraint is the closed and convex set

$$\mathcal{K} = \Big\{\omega \in C_c^0(U, \mathbf{\Lambda}^n \mathbf{R}^{n+N}) : \\ \Psi^*(\pi_1 z, \pi_2 z, \omega(z)) \leq 0, \forall z \in X \times Y\Big\}. \tag{31}$$

The final relaxed optimization problem for (1) reads

$$\inf_{T \in \mathbf{M}_n(U)} \mathbf{E}(T), \;\; \text{s.t.} \;\; \operatorname{spt} T \subset X \times Y, \; T \in \mathcal{C}. \tag{32}$$

Depending on the kind of problem one wishes to solve, a different convex constraint set $\mathcal{C}$ should be considered. For example, in the case of variational problems with Dirichlet boundary conditions, we set

$$\mathcal{C} = \{T \, : \, \pi_{1\sharp} T = [\![X]\!], \; \partial T = S\}, \tag{33}$$

where $S \in \mathbf{M}_{n-1}(U)$ is a given boundary datum. In case of Neumann boundary conditions, one constrains the support of the boundary to be zero inside the domain

$$\mathcal{C} = \{T \, : \, \pi_{1\sharp} T = [\![X]\!], \operatorname{spt} \partial T \subset (\partial X) \times Y\}, \tag{34}$$

to exclude surfaces with holes, but allow the boundary to be freely chosen on $(\partial X) \times Y$. In case $n = N$, one can also consider the constraint set

$$\mathcal{C} = \{T \, : \, \pi_{1\sharp} T = [\![X]\!], \pi_{2\sharp} T = [\![Y]\!], \\ \operatorname{spt} \partial T \subset \partial(X \times Y)\}, \tag{35}$$

where the additional pushforward constraint encourages bijectivity. Notice also the similarity of (32) together with (35) to the Kantorovich relaxation in optimal transport.

Existence of minimizing currents to a similar problem as (32) in a certain space of currents (*real flat chains*) is shown in [16, §3.9]. For dimension $n = 1$ or codimension $N = 1$, the infimum is actually realized by a surface (*integral flat chain*) [16, §5.10, §5.12]. An adaptation of such theoretical considerations to our setting and conditions under which the relaxation is tight in the scenario $n > 1$, $N > 1$ is a major open challenge and left for future work.

5. Discrete Formulation

In this section we present an implementation of the continuous model (32) using discrete exterior calculus [22]. We will base our discretization on cubes since they are easy to work with in high dimensions, but one could also use simplices. To define cubical meshes, we adopt some notations from computational homology [25].

Definition 3 (Elementary interval and cube)**.** *An **elementary interval** is an interval $I \subset \mathbf{R}$ of the form $I = [l, l+1]$ or $I = \{l\}$ for $l \in \mathbf{Z}$. Intervals that consist of a single point are **degenerate**. An **elementary cube** is given by a product $\kappa = I_1 \times \ldots \times I_d$, where each I_i is an elementary interval. The set of elementary cubes in $\mathbf{R}^d$ is denoted by K^d.*

For $\kappa \in K^d$, denote by $\dim \kappa \in \{1, \ldots, d\}$ the number of nondegenerate intervals. We denote $\mathbf{i}(\kappa) \in I(d, \dim \kappa)$ as the multi-index referencing the nondegenerate intervals.

Definition 4 (Cubical set)**.** *A set $Q \subset \mathbf{R}^d$ is a **cubical set** if it can be written as a finite union of elementary cubes.*

Let $K_k^d(Q) = \{\kappa \in K^d : \kappa \subset Q, \dim \kappa = k\}$ be the set of k-dimensional cubes contained in $Q \subset \mathbf{R}^d$. A map $\phi : Q \to X \times Y$ will transform the cubical set to our domain. As we work with images, it will just be a mesh spacing, i.e., we set $\phi(z) = (h_1 z_1, \ldots, h_d z_d)$.

Definition 5 (k-chains, k-cochains)**.** *We denote the space of finite formal sums of elements in $K_k^d(Q)$ with real coefficients as $\mathcal{C}_k(Q)$, called (real) **k-chains**. We denote the dual as $\mathcal{C}_k(Q)^* = \mathcal{C}^k(Q)$ and call the elements k-**cochains**.*

Definition 6 (Boundary)**.** *For $\kappa \in K_k^d(Q)$, denote the primary faces obtained by collapsing the j-th non-degenerate interval to the lower respectively upper boundary as $\kappa_j^-, \kappa_j^+ \in K_{k-1}^d$. The **boundary** of an elementary cube $\kappa \in K_k^d(Q)$ is the $(k-1)$-chain,*

$$\partial \kappa = \sum_{j=1}^{k} (-1)^{j-1} (\kappa_j^+ - \kappa_j^-) \in \mathcal{C}_{k-1}(Q). \tag{36}$$

*The **boundary operator** is given by the extension to a linear map $\partial : \mathcal{C}_k(Q) \to \mathcal{C}_{k-1}(Q)$.*

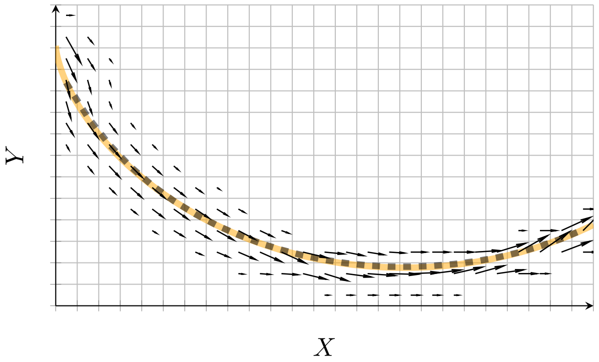


Figure 3: Minimization of the *Brachistochrone* energy on a 25×14 cubical set (gray squares). The proposed discretization yields a diffuse current (black vector field), whose center of mass (black, dashed) however is faithful to the analytical cycloid solution (orange) far beyond the mesh accuracy.

A k-chain $T = \sum_\kappa T_\kappa \kappa \in \mathcal{C}_k(Q)$ can be identified with a k-current $T' \in \mathcal{D}_k(U)$ by $T' = \sum_\kappa T_\kappa \phi_\sharp [\![\kappa]\!]$. The above discrete notion of boundary is defined in analogy to the continuous definition (24).

In our discretization, we will use the dual representation of the lifted energy from Prop. 2. To implement differential forms, we introduce an interpolation operator.

Definition 7 (Whitney map)**.** *The* ***Whitney map*** *extends a k-cochain ω to a k-form $(\mathbb{W}\omega) : X \times Y \to \mathbf{\Lambda}^k \mathbf{R}^d$,*

$$(\mathbb{W}\omega)(x) = \sum_{\kappa \in K_k^d(Q)} \omega_\kappa \widehat{\mathbb{W}}(\phi^{-1}(x), \kappa), \tag{37}$$

where $\omega_\kappa \in \mathbf{R}$ are the coefficients of the k-cochain,

$$\widehat{\mathbb{W}}(x, \kappa) = \mathrm{d}x_{\mathbf{i}(\kappa)} \prod_{i \in \overline{\mathbf{i}(\kappa)}} \max\{0, 1 - |x_i - I_i(\kappa)|\}, \tag{38}$$

and $I_i(\kappa) \in \mathbf{Z}$ is the element in the degenerate interval.

Interestingly, the Whitney map (for simplicial meshes) first appeared in [66, Eq. 27.12] but specializes to lowest-order Raviart-Thomas [49] ($k = 2$,$d = 3$) and Nédélec [45] elements (for $k = 1$, $d = 3$), see [3, 4]. Differential forms of type (37) are called Whitney forms.

We also define a weighted inner product $\langle \cdot, \cdot \rangle_\phi$ between chains and cochains by plugging the Whitney form associated to the k-cochain into the current corresponding to the k-chain. As both are constant on each k-cube, a quick calculation shows: $\langle T, \omega \rangle_\phi = \sum_\kappa T_\kappa \omega_\kappa \mathcal{H}^k(\phi(\kappa))$, where $\mathcal{H}^k(\phi(\kappa))$ is simply the volume of the k-cube under the mesh spacing ϕ.

Using the dual representation (30), and approximating the current by a k-chain and the differential forms with

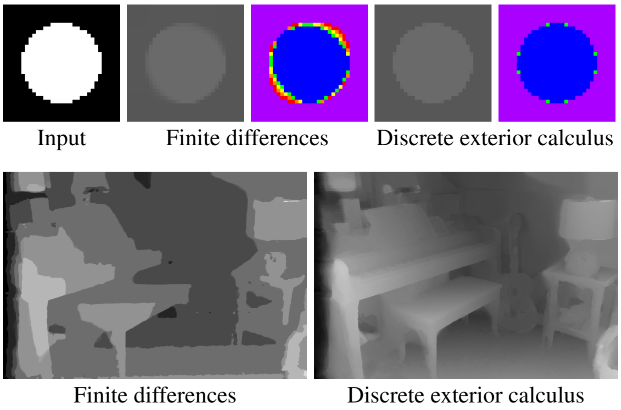

Input
Finite differences
Discrete exterior calculus
Finite differences
Discrete exterior calculus

Figure 4: Total variation minimization. **Top:** The proposed DEC discretization yields solutions with better isotropy and sharper discontinuities. **Bottom:** In that stereo matching example, we enforce the continuous constraints $\mathbb{W}\omega \in \mathcal{K}$ between the discretization points (here 8 *labels*), which leads to more precise (sublabel-accurate) solutions compared to the naive finite-difference approach.

k-cochains we arrive at the following finite-dimensional convex-concave saddle-point problem on $Q \subset \mathbf{R}^{n+N}$:

$$\begin{aligned} &\min_{T \in \mathcal{C}_n(Q)} \max_{\substack{\omega \in \mathcal{C}^n(Q) \\ \varphi \in \mathcal{C}^{n-1}(Q)}} \langle T, \omega \rangle_\phi + \langle \partial T - S, \varphi \rangle_\phi, \\ &\text{subject to} \quad \pi_{1\sharp} T = \mathbf{1}, \mathbb{W}\omega \in \mathcal{K}, \\ &\text{potentially} \quad \pi_{2\sharp} T = \mathbf{1} \text{ in case } n = N. \end{aligned} \tag{39}$$

$S \in \mathcal{C}_{n-1}(Q)$ is a given boundary datum, for free boundary conditions we replace the inner product $\langle S, \varphi \rangle$ with an indicator function $S : \mathcal{C}^{n-1} \to \overline{\mathbf{R}}$ forcing φ to be zero on the boundary. The pushforwards $\pi_{1\sharp}$, $\pi_{2\sharp}$ are linear constraints on the coefficients of the k-chain T.

6. Numerical Implementation

In practice we solve (39) with the first order primal-dual algorithm [8]. For the local constraints $\mathbb{W}\omega \in \mathcal{K}$ usually no closed form projection exists. In some situations ($N = 1$) they can be implemented exactly, see [41, 40]. In the general setting, we resort to implementing them at sampled points. To enforce the constraint $\mathbb{W}\omega \in \mathcal{K}$ at samples $Z = \{z_1, z_2, \dots\} \subset X \times Y$ we add another primal variable $\lambda : Z \to \mathbf{\Lambda}_n \mathbf{R}^{n+N}$ and the additional term $\sum_{z \in Z} \Psi^{**}(z, \lambda(z)) - \langle \lambda(z), (\mathbb{W}\omega)(z) \rangle$ to the saddle-point formulation (39).

Finally, one requires the proximal operator of the perspective function Ψ^{**}. These can be implemented using epigraphical projections as in [48]. For an overview over proximal operators of perspective functions we refer to [12].

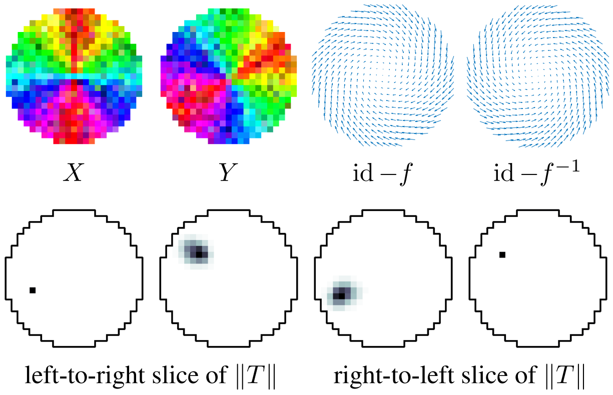


Figure 5: Global registration of X and Y. **Top:** Our method yields dense pointwise correspondences that are smooth in both directions and correspond to the correct transformation. **Bottom:** 2-D slices through the 4-D density $\|T\|$ at the single black pixel. We empirically verify (also at the other points) that the current concentrated near a surface, therefore the recovered solution is near the *global minimum* of the original nonconvex problem.

6.1. Properties of the Discretization

As a first example we solve the Brachistochrone [6], arguably the first variational approach. The cost is given by $c(x, y, \xi) = \sqrt{\frac{1+\xi^2}{2gy}}$ where $g > 0$ is the gravitational constant. Dirichlet boundary conditions are enforced using the boundary operator. In Fig. 3 we show the resulting current, which concentrates on the graph of the closed-form solution to the problem, which is a cycloid. The unlifted result is obtained by taking the center of mass of the first component $T^{\bar{0},0}$ of the current by summing over the horizontal edges in the 1-chain. The obtained result nearly coincides with the exact cycloid. Instead, solutions from MRF approaches would invariably be confined to the vertices or edges of the rather coarse grid.

In Fig. 4 we solve total variation regularized problems which corresponds to setting $c(x, y, \xi) = \rho(x, y) + |\xi|$ for some data ρ. The data is either a quadratic or a stereo matching cost in that example. The proposed approach based on discrete exterior calculus has better isotropy and leads to sharper discontinuities than the common forward difference approach used in literature. Furthermore, by enforcing the constraints $\mathbb{W}\omega \in \mathcal{K}$ also between the discretization points one can achieve "sublabel-accurate" results as demonstrated in the stereo matching example.

6.2. Global Registration

As an example of $n > 1$, $N > 1$ with polyconvex regularization, we tackle the problem of orientation preserving diffeomorphic registration between two shapes $X, Y \subset \mathbf{R}^2$ with boundary. We use the cost $c(x, y, \xi) = (\rho(x, y) + \varepsilon) \sqrt{\det(I + \xi^\top \xi)}$, which penalizes the surface area in the product space and favors local isometry. The parameter $\varepsilon > 0$ models the trade-off between data and smoothness. In the example considered in Fig. 5 the data is given by $\rho(x, y) = \|I_1(x) - I_2(y)\|$, where I_1, I_2 are the shown color images. A polyconvex extension of the above cost, which is large for non-simple vectors is given by the (weighted) mass norm (4). The 4-D cubical set Q is the product space between the two shapes X and Y, which are given as quads (pixels). We impose the constraints $\mathbb{W}\omega \in \mathcal{K}$ at the 16 vertices of each four dimensional hypercube. The proximal operator of the mass norm is computed as in [67]. Note that the required 4×4 real Schur decomposition can be reduced to a 2×2 SVD using a few Givens rotations, see [65]. We further impose $T^{\bar{0},0} \geq 0$ and $T^{0,\bar{0}} \geq 0$, and boundary conditions ensure that ∂X is matched to ∂Y. Bijectivity of the matching is encouraged by the pushforward constraints $\pi_{1\sharp} T = \mathbf{1}$, $\pi_{2\sharp} T = \mathbf{1}$. After solving (39) we obtain the final pointwise correspondences $f : X \to Y$ from the 2-chain $T \in \mathcal{C}_2(Q)$ by taking its center of mass.

In Fig. 5 we visualize $f(x) = \sum_y y \, |(\mathbb{W}T)(x, y)|$, $f^{-1}(y) = \sum_x x \, |(\mathbb{W}T)(x, y)|$. As one can see, the maps f and f^{-1} are smooth and inverse to each other. Despite $n > 1$, $N > 1$, the current apparently concentrated near a surface and the computed solutions are therefore near the *global optimum* of the original nonconvex problem.

7. Discussion and Limitations

In this work, we introduced a novel approach to vectorial variational problems based on geometric measure theory, along with a natural discretization using concepts from discrete exterior calculus. Though observed in practice, we do not have theoretical guarantees that the minimizing current will concentrate on a surface. In case of multiple global solutions, one might get a convex combination of minimizers. Some mechanism to select an extreme point of the convex solution set would therefore be desirable. The main drawback over MRFs, for which very efficient solvers exist [27], is that we had to resort to the generic algorithm [8] with $\mathcal{O}(1/\varepsilon)$ convergence. Yet, solutions with high numerical accuracy are typically not required in practice and the algorithm parallelizes well on GPUs. To conclude, we believe that the present work is a step towards making continuous approaches an attractive alternative to MRFs, especially in scenarios where faithfulness to certain geometric properties of the underlying continuous model is desirable.

Acknowledgements. The work was partially supported by the German Research Foundation (DFG); project 394737018 "*Functional Lifting 2.0 – Efficient Convexifications for Imaging and Vision*".

References

[1] G. Alberti, G. Bouchitté, and G. Dal Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Calc. Var. Partial Differential Equations*, 16(3):299–333, 2003. 2

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017. 1

[3] D. N. Arnold and G. Awanou. Finite element differential forms on cubical meshes. *Mathematics of Computation*, 83(288):1551–1570, 2014. 7

[4] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta numerica*, 15:1–155, 2006. 7

[5] P. Aviles and Y. Giga. Variational integrals on mappings of bounded variation and their lower semicontinuity. *Arch. Ration. Mech. Anal.*, 115(3):201–255, 1991. 2

[6] J. Bernoulli. Problema novum ad cujus solutionem mathematici invitantur. *Acta Eruditorum*, 18(269), 1696. 8

[7] H. Busemann, G. Ewald, and G. C. Shephard. Convex bodies and convexity on Grassmann cones. *Math. Ann.*, 151(1):1–41, 1963. 3

[8] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40:120–145, 2011. 7, 8

[9] Q. Chen and V. Koltun. Robust nonrigid registration by convex optimization. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[10] Q. Chen and V. Koltun. Full flow: Optical flow estimation by global optimization over regular grids. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[11] L. D. Cohen and R. Kimmel. Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision*, 24(1):57–78, 1997. 1

[12] P. L. Combettes and C. L. Müller. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *J. Math. Anal. Appl.*, 457(2):1283–1306, 2018. 7

[13] K. Crane. Discrete differential geometry: An applied introduction, 2019. 2

[14] C. Domokos, F. R. Schmidt, and D. Cremers. MRF optimization with separable convex prior on partially ordered labels. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[15] H. Federer. *Geometric Measure Theory*. Springer, 1969. 2

[16] H. Federer. Real flat chains, cochains and variational problems. *Indiana Univ. Math. J.*, 24(4):351–407, 1974. 2, 6

[17] A. Fix and S. Agarwal. Duality and the continuous graphical model. In *European Conference on Computer Vision (ECCV)*, 2014. 1

[18] M. Giaquinta, G. Modica, and J. Souček. *Cartesian currents in the calculus of variations I, II.*, volume 37-38 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3.* Springer, 1998. 4, 6

[19] B. Goldluecke, E. Strekalovskiy, and D. Cremers. Tight convex relaxations for vector-valued labeling. *SIAM J. Imaging Sciences*, 6(3):1626–1664, 2013. 2

[20] T. Goldstein, X. Bresson, and S. Osher. Global minimization of Markov random fields with applications to optical flow. *Inverse Problems & Imaging*, 6(4):623–644, 2012. 2

[21] D. Hilbert. Mathematische Probleme. *Nachrichten von der Königl. Gesellschaft der Wiss. zu Göttingen*, pages 253–297, 1900. 1

[22] A. N. Hirani. *Discrete exterior calculus*. PhD thesis, California Institute of Technology, 2003. 6

[23] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer, 2012. 3

[24] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1333–1336, 2003. 1

[25] T. Kaczynski, K. Mischaikow, and M. Mrozek. *Computational Homology*. Springer, 2006. 6

[26] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422, 1960. 1

[27] J. Kappes, B. Andres, F. Hamprecht, C. Schnorr, S. Nowozin, D. Batra, S. Kim, B. Kausler, J. Lellmann, N. Komodakis, et al. A comparative study of modern inference techniques for discrete energy minimization problems. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1335, 2013. 8

[28] R. Kimmel, R. Malladi, and N. Sochen. Images as embedded maps and minimal surfaces: Movies, color, texture, and volumetric medical images. *International Journal of Computer Vision 39(2)*, 2000. 1

[29] P. Kohli, A. Shekhovtsov, C. Rother, V. Kolmogorov, and P. Torr. On partial optimality in multi-label mrfs. In *International Conference on Machine learning (ICML)*, pages 480–487, 2008. 1

[30] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(10):1568–1583, 2006. 1

[31] V. Kolmogorov and C. Rother. Minimizing nonsubmodular functions with graph cuts - a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(7), 2007. 1

[32] S. G. Krantz and H. R. Parks. *Geometric Integration Theory*. Birkhäuser Boston, 2008. 2, 4

[33] E. Laude, T. Möllenhoff, M. Moeller, J. Lellmann, and D. Cremers. Sublabel-accurate convex relaxation of vectorial multilabel energies. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[34] J. Lellmann, E. Strekalovskiy, S. Koetter, and D. Cremers. Total variation regularization for functions with values in a manifold. In *International Conference on Computer Vision (ICCV)*, 2013. 2

[35] M. Li, A. Shekhovtsov, and D. Huber. Complexity of discrete energy minimization problems. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[36] B. Loewenhauser and J. Lellmann. Functional lifting for variational problems with higher-order regularization. *Imaging, Vision and Learning Based on Optimization and PDEs*, pages 101–120, 2018. 2

[37] F. Mémoli. On the use of Gromov-Hausdorff distances for shape comparison. In *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007. 2

[38] F. Mémoli. Spectral Gromov-Wasserstein distances for shape matching. In *International Conference on Computer Vision Workshops (ICCV Workshops)*, 2009. 2

[39] M. Menze, C. Heipke, and A. Geiger. Discrete optimization for optical flow. In *German Conference on Pattern Recognition (GCPR)*, 2015. 1

[40] T. Möllenhoff and D. Cremers. Sublabel-accurate discretization of nonconvex free-discontinuity problems. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 7

[41] T. Möllenhoff, E. Laude, M. Moeller, J. Lellmann, and D. Cremers. Sublabel-accurate relaxation of nonconvex energies. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 7

[42] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781. 1

[43] M. G. Mora. The calibration method for free-discontinuity problems on vector-valued maps. *J. Convex Anal.*, 9(1):1–29, 2002. 6

[44] F. Morgan. *Geometric Measure Theory: A Beginner's Guide*. Academic Press, 5th edition, 2016. 2

[45] J.-C. Nédélec. Mixed finite elements in $\mathbb{R}^3$. *Numerische Mathematik*, 35(3):315–341, 1980. 7

[46] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 31(4):30, 2012. 2

[47] G. Peyré and M. Cuturi. Computational optimal transport. *arXiv:1803.00567*, 2018. 2

[48] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM J. Imaging Sci.*, 3(4):1122–1145, 2010. 2, 7

[49] P.-A. Raviart and J.-M. Thomas. A mixed finite element method for 2nd order elliptic problems. In *Mathematical aspects of finite element methods*, pages 292–315. Springer, 1977. 7

[50] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996. 3

[51] E. Rodolà, Z. Lähner, A. M. Bronstein, M. M. Bronstein, and J. Solomon. Functional maps representation on product manifolds. In *Computer Graphics Forum*, volume 38, pages 678–689, 2019. 2

[52] F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, New York, 2015. 2

[53] T. Schoenemann and D. Cremers. A combinatorial solution for model-based image segmentation and real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(7):1153–1164, 2010. 1

[54] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[55] A. Shekhovtsov, I. Kovtun, and V. Hlaváč. Efficient MRF deformation model for non-rigid image matching. *Computer Vision and Image Understanding (CVIU)*, 112(1):91–99, 2008. 1

[56] J. Solomon, G. Peyré, V. G. Kim, and S. Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):72, 2016. 2

[57] M. Strecke and B. Goldluecke. Sublabel-accurate convex relaxation with total generalized variation regularization. In *German Conference on Pattern Recognition (GCPR)*, 2018. 2

[58] E. Strekalovskiy, A. Chambolle, and D. Cremers. A convex representation for the vectorial Mumford-Shah functional. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[59] E. Strekalovskiy, A. Chambolle, and D. Cremers. Convex relaxation of vectorial problems with coupled regularization. *SIAM J. Imaging Sci.*, 7(1):294–336, 2014. 2

[60] J. N. Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE Transactions on Automatic Control*, 40(9):1528–1538, 1995. 1

[61] M. Vestner, Z. Lähner, A. Boyarski, O. Litany, R. Slossberg, T. Remez, E. Rodola, A. Bronstein, M. Bronstein, R. Kimmel, and D. Cremers. Efficient deformable shape correspondence via kernel matching. In *International Conference on 3D Vision (3DV)*, 2017. 2

[62] M. Vestner, R. Litman, E. Rodolà, A. M. Bronstein, and D. Cremers. Product manifold filter: Non-rigid shape correspondence via kernel density estimation in the product space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6681–6690, 2017. 2

[63] C. Villani. *Optimal Transport: Old and New*. Springer, 2008. 2

[64] T. Vogt and J. Lellmann. Functional liftings of vectorial variational problems with Laplacian regularization. *arXiv:1904.00898*, 2019. 2

[65] R. C. Ward and L. J. Gray. Eigensystem computation for skew-symmetric matrices and a class of symmetric matrices. Technical report, Oak Ridge National Lab, 1976. 8

[66] H. Whitney. *Geometric Integration Theory*. Princeton University Press, 1957. 7

[67] T. Windheuser and D. Cremers. A convex solution to spatially-regularized correspondence problems. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 8

[68] T. Windheuser, U. Schlickewei, F. R. Schmidt, and D. Cremers. Geometrically consistent elastic matching of 3D shapes: A linear programming solution. In *International Conference on Computer Vision (ICCV)*, 2011. 2

[69] L. C. Young. *Lectures on the Calculus of Variations and Optimal Control Theory*. Chelsea Publishing Company, New York, second edition, 1980. 1

[70] C. Zach, C. Haene, and M. Pollefeys. What is optimized in tight convex relaxations for multi-label problems? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[71] C. Zach and P. Kohli. A convex discrete-continuous approach for Markov random fields. In *European Conference on Computer Vision (ECCV)*, 2014. 1

Flat Metric Minimization with Applications in Generative Modeling

Authors Thomas Möllenhoff[1] thomas.moellenhoff@tum.de
Daniel Cremers[1] cremers@tum.de

[1] Technische Universität München, Germany

Publication T. Möllenhoff, D. Cremers. Flat Metric Minimization with Applications in Generative Modeling. In: *Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97:4262-4635*, Long Beach, USA, 2019. `http://proceedings.mlr.press/v97/mollenhoff19a.html`, eprint: `https://arxiv.org/abs/1905.04730`. Reprinted with permission of the authors.

Contribution		
	Problem definition	*significantly contributed*
	Literature survey	*significantly contributed*
	Mathematical derivations	*significantly contributed*
	Numerical implementation	*significantly contributed*
	Experimental evaluation	*significantly contributed*
	Preparation of manuscript	*significantly contributed*

Notice Under permission of the authors, we include here the *accepted version* of the original publication [MC19b].

Flat Metric Minimization with Applications in Generative Modeling

Thomas Möllenhoff [1] **Daniel Cremers** [1]

Abstract

We take the novel perspective to view data not as a probability distribution but rather as a current. Primarily studied in the field of geometric measure theory, k-currents are continuous linear functionals acting on compactly supported smooth differential forms and can be understood as a generalized notion of oriented k-dimensional manifold. By moving from distributions (which are 0-currents) to k-currents, we can explicitly orient the data by attaching a k-dimensional tangent plane to each sample point. Based on the flat metric which is a fundamental distance between currents, we derive FlatGAN, a formulation in the spirit of generative adversarial networks but generalized to k-currents. In our theoretical contribution we prove that the flat metric between a parametrized current and a reference current is Lipschitz continuous in the parameters. In experiments, we show that the proposed shift to $k > 0$ leads to interpretable and disentangled latent representations which behave equivariantly to the specified oriented tangent planes.

1. Introduction

This work is concerned with the problem of representation learning, which has important consequences for many tasks in artificial intelligence, cf. the work of Bengio et al. (2013). More specifically, our aim is to learn representations which behave equivariantly with respect to selected transformations of the data. Such variations are often known beforehand and could for example describe changes in stroke width or rotation of a digit, changes in viewpoint or lighting in a three-dimensional scene but also the *arrow of time* (Pickup et al., 2014; Wei et al., 2018) in time-series, describing how a video changes from one frame to the next, see Fig. 1.

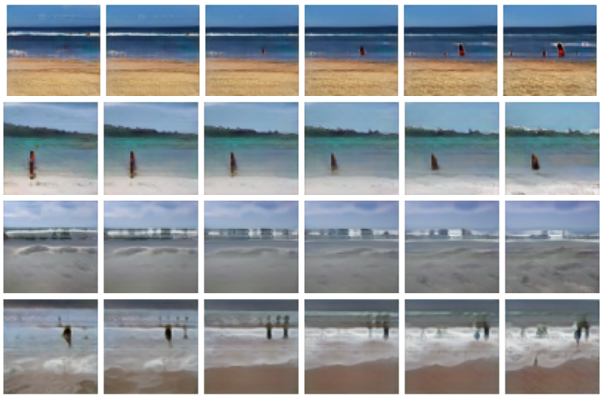


Figure 1. Discovering the arrow of time by training a generative model with the proposed formalism on the tinyvideos dataset (Vondrick et al., 2016). The approach we introduce allows one to learn latent representations which behave equivariantly to specified tangent vectors (here: difference of two successive video frames).

We tackle this problem by introducing a novel formalism based on *geometric measure theory* (Federer, 1969), which we find to be interesting in itself. To motivate our application in generative modeling, recall the manifold hypothesis which states that the distribution of real-world data tends to concentrate nearby a low-dimensional manifold, see Fefferman et al. (2016) and the references therein. Under that hypothesis, a possible unifying view on prominent methods in unsupervised and representation learning such as generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational auto-encoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014) is the following: both approaches aim to approximate the true distribution concentrating near the manifold with a distribution on some low-dimensional latent space $\mathcal{Z} \subset \mathbf{R}^l$ that is pushed through a decoder or generator $g : \mathcal{Z} \to \mathcal{X}$ mapping to the (high-dimensional) data space $\mathcal{X} \subset \mathbf{R}^d$ (Genevay et al., 2017; Bottou et al., 2017).

We argue that treating data as a distribution potentially ignores useful available geometric information such as orientation and tangent vectors to the data manifold. Such tangent vectors describe the aforementioned local variations or pertubations. Therefore we postulate that *data should not be viewed as a distribution but rather as a k-current*.

[1] Department of Informatics, Technical University of Munich, Garching, Germany. Correspondence to: Thomas Möllenhoff <thomas.moellenhoff@tum.de>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

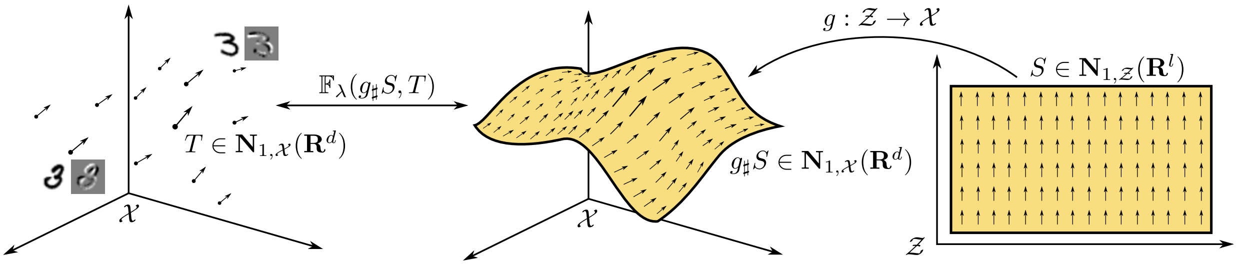

$g : \mathcal{Z} \to \mathcal{X}$
$\mathbb{F}_\lambda(g_\sharp S, T)$
$T \in \mathbf{N}_{1,\mathcal{X}}(\mathbf{R}^d)$
$g_\sharp S \in \mathbf{N}_{1,\mathcal{X}}(\mathbf{R}^d)$
$S \in \mathbf{N}_{1,\mathcal{Z}}(\mathbf{R}^l)$
$\mathcal{X}$
$\mathcal{X}$
$\mathcal{Z}$

Figure 2. **Illustration of the proposed idea.** We suggest the novel perspective to view observed data (here the MNIST dataset) as a k-current T, shown as the dots with attached arrows on the left. The arrows indicate the oriented tangent space, and we selected $k = 1$ to be rotational deformation. We propose to minimize the flat distance of T to the *pushforward* $g_\sharp S$ (shown in the middle) of a current S on a low-dimensional latent space $\mathcal{Z}$ (right) with respect to a "generator" map $g : \mathcal{Z} \to \mathcal{X}$. For 0-currents (no selected tangent vectors) and sufficiently large λ, the proposed "FlatGAN" formulation specializes to the Wasserstein GAN (Arjovsky et al., 2017).

We postpone the definition of k-currents (de Rham, 1955) to Sec. 3, and informally think of them as distributions over k-dimensional oriented planes. For the limiting case $k = 0$, currents simply reduce to distributions in the sense of Schwartz (1951, 1957) and positive 0-currents with unit mass are probability measures. A seminal work in the theory of currents was written by Federer & Fleming (1960), which established compactness theorems for subsets of currents (*normal* and *integral currents*). In this paper, we will work in the space of normal k-currents with compact support in $\mathcal{X} \subset \mathbf{R}^d$, denoted by $\mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$.

Similarly as probabilistic models build upon f-divergences (Csiszár et al., 2004), integral probability metrics (Sriperumbudur et al., 2012) or more general optimal transportation related divergences (Peyré & Cuturi, 2018; Feydy et al., 2018), we require a sensible notion to measure "distance" between k-currents.

In this work, we will focus on the flat norm[1] due to Whitney (1957). To be precise, we consider a scaled variant introduced and studied by Morgan & Vixie (2007); Vixie et al. (2010). This choice is motivated in Sec. 4, where we show that the flat norm enjoys certain attractive properties similar to the celebrated Wasserstein distances. For example, it metrizes the weak*-convergence for normal currents.

A potential alternative to the flat norm are kernel metrics on spaces of currents (Vaillant & Glaunès, 2005; Glaunès et al., 2008). These have been proposed for diffeomorphic registration, but kernel distances on distributions have also been sucessfully employed for generative modeling, see Li et al. (2017). Constructions similar to the Kantorovich relaxation in optimal transport but generalized to k-currents recently appeared in the context of convexifications for certain variational problems (Möllenhoff & Cremers, 2019).

2. Related Work

Our main idea is illustrated in Fig. 2, which was inspired from the optimal transportation point of view on GANs given by Genevay et al. (2017).

Tangent vectors of the data manifold, either prespecified (Simard et al., 1992; 1998; Fraser et al., 2003) or learned with a contractive autoencoder (Rifai et al., 2011), have been used to train classifiers that aim to be *invariant* to changes relative to the data manifold. In contrast to these works, we use tangent vectors to learn interpretable representations and a generative model that aims to be *equivariant*. The principled introduction of tangent k-vectors into probabilistic generative models is one of our main contributions.

Various approaches to learning informative or disentangled latent representations in a completely unsupervised fashion exist (Schmidhuber, 1992; Higgins et al., 2016; Chen et al., 2016; Kim & Mnih, 2018). Our approach is orthogonal to these works, as specifying tangent vectors further encourages informative representations to be learned. For example, our GAN formulation could be combined with a mutual information term as in InfoGAN (Chen et al., 2016).

Our work is more closely related to semi-supervised approaches on learning disentangled latent representations, which similarly also require some form of knowledge of the underlying factors (Hinton et al., 2011; Denton et al., 2017; Mathieu et al., 2016; Narayanaswamy et al., 2017) and also to conditional GANs (Mirza & Osindero, 2014; Odena et al., 2017). However, the difference is the connection to geometric measure theory which we believe to be completely novel, and our specific FlatGAN formulation that seamlessly extends the Wasserstein GAN (Arjovsky et al., 2017), cf. Fig. 2.

Since the concepts we need from geometric measure theory are not commonly used in machine learning, we briefly review them in the following section.

[1] The terminology "flat" carries no geometrical significance and refers to Whitney's use of musical notation flat $|\cdot|^\flat$ and sharp $|\cdot|^\sharp$.

3. Geometric Measure Theory

The book by Federer (1969) is still the formidable, definitive reference on the subject. As a more accessible introduction we recommend (Krantz & Parks, 2008) or (Morgan, 2016). While our aim is to keep the manuscript self-contained, we invite the interested reader to consult Chapter 4 in (Morgan, 2016), which in turn refers to the corresponding chapters in the book of Federer (1969) for more details.

3.1. Grassmann Algebra

Notation. Denote $\{e_1, \ldots, e_d\}$ a basis of $\mathbf{R}^d$ with dual basis $\{\mathrm{d}x_1, \ldots, \mathrm{d}x_d\}$ such that $\mathrm{d}x_i : \mathbf{R}^d \to \mathbf{R}$ is the linear functional that maps every $x = (x_1, \ldots, x_d)$ to the i-th component x_i. For $k \leq d$, denote $I(d,k)$ as the ordered multi-indices $\mathbf{i} = (i_1, \ldots, i_k)$ with $1 \leq i_1 < \ldots < i_k \leq d$.

One can multiply vectors in $\mathbf{R}^d$ to obtain a new object:

$$\xi = v_1 \wedge \ldots \wedge v_k, \tag{1}$$

called a k-vector ξ in $\mathbf{R}^d$. The wedge (or exterior) product $\wedge$ is characterized by multilinearity

$$\begin{aligned} &cv_1 \wedge v_2 = v_1 \wedge cv_2 = c(v_1 \wedge v_2), \quad \text{for } c \in \mathbf{R}, \\ &(u_1 + v_1) \wedge (u_2 + v_2) = \\ &\qquad u_1 \wedge u_2 + u_1 \wedge v_2 + v_1 \wedge u_2 + v_1 \wedge v_2, \end{aligned} \tag{2}$$

and it is alternating

$$u \wedge v = -v \wedge u, \quad u \wedge u = 0. \tag{3}$$

In general, any k-vector can be written as

$$\xi = \sum_{\mathbf{i} \in I(d,k)} a_{\mathbf{i}} \cdot e_{i_1} \wedge \ldots \wedge e_{i_k} = \sum_{\mathbf{i} \in I(d,k)} a_{\mathbf{i}} \cdot e_{\mathbf{i}}, \tag{4}$$

for coefficients $a_{\mathbf{i}} \in \mathbf{R}$. The vector space of k-vectors is denoted by $\mathbf{\Lambda}_k \mathbf{R}^d$ and has dimension $\binom{d}{k}$. We define for two k-vectors $v = \sum_{\mathbf{i}} a_{\mathbf{i}} e_{\mathbf{i}}$, $w = \sum_{\mathbf{i}} b_{\mathbf{i}} e_{\mathbf{i}}$ an inner product $\langle v, w \rangle = \sum_{\mathbf{i}} a_{\mathbf{i}} b_{\mathbf{i}}$ and the Euclidean norm $|v| = \sqrt{\langle v, v \rangle}$.

A simple (or decomposable) k-vector is any $\xi \in \mathbf{\Lambda}_k \mathbf{R}^d$ that can be written using products of 1-vectors. Simple k-vectors such as (1) are uniquely determined by the k-dimensional space spanned by the $\{v_i\}$, their orientation and the norm $|v|$ corresponding to the area of the parallelotope spanned by the $\{v_i\}$. Simple k-vectors with unit norm can therefore be thought of as oriented k-dimensional subspaces and the rules (2)-(3) can be thought of as equivalence relations.

It turns out that the inner product of two simple k-vectors can be computed by the $k \times k$-determinant

$$\langle w_1 \wedge \ldots \wedge w_k, v_1 \wedge \ldots \wedge v_k \rangle = \det \left(W^\top V \right), \tag{5}$$

where the columns of $W \in \mathbf{R}^{d \times k}$, $V \in \mathbf{R}^{d \times k}$ contain the individual 1-vectors. This will be useful later for our practical implementation.

Not all k-vectors are simple. An illustrative example is $e_1 \wedge e_2 + e_3 \wedge e_4 \in \mathbf{\Lambda}_2 \mathbf{R}^4$, which describes two 2-dimensional subspaces in $\mathbf{R}^4$ intersecting only at zero.

The dual space of $\mathbf{\Lambda}_k \mathbf{R}^d$ is denoted as $\mathbf{\Lambda}^k \mathbf{R}^d$, and its elements are called k-covectors. They are similarly represented as (4) but with dual basis $\mathrm{d}x_{\mathbf{i}}$. Analogously to the previous page, we can define an inner product between k-vectors and k-covectors. Next to the Euclidean norm $|\cdot|$, we define two additional norms due to Whitney (1957).

Definition 1 (Mass and comass)**.** *The comass norm defined for k-covectors $w \in \mathbf{\Lambda}^k \mathbf{R}^n$ is given by*

$$\|w\|^* = \sup \left\{ \langle w, v \rangle : v \textit{ is simple }, |v| = 1 \right\}, \tag{6}$$

and the mass norm for $v \in \mathbf{\Lambda}_k \mathbf{R}^n$ is given by

$$\begin{aligned} \|v\| &= \sup \left\{ \langle v, w \rangle : \|w\|^* \leq 1 \right\} \\ &= \inf \left\{ \sum_i |\xi_i| : \xi_i \textit{ are simple}, v = \sum_i \xi_i \right\}. \end{aligned} \tag{7}$$

The mass norm is by construction the largest norm that agrees with the Euclidean norm on simple k-vectors. For the non-simple 2-vector from before, we compute

$$\|e_1 \wedge e_2 + e_3 \wedge e_4\| = 2, \ |e_1 \wedge e_2 + e_3 \wedge e_4| = \sqrt{2}. \tag{8}$$

Interpreting the non-simple vector as two tangent planes, we see that the mass norm gives the correct area, while the Euclidean norm underestimates it. The comass $\|\cdot\|^*$ will be used later to define the mass of currents and the flat norm.

3.2. Differential Forms

In order to define currents, we first need to introduce differential forms. A differential k-form is a k-covectorfield $\omega : \mathbf{R}^d \to \mathbf{\Lambda}^k \mathbf{R}^d$. The support $\operatorname{spt} \omega$ is defined as the closure of the set $\{x \in \mathbf{R}^d : \omega(x) \neq 0\}$.

Differential forms allow one to perform coordinate-free integration over oriented manifolds. Given some manifold $\mathcal{M} \subset \mathbf{R}^d$, possibly with boundary, an *orientation* is a continuous map $\tau_{\mathcal{M}} : \mathcal{M} \to \mathbf{\Lambda}_k \mathbf{R}^d$ which assigns to each point a simple k-vector with unit norm that spans the tangent space at that point. Integration of a differential form over an oriented manifold $\mathcal{M}$ is then defined by:

$$\int_{\mathcal{M}} \omega = \int_{\mathcal{M}} \langle \omega(x), \tau_{\mathcal{M}}(x) \rangle \, \mathrm{d}\mathcal{H}^k(x), \tag{9}$$

where the second integral is the standard Lebesgue integral with respect to the k-dimensional Hausdorff measure $\mathcal{H}^k$ restricted to $\mathcal{M}$, i.e., $(\mathcal{H}^k \llcorner \mathcal{M})(A) = \mathcal{H}^k(A \cap \mathcal{M})$. The k-dimensional Hausdorff measure assigns to sets in $\mathbf{R}^d$ their k-dimensional volume, see Chapter 2 in Morgan (2016) for a nice illustration. For $k = d$ the Hausdorff measure coincides with the Lebesgue measure.

The exterior derivative of a differential k-form is the $(k+1)$-form $d\omega : \mathbf{R}^d \to \mathbf{\Lambda}^{k+1}\mathbf{R}^d$ defined by

$$\langle d\omega(x), v_1 \wedge \ldots \wedge v_{k+1}\rangle = \lim_{h \to 0} \frac{1}{h^{k+1}} \int_{\partial P} \omega, \qquad (10)$$

where ∂P is the oriented boundary of the parallelotope spanned by the $\{hv_i\}$ at point x. The above definition is for example used in the textbook of Hubbard & Hubbard (2015). To get an intuition, note that for $k = 0$ this reduces to the familiar directional derivative $\langle d\omega(x), v_1\rangle = \lim_{h\to 0} \frac{1}{h}\left(\omega(x + hv_1) - \omega(x)\right)$. In case $\omega : \mathbf{R}^d \to \mathbf{\Lambda}^k\mathbf{R}^d$ is sufficiently smooth, the limit in (10) is given by

$$\langle d\omega(x), v_1 \wedge \ldots \wedge v_{k+1}\rangle = \qquad (11)$$
$$\sum_{i=1}^{k+1} (-1)^{i-1} \nabla_x \langle \omega(x), v_1 \wedge \ldots \wedge \hat{v}_i \wedge \ldots \wedge v_k\rangle \cdot v_i,$$

where $\hat{v}_i$ means that the vector v_i is omitted. The formulation (11) will be used in the practical implementation. Interestingly, with (9) and (10) in mind, Stokes' theorem

$$\int_{\mathcal{M}} d\omega = \int_{\partial \mathcal{M}} \omega, \qquad (12)$$

becomes almost obvious, as (informally speaking) integrating (10) one obtains (12) since the oppositely oriented boundaries of neighbouring parallelotopes cancel each other out in the interior of $\mathcal{M}$.

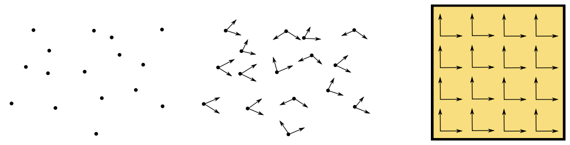
To define the pushforward of currents which is central to our formulation, we require the pullback of differential forms. The pullback $g^\sharp\omega : \mathbf{R}^l \to \mathbf{\Lambda}^k\mathbf{R}^l$ by a map $g : \mathbf{R}^l \to \mathbf{R}^d$ of the k-form $\omega : \mathbf{R}^d \to \mathbf{\Lambda}^k\mathbf{R}^d$ is given by

$$\langle g^\sharp\omega, v_1 \wedge .. \wedge v_k\rangle = \langle \omega \circ g, D_{v_1}g \wedge .. \wedge D_{v_k}g\rangle, \qquad (13)$$

where $D_{v_i}g := \nabla g \cdot v_i$ and $\nabla g \in \mathbf{R}^{d\times l}$ is the Jacobian. We will also require (13) for the practical implementation.

3.3. Currents

We have now the necessary tools to define currents and the required operations on them, which will be defined through duality with differential forms. Consider the space of compactly supported and smooth k-forms in $\mathbf{R}^d$ which we denote by $\mathcal{D}^k(\mathbf{R}^d)$. When furnished with an appropriate topology (cf. §4.1 in Federer (1969) for the details) this is a locally convex topological vector space. k-currents are continuous linear functionals on smooth, compactly supported differential forms, i.e., elements from the topological dual space $\mathcal{D}_k(\mathbf{R}^d) = \mathcal{D}^k(\mathbf{R}^d)'$. Some examples for currents are given in Fig. 3. The 0-current in **(a)** could be an empirical data distribution, and the 2-current in **(b)** represents the data distribution with a two dimensional oriented tangent space at each data point. The 2-current in **(c)** simply represents the set $[0,1]^2$ as an oriented manifold, its action on a differential form is given as in (9).



(a) $\sum_i \delta_{x_i}$ **(b)** $\sum_i \delta_{x_i} \wedge T_i$ **(c)** $\mathcal{H}^2 \llcorner [0,1]^2 \wedge e_{12}$

Figure 3. Example of a 0-current **(a)**, and 2-currents **(b)**, **(c)**.

A natural notion of convergence for currents is given by the weak* topology:

$$T_i \overset{*}{\rightharpoonup} T \text{ iff } T_i(\omega) \to T(\omega), \text{ for all } \omega \in \mathcal{D}^k(\mathbf{R}^d). \qquad (14)$$

The support of a current $T \in \mathcal{D}_k(\mathbf{R}^d)$, $\operatorname{spt} T$, is the complement of the largest open set, so that when testing T with compactly supported forms on that open set the answer is zero. Currents with compact support are denoted by $\mathcal{E}_k(\mathbf{R}^d)$. The boundary operator $\partial : \mathcal{D}_k(\mathbf{R}^d) \to \mathcal{D}_{k-1}(\mathbf{R}^d)$ is defined using exterior derivative

$$\partial T(\omega) = T(d\omega), \qquad (15)$$

and Stokes' theorem (12) ensures that this coincides with the intuitive notion of boundary for currents which are represented by integration over manifolds in the sense of (9).

The pushforward of a current is defined using the pullback

$$g_\sharp T(\omega) = T(g^\sharp\omega), \qquad (16)$$

where the intuition is that the pushforward transforms the current with the map g, see the illustration in Fig. 2.

The mass of a current $T \in \mathcal{D}_k(\mathbf{R}^d)$ is given by

$$\mathbb{M}(T) = \sup\left\{T(\omega) : \|\omega(x)\|^* \leq 1\right\}. \qquad (17)$$

If the current T is an oriented manifold then the mass $\mathbb{M}(T)$ is the *volume* of that manifold. One convenient way to construct k-currents, is by combining a smooth k-vectorfield $\xi : \mathbf{R}^d \to \mathbf{\Lambda}_k\mathbf{R}^d$ with a Radon measure μ:

$$(\mu \wedge \xi)(\psi) = \int \langle \xi, \psi\rangle \, \mathrm{d}\mu, \text{ for all } \psi \in \mathcal{D}^k(\mathbf{R}^d). \qquad (18)$$

A concrete example is illustrated in Fig. 3 **(b)**, where given samples $\{x_1, \ldots, x_N\} \subset \mathbf{R}^d$ and tangent 2-vectors $\{T_1, \ldots, T_N\} \subset \mathbf{\Lambda}_2\mathbf{R}^d$ a 2-current is constructed.

For currents with finite mass there is a measure $\|T\|$ and a map $\vec{T} : \mathbf{R}^d \to \mathbf{\Lambda}_k\mathbf{R}^d$ with $\|\vec{T}(\cdot)\| = 1$ almost everywhere so that we can represent it by integration as follows:

$$T(\omega) = \int \langle \omega(x), \vec{T}(x)\rangle \, \mathrm{d}\|T\|(x) = \|T\| \wedge \vec{T}(\omega). \qquad (19)$$

Another perspective is that finite mass currents are simply k-vector valued Radon measures. Currents with finite mass and finite boundary mass are called *normal currents* (Federer & Fleming, 1960). The space of normal currents with support in a compact set $\mathcal{X}$ is denoted by $\mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$.

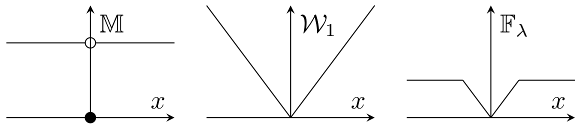


Figure 4. Illustration of distances between 0-currents on the example of two Dirac measures δ_x, δ_0. The flat metric $\mathbb{F}_\lambda$ has the following advantages: unlike the mass $\mathbb{M}$ it is continuous, and unlike Wasserstein-1 it easily generalizes to k-currents (see Fig. 5).

4. The Flat Metric

As indicated in Fig. 2, we wish to fit a current $g_\sharp S$ that is the pushforward of a low-dimensional latent current S to the current T given by the data. A more meaningful norm on currents than the mass $\mathbb{M}$ turns out to be the flat norm.

Definition 2 (Flat norm and flat metric)**.** *The flat norm with scale*[2] $\lambda > 0$ *is defined for any k-current $T \in \mathcal{D}_k(\mathbf{R}^d)$ as*

$$\mathbb{F}_\lambda(T) = \sup\{T(\omega) \mid \omega \in \mathcal{D}^k(\mathbf{R}^d), \textit{ with } \|\omega(x)\|^* \leq \lambda, \|d\omega(x)\|^* \leq 1, \textit{ for all } x\}. \tag{20}$$

For $\lambda = 1$ we simply write $\mathbb{F}(\cdot) \equiv \mathbb{F}_1(\cdot)$ and $\mathbb{F}_\lambda(S,T) = \mathbb{F}_\lambda(S-T)$ will be denoted as the flat metric.

The flat norm also has a primal formulation

$$\mathbb{F}_\lambda(T) = \min_{B \in \mathcal{E}_{k+1}(\mathbf{R}^d)} \lambda\mathbb{M}(T - \partial B) + \mathbb{M}(B) \tag{21}$$

$$= \min_{T = A + \partial B} \lambda\mathbb{M}(A) + \mathbb{M}(B), \tag{22}$$

where the minimum in (21)–(22) can be shown to exist, see §4.1.12 in Federer (1969). The flat norm is finite if T is a normal current and it can be verified that it is indeed a norm.

To get an intuition, we compare the flat norm to the mass (17) and the Wasserstein-1 distance in Fig. 4 on the example of Dirac measures δ_x, δ_0. The mass $x \mapsto \mathbb{M}(\delta_x - \delta_0)$ is discontinuous and has zero gradient and is therefore unsuitable as a distance between currents. While the Wasserstein-1 metric $x \mapsto \mathcal{W}(\delta_x, \delta_0)$ is continuous in x, it does not easily generalize from probability measures to k-currents. In contrast, the flat metric $x \mapsto \mathbb{F}_\lambda(\delta_x, \delta_0)$ has a meaningful geometric interpretation also for arbitrary k-currents. In Fig. 5 we illustrate the flat norm for two 1-currents. In that figure, if S and T are of length one and are ε apart, then $\mathbb{F}_\lambda(S,T) \leq (1+2\lambda)\varepsilon$ which converges to zero for $\varepsilon \to 0$.

Note that for 0-currents, the flat norm (20) is strongly related to the Wasserstein-1 distance except for the additional constraint on the dual variable $\|\omega(x)\|^* \leq \lambda$, which in the example of Fig. 4 controls the truncation cutoff. Notice also the similarity of (21) to the Beckmann formulation of the Wasserstein-1 distance (Beckmann, 1952; Santambrogio, 2015), with the difference being the implementation of the "divergence constraint" with a soft penalty $\lambda\mathbb{M}(T - \partial B)$. Considering the case $\lambda = \infty$ as in the Wasserstein distance is problematic in case we have $k > 0$, since not every current $T \in \mathcal{D}_k(\mathbf{R}^n)$ is the boundary of a $(k+1)$-current, see the example above in Fig. 5.

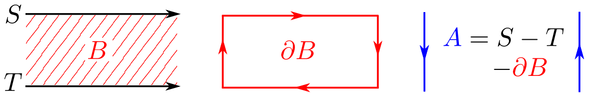


Figure 5. The flat metric $\mathbb{F}_\lambda(S,T)$ is given an optimal decomposition $S - T = A + \partial B$ into a k-current A and the boundary of a $(k+1)$-current B with minimal weighted mass $\lambda\mathbb{M}(A) + \mathbb{M}(B)$. An intuition is that $\lambda\mathbb{M}(A)$ is a penalty that controls how closely ∂B should approximate $S - T$, while $\mathbb{M}(B)$ is the $(k+1)$-dimensional volume of B.

The following proposition studies the effect of the scale parameter $\lambda > 0$ on the flat norm.

Proposition 1. *For any $\lambda > 0$, the following relation holds*

$$\min\{1,\lambda\} \cdot \mathbb{F}(T) \leq \mathbb{F}_\lambda(T) \leq \max\{1,\lambda\} \cdot \mathbb{F}(T), \tag{23}$$

meaning that $\mathbb{F}$ and $\mathbb{F}_\lambda$ are equivalent norms.

Proof. By a result of Morgan & Vixie (2007) we have the interesting relation

$$\mathbb{F}_\lambda(T) = \lambda^k\, \mathbb{F}(d_{\lambda^{-1}\sharp}T), \tag{24}$$

where d_λ is the λ-dilation. Using the bound $\mathbb{F}(f_\sharp T) \leq \sup\{\mathrm{Lip}(f)^k, \mathrm{Lip}(f)^{k+1}\}\mathbb{F}(T)$, §4.1.14 in Federer (1969), and the fact that $\mathrm{Lip}(d_{\lambda^{-1}}) = \lambda^{-1}$, one inequality directly follows. For the other side, notice that

$$\begin{aligned} \mathbb{F}(T) &= \mathbb{F}(d_{\lambda\sharp}d_{\lambda^{-1}\sharp}T) = \mathbb{F}_{\lambda^{-1}}(d_{\lambda^{-1}\sharp}T)\lambda^k \\ &\leq \max\{1,\lambda^{-1}\}\mathbb{F}(d_{\lambda^{-1}\sharp}T)\lambda^k \\ &= \max\{1,\lambda^{-1}\}\mathbb{F}_\lambda(T). \end{aligned} \tag{25}$$

and dividing by $\max\{1,\lambda^{-1}\}$ yields the result. □

The importance of the flat norm is due to the fact that it metrizes the weak*-convergence (14) on compactly supported normal currents with uniformly bounded mass and boundary mass.

Proposition 2. *Let $\mathcal{X} \subset \mathbf{R}^d$ be a compact set and $c > 0$ some fixed constant. For a sequence $\{T_j\} \subset \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ with $\mathbb{M}(T_j) + \mathbb{M}(\partial T_j) < c$ we have that:*

$$\mathbb{F}_\lambda(T, T_j) \to 0 \quad \textit{if and only if} \quad T_j \overset{*}{\rightharpoonup} T. \tag{26}$$

Proof. Due to Prop. 1 it is enough to consider the case $\lambda = 1$, which is given by Corollary 7.3 in the paper of Federer & Fleming (1960). □

[2] We picked a different convention for λ as in (Morgan & Vixie, 2007), where it bounds the other constraint, to emphasize the connection to the Wasserstein-1 distance.

5. Flat Metric Minimization

Motivated by the theoretical properties of the flat metric shown in the previous section, we consider the following optimization problem:

$$\min_{\theta \in \Theta} \mathbb{F}_\lambda(g_{\theta\sharp} S, T), \tag{27}$$

where $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ and $T \in \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$. We will assume that $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ is parametrized with parameters in a compact set $\Theta \subset \mathbf{R}^n$ and write $g_\theta : \mathcal{Z} \to \mathcal{X}$ to abbreviate $g(\cdot, \theta)$ for some $\theta \in \Theta$. We need the following assumption to be able to prove the existence of minimizers for the problem (27).

Assumption 1. *The map $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ is smooth in z with uniformly bounded derivative. Furthermore, we assume that $g(z, \cdot)$ is locally Lipschitz continuous and that the parameter set $\Theta \subset \mathbf{R}^n$ is compact.*

Under this assumption, we will show that the objective in (27) is Lipschitz continuous. This will in turn guarantee existence of minimizers, as the domain is assumed to be compact.

Proposition 3. *Let $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$, $T \in \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ be normal currents with compact support. If the pushforward map $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ fulfills Assumption 1, then the function $\theta \mapsto \mathbb{F}_\lambda(g_{\theta\sharp} S, T)$ is Lipschitz continuous and hence differentiable almost everywhere.*

Proof. In Appendix A. □

5.1. Application to Generative Modeling

We now turn towards our considered application illustrated in Fig. 2. There, we denote by $k \geq 0$ the number of tangent vectors we specify at each sample point. The latent current $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ is constructed by combining a probability distribution $\mu \in \mathbf{N}_{0,\mathcal{Z}}(\mathbf{R}^l)$, which could for example be the uniform distribution, with the unit k-vectorfield as follows:

$$S = \mu \wedge (e_1 \wedge \ldots \wedge e_k). \tag{28}$$

For an illustration, see the right side of Fig. 2 and Fig. 3. The data current $T \in \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ is constructed from the samples $\{x_i\}_{i=1}^N$ and tangent vectorfields $T_i : \mathcal{X} \to \mathbf{\Lambda}_k \mathbf{R}^d$.

$$T = \frac{1}{N} \sum_{i=1}^N \delta_{x_i} \wedge T_i, \tag{29}$$

The tangent k-vectorfields $T_i(x) = T_{i,1} \wedge \ldots \wedge T_{i,k}$ are given by individual tangent vectors to the data manifold $T_{i,j} \in \mathbf{R}^d$. For an illustration, see the left side of Fig. 2 or Fig. 3. After solving (27), the map $g_\theta : \mathcal{Z} \to \mathcal{X}$ will be our generative model, where changes in the latent space $\mathcal{Z}$ along the unit directions $e_1, \ldots, e_k$ are expected to behave equivariantly to the specified tangent directions $T_{i,1}, \ldots, T_{i,k}$ near $g(z)$.

5.2. FlatGAN Formulation

To get a primal-dual formulation (or two player zero-sum game) in the spirit of GANs, we insert the definition of the flat norm (20) into the primal problem (27):

$$\min_{\theta \in \Theta} \sup_{\substack{\omega \in \mathcal{D}^k(\mathbf{R}^d) \\ \|\omega\|^* \leq \lambda, \|d\omega\|^* \leq 1}} S(g_\theta{}^\sharp \omega) - T(\omega), \tag{30}$$

where $\theta \in \Theta$ are for example the parameters of a neural network. In the above equation, we also used the definition of pushforward (16). Notice that for $k = 0$ the exterior derivative in (30) specializes to the gradient. This yields a Lipschitz constraint, and as for sufficiently large λ the other constraint becomes irrelevant, the problem (30) is closely related to the Wasserstein GAN (Bottou et al., 2017). The novelty in this work is the generalization to $k > 0$.

Combining (28) and (29) into (30) we arrive at the objective

$$\begin{aligned} E(\theta, \omega) = &-\frac{1}{N} \sum_{i=1}^N \langle \omega(x_i), T_i \rangle \\ &+ \int \langle \omega \circ g_\theta, (\nabla_z g_\theta \cdot e_1) \wedge \ldots \wedge (\nabla_z g_\theta \cdot e_k) \rangle \, \mathrm{d}\mu. \end{aligned} \tag{31}$$

Interestingly, due to the pullback, the discriminator ω inspects not only the output of the generator, but also parts of its Jacobian matrix. As a remark, relations between the generator Jacobian and GAN performance have recently been studied by Odena et al. (2018).

The constraints in (30) are implemented using penalty terms. First notice that due to the definition of the comass norm (6), the first constraint is equivalent to imposing $|\langle \omega(x), v \rangle| \leq \lambda$ for all simple k-covectors with $|v| = 1$. We implement this with the a penalty term with parameter $\rho > 0$ as follows:

$$\rho \cdot \int_{\mathcal{X}} \int \max\{0, |\langle \omega(x), v \rangle| - \lambda\}^2 \, \mathrm{d}\gamma_{k,d}(v) \, \mathrm{d}x, \tag{32}$$

where $\gamma_{k,d}$ denotes the Haar measure on the Grassmannian manifold $\mathbf{Gr}(d, k) \subset \mathbf{\Lambda}_k \mathbf{R}^d$ of k-dimensional subspaces in $\mathbf{R}^d$, see Chapter 3.2 in Krantz & Parks (2008). Similarly, the constraint on the exterior derivative is implemented by another penalty term as follows:

$$\rho \cdot \int_{\mathcal{X}} \int \max\{0, |\langle d\omega(x), v \rangle| - 1\}^2 \, \mathrm{d}\gamma_{k+1,d}(v) \, \mathrm{d}x. \tag{33}$$

5.3. Implementation with Deep Neural Networks

For high dimensional practical problems it is completely infeasible to directly work with $\mathbf{\Lambda}_k \mathbf{R}^d$ due to the curse of dimensionality. For example, already for the MNIST dataset augmented with two tangent vectors ($k = 2$, $d = 28^2$), we have that $\dim(\mathbf{\Lambda}_k \mathbf{R}^d) \approx 3 \cdot 10^5$.

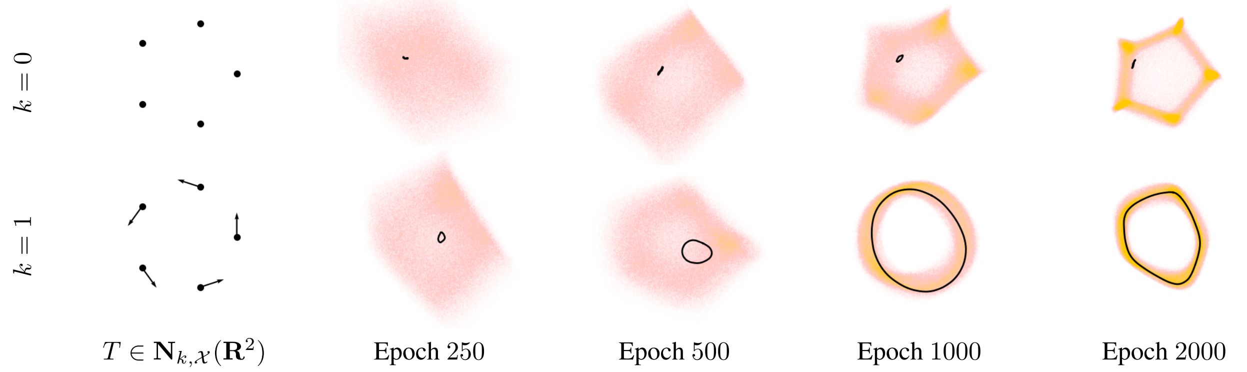


Figure 6. We illustrate the effect of moving from $k = 0$ to $k = 1$ and plot the measure $\|g_\sharp S\|$ of the pushforward of a k-current $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^5)$ (shown in orange) for different epochs. The black curve illustrates a walk along the first latent dimension z_1. For $k = 0$, which is similar to WGAN-GP (Gulrajani et al., 2017), the latent walk is not meaningful. The proposed approach ($k = 1$) allows to specify tangent vectors at the samples to which the first latent dimension behaves equivariantly, yielding an interpretable representation.

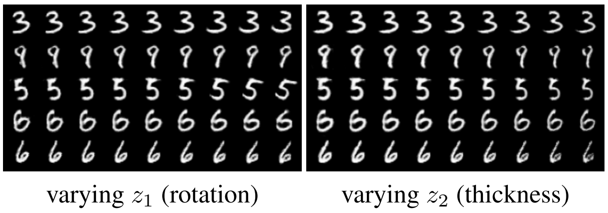


Figure 7. We show the effect of varying the first two components in 128-dimensional latent space, corresponding to the two selected tangent vectors which are rotation and thickness. As seen in the figure, varying the corresponding latent representation yields an interpretable effect on the output, corresponding to the specified tangent direction.

To overcome this issue, we unfortunately have to resort to a few heuristic approximations. To that end, we first notice that in the formulations the dual variable $\omega : \mathbf{R}^d \to \mathbf{\Lambda}^k\mathbf{R}^d$ only appears as an inner product with simple k-vectors, so we can implement it by implicitly describing its action, i.e., interpret it as a map $\omega : \mathbf{R}^d \times \mathbf{\Lambda}_k\mathbf{R}^d \to \mathbf{R}$:

$$\begin{aligned}&\omega(x, v_1 \wedge \ldots \wedge v_k) \\ &\quad = \omega^0(x) + \alpha\langle\omega^{1,1}(x) \wedge \ldots \wedge \omega^{1,k}(x), v_1 \wedge \ldots \wedge v_k\rangle,\end{aligned} \tag{34}$$

Theoretically, the "affine term" $\omega^0(x)$ is not fully justified as the map does not describe an inner product on $\mathbf{\Lambda}_k\mathbf{R}^d$ anymore, but we found it to improve the quality of the generative model. An attempt to justify this in the context of GANs is that the function $\omega^0 : \mathbf{R}^d \to \mathbf{R}$ is the usual "discriminator" while the $\omega^{1,i} : \mathbf{R}^d \to \mathbf{R}^d$ are combined to discriminate oriented tangent planes.

In practice, we parametrize ω^0, $\omega^{1,i}$ using deep neural networks. For efficiency reasons, the networks share their parameters up until the last few layers.

The inner product in (34) between the simple vectors is implemented by a $k \times k$-determinant, see (5). The reason we do this is to satisfy the properties of the Grassmann algebra (2) – (3). This is important, since otherwise the "discriminator" ω could distinguish between different representations of the same oriented tangent plane.

For the implementation of the penalty term (33), we use the definition of the exterior derivative (11) together with the "approximate form" (34). To be compatible with the affine term we use a seperate penalty on ω^0, which we also found to give better results:

$$\begin{aligned}&|d\omega(x, v_1 \wedge \ldots \wedge v_{k+1})| \approx (k+1)\|\nabla_x\omega^0(x)\| \\ &\quad + \alpha\left|\sum_{i=1}^{k+1}(-1)^{i-1}\nabla_x \det(W(x)^\top V_i) \cdot v_i\right|.\end{aligned} \tag{35}$$

In the above equation, $V_i \in \mathbf{R}^{d\times k}$ is the matrix with columns given by the vectors $v_1, \ldots, v_{k+1}$ but with v_i omitted and $W(x) \in \mathbf{R}^{d\times k}$ is the matrix with columns given by the $\omega^{1,i}(x)$. Another motivation for this implementation is, that in the case $k = 0$ the second term in (35) disappears and one recovers the well-known "gradient penalty" regularizer proposed by Gulrajani et al. (2017).

For the stochastic approximation of the penalty terms (32) – (33) we sample from the Haar measure on the Grassmannian (i.e., taking random k-dimensional and $(k+1)$-dimensional subspaces in $\mathbf{R}^d$) by computing singular value decomposition of random $k \times d$ Gaussian matrices. Furthermore, we found it beneficial in practice to enforce the penalty terms only at the data points as for example advocated in the recent work (Mescheder et al., 2018). The right multiplied Jacobian vector products (also referred to as "`rop`" in some frameworks) in (35) as well as in the loss function (31) are implemented using two additional backpropagations.

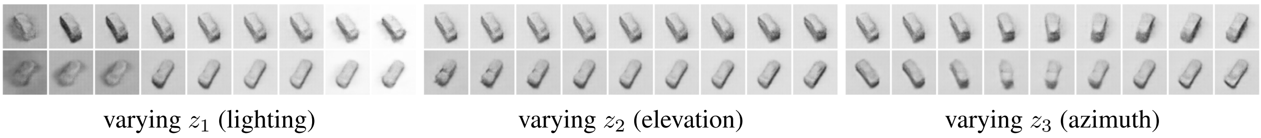


Figure 8. From left to right we vary the latent codes in $[-1, 1]$ after training on the smallNORB dataset (LeCun et al., 2004).

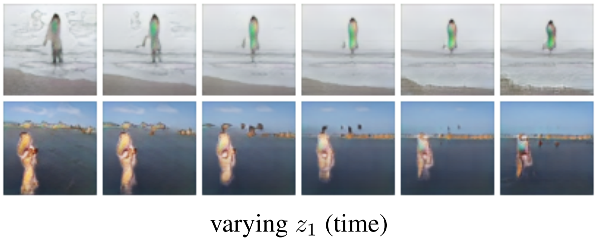


Figure 9. Varying the learned latent representation of time. The model captures behaviours such as people walking on the beach, see also the results shown in Fig. 1.

6. Experiments

The specific hyperparameters, architectures and tangent vector setups used in practice[3] are detailed in Appendix B.

6.1. Illustrative 2D Example

As a first proof of concept, we illustrate the effect of moving from $k = 0$ to $k = 1$ on a very simple dataset consisting of five points on a circle. As shown in Fig. 6, for $k = 0$ (corresponding to a WGAN-GP formulation) varying the first latent variable has no clear meaning. In contrast, with the proposed FlatGAN formulation, we can specify vectors tangent to the circle from which the data is sampled. This yields an interpretable latent representation that corresponds to an angular movement along the circle. As the number of epochs is increasing, both formulations tend to concentrate most of the probability mass on the five data points. However, since $g_\theta : \mathcal{Z} \to \mathcal{X}$ is continuous by construction an interpretable path remains.

6.2. Equivariant Representation Learning

In Fig. 7 and Fig. 8 we show examples for $k = 2$ and $k = 3$ on MNIST respectively the smallNORB dataset of LeCun et al. (2004). For MNIST, we compute the tangent vectors manually by rotation and dilation of the digits, similar as done by Simard et al. (1992; 1998). For the smallNORB example, the tangent vectors are given as differences between the corresponding images. As observed in the figures, the proposed formulation leads to interpretable latent codes which behave equivariantly with the generated images. We remark that the goal was not to achieve state-of-the-art image quality but rather to demonstrate that specifying tangent vectors yields disentangled representations. As remarked by Jaderberg et al. (2015), representing a 3D scene with a sequence of 2D convolutions is challenging and a specialized architecture based on a voxel representation would be more appropriate for the smallNORB example.

6.3. Discovering the Arrow of Time

In our last experiment, we set $k = 1$ and specify the tangent vector as the difference of two neighbouring frames in video data. We train on the tinyvideo beach dataset (Vondrick et al., 2016), which consists of more than 36 million frames. After training for about half an epoch, we can already observe a learned latent representation of time, see Fig. 1 and Fig. 9. We generate individual frames by varying the latent coordinate z_1 from -12.5 to 12.5.

Even though the model is trained on individual frames in random order, a somewhat coherent representation of time is discovered which captures phenomena such as ocean waves or people walking on the beach.

7. Discussion and Conclusion

In this work, we demonstrated that k-currents can be used introduce a notion of orientation into probabilistic models. Furthermore, in experiments we have shown that specifying partial tangent information to the data manifold leads to interpretable and equivariant latent representations such as the camera position and lighting in a 3D scene or the arrow of time in time series data.

The difference to purely unsupervised approaches such as InfoGAN or β-VAE is, that we can encourage potentially very complex latent representations to be learned. Nevertheless, an additional mutual information term as in (Chen et al., 2016) can be directly added to the formulation so that some representations could be encouraged through tangent vectors and the remaining ones are hoped to be discovered in an unsupervised fashion.

Generally speaking, we believe that geometric measure theory is a rather underexploited field with many possible application areas in probabilistic machine learning. We see this work as a step towards leveraging this potential.

[3] See `https://github.com/moellenh/flatgan` for a `PyTorch` implementation to reproduce Fig. 6 and Fig. 7.

Acknowledgements

We thank Kevin R. Vixie for his detailed feedback and comments on the manuscript. The work was partially supported by the German Research Foundation (DFG); project 394737018 "*Functional Lifting 2.0 – Efficient Convexifications for Imaging and Vision*".

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Beckmann, M. A continuous model of transportation. *Econometrica: Journal of the Econometric Society*, pp. 643–660, 1952.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Bottou, L., Arjovsky, M., Lopez-Paz, D., and Oquab, M. Geometrical insights for implicit generative modeling. *arXiv:1712.07822*, 2017.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

Csiszár, I., Shields, P. C., et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

de Rham, G. *Variétés différentiables, formes, courants, formes harmoniques*, volume 1222. Hermann, 1955.

Denton, E. L. et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, 2017.

Federer, H. *Geometric Measure Theory*. Springer, 1969.

Federer, H. and Fleming, W. H. Normal and integral currents. *Annals of Mathematics*, pp. 458–520, 1960.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-I., Trouvé, A., and Peyré, G. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. *arXiv:1810.08278*, 2018.

Fraser, A. M., Hengartner, N. W., Vixie, K. R., and Wohlberg, B. E. Incorporating invariants in Mahalanobis distance based classifiers: Application to Face Recognition. In *International Joint Conference on Neural Networks*, 2003.

Genevay, A., Peyré, G., and Cuturi, M. GAN and VAE from an optimal transport point of view. *arXiv:1706.01807*, 2017.

Glaunès, J., Qiu, A., Miller, M. I., and Younes, L. Large deformation diffeomorphic metric curve mapping. *International Journal of Computer Vision (IJCV)*, 80(3):317, 2008.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of Wasserstein GANs. *arXiv:1704.00028*, 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β–VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

Hinton, G. E., Krizhevsky, A., and Wang, S. D. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, 2011.

Hubbard, J. H. and Hubbard, B. B. *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*. Matrix Editions, 2015.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015.

Kim, H. and Mnih, A. Disentangling by factorising. *arXiv:1802.05983*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2014.

Krantz, S. G. and Parks, H. R. *Geometric Integration Theory*. Birkhäuser Boston, 2008.

LeCun, Y., Huang, F. J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017.

Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., and LeCun, Y. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, 2016.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually Converge? In *International Conference on Machine Learning*, 2018.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

Möllenhoff, T. and Cremers, D. Lifting vectorial variational problems: A natural formulation based on geometric measure theory and discrete exterior calculus. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Morgan, F. *Geometric Measure Theory: A Beginner's Guide*. Academic Press, 5th edition, 2016.

Morgan, S. P. and Vixie, K. R. L^1TV computes the flat norm for boundaries. In *Abstract and Applied Analysis*, 2007.

Narayanaswamy, S., Paige, T. B., Van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., and Torr, P. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, 2017.

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning*, 2017.

Odena, A., Buckman, J., Olsson, C., Brown, T. B., Olah, C., Raffel, C., and Goodfellow, I. Is generator conditioning causally related to GAN performance? In *International Conference on Machine Learning*, 2018.

Peyré, G. and Cuturi, M. Computational optimal transport. *arXiv:1803.00567*, 2018.

Pickup, L. C., Pan, Z., Wei, D., Shih, Y., Zhang, C., Zisserman, A., Schölkopf, B., and Freeman, W. T. Seeing the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv:1401.4082*, 2014.

Rifai, S., Dauphin, Y. N., Vincent, P., Bengio, Y., and Muller, X. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, 2011.

Santambrogio, F. *Optimal Transport for Applied Mathematicians*. Birkhäuser, New York, 2015.

Schmidhuber, J. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

Schwartz, L. *Théorie des distributions I, II*, volume 1245, 1122. Hermann, 1951, 1957.

Simard, P., Victorri, B., LeCun, Y., and Denker, J. Tangent prop – a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems*, 1992.

Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. Transformation invariance in pattern recognition – tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pp. 239–274, 1998.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G. R., et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Vaillant, M. and Glaunès, J. Surface matching via currents. In *Biennial International Conference on Information Processing in Medical Imaging*, 2005.

Vixie, K. R., Clawson, K., Asaki, T. J., Sandine, G., Morgan, S. P., and Price, B. Multiscale flat norm signatures for shapes and images. *Applied Mathematical Sciences*, 4 (14):667–680, 2010.

Vondrick, C., Pirsiavash, H., and Torralba, A. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, 2016.

Wei, D., Lim, J. J., Zisserman, A., and Freeman, W. T. Learning and using the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Whitney, H. *Geometric Integration Theory*. Princeton University Press, 1957.

List of Figures

Bibliography

[ABD03] G. Alberti, G. Bouchitté and G. Dal Maso. **The calibration method for the Mumford-Shah functional and free-discontinuity problems**. In: *Calculus of Variations and Partial Differential Equations* 16.3 (2003), pp. 299–333 (Cited on pages 36, 37, 44, 62, 76, 78, 79, 95).

[AM17] G. Alberti and A. Massaccesi. **On some geometric properties of currents and Frobenius theorem**. In: *arXiv:1705.09938* (2017) (Cited on pages 37, 135).

[AMS19] G. Alberti, A. Massaccesi and E. Stepanov. **On the geometric structure of currents tangent to smooth distributions**. In: *arXiv:1907.07456* (2019) (Cited on pages 37, 135).

[AB94] C. D. Aliprantis and K. C. Border. **Infinite Dimensional Analysis: A Hitchhiker's Guide**. Springer, 1994 (Cited on pages 11, 27).

[Ama16] S.-I. Amari. **Information geometry and its applications**. Vol. 194. Springer, 2016 (Cited on page 12).

[AFP00] L. Ambrosio, N. Fusco and D. Pallara. **Functions of bounded variation and free discontinuity problems**. The Clarendon Press Oxford University Press, 2000 (Cited on pages 9, 32–34, 36, 49, 67, 76, 78).

[AT05] B. Appleton and H. Talbot. **Globally minimal surfaces by continuous maximal flows**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 28.1 (2005), pp. 106–118 (Cited on page 21).

[ACB17] M. Arjovsky, S. Chintala and L. Bottou. **Wasserstein Generative Adversarial Networks**. In: *International Conference on Machine Learning (ICML)*. 2017 (Cited on pages 25, 94, 113, 114, 139).

[AA14] D. N. Arnold and G. Awanou. **Finite element differential forms on cubical meshes**. In: *Mathematics of Computation* 83.288 (2014), pp. 1551–1570 (Cited on page 105).

[AFW06] D. N. Arnold, R. S. Falk and R. Winther. **Finite element exterior calculus, homological techniques, and applications**. In: *Acta numerica* 15 (2006), pp. 1–155 (Cited on page 105).

[AG91] P. Aviles and Y. Giga. **Variational integrals on mappings of bounded variation and their lower semicontinuity**. In: *Arch. Ration. Mech. Anal.* 115.3 (1991), pp. 201–255 (Cited on pages 35, 95).

[Bac19] F. Bach. **Submodular functions: from discrete to continous domains**. In: *Mathematical Programming* (175 2019), pp. 419–459 (Cited on pages 77, 136).

[BDH96] C. B. Barber, D. P. Dobkin and H. Huhdanpaa. **The quickhull algorithm for convex hulls**. In: *ACM Transactions on Mathematical Software (TOMS)* 22.4 (1996), pp. 469–483 (Cited on pages 72, 161).

[Bar20a] S. Bartels. **Error estimates for a class of discontinuous Galerkin methods for nonsmooth problems via convex duality relations**. In: *arXiv:2004.09196* (2020) (Cited on page 136).

[Bar20b] S. Bartels. **Nonconforming discretizations of convex minimization problems and precise relations to mixed methods**. In: *arXiv:2002.02359* (2020) (Cited on page 136).

[BT70] E. Beale and J. Tomlin. **Special facilities in a general mathematical programming system for nonconvex problems using ordered sets of variables**. In: 1970 (Cited on page 46).

[Bec52] M. Beckmann. **A continuous model of transportation**. In: *Econometrica: Journal of the Econometric Society* (1952), pp. 643–660 (Cited on page 120).

[BPV91] G. Bellettini, M. Paolini and C. Verdi. **Convex approximations of functionals with curvature**. In: *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl.* 2.4 (1991), pp. 297–306 (Cited on page 17).

[BCV13] Y. Bengio, A. Courville and P. Vincent. **Representation learning: A review and new perspectives**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 35.8 (2013), pp. 1798–1828 (Cited on page 112).

[BER17] B. Berkels, A. Effland and M. Rumpf. **A posteriori error control for the binary Mumford-Shah model**. In: *Mathematics of computation* 86.306 (2017), pp. 1769–1791 (Cited on page 138).

[Ber96] J. Bernoulli. **Problema novum ad cujus solutionem mathematici invitantur**. In: *Acta Eruditorum* 18.269 (1696) (Cited on page 106).

[Bhu08] A. Bhusnurmath. **Applying convex optimization techniques to energy minimization problems in computer vision**. PhD thesis. University of Pennsylvania, 2008 (Cited on pages 19, 20).

[BT08] A. Bhusnurmath and C. J. Taylor. **Solving stereo matching problems using interior point methods**. In: *International Symposium on 3D Data Processing, Visualization and Transmission*. 2008 (Cited on pages 19, 20).

[BZ87] A. Blake and A. Zisserman. **Visual Reconstruction**. MIT Press, 1987 (Cited on page 76).

[BKM17] D. M. Blei, A. Kucukelbir and J. D. McAuliffe. **Variational inference: A review for statisticians**. In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877 (Cited on page 3).

[BPT12] G. Blekherman, P. A. Parrilo and R. R. Thomas. **Semidefinite Optimization and Convex Algebraic Geometry**. SIAM, 2012 (Cited on page 87).

[Bot+17] L. Bottou, M. Arjovsky, D. Lopez-Paz and M. Oquab. **Geometrical insights for implicit generative modeling**. In: *arXiv:1712.07822* (2017) (Cited on page 112).

[Bou98] G. Bouchitté. **Recent convexity arguments in the calculus of variations.** In: *Lecture notes from the 3rd Int. Summer School on the Calculus of Variations, Pisa* (1998) (Cited on pages 76, 78).

[BF15] G. Bouchitté and I. Fragalà. **Duality for non-convex variational problems**. In: *Comptes Rendus Mathematique* 353.4 (2015), pp. 375–379 (Cited on page 80).

[BP18] G. Bouchitté and M. Phan. **A duality recipe for non-convex variational problems**. In: *Comptes Rendus Mécanique* 346.3 (2018), pp. 206–221 (Cited on page 133).

[BK03] Y. Boykov and V. Kolmogorov. **Computing geodesics and minimal surfaces via graph cuts**. In: *International Conference on Computer Vision (ICCV)*. 2003 (Cited on page 20).

[BK04] Y. Boykov and V. Kolmogorov. **An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 26.9 (2004), pp. 1124–1137 (Cited on page 15).

[Bra95] K. A. Brakke. **Numerical solution of soap film dual problems**. In: *Experimental Mathematics* 4.4 (1995), pp. 269–287 (Cited on page 21).

[BKP10] K. Bredies, K. Kunisch and T. Pock. **Total generalized variation**. In: *SIAM J. Imaging Sci.* 3.3 (2010), pp. 492–526 (Cited on page 133).

[BPW13] K. Bredies, T. Pock and B. Wirth. **Convex relaxation of a class of vertex penalizing functionals**. In: *J. Math. Imaging Vis.* 47.3 (2013), pp. 278–302 (Cited on page 134).

[BPW15] K. Bredies, T. Pock and B. Wirth. **A convex, lower semicontinuous approximation of Euler's elastica energy**. In: *SIAM J. Math. Anal.* 47.1 (2015), pp. 566–613 (Cited on page 134).

[Bre03] Y. Brenier. **Extended Monge-Kantorovich theory**. In: *Optimal transportation and applications*. Springer, 2003, pp. 91–121 (Cited on page 13).

[BES63] H. Busemann, G. Ewald and G. C. Shephard. **Convex bodies and convexity on Grassmann cones**. In: *Math. Ann.* 151.1 (1963), pp. 1–41 (Cited on page 97).

[CC20] C. Caillaud and A. Chambolle. **Error estimates for finite differences approximations of the total variation**. In: (2020) (Cited on page 136).

[Car16] M. Carioni. **A Discrete Coarea-type Formula for the Mumford-Shah Functional in Dimension One**. In: *arXiv:1610.01846* (2016) (Cited on page 80).

[CR00] C. Carstensen and T. Roubicek. **Numerical approximation of young measuresin non-convex variational problems**. In: *Numerische Mathematik* 84.3 (2000), pp. 395–415 (Cited on page 134).

[CP11a] A. Chambolle and T. Pock. **A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging**. In: *J. Math. Imaging Vis.* 40 (2011), pp. 120–145 (Cited on pages 27, 37–39, 106, 109).

[Cha01] A. Chambolle. **Convex representation for lower semicontinuous envelopes of functionals in L^1**. In: *J. Convex Anal.* 8.1 (2001), pp. 149–170. ISSN: 0944-6532 (Cited on pages 76, 78, 80).

[Cha+10] A. Chambolle, V. Caselles, D. Cremers, M. Novaga and T. Pock. **An introduction to total variation for image analysis**. In: *Theoretical foundations and numerical methods for sparse recovery* 9.263-340 (2010), p. 227 (Cited on pages 7, 17, 43).

[CCP12] A. Chambolle, D. Cremers and T. Pock. **A convex approach to minimal partitions**. In: *SIAM J. Imaging Sci.* 5.4 (2012), pp. 1113–1158 (Cited on pages 17, 18, 49, 66, 77).

[CP16a] A. Chambolle and T. Pock. **An introduction to continuous optimization for imaging**. In: *Acta Numerica* 25 (2016), pp. 161–319 (Cited on page 3).

[CP16b] A. Chambolle and T. Pock. **On the ergodic convergence rates of a first-order primal–dual algorithm**. In: *Mathematical Programming* 159.1-2 (2016), pp. 253–287 (Cited on page 39).

[CP19] A. Chambolle and T. Pock. **Total roto-translational variation**. In: *Numerische Mathematik* 142.3 (2019), pp. 611–666 (Cited on page 134).

[CP20] A. Chambolle and T. Pock. **Crouzeix–Raviart approximation of the total variation on simplicial meshes**. In: *J. Math. Imaging Vis.* (2020), pp. 1–28 (Cited on page 136).

[CEN06] T. F. Chan, S. Esedoglu and M. Nikolova. **Algorithms for finding global minimizers of image segmentation and denoising models**. In: *SIAM J. Appl. Math.* 66.5 (2006), pp. 1632–1648 (Cited on page 17).

[CK15] Q. Chen and V. Koltun. **Robust nonrigid registration by convex optimization**. In: *International Conference on Computer Vision (ICCV)*. 2015 (Cited on page 93).

[CK16] Q. Chen and V. Koltun. **Full Flow: Optical Flow Estimation By Global Optimization Over Regular Grids**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (Cited on pages 8, 93).

[Che+16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel. **InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016 (Cited on pages 113, 114, 124).

[Cho69] G. Choquet. **Lectures on Analysis, Vol. 2: Representation Theory**. 1969 (Cited on page 11).

[CK97] L. D. Cohen and R. Kimmel. **Global minimum for active contour models: A minimal path approach**. In: *Int. J. Comput. Vis. (IJCV)* 24.1 (1997), pp. 57–78 (Cited on pages 15, 93).

[CM18] P. L. Combettes and C. L. Müller. **Perspective functions: Proximal calculus and applications in high-dimensional statistics**. In: *J. Math. Anal. Appl.* 457.2 (2018), pp. 1283–1306 (Cited on page 106).

[CP11b] P. L. Combettes and J.-C. Pesquet. **Proximal splitting methods in signal processing**. In: *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212 (Cited on page 19).

[Cra19] K. Crane. **Discrete differential geometry: An applied introduction**. 2019 (Cited on page 96).

[CS05] D. Cremers and S. Soatto. **Motion competition: A variational approach to piecewise parametric motion segmentation**. In: *Int. J. Comput. Vis. (IJCV)* 62.3 (2005), pp. 249–265 (Cited on page 9).

[CS13] D. Cremers and E. Strekalovskiy. **Total cyclic variation and generalizations**. In: *J. Math. Imaging Vis.* 47.3 (2013), pp. 258–277 (Cited on page 135).

[CS+04] I. Csiszár, P. C. Shields et al. **Information theory and statistics: A tutorial**. In: *Foundations and Trends® in Communications and Information Theory* 1.4 (2004), pp. 417–528 (Cited on page 112).

[Cut13] M. Cuturi. **Sinkhorn distances: Lightspeed computation of optimal transport**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2013 (Cited on page 140).

[DLS14] L. De Pascale, J. Louet and F. Santambrogio. **A first analysis of the Monge problem with vanishing gradient penalization**. In: *arXiv:1407.7022* (2014) (Cited on page 13).

[Den+17] E. L. Denton et al. **Unsupervised learning of disentangled representations from video**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 (Cited on page 114).

[DHK11] T. K. Dey, A. N. Hirani and B. Krishnamoorthy. **Optimal homologous cycles, total unimodularity, and linear programming**. In: *SIAM J. Comp.* 40.4 (2011), pp. 1026–1044 (Cited on page 135).

[DB16] S. Diamond and S. Boyd. **CVXPY: A Python-embedded modeling language for convex optimization**. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2909–2913 (Cited on page 39).

[DW20] C. Dirks and B. Wirth. **An adaptive finite element approach for lifted branched transport problems**. In: *arXiv:2003.13797* (2020) (Cited on page 138).

[DSC18] C. Domokos, F. R. Schmidt and D. Cremers. **MRF Optimization with Separable Convex Prior on Partially Ordered Labels**. In: *European Conference on Computer Vision (ECCV)*. 2018 (Cited on page 93).

[DE13] G. Dziuk and C. M. Elliott. **Finite element methods for surface PDEs**. In: *Acta Numerica* 22 (2013), pp. 289–396 (Cited on page 134).

[EZC10] E. Esser, X. Zhang and T. F. Chan. **A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science**. In: *SIAM J. Imaging Sci.* 3 (2010), pp. 1015–1046 (Cited on page 50).

[Ess10] J. E. Esser. **Primal dual algorithms for convex models and applications to image restoration, registration and nonlocal inpainting**. 2010 (Cited on page 27).

[Fed69] H. Federer. **Geometric measure theory**. Springer, 1969 (Cited on pages 13, 95, 96, 112, 114, 117, 119, 120, 160).

[Fed74] H. Federer. **Real flat chains, cochains and variational problems**. In: *Indiana Univ. Math. J.* 24.4 (1974), pp. 351–407 (Cited on pages 13, 95, 104, 135).

[FF60] H. Federer and W. H. Fleming. **Normal and integral currents**. In: *Annals of Mathematics* (1960), pp. 458–520 (Cited on pages 112, 118, 120).

[FMN16] C. Fefferman, S. Mitter and H. Narayanan. **Testing the manifold hypothesis**. In: *Journal of the American Mathematical Society* 29.4 (2016), pp. 983–1049 (Cited on pages 5, 112).

[Fey+18] J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé and G. Peyré. **Interpolating between Optimal Transport and MMD using Sinkhorn Divergences**. In: *arXiv:1810.08278* (2018) (Cited on page 112).

[FA14] A. Fix and S. Agarwal. **Duality and the Continuous Graphical Model**. In: *European Conference on Computer Vision (ECCV)*. 2014 (Cited on pages 16, 44, 63, 77, 84, 87, 94, 136, 138).

[For+18] D. Fortun, M. Storath, D. Rickert, A. Weinmann and M. Unser. **Fast Piecewise-Affine Motion Estimation Without Segmentation**. In: *IEEE Trans. Imag. Proc.* 27.11 (2018), pp. 5612–5624 (Cited on page 9).

[FB18] C. Fougner and S. Boyd. **Parameter selection and preconditioning for a graph form solver**. In: *Emerging Applications of Control and Systems Theory*. 2018, pp. 41–61 (Cited on pages 39, 40).

[Fra+03] A. M. Fraser, N. W. Hengartner, K. R. Vixie and B. E. Wohlberg. **Incorporating invariants in Mahalanobis distance based classifiers: Application to Face Recognition**. In: *International Joint Conference on Neural Networks*. 2003 (Cited on page 113).

[GM18] J. Geiping and M. Moeller. **Composite optimization by nonconvex majorization-minimization**. In: *SIAM J. Imaging Sci.* 11.4 (2018), pp. 2494–2528 (Cited on page 134).

[Gen+16] A. Genevay, M. Cuturi, G. Peyré and F. Bach. **Stochastic Optimization for Large-scale Optimal Transport**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016 (Cited on page 140).

[GPC17] A. Genevay, G. Peyré and M. Cuturi. **GAN and VAE from an optimal transport point of view**. In: *arXiv:1706.01807* (2017) (Cited on pages 112, 113).

[Gho+20] N. Ghoussoub, Y.-H. Kim, H. Lavenant and A. Z. Palmer. **Hidden convexity in a problem of nonlinear elasticity**. In: *arXiv:2004.10287* (2020) (Cited on page 13).

[GMS98] M. Giaquinta, G. Modica and J. Souček. **Cartesian currents in the calculus of variations I, II.** Vol. 37-38. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Springer, 1998 (Cited on pages 35, 46, 62, 99, 103).

[Gla+08] J. Glaunès, A. Qiu, M. I. Miller and L. Younes. **Large deformation diffeomorphic metric curve mapping**. In: *Int. J. Comput. Vis. (IJCV)* 80.3 (2008), p. 317 (Cited on page 113).

[GSC12] B. Goldluecke, E. Strekalovskiy and D. Cremers. **The natural total variation which arises from geometric measure theory**. In: *SIAM J. Imaging Sci.* 5.2 (2012), pp. 537–563 (Cited on pages 69, 130).

[GSC13] B. Goldluecke, E. Strekalovskiy and D. Cremers. **Tight convex relaxations for vector-valued labeling**. In: *SIAM J. Imaging Sci.* 6.3 (2013), pp. 1626–1664 (Cited on pages 21, 63, 72, 73, 77, 95, 136, 161).

[GBO12] T. Goldstein, X. Bresson and S. Osher. **Global minimization of Markov random fields with applications to optical flow**. In: *Inverse Problems & Imaging* 6.4 (2012), pp. 623–644 (Cited on pages 21, 95).

[Gol+13] T. Goldstein, M. Li, X. Yuan, E. Esser and R. Baraniuk. **Adaptive primal-dual hybrid gradient methods for saddle-point problems**. In: *arXiv:1305.0546* (2013) (Cited on pages 40, 50–52).

[Goo+14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. **Generative adversarial nets**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2014 (Cited on page 112).

[GGK19] A. Görlitz, J. Geiping and A. Kolb. **Piecewise Rigid Scene Flow with Implicit Motion Segmentation**. In: *International Conference on Intelligent Robots and Systems (IROS)*. 2019 (Cited on page 9).

[Gul+17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville. **Improved training of Wasserstein GANs**. In: *arXiv:1704.00028* (2017) (Cited on pages 122, 163).

[HZ03] R. Hartley and A. Zisserman. **Multiple view geometry in computer vision**. Cambridge university press, 2003 (Cited on page 8).

[HY12] B. He and X. Yuan. **Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective**. In: *SIAM J. Imaging Sci.* 5.1 (2012), pp. 119–149 (Cited on page 39).

[Hei+13] P. Heise, S. Klose, B. Jensen and A. Knoll. **PM-Huber: PatchMatch with Huber regularization for stereo matching**. In: *International Conference on Computer Vision (ICCV)*. 2013 (Cited on page 8).

[Her+19] M. Herrmann, R. Herzog, S. Schmidt, J. Vidal-Núñez and G. Wachsmuth. **Discrete total variation with finite elements and applications to imaging**. In: *J. Math. Imaging Vis.* 61.4 (2019), pp. 411–431 (Cited on page 136).

[Hig+16] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed and A. Lerchner. **β–VAE: Learning basic visual concepts with a constrained variational framework**. In: *International Conference on Learning Representations*. 2016 (Cited on page 113).

[Hil00] D. Hilbert. **Mathematische Probleme**. In: *Nachrichten von der Königl. Gesellschaft der Wiss. zu Göttingen* (1900), pp. 253–297 (Cited on page 94).

[HIK02] M. Hintermüller, K. Ito and K. Kunisch. **The primal-dual active set strategy as a semismooth Newton method**. In: *SIAM J. Optim.* 13.3 (2002), pp. 865–888 (Cited on page 139).

[HKW11] G. E. Hinton, A. Krizhevsky and S. D. Wang. **Transforming autoencoders**. In: *International Conference on Artificial Neural Networks*. 2011 (Cited on page 114).

[Hir03] A. N. Hirani. **Discrete exterior calculus**. PhD thesis. California Institute of Technology, 2003 (Cited on pages 80, 104).

[HL12] J.-B. Hiriart-Urruty and C. Lemaréchal. **Fundamentals of convex analysis**. Springer, 2012 (Cited on pages 27–31, 96).

[HS81] B. K. Horn and B. G. Schunck. **Determining optical flow**. In: *Artificial Intelligence* 17.1-3 (1981), pp. 185–203 (Cited on pages 3, 8).

[HH15] J. H. Hubbard and B. B. Hubbard. **Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach**. Matrix Editions, 2015 (Cited on page 116).

[IH81] K. Ikeuchi and B. K. Horn. **Numerical shape from shading and occluding boundaries**. In: *Artificial Intelligence* 17.1-3 (1981), pp. 141–184 (Cited on page 3).

[Ish03] H. Ishikawa. **Exact optimization for Markov random fields with convex priors**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 25.10 (2003), pp. 1333–1336 (Cited on pages 15, 44, 61, 77, 79, 93).

[IK03] K. Ito and K. Kunisch. **Semi–smooth Newton methods for variational inequalities of the first kind**. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 37.1 (2003), pp. 41–62 (Cited on page 139).

[Jad+15] M. Jaderberg, K. Simonyan, A. Zisserman and K. Kavukcuoglu. **Spatial transformer networks**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2015 (Cited on page 124).

[Jai+15] M. Jaimez, M. Souiai, J. Stückler, J. Gonzalez-Jimenez and D. Cremers. **Motion cooperation: Smooth piece-wise rigid scene flow from RGB-D images**. In: *International Conference on 3D Vision (3DV)*. 2015 (Cited on page 9).

[KMM06] T. Kaczynski, K. Mischaikow and M. Mrozek. **Computational Homology**. Springer, 2006 (Cited on page 104).

[Kak02] S. M. Kakade. **A natural policy gradient**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2002, pp. 1531–1538 (Cited on page 12).

[Kan60] L. V. Kantorovich. **Mathematical methods of organizing and planning production**. In: *Management Science* 6.4 (1960), pp. 366–422 (Cited on pages 12, 13, 94).

[Kap+13] J. Kappes, B. Andres, F. Hamprecht, C. Schnorr, S. Nowozin, D. Batra, S. Kim, B. Kausler, J. Lellmann, N. Komodakis et al. **A comparative study of modern inference techniques for discrete energy minimization problems**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013 (Cited on pages 3, 109).

[KM18] H. Kim and A. Mnih. **Disentangling by factorising**. In: *arXiv:1802.05983* (2018) (Cited on page 113).

[KMS00] R. Kimmel, R. Malladi and N. Sochen. **Images as Embedded Maps and Minimal Surfaces: Movies, Color, Texture, and Volumetric Medical Images**. In: *Int. J. Comput. Vis. (IJCV)* 39.2 (2000) (Cited on pages 94, 131).

[KW14] D. P. Kingma and M. Welling. **Auto-encoding variational Bayes**. In: *arXiv:1312.6114* (2014) (Cited on page 112).

[KB14] D. P. Kingma and J. Ba. **Adam: A method for stochastic optimization**. In: *arXiv:1412.6980* (2014) (Cited on page 163).

[KG04] D. Kirsanov and S. J. Gortler. **A discrete global minimization algorithm for continuous variational problems**. In: *Harvard CS Technical Report TR-14-04* (2004) (Cited on pages 20, 21).

[Klo+08] M. Klodt, T. Schoenemann, K. Kolev, M. Schikora and D. Cremers. **An experimental comparison of discrete and continuous shape optimization methods**. In: *European Conference on Computer Vision (ECCV)*. Springer. 2008 (Cited on page 14).

[Koh+08] P. Kohli, A. Shekhovtsov, C. Rother, V. Kolmogorov and P. Torr. **On partial optimality in multi-label MRFs**. In: *International Conference on Machine Learning (ICML)*. 2008, pp. 480–487 (Cited on page 93).

[Kol06] V. Kolmogorov. **Convergent tree-reweighted message passing for energy minimization**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 28.10 (2006), pp. 1568–1583 (Cited on page 93).

[KPR16] V. Kolmogorov, T. Pock and M. Rolinek. **Total variation on a tree**. In: *SIAM J. Imaging Sci.* 9.2 (2016), pp. 605–636 (Cited on page 16).

[KR07] V. Kolmogorov and C. Rother. **Minimizing nonsubmodular functions with graph cuts - a review**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 29.7 (2007) (Cited on page 93).

[KP08] S. G. Krantz and H. R. Parks. **Geometric Integration Theory**. Birkhäuser Boston, 2008 (Cited on pages 34, 96, 100, 114).

[KC13] G. Kuschk and D. Cremers. **Fast and accurate large-scale stereo reconstruction using variational methods**. In: *International Conference on Computer Vision Workshops (ICCV Workshops)*. 2013 (Cited on page 19).

[Las00] J. B. Lasserre. **Global Optimization with Polynomials and the Problem of Moments**. In: *SIAM J. on Optimization* 11.3 (2000), pp. 796–817 (Cited on page 12).

[LHB04] Y. LeCun, F. J. Huang and L. Bottou. **Learning methods for generic object recognition with invariance to pose and lighting**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2004 (Cited on pages 122, 123).

[LS11] J. Lellmann and C. Schnörr. **Continuous Multiclass Labeling Approaches and Algorithms**. In: *SIAM J. Imaging Sci.* 4.4 (2011), pp. 1049–1096 (Cited on pages 44, 48, 66, 77).

[Lel+13a] J. Lellmann, E. Strekalovskiy, S. Koetter and D. Cremers. **Total Variation Regularization for Functions with Values in a Manifold**. In: *International Conference on Computer Vision (ICCV)*. 2013 (Cited on pages 18, 21, 44, 62, 63, 66, 68, 70–73, 77, 95, 134).

[Lel+09] J. Lellmann, J. Kappes, J. Yuan, F. Becker and C. Schnörr. **Convex multi-class image labeling by simplex-constrained total variation**. In: *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*. Springer. 2009, pp. 150–162 (Cited on pages 17, 18).

[Lel+13b] J. Lellmann, B. Lellmann, F. Widmann and C. Schnörr. **Discrete and continuous models for partitioning problems**. In: *Int. J. Comput. Vis. (IJCV)* 104.3 (2013), pp. 241–269 (Cited on pages 14, 17, 135).

[Li+17] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang and B. Póczos. **MMD GAN: Towards deeper understanding of moment matching network**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 (Cited on page 113).

[LSH16] M. Li, A. Shekhovtsov and D. Huber. **Complexity of discrete energy minimization problems**. In: *European Conference on Computer Vision (ECCV)*. 2016 (Cited on pages 15, 93).

[LB14] Y. Li and M. S. Brown. **Single image layer separation using relative smoothness**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014 (Cited on page 9).

[LL18] B. Loewenhauser and J. Lellmann. **Functional Lifting for Variational Problems with Higher-Order Regularization**. In: *Imaging, Vision and Learning Based on Optimization and PDEs* (2018), pp. 101–120 (Cited on pages 95, 134).

[Lou14] J. Louet. **Optimal transport problems with gradient penalization**. PhD thesis. Université Paris-Sud, 2014 (Cited on page 13).

[Mat+16] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann and Y. LeCun. **Disentangling factors of variation in deep representation using adversarial training**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016 (Cited on page 114).

[Mat99] P. Mattila. **Geometry of sets and measures in Euclidean spaces: fractals and rectifiability**. 44. Cambridge University Press, 1999 (Cited on page 33).

[Mei+18] J.-J. Mei, Y. Dong, T.-Z. Huang and W. Yin. **Cauchy noise removal by nonconvex ADMM with convergence guarantees**. In: *Journal of Scientific Computing* 74.2 (2018), pp. 743–766 (Cited on page 7).

[Mémo7] F. Mémoli. **On the use of Gromov-Hausdorff distances for shape comparison**. In: *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007 (Cited on page 95).

[Mémo9] F. Mémoli. **Spectral Gromov-Wasserstein distances for shape matching**. In: *International Conference on Computer Vision Workshops (ICCV Workshops)*. 2009 (Cited on page 95).

[MHG15] M. Menze, C. Heipke and A. Geiger. **Discrete optimization for optical flow**. In: *German Conference on Pattern Recognition (GCPR)*. 2015 (Cited on page 93).

[MGN18] L. Mescheder, A. Geiger and S. Nowozin. **Which training methods for GANs do actually Converge?** In: *International Conference on Machine Learning (ICML)*. 2018 (Cited on page 164).

[MO14] M. Mirza and S. Osindero. **Conditional generative adversarial nets**. In: *arXiv:1411.1784* (2014) (Cited on page 114).

[Mon81] G. Monge. **Mémoire sur la théorie des déblais et des remblais**. In: *Histoire de l'Académie Royale des Sciences de Paris* (1781) (Cited on pages 12, 94).

[Moro2] M. G. Mora. **The calibration method for free-discontinuity problems on vector-valued maps**. In: *Journal of Convex Analysis* 9.1 (2002), pp. 1–30 (Cited on pages 35, 103).

[Mor16] F. Morgan. **Geometric Measure Theory: A Beginner's Guide**. 5th. Academic Press, 2016 (Cited on pages 32, 33, 94, 96, 114, 116).

[MVo7] S. P. Morgan and K. R. Vixie. **L^1TV computes the flat norm for boundaries**. In: *Abstract and Applied Analysis*. 2007 (Cited on pages 112, 119, 120).

[Mum94] D. Mumford. **The Bayesian rationale for energy functionals**. In: *Geometry-driven diffusion in Computer Vision* 1 (1994), pp. 141–153 (Cited on page 6).

[MS89] D. Mumford and J. Shah. **Optimal approximations by piecewise smooth functions and associated variational problems**. In: *Communications on Pure and Applied Mathematics* 42.5 (1989), pp. 577–685 (Cited on pages 9, 10, 76).

[Nar+17] S. Narayanaswamy, T. B. Paige, J.-W. Van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood and P. Torr. **Learning disentangled representations with semi-supervised deep generative models**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 (Cited on page 114).

[NN94] S. K. Nayar and Y. Nakagawa. **Shape from Focus**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 16.8 (Aug. 1994), pp. 824–831 (Cited on page 59).

[Néd80] J.-C. Nédélec. **Mixed finite elements in $\mathbb{R}^3$**. In: *Numerische Mathematik* 35.3 (1980), pp. 315–341 (Cited on page 105).

[NLD11] R. A. Newcombe, S. J. Lovegrove and A. J. Davison. **DTAM: Dense tracking and mapping in real-time**. In: *International Conference on Computer Vision (ICCV)*. IEEE. 2011 (Cited on page 19).

[NB15] R. M. Nguyen and M. S. Brown. **Fast and effective Lo gradient minimization by region fusion**. In: *International Conference on Computer Vision (ICCV)*. 2015 (Cited on page 9).

[Nie+14] C. Nieuwenhuis, S. Hawe, M. Kleinsteuber and D. Cremers. **Co-sparse textural similarity for interactive segmentation**. In: *European Conference on Computer Vision (ECCV)*. 2014, pp. 285–301 (Cited on page 9).

[NTC13] C. Nieuwenhuis, E. Töppe and D. Cremers. **A survey and comparison of discrete and continuous multi-label optimization approaches for the Potts model**. In: *Int. J. Comput. Vis. (IJCV)* 104.3 (2013), pp. 223–240 (Cited on page 17).

[NW06] J. Nocedal and S. J. Wright. **Numerical Optimization**. 2nd. New York: Springer, 2006 (Cited on page 71).

[OOS17] A. Odena, C. Olah and J. Shlens. **Conditional image synthesis with auxiliary classifier GANs**. In: *International Conference on Machine Learning (ICML)*. 2017 (Cited on page 114).

[Ovs+12] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher and L. Guibas. **Functional maps: a flexible representation of maps between shapes**. In: *ACM Transactions on Graphics (TOG)* 31.4 (2012), p. 30 (Cited on page 95).

[PS13] E. Paolini and E. Stepanov. **Structure of metric cycles and normal one-dimensional currents**. In: *Journal of Functional Analysis* 264.6 (2013), pp. 1269–1295 (Cited on page 135).

[PB13] N. Parikh and S. Boyd. **Proximal Algorithms**. In: *Foundations and Trends in Optimization* 1 (2013), pp. 123–231 (Cited on page 27).

[Par92] H. R. Parks. **Numerical approximation of parametric oriented area-minimizing hypersurfaces**. In: *SIAM J. Sci. Stat. Comp.* 13.2 (1992), pp. 499–511 (Cited on page 20).

[Ped99] P. Pedregal. **Optimization, relaxation and Young measures**. In: *Bulletin of the American Mathematical Society* 36.1 (1999), pp. 27–58 (Cited on page 134).

[Pen+11] J. Peng, T. Hazan, D. McAllester and R. Urtasun. Convex max-product algorithms for continuous MRFs with applications to protein folding. In: *International Conference on Machine Learning (ICML)*. 2011 (Cited on page 16).

[PS08] J. Peters and S. Schaal. Natural actor-critic. In: *Neurocomputing* 71.7-9 (2008), pp. 1180–1190 (Cited on page 12).

[PC18] G. Peyré and M. Cuturi. Computational Optimal Transport. In: *arXiv:1803.00567* (2018) (Cited on pages 13, 94, 112).

[Pic+14] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Schölkopf and W. T. Freeman. Seeing the arrow of time. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014 (Cited on page 112).

[PC11] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: *International Conference on Computer Vision (ICCV)*. 2011 (Cited on pages 39, 40, 50).

[Poc+09a] T. Pock, D. Cremers, H. Bischof and A. Chambolle. An Algorithm for Minimizing the Piecewise Smooth Mumford-Shah Functional. In: *International Conference on Computer Vision (ICCV)*. 2009 (Cited on pages 38, 44, 50, 62, 69, 77, 84, 89, 91, 154).

[Poc+08] T. Pock, T. Schoenemann, G. Graber, H. Bischof and D. Cremers. A Convex Formulation of Continuous Multi-Label Problems. In: *ECCV*. 2008 (Cited on pages 21, 43–45, 48, 62).

[Poc+09b] T. Pock, A. Chambolle, D. Cremers and H. Bischof. A convex relaxation approach for computing minimal partitions. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009 (Cited on pages 17, 18, 44, 48).

[Poc+10] T. Pock, D. Cremers, H. Bischof and A. Chambolle. Global solutions of variational models with convex regularization. In: *SIAM J. Imaging Sci.* 3.4 (2010), pp. 1122–1145 (Cited on pages 21, 34–37, 43–45, 47, 48, 51–53, 62, 66, 80, 83, 88, 89, 95, 106, 135).

[PZB07] T. Pock, C. Zach and H. Bischof. Mumford-Shah Meets Stereo: Integration of Weak Depth Hypotheses. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007 (Cited on page 90).

[QDA18] Y. Quéau, J.-D. Durou and J.-F. Aujol. Variational methods for normal integration. In: *J. Math. Imaging Vis.* 60.4 (2018), pp. 609–632 (Cited on page 9).

[RMC15] A. Radford, L. Metz and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In: *arXiv:1511.06434* (2015) (Cited on page 163).

[RPB13] R. Ranftl, T. Pock and H. Bischof. **Minimizing TGV-based variational models with non-convex data terms**. In: *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*. 2013 (Cited on page 134).

[RT77] P.-A. Raviart and J.-M. Thomas. **A mixed finite element method for 2nd order elliptic problems**. In: *Mathematical aspects of finite element methods*. Springer, 1977, pp. 292–315 (Cited on page 105).

[RMW14] D. J. Rezende, S. Mohamed and D. Wierstra. **Stochastic backpropagation and approximate inference in deep generative models**. In: *arXiv:1401.4082* (2014) (Cited on page 112).

[Rha55] G. de Rham. **Variétés différentiables, formes, courants, formes harmoniques**. Vol. 1222. Paris: Hermann, 1955 (Cited on page 112).

[Rif+11] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio and X. Muller. **The manifold tangent classifier**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2011 (Cited on page 113).

[Roc74] R. T. Rockafellar. **Conjugate duality and optimization**. Vol. 16. SIAM, 1974 (Cited on pages 27, 139).

[Roc76] R. T. Rockafellar. **Monotone operators and the proximal point algorithm**. In: *SIAM J. Control Optim.* 14.5 (1976), pp. 877–898 (Cited on page 39).

[Roc96] R. T. Rockafellar. **Convex Analysis**. Princeton University Press, 1996 (Cited on pages 27, 45, 86, 96, 157, 158).

[RWW98] R. T. Rockafellar, R. J.-B. Wets and M. Wets. **Variational analysis**. Springer, 1998 (Cited on pages 3, 69, 87, 159).

[Rod+19] E. Rodolà, Z. Lähner, A. M. Bronstein, M. M. Bronstein and J. Solomon. **Functional Maps Representation On Product Manifolds**. In: *Computer Graphics Forum*. Vol. 38. 2019 (Cited on page 95).

[ROF92] L. I. Rudin, S. Osher and E. Fatemi. **Nonlinear total variation based noise removal algorithms**. In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268 (Cited on pages 7, 19, 51, 71).

[Ruo15] N. Ruozzi. **Exactness of approximate MAP inference in continuous MRFs**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2015 (Cited on page 16).

[Sal13] M. Salzmann. **Continuous inference in graphical models with polynomial energies**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013 (Cited on page 16).

[San15] F. Santambrogio. **Optimal Transport for Applied Mathematicians**. New York: Birkhäuser, 2015 (Cited on pages 13, 94, 120).

[SR96] G. Sapiro and D. Ringach. **Anisotropic diffusion of multivalued images with applications to color filtering**. In: *IEEE Trans. Img. Proc.* 5.11 (1996), pp. 1582–1586 (Cited on page 61).

[Sch+14] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang and P. Westling. **High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth**. In: *German Conference on Pattern Recognition (GCPR)*. Vol. 8753. 2014, pp. 31–42 (Cited on page 56).

[Sch+09] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier and F. Lenzen. **Variational methods in imaging**. Springer, 2009 (Cited on page 3).

[Sch76] M. Schlesinger. **Sintaksicheskiy analiz dvumernykh zritelnikh signalov v usloviyakh pomekh (Syntactic analysis of two-dimensional visual signals in noisy conditions)**. In: *Kibernetika* 4 (1976), pp. 113–130 (Cited on pages 44, 77).

[Sch92] J. Schmidhuber. **Learning factorial codes by predictability minimization**. In: *Neural Computation* 4.6 (1992), pp. 863–879 (Cited on page 113).

[SC10] T. Schoenemann and D. Cremers. **A combinatorial solution for model-based image segmentation and real-time tracking**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 32.7 (2010), pp. 1153–1164 (Cited on pages 15, 93).

[SKC09] T. Schoenemann, F. Kahl and D. Cremers. **Curvature regularity for region-based image segmentation and inpainting: A linear programming relaxation**. In: *International Conference on Computer Vision (ICCV)*. IEEE. 2009 (Cited on page 134).

[Sch51] L. Schwartz. **Théorie des distributions I, II**. Vol. 1245, 1122. Paris: Hermann, 1951 (Cited on page 112).

[Seg+18] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet and M. Blondel. **Large-scale optimal transport and mapping estimation**. In: *International Conference on Learning Representations (ICLR)*. 2018 (Cited on pages 94, 139).

[SKH08] A. Shekhovtsov, I. Kovtun and V. Hlaváč. **Efficient MRF deformation model for non-rigid image matching**. In: *Computer Vision and Image Understanding (CVIU)* 112.1 (2008), pp. 91–99 (Cited on page 93).

[Shl76] M. Shlezinger. **Syntactic analysis of two-dimensional visual signals in the presence of noise**. In: *Cybernetics* 12.4 (1976), pp. 612–628 (Cited on page 15).

[Sim+98] P. Y. Simard, Y. A. LeCun, J. S. Denker and B. Victorri. **Transformation invariance in pattern recognition – tangent distance and tangent propagation**. In: *Neural networks: tricks of the trade*. 1998 (Cited on pages 113, 122).

[Sim+92] P. Simard, B. Victorri, Y. LeCun and J. Denker. **Tangent prop – a formalism for specifying selected invariances in an adaptive network**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 1992 (Cited on pages 113, 122).

[Smi93] S. K. Smirnov. **Decomposition of solenoidal vector charges into elementary solenoids, and the structure of normal one-dimensional flows**. In: *Algebra i Analiz* 5.4 (1993), pp. 206–238 (Cited on page 135).

[SSC18] Y. Soliman, D. Slepčev and K. Crane. **Optimal cone singularities for conformal flattening**. In: *ACM Transactions on Graphics (TOG)* 37.4 (2018), pp. 1–17 (Cited on page 139).

[Sol+16] J. Solomon, G. Peyré, V. G. Kim and S. Sra. **Entropic metric alignment for correspondence problems**. In: *ACM Transactions on Graphics (TOG)* 35.4 (2016), p. 72 (Cited on page 95).

[Sri+12] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet et al. **On the empirical estimation of integral probability metrics**. In: *Electronic Journal of Statistics* 6 (2012), pp. 1550–1599 (Cited on page 112).

[SPC09] F. Steinbrücker, T. Pock and D. Cremers. **Large displacement optical flow computation without warping**. In: *International Conference on Computer Vision (ICCV)*. 2009 (Cited on page 19).

[SW14] M. Storath and A. Weinmann. **Fast partitioning of vector-valued images**. In: *SIAM J. Imaging Sci.* 7.3 (2014), pp. 1826–1852 (Cited on pages 9, 16).

[Str83] G. Strang. **Maximal flow through a domain**. In: *Mathematical Programming* 26.2 (1983), pp. 123–143 (Cited on page 21).

[Str09] G. Strang. **Maximum flows and minimum cuts in the plane**. In: *Advances in Applied Mathematics and Global Optimization*. Springer, 2009, pp. 1–11 (Cited on page 21).

[SG18] M. Strecke and B. Goldluecke. **Sublabel-accurate Convex Relaxation with Total Generalized Variation Regularization**. In: *German Conference on Pattern Recognition (GCPR)*. 2018 (Cited on pages 95, 134).

[SCC12] E. Strekalovskiy, A. Chambolle and D. Cremers. **A Convex Representation for the Vectorial Mumford-Shah Functional**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012 (Cited on pages 21, 22, 75, 77, 78, 80, 88–91, 95).

[SCC14] E. Strekalovskiy, A. Chambolle and D. Cremers. **Convex Relaxation of Vectorial Problems with Coupled Regularization**. In: *SIAM J. Imaging Sci.* 7.1 (2014), pp. 294–336 (Cited on pages 21, 24, 51, 63, 69, 77, 95, 136).

[SC14] E. Strekalovskiy and D. Cremers. **Real-Time Minimization of the Piecewise Smooth Mumford-Shah Functional**. In: *European Conference on Computer Vision (ECCV)*. 2014 (Cited on page 9).

[SGC10] J. Stühmer, S. Gumhold and D. Cremers. **Real-time dense geometry from a handheld camera**. In: *Joint Pattern Recognition Symposium*. 2010 (Cited on page 19).

[Sul90] J. M. Sullivan. **A crystalline approximation theorem for hypersurfaces**. PhD thesis. Princeton University, 1990 (Cited on page 20).

[Tsi95] J. N. Tsitsiklis. **Efficient algorithms for globally optimal trajectories**. In: *IEEE Transactions on Automatic Control* 40.9 (1995), pp. 1528–1538 (Cited on page 93).

[Ulb02] M. Ulbrich. **Nonsmooth Newton-like methods for variational inequalities and constrained optimization problems in function spaces**. PhD thesis. Habilitation thesis, Fakultät für Mathematik, Technische Universität München, 2002 (Cited on page 139).

[Ulb11] M. Ulbrich. **Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces**. Vol. 11. Society for Industrial and Applied Mathematics, 2011 (Cited on page 139).

[VG05] M. Vaillant and J. Glaunès. **Surface matching via currents**. In: *Biennial International Conference on Information Processing in Medical Imaging*. 2005 (Cited on page 113).

[Ves+17a] M. Vestner, Z. Lähner, A. Boyarski, O. Litany, R. Slossberg, T. Remez, E. Rodola, A. Bronstein, M. Bronstein, R. Kimmel and D. Cremers. **Efficient deformable shape correspondence via kernel matching**. In: *International Conference on 3D Vision (3DV)*. 2017 (Cited on page 95).

[Ves+17b] M. Vestner, R. Litman, E. Rodolà, A. M. Bronstein and D. Cremers. **Product Manifold Filter: Non-Rigid Shape Correspondence via Kernel Density Estimation in the Product Space.** In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (Cited on page 95).

[Vil08] C. Villani. **Optimal Transport: Old and New**. Springer, 2008 (Cited on page 94).

[Vix+10] K. R. Vixie, K. Clawson, T. J. Asaki, G. Sandine, S. P. Morgan and B. Price. **Multiscale flat norm signatures for shapes and images**. In: *Applied Mathematical Sciences* 4.14 (2010), pp. 667–680 (Cited on page 112).

[Vog20] T. Vogt. **Measure-Valued Variational Models with Applications in Image Processing**. PhD thesis. Institute of Mathematics and Image Computing, University of Lübeck, 2020 (Cited on page 134).

[VL19] T. Vogt and J. Lellmann. **Functional Liftings of Vectorial Variational Problems with Laplacian Regularization**. In: *arXiv:1904.00898* (2019) (Cited on pages 95, 134).

[Vog+19] T. Vogt, E. Strekalovskiy, D. Cremers and J. Lellmann. **Lifting methods for manifold-valued variational problems**. In: *arXiv:1908.03776* (2019) (Cited on pages 24, 134).

[VPT16] C. Vondrick, H. Pirsiavash and A. Torralba. **Generating videos with scene dynamics**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016 (Cited on pages 111, 124).

[WJ+08] M. J. Wainwright, M. I. Jordan et al. **Graphical models, exponential families, and variational inference**. In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305 (Cited on pages 3, 15).

[WG14] Y. Wald and A. Globerson. **Tightness Results for Local Consistency Relaxations in Continuous MRFs.** In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. 2014 (Cited on page 16).

[WSU14] S. Wang, A. Schwing and R. Urtasun. **Efficient inference of continuous Markov random fields with polynomial potentials**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2014 (Cited on page 16).

[WG76] R. C. Ward and L. J. Gray. **Eigensystem computation for skew-symmetric matrices and a class of symmetric matrices**. Tech. rep. Oak Ridge National Lab, 1976 (Cited on page 108).

[Wed+09] A. Wedel, T. Pock, C. Zach, H. Bischof and D. Cremers. **An improved algorithm for TV-L1 optical flow**. In: *Statistical and geometrical approaches to visual motion analysis*. Springer, 2009, pp. 23–45 (Cited on pages 8, 19).

[Wei+18] D. Wei, J. J. Lim, A. Zisserman and W. T. Freeman. **Learning and using the arrow of time**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (Cited on page 112).

[WF00] Y. Weiss and W. T. Freeman. **Correctness of belief propagation in Gaussian graphical models of arbitrary topology**. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2000 (Cited on page 15).

[Wer07] T. Werner. **A linear programming approach to max-sum problem: A review**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 29.7 (2007), pp. 1165–1179 (Cited on pages 15, 44, 77).

[Whi57] H. Whitney. **Geometric Integration Theory**. Princeton University Press, 1957 (Cited on pages 105, 112, 115).

[Wie+08] D. Wierstra, T. Schaul, J. Peters and J. Schmidhuber. **Natural evolution strategies**. In: *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*. IEEE. 2008, pp. 3381–3387 (Cited on page 12).

[Wil92] R. J. Williams. **Simple statistical gradient-following algorithms for connectionist reinforcement learning**. In: *Machine learning* 8.3-4 (1992), pp. 229–256 (Cited on page 12).

[WC16] T. Windheuser and D. Cremers. **A Convex Solution to Spatially-Regularized Correspondence Problems**. In: *European Conference on Computer Vision (ECCV)*. 2016 (Cited on pages 21, 77, 95, 108).

[Win+11] T. Windheuser, U. Schlickewei, F. R. Schmidt and D. Cremers. **Geometrically consistent elastic matching of 3D shapes: A linear programming solution**. In: *International Conference on Computer Vision (ICCV)*. 2011 (Cited on page 95).

[Xu+11] L. Xu, C. Lu, Y. Xu and J. Jia. **Image Smoothing via L_0 Gradient Minimization**. In: 2011 (Cited on page 9).

[Yam+12] K. Yamaguchi, T. Hazan, D. McAllester and R. Urtasun. **Continuous Markov random fields for robust stereo estimation**. In: *European Conference on Computer Vision (ECCV)*. 2012 (Cited on page 16).

[You80] L. C. Young. **Lectures on the Calculus of Variations and Optimal Control Theory**. Second edition. New York: Chelsea Publishing Company, 1980 (Cited on page 94).

[YBT10] J. Yuan, E. Bae and X.-C. Tai. **A study on continuous max-flow and min-cut approaches**. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010 (Cited on page 17).

[Yua+10] J. Yuan, E. Bae, X.-C. Tai and Y. Boykov. **A continuous max-flow approach to potts model**. In: *European Conference on Computer Vision (ECCV)*. 2010 (Cited on page 17).

[Zac13] C. Zach. **Dual decomposition for joint discrete-continuous optimization**. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2013 (Cited on pages 16, 44, 45, 63, 77).

[Zac+08] C. Zach, D. Gallup, J.-M. Frahm and M. Niethammer. **Fast Global Labeling for Real-Time Stereo Using Multiple Plane Sweeps.** In: *Proceedings of the Vision, Modeling and Visualization Workshop*. 2008 (Cited on pages 17, 18).

[ZH17] C. Zach and C. Häne. **Discretized convex relaxations for the piecewise smooth Mumford-Shah model**. In: *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. 2017, pp. 548–563 (Cited on page 136).

[ZHP13] C. Zach, C. Häne and M. Pollefeys. **What is optimized in convex relaxations for multilabel problems: Connecting discrete and continuously inspired MAP inference**. In: *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* 36.1 (2013), pp. 157–170 (Cited on pages 17, 44, 93).

[ZK12] C. Zach and P. Kohli. **A Convex Discrete-Continuous Approach for Markov random fields**. In: *European Conference on Computer Vision (ECCV)*. 2012 (Cited on pages 16, 24, 44, 45, 54, 55, 63, 77, 85, 86, 93, 138, 157).

[ZPB07] C. Zach, T. Pock and H. Bischof. **A duality based approach for realtime TV-L1 optical flow**. In: *Joint Pattern Recognition Symposium*. 2007, pp. 214–223 (Cited on page 19).

[ZB14] D. Zosso and A. Bustin. **A primal-dual projected gradient algorithm for efficient Beltrami regularization**. In: *Computer Vision and Image Understanding* (2014), pp. 14–52 (Cited on page 130).