FAKULTÄT FÜR INFORMATIK

DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Lower Envelopes and Lifting for Structured Nonconvex Optimization

Emanuel Laude

Technische Universität München

Fakultät für Informatik

Lower Envelopes and Lifting for Structured Nonconvex Optimization

Emanuel Laude

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. Michael G. Bader
Prüfer der Dissertation:	1. Prof. Dr. Daniel Cremers
	2. Prof. Dr. Peter Ochs Universität Tübingen

Die Dissertation wurde am 07.06.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 30.08.2021 angenommen.

Abstract

Many important problems in machine learning, signal processing and beyond can be cast as the minimization of an additive composite objective. However, due to their coupled structure, composite problems are notoriously difficult to optimize. As a remedy, this thesis considers two complementary approaches for decoupling in composite problems by lower relaxations, which allows one to solve the problem in a distributed fashion:

The first approach is based on component-wise infimal convolution or inf-projection wrt a suitable proximity measure. This yields a generally nonconvex but often smooth lower envelope to the original objective. For a squared distance, this yields the well-known Moreau envelope with its associated proximal mapping. A desirable property in this context is the single-valuedness and continuity of the proximal mapping, which is crucial to guarantee the smoothness of the Moreau envelope. A sufficient condition for this to hold locally is prox-regularity of functions due to Poliquin and Rockafellar, which is closely related to Federers concept of a set of positive reach. However, there exist simple even smooth functions which are not prox-regular. As a remedy, we propose to alter the Euclidean geometry in the proximal mapping and the definition of prox-regularity in a nonlinear way: More precisely we consider two different "nonlinear geometries" based on Legendre functions: The first one is based on an anisotropic generalization of the proximal mapping where the quadratic penalty is replaced by a Legendre function. The second geometry is obtained by replacing the squared distance with a Bregman distance. Both extensions are strict generalizations of the Euclidean case, while the gradient of the involved Legendre function can be seen as a certain nonlinear preconditioner. A major goal of this thesis is to study the single-valuedness and continuity of the non-Euclidean proximal mapping and the smoothness of the associated envelope function under non-Euclidean extensions of prox-regularity. The continuity and single-valuedness of the proximal mapping is leveraged in the analysis of inexact and stochastic alternating proximal point methods. In particular, a stochastic averaged proximal point method is derived and numerically applied to the problem of federated learning. We also discuss the application of alternating Bregman Proximal Point in semisupervised and transductive learning where the Euclidean geometry is altered by an entropic geometry which leads to the KL-divergence.

The second approach for decoupling is based on Lagrangian relaxations which is a convex relaxation method. Here, the focus is on MAP-inference in a continuous *Markov Random Field* (MRF) and spatially continuous, total variation regularized variational problems. Since direct Lagrangian relaxations of a nonconvex problem in general suffer from potentially large duality gaps, infinite-dimensional reformulations over the space of probability measures are considered. This can be seen as a certain lifting of the optimization variable to the space of Radon measures. As these programs are intractable, a family of semi-infinite dual programs is considered which is obtained by piecewise polynomial subspace approximations. In the primal this corresponds to a discretization of the optimization variable with moments. A geometric intuition of this lifting in the primal is provided and connections to relaxations by inf-projection are identified under the light of generalized conjugate functions. A special case is the component-wise convex envelope, which corresponds to the classical Lagrangian relaxation. A tractable cone programming formulation is derived using tools from convex algebraic geometry which is solved on a GPU using a concretization of a first-order primal-dual algorithm. Experimentally, the approach is applied to stereo matching, optical flow estimation and robust image denoising showing merits over standard discretizations which suffer from a sampling bias.

Zusammenfassung

Viele wichtige Probleme im maschinellen Lernen, in der Signalverarbeitung und darüber hinaus lassen sich als Minimierung einer additiv zusammengesetzten Kostenfunktion beschreiben. Aufgrund ihrer gekoppelten Struktur sind additiv zusammengesetzte Probleme jedoch bekanntermaßen schwer zu optimieren. Als Lösungsansatz werden in dieser Arbeit zwei komplementäre Ansätze zur Entkopplung in zusammengesetzten Problemen mittels unterer Relaxationen betrachtet, die es ermöglichen das Problem auf verteilte Weise zu lösen:

Der erste Ansatz basiert auf einer komponentenweisen Infimalfaltung oder Inf-Projektion mit einem geeigneten Abstandsmaß. Dies führt zu einer einer, im Allgemeinen, nicht-konvexen aber häufig glatten unteren Hüllkurve der ursprünglichen Kostenfunktion. Für ein quadratisches Abstandsmaß ergibt dies die bekannte Moreausche Hüllkurve mit dem dazugehörigen proximalen Operator oder kurz Prox-Operator. Eine wünschenswerte Eigenschaft in diesem Zusammenhang ist die Einelementigkeit und Stetigkeit des Prox-Operators, welche entscheidend ist, um die Glattheit der Moreauschen Hüllkurve zu gewährleisten. Eine hinreichende Bedingung, um dies lokal sicherzustellen, ist Prox-Regularität von Funktionen durch Poliquin und Rockafellar, die eng zu Federers Konzept einer Menge mit positivem Reach verwandt ist. Es gibt jedoch einfache sogar glatte Funktionen, die nicht prox-regulär sind. Als Lösungsansatz schlagen wir daher vor, die Euklidische Geometrie des Prox-Operators und der Definition von Prox-Regularität nichtlinear abzuändern: Genauer gesagt betrachten wir zwei verschiedene "nichtlineare Geometrien", die auf Legendre-Funktionen basieren: Die erste basiert auf einer anisotropen Verallgemeinerung des Prox-Operators, bei der der Quadratabstand durch eine Legendre-Funktion ersetzt wird. Die zweite Geometrie erhalten wir, indem wir den Quadratabstand durch einen Bregman-Abstand ersetzen. Beide Erweiterungen sind strikte Verallgemeinerungen des Euklidischen Falls, wobei der Gradient der beteiligten Legendre-Funktion als ein nichtlinearer Vorkonditionierer verstanden werden kann. Ein Hauptziel dieser Arbeit ist es, die Einelementigkeit und Stetigkeit des nicht-Euklidischen Prox-Operators und die Glattheit der zugehörigen Moreauschen Hüllkurvenfunktion unter nicht-Euklidischen Erweiterungen von Prox-Regularität zu untersuchen. Die Stetigkeit und Einelementigkeit des Prox-Operators wird in der Analyse von inexakten und stochastischen alternierenden Proximal Point Methoden ausgenutzt. Insbesondere wird eine stochastische alternierende Proximal Point Methode entwickelt und numerisch auf das Problem des federated learning angewendet. Ferner diskutieren wir auch die Anwendung des alternierenden Bregman Proximal Point Verfahrens im halbüberwachten und transduktiven Lernen, bei der die Euklidische Geometrie durch eine entropische Geometrie abgeändert wird, was zur KL-Divergenz führt.

Der zweite Ansatz zur Entkopplung basiert auf Lagrange-Relaxationen, einer konvexen Relaxationsmethode. Hier liegt der Fokus auf der MAP-Inferenz in einem kontinuierlichen Markov Random Field (MRF) und räumlich kontinuierlichen Variationsproblemen mit totaler Variationsregularisierung. Da direkte Lagrange-Relaxationen eines nicht-konvexen Problems im Allgemeinen zu großen Dualitätslücken führen, werden unendlich-dimensionale Umformulierungen über dem Raum der Wahrscheinlichkeitsmaßen betrachtet. Dies kann als ein Lifting der Optimierungsvariablen in den Raum der

Radonmaße angesehen werden. Da diese Programme nicht handhabbar sind, wird eine Familie von semi-infiniten dualen Programmen betrachtet, die sich durch stückweise polynomielle Dualraumapproximationen herleiten lassen. Im Primalen entspricht dies der Diskretisierung der Optimierungsvariablen mit Momenten. Dieses Lifting besitzt eine interessante geometrische Intuition im Primalen. Verbindungen zu Relaxationen durch Inf-Projektion werden im Lichte verallgemeinerter Konjugierter identifiziert. Ein Sonderfall ist die komponentenweise konvexe Hüllkurve, die der klassischen Lagrange-Relaxation entspricht. Mittels Werkzeugen aus der konvexen algebraischen Geometrie wird ein Cone Programm hergeleitet, welches wir auf einer GPU mittels einer Konkretisierung eines primal-dualen first-order Algorithmus lösen. Experimentell wird der Ansatz auf Stereomatching, optische Flussschätzung und robustes Bildentrauschen angewendet, wobei Vorteile gegenüber Standarddiskretisierungen gezeigt werden, die einen Sampling-Bias aufweisen.

Acknowledgments

First and foremost I would like to thank Daniel Cremers for accepting me as a PhD student in his chair. I would like to thank Daniel for his loyalty, his constant support and his trust as well as the opportunity to enjoy some academic freedom to explore some own, not always matured, ideas.

This thesis wouldn't have been possible without the help and input of other great people and researchers:

Therefore I also would like to thank Tao Wu for the collaborations on ADMM and infimal convolutions which form a substantial part of this thesis. During this journey Tao taught me many useful things related to nonconvex optimization which I am very grateful for. I also thank him for the great time we had on our journey to Japan.

I would like to thank Peter Ochs for the collaboration on Bregman proximal operators and beyond. In particular I would like to thank Peter for being available for (scientific) questions not always related to our collaborations and for the numerous insightful discussions. Thank you Peter also for your honesty. You are truly a professional.

I also would like to thank Thomas Möllenhoff for the uncountable, insightful and long discussions which have informed some directions in this thesis. Also thank you Thomas for the journey through the deep waters of functional lifting which started off during my master's thesis. Recently this journey was continued with a new crew member on board: Hartmut Bauermeister. I thank Hartmut, who was an indispensable part of the polynomial episode of the lifting journey, for the great collaboration. During this collaboration, which took place during the pandemic, Thomas, Hartmut and I had numerous scientifically insightful but also funny and sometimes philosophically deep zoom sessions, which I really enjoyed. Thank you guys. Very special thanks also goes to Jan Lellmann for his indispensable input in the collaboration on sublabel accurate lifting. Without Jan, who is a true lifting expert this project wouldn't have been possible. Also thanks to Michael Möller for the collaboration on sublabel accurate lifting.

I would like to thank Matthias Vestner for his support in a mathematical problem in one of my projects. Also thank you Matthias for our bike tours which I really enjoyed.

I also thank my former roommate Mo for the great times in Munich and Robert for his artistic and entertaining moonwalk performances in front of my office door. Thank you Robert. Those were fun times. :)

These funny moments and the harmonic atmosphere in Daniels group with its great members (a shoutout goes to all of you guys) made my stay in Munich really enjoyable and helped to overcome phases of frustration or lack of motivation.

I also would like to thank Mahesh, Jan-Hendrik, Csaba, Laura, Frank Schmidt, Björn Andres, Niko and Maolin for the great scientific discussions and other collaborations.

Also many thanks to Thomas Vogt for highly insightful discussions.

Finally, a big shoutout goes to my girlfriend Shantha and my family. Thank you so much for your patience and your warm support. Without you this journey wouldn't have been possible.

Contents

Chapter 1.

Introduction

1.1. Lower envelopes and lifting in optimization

1.1.1. Composite optimization problems

Continuous optimization lies at the heart of machine learning and signal processing and many other fields. In fact, many important problems in these areas and beyond can be cast as the minimization of a cost function:

$$\min_{x \in X} \; f(x). \tag{1.1}$$

The idea is to set up a cost function $f : X \to \mathbb{R} \cup \{\infty\}$ that assigns high values (including infinity "∞") to points x that are undesirable solutions and low cost to points that correspond to preferable solutions. Via the choice $f(x) = \infty$ one can exclude points from the feasible set, i.e., the set of solution candidates, which is called a constraint in optimization.

For example, in image processing one may ask for a denoised version of a noisy input image, which is not too far from the input image but contains less noise. This can be cast as an optimization problem where the optimization variable x corresponds to the unknown output image and the cost function assigns small values to images x which are close to the input image and are less noisy. In many applications these cost functions do not come out of nowhere but can often be derived through a Bayesian or physics based approach, which, however, is not the focus of this thesis.

In logistics, given a list of cities and the pairwise distances between them, one may ask for the shortest possible route that visits all cities exactly once and returns to the origin. This is called the *travelling salesman problem* (TSP). There, the variable x encodes a possible route. The cost function returns the length of a feasible route x while it returns the cost ∞ if x is not a roundtrip or if x is not a trip that visits each city exactly once.

In aircraft design one seeks to find an airfoil x that minimizes drag under a minimal lift constraint.

These, at first sight, very different optimization problems shall illustrate what many optimization problems have in common and what makes them difficult: The challenge is to balance *competing goals* which are often easy to achieve when considered separately. In image denoising the goal to find an image which is close to the input image is trivial: It is the input image. The image which among all possible images contains least noise is the constant image. Both results represent the extreme cases in a coupled problem which is to find a tradeoff between the two. In the travelling salesman problem the situation is considerably easier if one drops the "visit exactly once"-constraint in the formulation.

Therefore, the true challenge in many optimization problems is to find a consensus or tradeoff between these competing goals.

This is also reflected in the structure of the cost functions which in many cases amounts to an additive composition of two or more simple terms f_i, which represent the individual goals, and positive weights π_i:

$$\min_{x \in X} \sum_{i=1}^{N} \pi_i f_i(x). \tag{1.2}$$

Such an additive composite objective can therefore also be regarded as a linearly scalarized multiobjective program, whose minima are *Pareto* optimal in the sense that one objective f_i cannot be improved without degrading at least one of the other objectives f_j, $i \neq j$. While in true multiobjective optimization one often seeks to characterize the set of Pareto optimal solutions the focus in this thesis is purely on strategies for minimizing a single composite objective with known weights π_i.

1.1.2. Relaxation by inf-projection

The coupled structure in composite optimization problems introduces a major challenge for efficient optimization. In some situations, however, the coupling can be relaxed: This is very common in so-called feasibility problems: In a feasibility problem one seeks to find a point x in the intersection $A \cap B$ of two sets $A \subset X$ and $B \subset X$, which can be thought of as competing constraints. Feasibility problems and optimization problems are actually closely related to each other: The travelling salesman problem, for instance, is routinely formulated in terms of a decision problem in theoretical computer science: There, in addition to the list of cities and the list of pairwise distances, one is given a cost parameter C, and one asks whether there exists a roundtrip that visits all cities exactly once with length at most C. Once we can solve the decision variant the corresponding optimization problem is solved too, e.g., via a binary search over the cost parameter C. To formulate TSP in terms of a feasibility problem one can define A as the set of all roundtrips with cost at most C and the set B is the set of paths (not necessarily round trips) that visit each city exactly once. To formulate an optimization problem in the form of Problem (1.2) one introduces corresponding indicator functions $f_1 = \iota_A$ and $f_2 = \iota_B$. The indicator function ι_A attains the value ∞ at points x which lie outside the set and 0 whenever $x \in A$. A common strategy to untangle the coupling in feasibility problems is to relax one of the indicator functions by means of a (squared) distance function $\operatorname{dist}^2(x, A) = \inf_{x' \in A} \|x' - x\|^2 = \inf_{x' \in X} \|x' - x\|^2 + \iota_A(x')$. A distance function to a set $A \subset X$ can actually be regarded as a certain *lower envelope* of the indicator function of the set: In contrast to the indicator function, which is ∞ outside the set, the distance function attains finite values everywhere while it coincides with the indicator function inside the set. The relaxed optimization problem reads

$$\min_{x \in X} \operatorname{dist}^2(x, A) + \iota_B(x).$$

It is easy to see, that whenever there exists a point in the intersection $x \in A \cap B$ the set of global solutions to the relaxation equals the set of global solutions to the original problem (1.2). Substituting the definition of the distance function one obtains

$$\min_{x \in X} \min_{x' \in X} \iota_A(x') + \|x' - x\|^2 + \iota_B(x),$$

where x', x are copies associated to the individual constraints ι_A and ι_B and $\|x' - x\|^2$ is a relaxation of the equality constraint $x' = x$. This problem is certainly less coupled and therefore a viable approach to find a solution is to minimize the cost function in an alternating fashion wrt x and x' while one or the other variable is fixed. The minimizations wrt x and x' are the projections onto the sets A and B which are typically simple problems. Indeed, this is the alternating projections algorithm which dates back to John von Neumann [VN50].

Relaxations of constraints by means of the squared distances of the feasible set is actually based upon a more general approach to obtain (lower) envelopes in optimization, called infimal convolution:

By definition, the infimal convolution $f \oplus g$ of a function f with a kernel function g at a point y is the infimum over all additive decompositions of the input $y = x + z$:

$$(f \oplus g)(y) = \inf_{\substack{z,x \in X,\\ y=x+z}} f(x) + g(z) = \inf_{x \in X} f(x) + g(y - x). \tag{1.3}$$

This translates to the following intuitive geometric interpretation of the epigraph of $f \oplus g$ in terms of the Minkowski sum of the individual epigraphs: $\operatorname{epi}(f \oplus g) = \operatorname{epi} f + \operatorname{epi} g$, where $\operatorname{epi} f = \{(x, \alpha) \in X \times \mathbb{R} : f(x) \leq \alpha\}$. Therefore, in an epigraphical sense, the infimal convolution of f with g is the "dilation" of the epigraph of f with the epigraph of g. One recovers the squared distance function $\operatorname{dist}^2(\cdot, A) = f \oplus g$ for the choice $f = \delta_A$ and $g = \|\cdot\|^2$.

The infimal convolution can be further generalized under the concept of inf-projection,

$$(f \triangle g)(y) = \inf_{x \in X} f(x) + g(x, y), \tag{1.4}$$

where g is a coupling function. The epigraph of the inf-projection enjoys a rich geometric intuition too: If the infimum above is attained when finite, then $\operatorname{epi}(f \triangle g)$ is the image of the epigraph of $(x, y) \mapsto f(x) + g(x, y)$ under the projection $(x, y, \alpha) \mapsto (y, \alpha)$, see [RW98, Proposition 1.18].

To obtain a meaningful relaxation and in particular a *lower* approximation to the original function, the coupling term g must be a *proximity measure* or *discrepancy measure*: This means we request that $g(x, y) \geq 0$ and $g(x, y) = 0$ if and only if $x = y$. Then it is easy to see that $\inf f = \inf f \triangle g$ while $f \triangle g$ is a lower approximation to f, i.e., $f \triangle g \leq f$. An obvious choice for $g(x, y)$ in that regards is

$$g(x, y) = \frac{1}{2\lambda}\|x - y\|^2,$$

for some $\lambda > 0$, which gives rise to the well known Moreau envelope $e_\lambda f = f \triangle g$.

Among other things, this thesis is about decoupling additive composite problems by inf-projection, where f_i are not necessarily indicator functions but more general extended real-valued functions, i.e., functions that attain values in the extended real line $\mathbb{R} \cup \{\infty\}$. In this thesis we will consider the problem of federated learning [McM+17], for which this approach is particularly well-suited: In federated learning the goal is to train a machine learning model in a collaborative fashion by a set of clients to which a private set of training data is associated. The problem can be cast as the minimization of a weighted finite sum $\sum_{i=1}^N \pi_i f_i$ of private empirical risks f_i, which are associated to the clients. The weights π_i are positive and sum to 1. To solve the problem in a distributed fashion with low communication between the clients one needs to decouple the problem. To this end copies x_i of the parameters of the model are introduced for each client and the discrepancy between them is penalized with a quadratic distance. Then, the clients attempt to solve the associated proximal mappings, i.e., the minimization of the sum of the associated partial empirical risk and a quadratic damping term which softly enforces consensus between the clients. Due to separability this can be carried out in parallel. Similar algorithms have been considered before in the context of federated and distributed learning [ZCL15; Li+20; Bor+21], however, without exploring relations to Moreau envelopes and feasibility problems, which opens up new perspectives and techniques for their analysis: Indeed, these algorithms attempt to solve a lower relaxation which is equivalent to replacing the individual risks with the corresponding Moreau envelopes and corresponds to the lower relaxation $\sum_{i=1}^N \pi_i e_\lambda f_i \leq \sum_{i=1}^N \pi_i f_i$. Since modern deep learning models are often over-parameterized, the intersection $\bigcap_{i=1}^N \operatorname{arg\,min} f_i \neq \emptyset$ is nonempty. This resembles the situation of a nonempty intersection of sets in the feasibility problem. Indeed, if this condition holds, it is easy to see that the relaxation is exact in the sense that the sets of global minimizers of $\sum_{i=1}^N \pi_i e_\lambda f_i$ and $\sum_{i=1}^N \pi_i f_i$ both coincide with $\bigcap_{i=1}^N \operatorname{arg\,min} f_i$. In this context the study of the associated parametric minimization problem or solution mapping

$$y \mapsto \operatorname*{arg\,min}_{x \in \mathbb{R}^m} f(x) + g(x, y),$$

is particularly useful, which, for the quadratic choice of g, specializes to the well-known proximal mapping. If g is a proximity measure, global minima of f are fixed points of the solution mapping. However, in general the converse is false: Not every fixed point of the solution mapping is a minimizer to f. The single-valuedness of this mapping, i.e., the existence and uniqueness of the minimizer of the parametric minimization problem is crucial to achieve another important goal in infimal convolutions: In relaxations by inf-projection a desirable property is the smoothness of the resulting function even if the original function is nonsmooth: This can be achieved if, among other properties such as single-valuedness of the associated solution mapping, the coupling functional g is smooth itself. Then the envelope function $f \mathbin{\triangle} g$ eventually inherits the smoothness from g. In the context of optimization this allows one to apply fast algorithms for smooth problems such as Newton's method or inertial type first-order algorithms. Indeed, smoothing approaches for nonsmooth problems are classical in optimization: Typically, there is an additional parameter λ to control the degree of smoothing, which is decreased within an outer loop. Therefore, a desirable property in this context is that the lower envelope, in an appropriate topology, converges to the original cost function. Then, the ultimate goal is to ensure that any limit point of the sequence of iterates produced by such a nested loop smoothing method is a stationary point of the original function. However, in this thesis the focus rather is on the relaxation aspects induced by a fixed smoothing. Beyond optimization this is related to a more fundamental question: Given a very irregular original function one seeks to find a surrogate which enjoys favorable regularity properties, while preserving certain important characteristics of the original function. In the example from above one can formulate the goal as follows: We seek to find a surrogate to some original function which is smooth but whose global minima are near the minima of the original function. This leads us to draw a connection to integral convolution which is perhaps more familiar to many readers than infimal convolution and sometimes used to achieve similar goals: The integral convolution $f * g$ of some input function $f : \mathbb{R}^m \to \mathbb{R}$ with a convolution kernel $g : \mathbb{R}^m \to \mathbb{R}$ is defined by

$$(f * g)(y) = \int_{\mathbb{R}^m} f(x) g(x - y) \, \mathrm{d}x. \tag{1.5}$$

While integral and infimal convolution are different, there still exist a couple of remarkable algebraic similarities between both: Aside from associativity or commutativity, the integral convolution (extended to measures) of f with a Dirac measure δ_0 centered at 0 yields the original function, while the indicator function $g := \iota_{\{0\}}$ of the singleton $\{0\}$ can be regarded as the neutral element in the inf-convolution operation. In the same way the Fourier transform of the integral convolution $f * g$ is identical to the product of the Fourier transforms of the individual functions, the Legendre–Fenchel transform $(f \oplus g)^*$ of the inf-convolution $f \oplus g$ of the convex proper lower semicontinuous functions f and g is the sum $f^* + g^*$ of the Legendre–Fenchel transforms f^* and g^* of the individual functions f and g.

1.1.3. Convex relaxation by lifting and dual discretization

As we have seen, a possible strategy to relax the coupling in additive composite problems is to introduce auxiliary variables associated to the individual functions and to penalize the discrepancy between them with a squared distance. This corresponds to a component-wise Moreau envelope.

An alternative strategy for decoupling is to derive a Lagrangian relaxation to the problem which results in a convex lower relaxation. Lagrangian relaxations can be applied to the partially separable problem

$$\min_{x \in X^{\mathcal{V}}} f(x) + g(Ax), \qquad f(x) = \sum_{u \in \mathcal{V}} f_u(x_u), \tag{1.6}$$

which is an additive composition of a separable part $f(x)$ and a coupling part $g(Ax)$ for a linear mapping $A : X^{\mathcal{V}} \to Y$ and an extended real-valued function $g : X \to \mathbb{R} \cup \{\infty\}$ and X, Y are Euclidean spaces and $\mathcal{V}$ is a finite index set.

It is instructive to discuss important specializations of $g(Ax)$ for different applications:

- For the choice

  $$g(x) = \iota_{\{x : x_u = x_v = \cdots\}}(x), \tag{1.7}$$

  being the indicator function of a subspace of $X^{\mathcal{V}}$ that enforces the individual components of x to be identical and $A = I$ the identity mapping one recovers the finite sum problem.

- For $Ax = \sum_{u \in \mathcal{V}} A_u x_u$ and linear mappings $A_u : X \to Y$ one recovers the sharing problem, see, e.g., [Boy+11, Section 7.3].

- Another important specialization is MAP-inference in a pairwise continuous *Markov Random Field* (MRF), which will be the main focus of this thesis. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. Then MAP-inference in a continuous MRF amounts to the following optimization problem:

  $$\min_{x \in X^{\mathcal{V}}} \sum_{u \in \mathcal{V}} f_u(x_u) + \sum_{uv \in \mathcal{E}} f_{uv}(x_u, x_v),$$

  for unary terms $f_u(x_u)$ and pairwise symmetric terms $f_{uv}(x_u, x_v)$, i.e., $f_{uv}(x_u, x_v) = f_{uv}(x_v, x_u)$. As the name suggests, the pairwise terms f_{uv} model pairwise relations between the variables x_u and x_v with $uv \in \mathcal{E}$. For instance, a typical choice is $f_{uv}(x_u, x_v) = |x_u - x_v|$, where the graph $\mathcal{G}$ represents a pixel grid. Then, the pairwise terms correspond to a *total variation*-like regularization which favors solutions $x \in X^{\mathcal{V}}$ that are spatially smooth. In that case, we choose $(Ax)_{uv} = (x_u, x_v)$ and $g(y) = \sum_{uv \in \mathcal{E}} f_{uv}(y_{uv})$.

- The pairwise MRF can also be generalized to a higher-order MRF where $g(Ax) = \sum_{W \in \mathcal{W}} f_W(x_W)$ models relations over subsets $\mathcal{W} \subset 2^{\mathcal{V}}$ of variables.

The Lagrangian relaxation to the problem (1.6) is obtained by introducing auxiliary variables $y \in Y$ and linear equality constraints $Ax = y$. Dualizing the linear constraints with Lagrange multipliers $\lambda \in Y^*$ one obtains the following Lagrangian dual problem:

$$\sup_{\lambda \in Y^*} \min_{x \in X^{\mathcal{V}}, y \in Y} \langle \lambda, Ax - y \rangle + f(x) + g(y),$$

which can be written in terms of the convex conjugates f^* and g^* of f and g:

$$\max_{\lambda \in Y^*} -f^*(-A^*\lambda) - g^*(\lambda). \tag{1.8}$$

This is a convex optimization problem which can potentially be solved by highly parallelizable convex optimization tools such as the *alternating direction method of multipliers* (ADMM). However, the approach often suffers form large duality gaps. Indeed, going back to the primal the Lagrangian relaxation is equivalent to:

$$\max_{\lambda \in Y^*} -f^*(-A^*\lambda) - g^*(\lambda) = \min_{x \in X^{\mathcal{V}}} \sum_{u \in \mathcal{V}} f_u^{**}(x_u) + g^{**}(Ax), \tag{1.9}$$

where f_u^{**} are the convex biconjugates, i.e., the largest lower semicontinuous convex functions below f_u. For nonconvex f_u such component-wise convex envelopes are often inaccurate approximations to the original objective and are trivial in some cases.

As a remedy we borrow a strategy which is common in combinatorial optimization: A convex relaxation is to reformulate the problem in terms of an *integer linear program* (ILP) and to relax the integrality constraint which yields a linear program. For instance in the TSP, one introduces a matrix x of binary values where the entry $x_{ij} \in \{0, 1\}$ encodes whether the connection between i^{th} and the j^{th} city is contained in the tour ($x_{ij} = 1$) or not ($x_{ij} = 0$). Then, the cost function becomes linear and the restriction to the set of feasible roundtrips can be formulated in terms of linear constraints. This is the Miller–Tucker–Zemlin formulation [MTZ60]. In the well-known Sudoku puzzle one is asked to fill a 9×9 grid with natural numbers between 1 and 9 such that each row, each column and each of the 9 subgrids with dimensions 3×3 contains all numbers between 1 and 9. The puzzle is set up with a partially completed grid, such that in many cases there exists a unique solution. In an ILP formulation of the puzzle the optimization variable x_{ij} for the cell in the i^{th} row and the j^{th} column is not a number between 1 and 9, which would correspond to direct optimization, but rather represented in terms of a binary *one-hot* or unit vector $x_{ij} \in \{0, 1\}^9$ that satisfies a sum-to-one constraint. After relaxation of the integrality constraint the variable for each cell can be interpreted as a probability vector of the discrete space $\{1, 2, \ldots, 9\}$.

To adopt the ILP approach for continuous optimization we restrict ourselves to the MAP-inference problem in a pairwise continuous MRF. The problem has a counterpart in the discrete setting where X is a finite set of labels. Then the problem is called MAP-inference in a (discrete) MRF, which is routinely formulated in terms of the linear *local marginal polytope relaxation*. In the continuous case the local marginal polytope relaxation is the following infinite-dimensional linear program:

$$\min_{\mu \in \mathcal{P}(X)^{\mathcal{V}}} \sum_{u \in \mathcal{V}} \int_X f_u(x) \, \mathrm{d}\mu(x) + \sum_{uv \in \mathcal{E}} \mathrm{OT}_{f_{uv}}(\mu_u, \mu_v), \tag{1.10}$$

where $\mathcal{P}(X)$ is the space of probability measures over X and $\mathrm{OT}_{f_{uv}}$ is the optimal transportation between μ_u and μ_v with cost f_{uv}. Note that in that case X need not even be Euclidean but can rather be a manifold.

A strategy to derive a Lagrangian relaxation is to dualize the infinite-dimensional marginalization constraints $M_u\mu_{uv} = \mu_v$ and $M_v\mu_{uv} = \mu_u$ in the optimal transportation, see Equation (5.6), with Lagrange multipliers $\lambda_{uv} \in \mathcal{C}(X) \times \mathcal{C}(X)$, which are continuous functions. This dual program compares favorably to the classical Lagrangian relaxation: In particular it doesn't suffer from a duality gap since the local marginal polytope relaxation is convex. However, the program is infinite-dimensional and therefore intractable. As a remedy, we consider families of semi-infinite programs obtained by subspace approximations. Using tools from convex algebraic geometry [BPT12] the semi-infinite program can be reduced to a finite one. If the Lagrange multipliers are restricted to the space of affine functions, one recovers the classical Lagrangian relaxation (1.9), while in the more general case, if, e.g. the Lagrange multipliers are quadratic functions, one obtains a tighter convex relaxation. Indeed, if one considers a polynomial hierarchy of dual subspaces with increasing degree, the duality gap between the infinite-dimensional LP-relaxation and the discretized dual problem will eventually vanish in the limit.

1.1.4. Relaxation by inf-projection vs. relaxation by lifting: A generalized conjugacy perspective

As we have seen, decoupling by inf-projection and dual decompostion (plus lifting) are two disctinct approaches to derive *lower envelopes* for composite optimization problems. Both decoupling stategies have in common that the resulting lower envelopes involve additive compositions of elementary lower envelopes, the Moreau envelope or the convex envelope. Both of which are intractable if applied to the composite objective.

Another connection can be observed under the light of *generalized conjugate functions* [RW98, Chapter 11L*]. As we have seen, one possible strategy to obtain an elementary lower envelope of a function f is to consider the inf-projection $(f \bigtriangleup g)(y) = \inf_{x \in X} f(x) + g(x, y)$ wrt to a proximity measure g. An important special case is the well-known Moreau envelope.

A complementary approach to obtain elementary lower envelopes is by pointwise lower approximation with a parametric family of functions, which specializes to the convex envelope in dual decomposition approaches: Given a familiy of functions the lower envelope of f wrt that family is defined as the largest function below f which, up to constant translation, can be written in terms of a pointwise supremum over elements in the family. Specializing to the family of linear functions, one obtains the convex biconjugate, which is the largest lower semicontinuous lower convex envelope of f. More generally, this can be expressed within the framework of generalized conjugacy: Given a coupling function $\Phi : X \times Y \to \mathbb{R} \cup \{+\infty, -\infty\}$ one defines the Φ-conjugate as $f^{\Phi}(y) = \sup_{x \in X} \Phi(x, y) - f(x)$ and the Φ-biconjugate as $f^{\Phi\Phi}(x) = \sup_{y \in Y} \Phi(x, y) - f^{\Phi}(y)$. Specializing $X = Y = \mathbb{R}^m$ and $\Phi(x, y) = \langle x, y \rangle$ one recovers the classical convex conjugate and convex biconjugate. In the general case, the Φ-biconjugate is the pointwise supremum over all elementary functions $(y, \beta) \mapsto \Phi(x, y) - \beta$ such that $\Phi(x, y) - \beta$ is majorized by f and (y, β) is the parameter element. Any function that can be expressed as such a pointwise supremum is called a Φ-envelope. This also sets up a connection to the envelopes obtained via inf-projection, where the Moreau envelope rather takes the role of a generalized conjugate function: Indeed, if $\Phi(x, y) = -g(x, y)$ the Φ-conjugate is the negative inf-projection $f^{\Phi}(y) = -(f \bigtriangleup g)(y)$ wrt g. In particular, if

$$\Phi(x, y) = -g(x, y) = -\frac{1}{2\lambda}\|x - y\|^2,$$

the Φ-conjugate $f^{\Phi}(y) = -e_{\lambda}f(y)$ is the negative Moreau envelope. Then, the corresponding Φ-biconjugate is the largest function below f that can be written as a pointwise supremum over concave quadratics with uniform curvature $1/\lambda$. This is also called the *proximal transform* [RW98, Example 11.64]. Equivalently, $f^{\Phi\Phi}$ is the largest lower semicontinuous $1/\lambda$-hypoconvex function below f in the same way the convex biconjugate is the largest convex lower semicontinuous function below f. We say h is r-hypoconvex if $h + (r/2)\|\cdot\|^2$ is convex. Such functions are also called r-semiconvex. This shows that all lower semicontinuous $1/\lambda$-hypoconvex functions are Φ-envelopes whose dual representation amounts to the negative Moreau envelope. Indeed, such functions can be recovered from their Moreau envelopes by inf-deconvolution.

Φ-conjugacy also appears in dual discretization plus lifting for Lagrangian relaxations which was discussed above. As we have seen, if the Lagrange multipliers are restricted to an affine subspace, one recovers the classical Lagrangian relaxation and the dual problem involves the convex conjugates f_u^* for each vertex $u \in \mathcal{V}$. More generally, the dual problem involves the Φ-conjugates f_u^{Φ} instead of the convex conjugates f_u^*. In the infinite-dimensional LP formulation, the function $\Phi(x, \lambda) = \lambda(x)$ couples the Lagrange multiplier with a point $x \in X$ via point evaluation. If $X = \mathbb{R}^m$ and the Lagrange multipliers are quadratic functions, rather than affine functions, one has $\Phi(x, p) = \langle p_1, x\rangle + p_2\|x\|^2$, where p denotes the vector of coefficients of λ. Therefore, the Φ-biconjugate of f_u is the largest function below f_u which can be written (up to constant translation) as a pointwise supremum over quadratic functions parametrized by coefficients p. This is the *basic quadratic transform* [RW98, Example 11.66]. In contrast to the proximal transform in which the curvature of the lower supporting quadratics is fixed, in the basic quadratic transform the pointwise supremum is taken over all concave lower quadratics. Therefore one can show that the basic quadratic transform of any function f (which is bounded from below by a concave quadratic) yields the lower semicontinuous closure of f, see [RW98, Example 11.66], while the proximal transform only yields the largest lower semicontinuous $1/\lambda$-hypoconvex under-approximation.

Overall, Φ-conjugacy and lifting are helpful tools to understand what happens in the primal if one discretizes the dual program of the infinite-dimensional LP-relaxation for MRF-inference.

1.2. Preliminaries

1.2.1. Notation

The following sections shall clarify the notation and introduce some required basic concepts such as subdifferential calculus and basics of measure theory as well as Fenchel–Rockafellar duality in infinite dimensions. The section can be skipped and rather used as a reference when concepts are unknown to the reader.

If not stated otherwise we adopt the notation from [RW98]. In particular we abbreviate lower semicontinuous by lsc. For a set $C \subset \mathbb{R}^m$ we denote by $\sigma_C(x) = \sup_{y \in C}\langle x, y\rangle$ the support function of C at $x \in \mathbb{R}^m$. For a subset $C \subset \mathbb{R}^m$ we denote by $C^* = \{y \in \mathbb{R}^m : \langle y, x\rangle \geq 0, \forall x \in C\}$ the dual cone of C. The convex hull $\operatorname{con} C$ of a set $C \subset \mathbb{R}^m$ is the smallest convex set that contains C. Equivalently, $\operatorname{con} C$ is the set of all finite convex combinations of points in C. With some abuse of notation we write $\operatorname{con} f$ for a function f to denote the largest convex function below f. If not stated otherwise, $\|\cdot\|$ refers to the Euclidean norm and $\langle\cdot,\cdot\rangle$ to the Euclidean inner product.

1.2.2. Extended real-valued functions and extended arithmetic

In the course of this thesis we work with improper (not proper) extended real-valued functions $f \colon \mathbb{R}^m \to \mathbb{R} \cup \{-\infty, +\infty\} =: \overline{\mathbb{R}}$. In accordance with [Mor66] we therefore have to extend the classical arithmetic on $\mathbb{R}$ to the extended real line $\overline{\mathbb{R}}$. We define upper and lower addition:

$$-\infty \dotplus \infty = \infty, \qquad -\infty \underset{\cdot}{+} \infty = -\infty, \tag{1.11}$$

and accordingly upper and lower subtraction:

$$\infty \overset{\cdot}{-} \infty = \infty, \qquad \infty \underset{\cdot}{-} \infty = -\infty. \tag{1.12}$$

A particularly important class of extended real-valued functions are indicator functions: For a set $C \subset \mathbb{R}^m$ we denote by $\iota_C : \mathbb{R}^m \to \overline{\mathbb{R}}$ the indicator function with

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

1.2.3. Subgradients, subdifferential calculus and Fermat's rule

We define regular, limiting and horizon subgradients according to [RW98, Definition 8.3]:

Definition 1.1 (subgradients)**.** *Consider a function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ and a point $\bar{x}$ with $f(\bar{x})$ finite. For a vector $v \in \mathbb{R}^m$, one says that*

(i) *v is a regular subgradient of f at $\bar{x}$, written $v \in \widehat{\partial} f(\bar{x})$, if*

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x}\rangle + o(\|x - \bar{x}\|).$$

(ii) *v is a limiting subgradient of f at $\bar{x}$, written $v \in \partial f(\bar{x})$, if there are sequences $x^\nu \to \bar{x}$ with $f(x^\nu) \to f(\bar{x})$ and $v^\nu \in \widehat{\partial} f(x^\nu)$ with $v^\nu \to v$.*

(iii) *v is a horizon subgradient of f at $\bar{x}$, written $v \in \partial^\infty f(\bar{x})$, if the same holds as in* (ii)*, except that instead of $v^\nu \to v$ one has $\lambda^\nu v^\nu \to v$ for some sequence $\lambda \to 0^+$ (or $v = 0$).*

Note that the definition of regular subgradients is parallel to the definition of a subgradient of a convex function, except for the error term $o(|x - \bar{x}|)$. If that error term specializes to a negative quadratic $-r/2|x - \bar{x}|$ we say that v is a proximal subgradient. For proper convex functions, indeed, $\partial f(\bar{x}) = \widehat{\partial} f(\bar{x})$, while $\partial^\infty f(\bar{x}) \subset N_{\operatorname{dom} f}(\bar{x})$. The inclusion holds with equality if, in addition, f is locally lsc at $\bar{x}$. Intuitively, $0 \neq v \in \partial^\infty f(\bar{x})$ is a horizon subgradient of f at $\bar{x}$ if f is infinitely steep at $\bar{x}$.

Subgradients enjoy a rich calculus. An important example is the sum-rule specialized from [RW98, Corollary 10.9]:

Lemma 1.2 (sum-rule of subdifferentials)**.** *Suppose $f = f_1 + \cdots + f_N$ for proper, lsc functions $f_i : \mathbb{R}^m \to \overline{\mathbb{R}}$, and let $\bar{x} \in \operatorname{dom} f$. Then*

$$\widehat{\partial} f(\bar{x}) \supset \widehat{\partial} f_1(\bar{x}) + \cdots + \widehat{\partial} f_N(\bar{x}).$$

Under the condition that the only combination of vectors $v_i \in \partial^\infty f_i(\bar{x})$ with $v_1+\cdots+v_N = 0$ is $v_1 = v_2 = \cdots = v_N = 0$ (this being true in the case of convex functions f_1, f_2 when $\operatorname{dom} f_1$ *and* $\operatorname{dom} f_2$ *cannot be separated), one also has that*

$$\begin{aligned}\partial f(\bar{x}) &\subset \partial f_1(\bar{x}) + \cdots + \partial f_N(\bar{x}),\\ \partial^\infty f(\bar{x}) &\subset \partial^\infty f_1(\bar{x}) + \cdots + \partial^\infty f_N(\bar{x}).\end{aligned}$$

If also each f_i is regular at $\bar{x}$, then f is regular at $\bar{x}$ and

$$\begin{aligned}\partial f(\bar{x}) &= \partial f_1(\bar{x}) + \cdots + \partial f_N(\bar{x}),\\ \partial^\infty f(\bar{x}) &= \partial^\infty f_1(\bar{x}) + \cdots + \partial^\infty f_N(\bar{x}).\end{aligned}$$

We specialize [RW98, Theorem 10.1], a generalization of Fermat's rule, which constitutes a *first order necessary local optimality condition* in the nonsmooth setting.

Lemma 1.3 (Fermat's rule generalized)**.** *If a proper function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ has a local minimum at $\bar{x}$, then*

$$\widehat{\partial} f(\bar{x}) \ni 0, \quad \partial f(\bar{x}) \ni 0,$$

where the first condition implies the second.

If f is subdifferentially regular or in particular convex, these conditions are equivalent. In the convex case they are not just necessary for a local minimum but sufficient for a global minimum, i.e., for having $\bar{x} \in \operatorname{arg\,min} f$.

If $f = f_0 + g$ with f_0 smooth, $\partial f(\bar{x}) \ni 0$ comes out as $-\nabla f_0(\bar{x}) \in \partial g(\bar{x})$.

1.2.4. Set-valued mappings and graphical localizations

In the course of this thesis we need the concept of a set valued mapping $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$. A set valued mapping maps individual points $x \in \mathbb{R}^m$ to subsets $F(x) \subset \mathbb{R}^n$ and therefore generalizes the notion of a classical function.

For a set-valued mapping $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ let

$$\operatorname{dom} F := \{x \in \mathbb{R}^m : F(x) \neq \emptyset\} \tag{1.13}$$

denote the domain of F.

An important example that is considered here is the set-valued subdifferential mapping $x \mapsto \partial f(x)$ which maps a point x to the set of (limiting) subgradients $\partial f(x)$ of f at x. By definition, subgradients only exist at points x where f is finite. Therefore we have the relation $\operatorname{dom} \partial f \subset \operatorname{dom} f$.

We define the graph of F as

$$\operatorname{gph} F := \{(x, v) \in \mathbb{R}^m \times \mathbb{R}^n : v \in F(x)\}. \tag{1.14}$$

We define the range of F as

$$\operatorname{rge} F := \{v \in \mathbb{R}^n : v \in F(x) \text{ for some } x \in \mathbb{R}^m\}. \tag{1.15}$$

There always exists an inverse of F, written $F^{-1} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ which is defined via

$$F^{-1}(v) := \{x \in \mathbb{R}^m : v \in F(x)\}. \tag{1.16}$$

We have the relation $\operatorname{dom} F = \operatorname{rge} F^{-1}$.

Let I denote the identity mapping with $I(x) = x$.

In the course of this thesis we work with graphical localizations of set-valued mappings, see [DR09], which are constructed graphically by intersecting the graph of F with some neighborhood of some reference point $(\bar{x}, \bar{v}) \in \operatorname{gph} F$:

Definition 1.4 (graphical localization). *For $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ and a pair $(x, v) \in \operatorname{gph} F$, a graphical localization of F at $\bar{x}$ for $\bar{v}$ is a set-valued mapping T such that $\operatorname{gph} T = (U \times V) \cap \operatorname{gph} F$ for some neighborhoods U of $\bar{x}$ and V of $\bar{v}$, so that*

$$T(x) := \begin{cases} F(x) \cap V & \text{if } x \in U, \\ \emptyset & \text{otherwise.} \end{cases} \tag{1.17}$$

The inverse of T then has

$$T^{-1}(v) := \begin{cases} F^{-1}(v) \cap U & \text{if } v \in V, \\ \emptyset & \text{otherwise,} \end{cases} \tag{1.18}$$

and is thus a graphical localization of the set-valued mapping F^{-1} at $\bar{v}$ for $\bar{x}$. By a single-valued localization of F at $\bar{x}$ for $\bar{v}$ will be meant a graphical localization that is a function, its domain not necessarily being a neighborhood of $\bar{x}$. The case where the domain is indeed a neighborhood of $\bar{x}$ will be indicated by referring to a single-valued localization of F around $\bar{x}$ for $\bar{v}$ instead of just at $\bar{x}$ for $\bar{v}$.

We state [RW98, Definition 1.33]:

Definition 1.5 (local semicontinuity). *A function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is locally lower semicontinuous (lsc) at $\bar{x}$, a point where $f(\bar{x})$ is finite, if there is an $\varepsilon > 0$ such that all sets of the form $\{x \in \mathbb{R}^m : \|x - \bar{x}\| \leq \varepsilon, f(x) \leq \alpha\}$ with $\alpha \leq f(\bar{x}) + \varepsilon$ are closed.*

In this thesis, a particularly important example of localizations of set-valued mappings are localizations of the subdifferential mapping of some locally lsc function f, where the neighborhoods around the reference point are taken in the f-attentive topology:

Definition 1.6 (f-attentive localization of limiting subdifferential). *Let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be locally lsc at $\bar{x}$, a point where $f(\bar{x})$ is finite. Let $\bar{v} \in \partial f(\bar{x})$. Then for some $\varepsilon > 0$ the f-attentive ε-localization $T : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ of ∂f at $\bar{x}$ for $\bar{v}$ is defined by*

$$T(x) := \begin{cases} \{v \in \partial f(x) : \|v - \bar{v}\| < \varepsilon\} & \text{if } \|x - \bar{x}\| < \varepsilon \text{ and } f(x) < f(\bar{x}) + \varepsilon, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{1.19}$$

The f-attentive localization of the limiting subdifferential coincides with the classical localization of the limiting subdifferential if nearness of points $\|x-\bar{x}\| < \varepsilon$ and subgradients $\|v - \bar{v}\| < \varepsilon$ implies nearness of function values $f(x) < f(\bar{x}) + \varepsilon$. This is guaranteed if the function f is in addition subdifferentially continuous [RW98, Definition 13.28]:

Definition 1.7 (subdifferential continuity). *A function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is called subdifferentially continuous at $\bar{x}$ for $\bar{v}$ if $\bar{v} \in \partial f(\bar{x})$ and, whenever $(x^\nu, v^\nu) \to (\bar{x}, \bar{v})$ with $v^\nu \in \partial f(x^\nu)$, one has $f(x^\nu) \to f(\bar{x})$. If this holds for all $\bar{v} \in \partial f(\bar{x})$, f is said to be subdifferentially continuous at $\bar{x}$.*

We state [RW98, Definition 12.1]:

Definition 1.8 (monotonicity)**.** *A set-valued mapping $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ is called* monotone *if it has the property*

$$\langle y_1 - y_2, x_1 - x_2 \rangle \geq 0,$$

whenever $y_1 \in F(x_1)$, $y_2 \in F(x_2)$, and strictly monotone *if the inequality is strict when $x_1 \neq x_2$.*

We state [RW98, Definition 12.53]:

Definition 1.9 (strong monotonicity)**.** *A set-valued mapping $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ is called* strongly monotone *if there is $\sigma > 0$ such that $F - \sigma I$ is monotone, or equivalently*

$$\langle y_1 - y_2, x_1 - x_2 \rangle \geq \sigma \|x_1 - x_2\|^2,$$

whenever $y_1 \in F(x_1)$, $y_2 \in F(x_2)$.

Lemma 1.10. *Let $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ be strictly monotone. Then $F^{-1} : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ is at most single-valued. If, in addition, F is σ-strongly monotone, F^{-1} is also $1/\sigma$-Lipschitz continuous on* $\operatorname{rge} F = \operatorname{dom} F^{-1}$*.*

Proof. Suppose F^{-1} is not at most single-valued. Then there exists $y \in \mathbb{R}^m$ with $x_1 \in F^{-1}(y)$ and $x_2 \in F^{-1}(y)$ with $x_1 \neq x_2$. This means $y = y_1 \in F(x_1)$ and $y = y_2 \in F(x_2)$, and therefore $\langle y_1 - y_2, x_1 - x_2 \rangle = 0$, which contradicts strict monotonicity.

Now let $y_1, y_2 \in \operatorname{dom} F^{-1} = \operatorname{rge} F$. We know that $y_1 \in F(x_1)$ and $y_2 \in F(x_2)$ for some $x_1, x_2 \in \mathbb{R}^m$ as well as $x_1 \in F^{-1}(y_1)$, $x_2 \in F^{-1}(y_2)$. Since F is σ-strongly monotone, by Cauchy–Schwarz we have:

$$\|y_1 - y_2\| \cdot \|x_1 - x_2\| \geq \langle y_1 - y_2, x_1 - x_2 \rangle \geq \sigma \|x_1 - x_2\|^2,$$

which implies

$$\|x_1 - x_2\| \leq \frac{1}{\sigma} \|y_1 - y_2\|. \qquad \square$$

We define the following notion of semicontinuity of set-valued mappings specialized from [RW98, Definition 5.4]:

Definition 1.11 (outer semicontinuity)**.** *A set-valued mapping $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ is* outer semicontinuous *(osc) at $\bar{x}$ if*

$$\{u \in \mathbb{R}^n : \exists x^\nu \to \bar{x}, u^\nu \to u \textit{ with } u^\nu \in F(x^\nu)\} \subset F(\bar{x}).$$

1.2.5. Parametric minimization

In certain parametric minimization problems, such as encountered in the definition of the proximal mapping, a sufficient condition for the outer semicontinuity Definition 1.11 of the solution mapping is uniform level boundedness [RW98, Definition 1.16].

We state the following pointwise version of [RW98, Definition 1.16] adopted from [Bau+09, Definition 4.1]:

Definition 1.12 (uniform level boundedness)**.** *We say a function $h : \mathbb{R}^m \times \mathbb{R}^n \to \overline{\mathbb{R}}$ with values $h(x, y)$ is level-bounded in x locally uniformly at $\bar{y} \in \mathbb{R}^n$ if for each $\alpha \in \mathbb{R}$ there is a neighborhood V of $\bar{y}$ along with a bounded set $X \subset \mathbb{R}^m$ such that*

$$\{x \in \mathbb{R}^m : h(x, y) \leq \alpha\} \subset X$$

for all $y \in V$. Equivalently, there is a neighborhood V of $\bar{y}$ such that the set

$$\big\{(x,y) \in \mathbb{R}^m \times V : h(x,y) \le \alpha\big\}$$

is bounded in $\mathbb{R}^m \times \mathbb{R}^n$. We say a function $h : \mathbb{R}^m \times \mathbb{R}^n \to \overline{\mathbb{R}}$ with values $h(x,y)$ is level-bounded in x locally uniformly in y if this holds for all $\bar{y} \in \mathbb{R}^n$.

We state [RW98, Theorem 1.17]:

Theorem 1.13 (parametric minimization)**.** *Consider*

$$p(y) := \inf_{x\in\mathbb{R}^m} h(x,y), \qquad P(y) := \operatorname*{arg\,min}_{x\in\mathbb{R}^m} h(x,y),$$

in the case of a proper, lsc function $h : \mathbb{R}^m \times \mathbb{R}^n \to \overline{\mathbb{R}}$ such that $h(x,y)$ is level-bounded in x locally uniformly in y.

(i) *The function p is proper and lsc on $\mathbb{R}^n$, and for each $y \in \operatorname{dom} p$ the set $P(y)$ is nonempty and compact, whereas $P(y) = \emptyset$ when $y \notin \operatorname{dom} p$.*

(ii) *If $x^\nu \in P(y^\nu)$, and if $y^\nu \to y$ in such a way that $p(y^\nu) \to p(\bar{y})$ (as when p is continuous at $\bar{y}$ relative to a set V containing $\bar{y}$ and y^ν), then the sequence $\{x^\nu\}_{\nu\in\mathbb{N}}$ is bounded, and all its cluster points lie in $P(\bar{y})$.*

(iii) *For p to be continuous at a point $\bar{y}$ relative to a set V containing $\bar{y}$, a sufficient condition is the existence of some $\bar{x} \in P(\bar{y})$ such that $h(\bar{x}, y)$ is continuous in y at $\bar{y}$ relative to V.*

We state the *implicit function theorem for generalized equations*, adopted from [DR09, Theorem 2B.7] which generalizes [Rob80], for analyzing the solution mappings in parametric minimization problems.

Theorem 1.14 (implicit function theorem for generalized equations)**.** *Consider a function $G : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ and a set-valued map $T : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ with $(\bar{x}, \bar{y}) \in \operatorname{int}(\operatorname{dom} G)$ and $0 \in G(\bar{x}, \bar{y}) + T(\bar{x})$, and suppose that*

$$\widehat{\operatorname{lip}}_y(G; (\bar{x}, \bar{y})) := \limsup_{\substack{y,y'\to\bar{y}\\ x\to\bar{x}\\ y\neq y'}} \frac{\|G(x,y) - G(x,y')\|}{\|y - y'\|} \le \gamma < \infty.$$

Let $H : \mathbb{R}^m \to \mathbb{R}^m$ be a strict estimator of G wrt x uniformly in y at $(\bar{x}, \bar{y})$ with constant μ, i.e.,

$$\widehat{\operatorname{lip}}_x(e; (\bar{x}, \bar{y})) \le \mu < \infty \quad \textit{for } e(x,y) = G(x,y) - H(x).$$

Suppose that $(H + T)^{-1}$ has a Lipschitz continuous single-valued localization around 0 *for $\bar{x}$ with constant κ and $\kappa\mu < 1$. Then the solution mapping*

$$S : y \in \mathbb{R}^m \mapsto \{x \in \mathbb{R}^m : 0 \in G(x,y) + T(x)\}$$

has a Lipschitz continuous single-valued localization around $\bar{y}$ for $\bar{x}$ with constant $\frac{\kappa\gamma}{1-\kappa\mu}$.

1.2.6. Legendre functions and Bregman distances

In this thesis we consider Legendre functions to generate discrepancy measures for generalized proximal operators. To some extent the following survey of results is adapted from [LOC20].

A Legendre function $\phi \in \Gamma_0(X)$ is defined according to [Roc70, Section 26]:

Definition 1.15 (Legendre function). *The function $\phi \in \Gamma_0(\mathbb{R}^m)$ is*

(i) essentially smooth*, if* $\operatorname{int}(\operatorname{dom}\phi) \neq \emptyset$ *and ϕ is differentiable on* $\operatorname{int}(\operatorname{dom}\phi)$ *such that $\|\nabla\phi(x^\nu)\| \to \infty$, whenever* $\operatorname{int}(\operatorname{dom}\phi) \ni x^\nu \to x \in \operatorname{bdry}\operatorname{dom}\phi$*, and*

(ii) essentially strictly convex*, if ϕ is strictly convex on every convex subset of* $\operatorname{dom}\partial\phi$*, and*

(iii) Legendre*, if ϕ is both essentially smooth and essentially strictly convex.*

We list some basic properties of Legendre functions:

Lemma 1.16. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. Then ϕ has the following properties:*

(i) $\operatorname{dom}\partial\phi = \operatorname{int}(\operatorname{dom}\phi)$*,* [Roc70, Theorem 26.1]*.*

(ii) *ϕ^* is Legendre,* [Roc70, Theorem 26.3]*.*

(iii) $\nabla\phi\colon \operatorname{int}(\operatorname{dom}\phi) \to \operatorname{int}(\operatorname{dom}\phi^*)$ *is a homeomorphism between* $\operatorname{int}(\operatorname{dom}\phi)$ *and* $\operatorname{int}(\operatorname{dom}\phi^*)$*, i.e., $\nabla\phi$ is bijective with inverse* $\nabla\phi^*\colon \operatorname{int}(\operatorname{dom}\phi^*) \to \operatorname{int}(\operatorname{dom}\phi)$ *and $\nabla\phi$ and $\nabla\phi^*$ are both continuous on* $\operatorname{int}(\operatorname{dom}\phi)$ *resp.* $\operatorname{int}(\operatorname{dom}\phi^*)$*,* [Roc70, Theorem 26.5]*.*

(iv) *ϕ is* super-coercive*, i.e., $\|\phi(w)\|/\|w\| \to \infty$ whenever $\|w\| \to \infty$, if and only if* $\operatorname{dom}\phi^* = \mathbb{R}^m$*,* [BB97, Proposition 2.16]*.*

Even though our main focus is on general Legendre functions, many classical Legendre functions satisfy the following additional property, which we adopt from [BL00, Definition 2.8]:

Definition 1.17 (very strictly convex functions). *Suppose $\phi \in \Gamma_0(\mathbb{R}^m)$ is $\mathcal{C}^2$ on* $\operatorname{int}(\operatorname{dom}\phi) \neq \emptyset$ *and $\nabla^2\phi(x)$ is positive definite for all $x \in$* $\operatorname{int}(\operatorname{dom}\phi)$*. Then we say ϕ is* very strictly convex*.*

Lemma 1.18. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and very strictly convex. Then ϕ^* is Legendre and very strictly convex. Moreover, for any conjugate pair $x \in$* $\operatorname{int}(\operatorname{dom}\phi)$ *and $\nabla\phi(x) \in$* $\operatorname{int}(\operatorname{dom}\phi^*)$ *the Hessian matrices $\nabla^2\phi(x)$ and $\nabla^2\phi^*(\nabla\phi(x))$ are inverse to each other.*

Proof. By Lemma 1.16 we know that $\phi^* \in \Gamma_0(\mathbb{R}^m)$ is Legendre. By assumption $\nabla\phi$ is continuously differentiable on $\operatorname{int}(\operatorname{dom}\phi)$ with derivative $\nabla^2\phi(x)$ invertible for any $x \in \operatorname{int}(\operatorname{dom}\phi)$. Thus, by the inverse function theorem for any $x \in \operatorname{int}(\operatorname{dom}\phi)$ there exist open neighborhoods V of x and U of $\nabla\phi(x)$ such that locally $(\nabla\phi)^{-1}\colon U \to V$ is continuously differentiable with derivative $\nabla((\nabla\phi)^{-1})(\nabla\phi(x)) = (\nabla^2\phi(x))^{-1}$. Since $(\nabla\phi)^{-1} = \nabla\phi^*$ and $(\nabla^2\phi(x))^{-1}$ is positive definite, the assertion follows. □

For examples of typical Legendre functions (e.g. Boltzmann–Shannon, Burg's or Fermi–Dirac entropy, Hellinger, Fractional Power) as well as their convex conjugates and derivatives we refer to [Bau+19, Example 2.2]. More examples can be found in [Bre67; Teb92; Eck93; BB97; BBC01]. In particular, we highlight that the Legendre function $\phi(x) = (1/p)|x|^p$, $p > 1$, is not very strictly convex for $p \neq 2$ and not even $\mathcal{C}^2$ if $p \in (1,2)$. The class of Legendre functions induces favorable properties for the following generalized distance-like measure.

Definition 1.19 (Bregman distance). *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. Then, the Bregman distance $D_\phi \colon \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$ generated by ϕ is defined by*

$$D_\phi(x,y) = \begin{cases} \phi(x) - \phi(y) - \langle \nabla\phi(y), x-y \rangle\,, & \text{if } y \in \operatorname{int}(\operatorname{dom}\phi), \\ +\infty, & \text{otherwise.} \end{cases} \tag{1.20}$$

Lemma 1.20. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. Then the following properties hold for the Bregman distance $D_\phi(\cdot,\cdot)$ induced by ϕ:*

(i) *For all $x \in \mathbb{R}^m$ and $y \in \operatorname{int}(\operatorname{dom}\phi)$ we have $D_\phi(x,y) = 0 \iff x = y$,* [BB97, Theorem 3.7(iv)].

(ii) *For all $x, y \in \operatorname{int}(\operatorname{dom}\phi)$ we have $D_\phi(x,y) = D_{\phi^*}(\nabla\phi(y), \nabla\phi(x))$,* [BB97, Theorem 3.7(v)].

(iii) *If ϕ is very strictly convex for any compact and convex $K \subset \operatorname{int}(\operatorname{dom}\phi)$ there exist positive scalars $+\infty > \Theta$ and $\theta > 0$ such that*

$$\frac{\theta}{2}\|x-y\|^2 \leq D_\phi(x,y), \qquad \|\nabla\phi(x) - \nabla\phi(y)\| \leq \Theta\|x-y\|,$$

for any $x, y \in K$, [BL00, Proposition 2.10].

1.2.7. Measures, convex functions and Fenchel–Rockafellar duality in infinite dimensions

In Chapter 5 we need the following basic measure theoretic notations as well as convexity and duality in infinite dimensions. For self-containedness we present a short survey of the required concepts which is based on the lecture notes of Bernhard Schmitzer[1]. For further details also see the references therein. In particular consult [AFP00] and [KZ06].

Definition 1.21 (σ-algebra). *A collection Σ of subsets of a set X is called a σ-algebra if:*

(i) $\emptyset \in \Sigma$ *and* $A \in \Sigma \implies X \setminus A \in \Sigma$

(ii) $A_n \in \Sigma \implies \bigcup_{n=0}^\infty A_n \in \Sigma$.

The σ-algebra is closed under countable unions and intersections. We call the elements of Σ *measureable sets* while we call the pair (X, Σ) measure space.

The *Borel* σ-algebra is the smallest σ-algebra containing all open sets of a topological space.

Definition 1.22 (nonnegative measures). *For a measure space (X,Σ) a function $\mu : \Sigma \to [0,+\infty]$ is called a* nonnegative measure *if*

(i) $\mu(\emptyset) = 0$

(ii) $A_n \in \Sigma$ *are pairwise disjoint* $\implies \mu(\bigcup_{n=0}^\infty A_n) = \sum_{n=0}^\infty \mu(A_n)$.

[1] `https://www.uni-muenster.de/AMM/num/Vorlesungen/OptTransp_SS17/ss2017_OTDataAnalysis_2017-05-02.pdf`

Definition 1.23 (total variation). *For a measure μ on (X, Σ) the total variation $|\mu|$ of $A \in \Sigma$ is*

$$|\mu|(A) = \sup \left\{ \sum_{n=0}^{\infty} |\mu(A_n)| : A_n \in \Sigma, \text{ pairwise disjoint, } \bigcup_{n=0}^{\infty} A_n = A \right\}.$$

Then $|\mu|$ is a finite, nonnegative measure on (X, Σ).

With some abuse of terminology, in Chapter 5, we will introduce the total variation $TV(u)$ of a function $u : (\Omega \subset \mathbb{R}^d) \to \mathbb{R}$, Ω nonempty, open and bounded, which for smooth u takes the form $\int_\Omega \|\nabla u(x)\| \,\mathrm{d}x$. More generally, for functions with bounded variation, $TV(u)$ is related to the measure theoretic total variation from above: $TV(u)$ is the measure theoretic total variation of the distributional derivative Du which is a Radon measure. For details see [AFP00].

Definition 1.24 (dual space). *For a normed vector space $(X, \|\cdot\|_X)$ its topological dual space is given by*

$$X^* = \{y : X \to \mathbb{R} : y \text{ linear, continuous, i.e., } \exists\, C < \infty : |y(x)| \leq C\|x\|_X, \forall x \in X\}.$$

This induces a norm on X^*:

$$\|y\|_{X^*} = \sup\{|y(x)| : x \in X, \|x\|_X \leq 1\}.$$

Then $(X^*, \|\cdot\|_{X^*})$ is a Banach space. In place of $y(x)$ we will write $\langle y, x\rangle$.

We have the following notions of convergence:

Definition 1.25 (weak convergence). *A sequence x^n in X converges weakly to $x \in X$ if $y(x^n) \to y(x)$ for all $y \in X^*$. We write $x^n \rightharpoonup x$.*

Definition 1.26 (weak* convergence). *A sequence y^n in X^* converges weakly to $y \in X^*$ if $y^n(x) \to y(x)$ for all $x \in X$. We write $y^n \overset{*}{\rightharpoonup} y$.*

Definition 1.27 (Radon measure). *Let (Ω, d) be a compact metric space and let Σ be the Borel σ-algebra. A finite measure (nonnegative or vector-valued) is called a* Radon measure. *We will write:*

- *$\mathcal{M}_+(\Omega)$ for the set of nonnegative Radon measures,*
- *$\mathcal{P}(\Omega) \subset \mathcal{M}_+(\Omega)$ for the set of Radon probability measures with total mass 1,*
- *$\mathcal{M}(\Omega)^m$ for the set of (vector-valued) Radon measures.*

Theorem 1.28 (duality). *Let (Ω, d) be a compact metric space. Let $\mathcal{C}(\Omega)$ be the space of continuous functions from Ω to $\mathbb{R}$, equipped with the* sup*-norm. Then, the topological dual space of $\mathcal{C}(\Omega)$ can be identified with the space $\mathcal{M}(\Omega)$ equipped with the total variation norm $\|\mu\|_{\mathcal{M}} = |\mu|(\Omega)$. In addition, we have a duality pairing for $\mu \in \mathcal{M}(\Omega)$ and $f \in \mathcal{C}(\Omega)$:*

$$\langle \mu, f\rangle = \int_\Omega f(x)\,\mathrm{d}\mu(x).$$

Two measures $\mu, \nu \in \mathcal{M}(\Omega)$ with $\mu(f) = \nu(f)$ for all $f \in \mathcal{C}(\Omega)$ coincide.

Theorem 1.29 (Banach–Alaoglu). *Let X be a separable normed space. Any bounded sequence in X^* has a weak* convergent subsequence.*

We have the following: Let (Ω, d) be a compact metric space, then any bounded sequence in $\mathcal{M}(\Omega)$ has a weak* convergent subsequence.

Definition 1.30 (topologically paired spaces). *Two vector spaces X, X^* with the locally convex Hausdorff topology are called topologically paired spaces if all continuous linear functionals on one space can be identified with all elements of the other.*

Example 1.31. Let (Ω, d) be a compact metric space. $\mathcal{C}(\Omega)$ and $\mathcal{M}(\Omega)$ with the sup-norm topology and the weak* topology are topologically paired spaces.

Definition 1.32 (Legendre–Fenchel conjugates). *Let X, X^* be topologically paired spaces. Let $f : X \to \mathbb{R} \cup \{+\infty\}$. Its Legendre–Fenchel conjugate $f^* : X^* \to \mathbb{R} \cup \{+\infty\}$ is given by $f^*(y) = \sup_{x \in X} \langle x, y\rangle - f(x)$. Likewise, for $g : X^* \to \mathbb{R} \cup \{+\infty\}$, we have $g^*(x) = \sup_{y \in X^*} \langle x, y\rangle - g(y)$.*

If f, g are convex, lsc we have $f = f^{**}$ and $g = g^{**}$. We state the Fenchel–Rockafellar duality theorem adapted from [Roc67]:

Theorem 1.33 (Fenchel–Rockafellar duality). *Let $(X, X^*), (Y, Y^*)$ be two pairs of topologically paired spaces. Let $f : X \to \mathbb{R} \cup \{+\infty\}, g : Y \to \mathbb{R} \cup \{+\infty\}$, f, g convex, $A : X \to Y$ linear, continuous. Assume there is $x \in X$ such that f is finite at x, and g finite and continuous at Ax. Then*

$$\inf\{f(x) + g(Ax) : x \in X\} = \max\{-f^*(-Az) - g^*(z) : z \in Y^*\}.$$

In particular a maximizer of the problem on the right exists. $A^* : Y^* \to X^*$ is the adjoint of A.

1.3. Outlook and Summary of Results

The results in this thesis are to a substantial extent based on own publications. All of which are collaborative works. As mentioned in the introduction, in this thesis we consider two complementary approaches to obtain relaxations to problems with composite structure based on lower envelopes. Part I considers relaxation by inf-projection and is based on [LWC18], [LWC19], [LOC20] and [LOC21]. Part II considers lifting to measures in Lagrangian relaxations and is based on [Lau+16] and [Bau+21]. Another major goal of this thesis is to identify connections between these two approaches under the light of the framework of generalized conjugacy.

Chapter 2 is based on [LWC18], [LWC19] and [LOC20]. In that chapter we study two complementary generalizations of the Euclidean proximal mapping and Moreau envelope in a nonconvex setting, which are based on Legendre functions: In Section 2.2 which expands upon [LWC19] we consider the anisotropic proximal mapping which is obtained by replacing the quadratic penalty $\|x - y\|^2$ in the proximal mapping with $\phi(x - y)$ for a certain Legendre function ϕ, see Definition 2.1. In Section 2.3, which is based on [LOC20] we consider the Bregman proximal mapping which is obtained when $\|x - y\|^2$ is replaced by a Bregman distance generated by a Legendre function. Since Bregman distances are asymmetric in general we consider a left and a right Bregman proximal mapping and Moreau envelope depending on the order of arguments. A major goal of these sections is to study the (local) single-valuedness and (Lipschitz) continuity of the proximal mapping and the smoothness of the Moreau envelope function along with a gradient formula. To this end we alter the geometry in Euclidean prox-regularity which leads to two distinct extensions thereof, which we call anisotropic and relative prox-regularity respectively. In Section 2.2.4 we also consider a small toy feasibility problem which uses anisotropic prox-potentials in place of standard quadratics. In Section 2.4, which expands upon [LWC18] we study (proximal mappings of) pointwise minima over a finite collection of functions. These functions are in particular upper-$\mathcal{C}^1$ but not prox-regular everywhere. A goal of this section is to identify a sufficient condition, based on the linear independence constraint qualification, that guarantees the gradient formula of the Moreau envelope to hold in terms of the limiting subdifferential.

The goal of the short Chapter 3 is to discuss and survey some existing results on generalized conjugate functions and the proximal transform [RW98, Example 11.64] as well as a recent generalization of the proximal average to nonconvex functions [CWP20]. The purpose of that chapter is to lay down the theoretical framework to identify connections between relaxation by inf-projection and relaxation by lifting to measures and dual discretization. Beyond these existing results we discuss an unknown (to our knowledge) duality relation between Proximal Point and gradient descent on a Lipschitz differentiable function invoking the proximal transform based on [LOC21].

The proximal average is revisited in Chapter 4. This chapter is based on [LOC20] and [LOC21] and partially on [LWC19]: Here we consider different variants of inexact alternating and averaged (Bregman) Proximal Point including stochastic variants with applications to federated learning and semisupervised learning. There we will leverage the gradient formulas for the proximal mapping in the Euclidean and the Bregmanian case under prox-regularity to characterize stationary points of a certain relaxation based on Moreau envelopes. In particular, relative prox-regularity turns out to be a sufficient condition to show that inexact alternating Bregman Proximal Point converges (locally) to a stationary point of the Moreau regularized problem. This cannot be taken for granted. We derive an inexact stochastic averaged proximal point method for nonconvex federated learning and invoke the theory from Chapter 2 to characterize the stationary points computed by the algorithm. In the federated learning case we use the gradient formula of the Moreau envelope to argue that such a stationary point is near stationarity wrt the original problem implying almost consensus between the individual clients, see Corollary 4.20. Leveraging the duality relation between Proximal Point and gradient descent, explored in the previous chapter, the proposed stochastic averaged Proximal Point can be specialized to the Finito/MISO algorithm [DDC14; Mai15]. In that sense one obtains a novel convergence proof for the Finito/MISO algorithm in the nonconvex setting under very general sampling strategies.

Chapter 5 is based on [Lau+16] and [Bau+21]. In this chapter we consider convex relaxations for partially separable problems and in particular the MAP-inference problem in a MRF. In contrast to relaxations by inf-projection, which is a nonconvex relaxation in general, the approach considered in this section is based on dual discretizations of infinite-dimensional linear programming relaxations and yields a convex problem. Expanding upon [Bau+21] a goal of this section is to extend the approach for MRFs to spatially continuous domains and variational problems. In particular this sets up a connection to the convex relaxation for vectorial variational problems considered in [Lau+16], which can be interpreted as a piecewise linear dual discretization of an infinite-dimensional variational problem formulated over measures which was studied in [VL18].

Part I.

Relaxation by inf-projection

Chapter 2.

Generalized Moreau envelopes and proximal mappings: A local perspective

2.1. Why it matters

A major goal of this chapter is to study two generalizations of the classical proximal mapping and Moreau envelope based on Legendre functions. This chapter is based on [LWC18], [LWC19] and [LOC20].

The Euclidean proximal mapping and Moreau envelope date back to the seminal work of [Mor62; Mor65]. Their systematic study was initiated by Attouch [Att77; Att84]. A classical result is that the Moreau envelope of a convex function is smooth with a Lipschitz continuous gradient mapping and the associated proximal mapping is single-valued and Lipschitz continuous. In the nonconvex setting, however, this is in general not the case. An exception is the class of nonconvex prox-regular functions introduced by Poliquin and Rockafellar, for which this is true at least locally [PR96] (see [Bač+10; JTZ14] for the infinite dimensional setting). Prox-regular functions comprise several widely used classes of functions, such as primal-lower-nice functions [Pol91], subsmooth functions, strongly amenable functions [PR96; RW98], and proper lower semicontinuous convex functions. Prox-regularity of indicator functions is closely related to Federers concept of a set with positive reach [Fed59] as studied in [PRT00]. In the nonconvex setting, however, there exist situations where the Euclidean proximal mapping of f is not single-valued and not even single-valued in a local neighborhood of a point even though f is actually a smooth function, see Example 2.32. As explored in Chapter 4 single-valuedness of the proximal mapping is actually an important property that has practical implications. A remedy is to replace the Euclidean proximity measure with alternative "nonlinear geometries" which eventually helps to recover the single-valuedness property for some functions which do not admit a (locally) single-valued Euclidean proximal mapping at some points. We therefore consider two alternative "nonlinear geometries": In Section 2.2 we consider the anisotropic proximal mapping which is obtained by replacing the quadratic penalty $\|x - y\|^2$ in the proximal mapping with $\phi(y - x)$ for an anisotropic prox-potential ϕ, see Definition 2.1, while Section 2.3 considers the Bregman proximal mapping which is obtained when $\|x - y\|^2$ is replaced by a Bregman distance [Bre67]. Expanding upon [LWC19], in Section 2.2, we identify a generalization of prox-regularity, called anisotropic prox-regularity, see Definition 2.13, which admits a (locally) single-valued anisotropic proximal mapping, see Theorem 2.14. An analogous result is proved for the Bregman proximal mapping. Both results can be seen as generalizations of [RW98, Proposition 13.37] for the Euclidean proximal mapping under prox-regularity.

Expanding upon [LOC20], in Section 2.3, we reveal an interesting connection between our notion of a relatively proximal subgradient, see Definition 2.30, and a *variational description of regular subgradients* due to Mordukhovich [Mor18, Theorem 1.27], showing that every regular subgradient of a relatively prox-bounded function can be expressed in terms of the Bregman proximal mapping, see Proposition 2.36.

In Section 2.4, we consider pointwise minima over a finite collection of $\mathcal{C}^1$ functions. Such functions are in particular upper-$\mathcal{C}^1$ but not prox-regular everywhere. Expanding upon [LWC18], we show a refinement of [RW98, Theorem 10.31] for finitely many pieces, showing that the limiting subdifferential can be expressed in terms of the gradient of the active pieces, if a *linear independence constraint qualification* wrt the hypograph of this function holds true.

2.2. Anisotropic Moreau envelope and proximal mapping

2.2.1. Definition and continuity properties

We begin with an anisotropic generalization of the classical Moreau envelope and associated proximal mapping under nonconvexity which has been considered before in the convex setting [Les67; Wex73; CR13]. The term *anisotropic proximal mapping* was coined by [CR13] who revealed an interesting connection to the Bregman proximal mapping via a generalization of Moreau's decomposition which holds under convexity.

The anisotropic proximal mapping is constructed by replacing the quadratic function $\|\cdot\|^2$ in the classical proximal mapping with a certain strictly convex potential function ϕ. This results in an infimal convolution type Moreau envelope as discussed in the introduction. For this formulation to yield a proper proximity measure we require $\phi(y-x)=0$ if and only if $y=x$ and $\phi(y-x)\geq 0$. This can be guaranteed if ϕ is essentially strictly convex and $\phi(0)=0$ and $\nabla\phi(0)=0$.

More precisely we will assume that ϕ satisfies the following properties:

Definition 2.1 (anisotropic prox-potential)**.** *Let $\phi\in\Gamma_0(\mathbb{R}^m)$. We call ϕ an anisotropic prox-potential if it satisfies the following assumptions:*

(i) *ϕ is essentially strictly convex,*

(ii) *ϕ is differentiable on* $\operatorname{dom}\phi$ *open,*

(iii) *ϕ is* super-coercive*,*

(iv) *and $\phi(0)=0$ and $\nabla\phi(0)=0$.*

The following univariate functions $\phi:\mathbb{R}\to\overline{\mathbb{R}}$ are anisotropic prox-potentials: $\phi(x)=|x|^3$, $\phi(x)=|x|^{3/2}$ both of which have full domain or $\phi(x)=-\log(1-x^2)$ with domain $(-1,1)$. The assumption $\operatorname{dom}\phi$ is open guarantees that ϕ has a certain barrier behavior at boundary points $x\in\operatorname{bdry}\operatorname{dom}\phi$: For a sequence $\operatorname{dom}\phi\ni x^\nu\to x\in\operatorname{bdry}\operatorname{dom}\phi$, since $\phi(x)=\infty$ and ϕ is lsc we know that $\phi(x^\nu)\to\infty$. This turns out helpful in algorithms when we linearize the prox-term and overall simplifies our study. The property $\operatorname{dom}\phi$ open also implies that ϕ is essentially smooth and therefore Legendre:

Lemma 2.2. *Let $\phi\in\Gamma_0(\mathbb{R}^m)$. Let* $\operatorname{dom}\phi$ *open and ϕ differentiable on* $\operatorname{dom}\phi$*. Then ϕ is essentially smooth.*

Proof. Since ϕ is proper and $\operatorname{dom}\phi$ open we have that $\emptyset \neq \operatorname{dom}\phi = \operatorname{int}(\operatorname{dom}\phi)$. Now consider a sequence $\operatorname{dom}\phi \ni x^\nu \to x \in \operatorname{bdry}\operatorname{dom}\phi$. Take $y \in \operatorname{dom}\phi$. Since $\phi(x) = \infty$ and ϕ is lsc we know that $\phi(x^\nu) \to \infty$. Since ϕ is convex we have

$$\phi(y) \geq \phi(x^\nu) + \langle \nabla\phi(x^\nu), y - x^\nu \rangle.$$

Cauchy–Schwartz implies for $y \neq x^\nu$:

$$\|\nabla\phi(x^\nu)\| \geq \frac{\phi(x^\nu) - \phi(y)}{\|y - x^\nu\|}.$$

Passing $\nu \to \infty$ we have $\|\nabla\phi(x^\nu)\| \to \infty$. We conclude that ϕ is essentially smooth. □

The Legendre property of ϕ turns out helpful in the study of the proximal mapping and allows us to derive an interesting resolvent expression for the proximal mapping, where $\nabla\phi^*$ appears in the form of a *nonlinear preconditioner*, see Theorem 2.11.

We are now ready to formally define the anisotropic proximal mapping and associated Moreau envelope:

Definition 2.3 (anisotropic proximal mapping and Moreau envelope)**.** *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper and let $\phi \in \Gamma_0(\mathbb{R}^m)$ be an anisotropic prox-potential. Then for a parameter $\lambda > 0$ and $y \in \mathbb{R}^m$*

$$e_\lambda^\phi f(y) := \inf_{x \in \mathbb{R}^m} f(x) + \frac{1}{\lambda}\phi(y - x), \tag{2.1}$$

is the anisotropic Moreau envelope of f at y wrt ϕ, and

$$P_\lambda^\phi f(y) := \operatorname*{arg\,min}_{x \in \mathbb{R}^m} f(x) + \frac{1}{\lambda}\phi(y - x), \tag{2.2}$$

the anisotropic proximal mapping of f at y wrt ϕ.

We define anisotropic prox-boundedness generalizing [RW98, Definition 1.23]. It is used as a sufficient condition for the uniform level-boundedness, see Definition 1.12, which yields a sufficient condition for the outer semicontinuity of the proximal mapping.

Definition 2.4 (anisotropic prox-boundedness)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be an anisotropic prox-potential. We say $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is anisotropically prox-bounded relative to ϕ if there exists $\lambda > 0$ such that for any $\bar{y} \in \mathbb{R}^m$ there exists $\varepsilon > 0$ and a constant $\beta > -\infty$ such that*

$$e_\lambda^\phi f(y) \geq \beta, \tag{2.3}$$

for any y with $\|y - \bar{y}\| \leq \varepsilon$. The supremum of the set of all such λ is the threshold λ_f of the anisotropic prox-boundedness.

When f is bounded from below it is anisotropically prox-bounded with threshold $\lambda_f = \infty$. Notably, in the classical case (when ϕ is quadratic) the definition can be made minimalistic, cf. [RW98, Definition 1.23]: It suffices to assume the existence of some $\bar{y} \in \mathbb{R}^m$ so that $e_\lambda^\phi f(\bar{y}) > -\infty$:

Anisotropic prox-boundedness is implied by classical prox-boundedness whenever ϕ is strongly convex:

Lemma 2.5. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded with threshold λ_f and let $\phi \in \Gamma_0(\mathbb{R}^m)$ be an anisotropic prox-potential and strongly convex with constant $\theta > 0$. Then f is anisotropically prox-bounded with threshold $\lambda_f \theta$.*

Proof. Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded with threshold λ_f. In view of [RW98, Theorem 1.25] for any $\lambda' \in (0, \lambda_f)$ and any $\bar{y} \in \mathbb{R}^m$, the classical Moreau envelope $e_{\lambda'} f$ of f is finite and continuous and therefore uniformly bounded from below on a compact neighborhood of $\bar{y}$, i.e., there is $\varepsilon > 0$ and $\beta > -\infty$ so that

$$e_{\lambda'} f(y) \geq \beta,$$

for all $\|y - \bar{y}\| \leq \varepsilon$. By strong convexity of ϕ with constant $\theta > 0$ and since ϕ is an anisotropic prox-potential we have the following estimate:

$$\frac{1}{\lambda'}\phi(y - x) \geq \frac{1}{\lambda'}\phi(0) + \frac{1}{\lambda'}\langle \nabla\phi(0), y - x\rangle + \frac{\theta}{2\lambda'}\|y - x\|^2 = \frac{\theta}{2\lambda'}\|y - x\|^2. \tag{2.4}$$

Then we have for $\lambda := \theta\lambda'$:

$$e_\lambda^\phi f(y) = \inf_{x \in \mathbb{R}^m} f(x) + \frac{1}{\theta\lambda'}\phi(y - x) \geq \inf_{x \in \mathbb{R}^m} f(x) + \frac{1}{2\lambda'}\|y - x\|^2 = e_{\lambda'} f(y) \geq \beta. \qquad \square$$

In the next lemma we establish the uniform level boundedness wrt x of the map $h : (x, y, u) \mapsto f(x) + \frac{1}{\lambda}\phi(y - x) - \langle u, x\rangle$ from anisotropic prox-boundedness so that Theorem 1.13 can be invoked to assert the outer semicontinuity Definition 1.11 of the anisotropic proximal mapping and the continuity of the anisotropic Moreau envelope. The lemma is stated in a more general form which involves an additional tilt perturbation $\langle u, x\rangle$, which turns out helpful in the proof of Theorem 2.11.

Lemma 2.6. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and anisotropically prox-bounded relative to the anisotropic prox-potential $\phi \in \Gamma_0(\mathbb{R}^m)$ with threshold $\lambda_f > 0$. Then for any $\lambda \in (0, \lambda_f)$, the function $h : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$, defined via*

$$h(x, y, u) := f(x) + \frac{1}{\lambda}\phi(y - x) - \langle u, x\rangle,$$

is level-bounded in x locally uniformly in (y, u).

Proof. We assume the contrary: More precisely let $\lambda \in (0, \lambda_f)$ and assume that h is not level-bounded in x locally uniformly in (y, u). On the one hand, this means that there exists $(\bar{y}, \bar{u}) \in \mathbb{R}^m \times \mathbb{R}^m$, $\alpha \in \mathbb{R}$ and sequences $y^\nu \to \bar{y}$, $u^\nu \to \bar{u}$ and x^ν, $\|x^\nu\| \to \infty$ such that

$$f(x^\nu) + \frac{1}{\lambda}\phi(y^\nu - x^\nu) - \langle u^\nu, x^\nu\rangle \leq \alpha.$$

On the other hand, we know that

$$f(x^\nu) + \frac{1}{\lambda'}\phi(y^\nu - x^\nu) \geq \beta,$$

for some $\lambda' > \lambda$, with $\lambda' \in (0, \lambda_f)$ and ν sufficiently large. Summing the inequalities yields:

$$\left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right)\phi(y^\nu - x^\nu) - \langle u^\nu, x^\nu\rangle \leq \alpha - \beta.$$

We divide the inequality by $\|y^\nu - x^\nu\|$ and obtain

$$\left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right) \frac{\phi(y^\nu - x^\nu)}{\|y^\nu - x^\nu\|} + \frac{\langle u^\nu, y^\nu - x^\nu\rangle}{\|y^\nu - x^\nu\|} - \frac{\langle u^\nu, y^\nu\rangle}{\|y^\nu - x^\nu\|} \leq \frac{\alpha - \beta}{\|y^\nu - x^\nu\|}.$$

Applying Cauchy–Schwarz we obtain

$$\left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right) \frac{\phi(y^\nu - x^\nu)}{\|y^\nu - x^\nu\|} - \frac{\|u^\nu\|\|y^\nu - x^\nu\|}{\|y^\nu - x^\nu\|} - \frac{\|u^\nu\|\|y^\nu\|}{\|y^\nu - x^\nu\|} \leq \frac{\alpha - \beta}{\|y^\nu - x^\nu\|}.$$

Passing $\nu \to \infty$ we obtain due to the super-coercivity of ϕ:

$$\infty \leq 0,$$

a contradiction. □

Now we are ready to prove the following lemma invoking Theorem 1.13 which sets up the aforementioned continuity properties:

Lemma 2.7. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and anisotropically prox-bounded relative to the anisotropic prox-potential $\phi \in \Gamma_0(\mathbb{R}^m)$ with threshold $\lambda_f > 0$. Then for any $\lambda \in (0, \lambda_f)$, $P^\phi_\lambda f$ and $e^\phi_\lambda f$ have the following properties:*

(i) *$P^\phi_\lambda f(y) \neq \emptyset$ is compact for all $y \in \operatorname{dom} e^\phi_\lambda f = \operatorname{dom} f + \operatorname{dom} \phi$, whereas $P^\phi_\lambda f(y) = \emptyset$ for $y \notin \operatorname{dom} e^\phi_\lambda f$.*

(ii) *The envelope $e^\phi_\lambda f$ is continuous relative to $\operatorname{dom} e^\phi_\lambda f$.*

(iii) *For any sequence $y^\nu \to y^* \in \operatorname{dom} e^\phi_\lambda f$ contained in $\operatorname{dom} e^\phi_\lambda f$ and $x^\nu \in P^\phi_\lambda f(y^\nu)$ we have $\{x^\nu\}_{\nu\in\mathbb{N}}$ is bounded and all its cluster points x^* lie in $P^\phi_\lambda f(y^*)$.*

Proof. Obviously it holds for the domain that $\operatorname{dom} e^\phi_\lambda f = \operatorname{dom} f + \operatorname{dom} \phi$. In view of Lemma 2.6 (with $u = 0$) we assert that $h : (x, y) \mapsto f(x) + \frac{1}{\lambda}\phi(y - x)$ is level-bounded in x locally uniformly in y. Then we may invoke Theorem 1.13 to assert that $P^\phi_\lambda f(y) \neq \emptyset$ is compact for any $y \in \operatorname{dom} e^\phi_\lambda f$ whereas $P^\phi_\lambda f(y) = \emptyset$ for $y \notin \operatorname{dom} e^\phi_\lambda f$ and in addition for any $y^* \in \operatorname{dom} e^\phi_\lambda f$ and any sequence $x^\nu \in P^\phi_\lambda f(y^\nu)$ with $y^\nu \to y^*$ contained in $\operatorname{dom} e^\phi_\lambda f$, that $\{x^\nu\}_{\nu\in\mathbb{N}}$ is bounded. Furthermore, as ϕ is continuous relative to its domain, we know for some $x \in P^\phi_\lambda f(y)$ that $h(x, \cdot)$ is continuous relative to $x + \operatorname{dom} \phi$ containing y. Through Theorem 1.13 all cluster points of the sequence $x^\nu \in P^\phi_\lambda f(y^\nu)$ lie in $P^\phi_\lambda f(y^*)$ and $e^\phi_\lambda f(y^\nu) \to e^\phi_\lambda f(y^*)$ and therefore $e^\phi_\lambda f$ is continuous at y relative to $\operatorname{dom} e^\phi_\lambda f$. Since this holds for all $y \in \operatorname{dom} e^\phi_\lambda f$, $e^\phi_\lambda f$ is continuous relative to $\operatorname{dom} e^\phi_\lambda f$. □

2.2.2. Single-valuedness and Lipschitz continuity of the anisotropic proximal mapping under prox-regularity

Our next goal is to establish the local single-valuedness and Lipschitz continuity of the anisotropic proximal mapping under prox-regularity. We define prox-regularity of functions, according to [RW98, Definition 13.27]:

Definition 2.8 (prox-regularity of functions)**.** *A function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is prox-regular at $\bar{x}$ for $\bar{v}$ if f is finite and locally lsc at $\bar{x}$ with $\bar{v} \in \partial f(\bar{x})$, and there exist $\varepsilon > 0$ and $r \geq 0$*

such that for all $\|x' - \bar{x}\| < \varepsilon$

$$f(x') \geq f(x) + \langle v, x' - x \rangle - \frac{r}{2}\|x' - x\|^2, \tag{2.5}$$

whenever $\|x - \bar{x}\| < \varepsilon$*,* $f(x) - f(\bar{x}) < \varepsilon$*,* $v \in \partial f(x)$*,* $\|v - \bar{v}\| < \varepsilon$*. When this holds for all* $\bar{v} \in \partial f(\bar{x})$*,* f *is said to be prox-regular at* $\bar{v}$*.*

In addition to the single-valuedness of the prox, exploiting the Legendre property of ϕ, we obtain an interesting expression for the proximal mapping in terms of the resolvent $(I + \nabla\phi^* \circ \lambda T)^{-1}$ of the map $\nabla\phi^* \circ \lambda T$ where $\nabla\phi^*$ can be interpreted as a nonlinear preconditioner for the classical proximal mapping of f with resolvent $(I + \lambda T)^{-1}$ and T is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$.

To this end we propose to invoke the *implicit function theorem for generalized equations*, Theorem 1.14, to prove the following proposition which is applicable in a more general setup:

Proposition 2.9. *Let* $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ *be proper lsc and prox-regular with constant* r *at* $\bar{x}$*, a point where* f *is finite, for* $\bar{v} \in \partial f(\bar{x})$*. Let* $g : \mathbb{R}^m \times \mathbb{R}^n \to \overline{\mathbb{R}}$ *be* $\mathcal{C}^2$ *on* $\operatorname{int}(\operatorname{dom} g)$*. Let* $(\bar{x}, \bar{y}) \in \operatorname{int}(\operatorname{dom} g)$ *and assume that* $\{\bar{x}\} = \operatorname{arg\,min}_{x\in\mathbb{R}^m} f(x) + g(x, \bar{y})$ *and* $\bar{v} = -\nabla_x g(\bar{x}, \bar{y})$*. Furthermore assume that* $\nabla^2_{xx} g(\bar{x}, \bar{y}) \succeq \sigma I$ *with* $r < \sigma$ *and that* $h : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ *defined via* $h(x, y, u) := f(x) + g(x, y) - \langle u, x \rangle$ *is level-bounded in* x *locally uniformly in* (y, u)*. Then, the solution mapping* $y \mapsto \operatorname{arg\,min}_{x\in\mathbb{R}^m} f(x) + g(x, y)$ *is a single-valued Lipschitz map in a neighborhood of* $\bar{y}$ *and satisfies*

$$y \mapsto \operatorname*{arg\,min}_{x\in\mathbb{R}^m} f(x) + g(x, y) = \{x \in \mathbb{R}^m : 0 \in T(x) + \nabla_x g(x, y)\},$$

where T *is an* f*-attentive localization of* ∂f *at* $\bar{x}$ *for* $\bar{v}$*.*

Proof. We show that $y \mapsto \{x \in \mathbb{R}^m : 0 \in \partial f(x) + \nabla_x g(x, y)\}$ has a single-valued Lipschitz localization around $\bar{y}$ for $\bar{x}$ which we denote by M. Then we prove that $\operatorname{arg\,min}_{x\in\mathbb{R}^m} f(x) + g(x, y) = M(y)$ for y near $\bar{y}$. To this end we first show that $u \mapsto (T + \nabla_x g(\cdot, \bar{y}))^{-1}(u)$ is a single-valued, Lipschitz localization of $(\partial f + \nabla_x g(\cdot, \bar{y}))^{-1}$ around 0 for $\bar{x}$ where T is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$. The desired result is then obtained invoking the generalized implicit function theorem.

Since $h(x, y, u)$ is level-bounded in x locally uniformly in (y, u), in view of Theorem 1.13 we know that for any sequence $u^\nu \to 0$ with $\inf_{x\in\mathbb{R}^m} h(x, \bar{y}, u^\nu) < \infty$ there is $x^\nu \in \operatorname{arg\,min}_{x\in\mathbb{R}^m} h(x, \bar{y}, u^\nu) \neq \emptyset$ with $x^\nu \to \bar{x} = \operatorname{arg\,min}_{x\in\mathbb{R}^m} h(x, \bar{y}, 0)$ and $\inf_{x\in\mathbb{R}^m} h(x, \bar{y}, u^\nu) \to \inf_{x\in\mathbb{R}^m} h(x, \bar{y}, 0)$.

From applying Fermat's rule Lemma 1.3 to the minimization problem above we know that $u^\nu - \nabla_x g(x^\nu, \bar{y}) \in \partial f(x^\nu)$ and for ν sufficiently large we have that $\|u^\nu - \nabla_x g(x^\nu, \bar{y}) - \bar{v}\| \leq \varepsilon$ due to the continuity of $\nabla_x g(x^\nu, \bar{y})$ on a neighborhood of $\bar{x}$. In addition, we have $f(x^\nu) \to f(\bar{x})$ as $\inf_{x\in\mathbb{R}^m} h(x, \bar{y}, u^\nu) = f(x^\nu) + g(x^\nu, \bar{y}) - \langle u^\nu, x^\nu \rangle \to \inf_{x\in\mathbb{R}^m} h(x, \bar{y}, 0) = f(\bar{x}) + g(\bar{x}, \bar{y})$. Overall this means that for any u sufficiently near 0 we have:

$$u \in S(x) + \nabla_x g(x, \bar{y}),$$

for some x near $\bar{x}$, where S is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$.

Now pick any $(x_1, v_1), (x_2, v_2) \in \operatorname{gph} S$. Then it follows from prox-regularity that

$$f(x_2) \geq f(x_1) + \langle v_1, x_2 - x_1 \rangle - \frac{r}{2}\|x_2 - x_1\|^2, \tag{2.6}$$

$$f(x_1) \geq f(x_2) + \langle v_2, x_1 - x_2 \rangle - \frac{r}{2}\|x_1 - x_2\|^2. \tag{2.7}$$

By assumption $\nabla^2_{xx} g(\bar{x}, \bar{y}) \succeq \sigma I$, $\sigma > r$. Since $\nabla^2_{xx} g(\cdot, \bar{y})$ is continuous and the smallest eigenvalue map is continuous as well there is $\varepsilon > 0$ sufficiently small so that $\nabla^2_{xx} g(x, \bar{y}) \succeq \gamma I$ with $r < \gamma < \sigma$ for $\|x - \bar{x}\| \leq \varepsilon$ and therefore $g(\cdot, \bar{y})$ is γ-strongly convex on that neighborhood. Then we have for the chosen x_1, x_2

$$g(x_2, \bar{y}) \geq g(x_1, \bar{y}) + \langle \nabla_x g(x_1, \bar{y}), x_2 - x_1 \rangle + \frac{\gamma}{2}\|x_2 - x_1\|^2, \tag{2.8}$$

$$g(x_1, \bar{y}) \geq g(x_2, \bar{y}) + \langle \nabla_x g(x_2, \bar{y}), x_1 - x_2 \rangle + \frac{\gamma}{2}\|x_1 - x_2\|^2. \tag{2.9}$$

Summing the four inequalities yields for $u_1 := v_1 + \nabla_x g(x_1, \bar{y})$ and $u_2 := v_2 + \nabla_x g(x_2, \bar{y})$:

$$\langle x_1 - x_2, u_1 - u_2 \rangle \geq (\gamma - r)\|x_1 - x_2\|^2. \tag{2.10}$$

Consequently, the set-valued mapping $x \mapsto S(x) + \nabla_x g(x, \bar{y})$ is a $(\gamma - r)$-strongly monotone localization of $\partial f + \nabla_x g(\cdot, \bar{y})$ at $\bar{x}$ for 0.

Define $F(x) = \partial f(x)$, $H(x) := \nabla_x g(x, \bar{y})$, $G(x, y) := \nabla_x g(x, y)$ and $e(x, y) := G(x, y) - H(x)$. Then the above argument implies via Lemma 1.10 that for $\kappa := (\gamma - r)^{-1}$ the map $(S + H)^{-1}$ is a single-valued, κ-Lipschitz localization of $(F + H)^{-1}$ at 0 for $\bar{x}$. Since $\operatorname{dom} (S + H)^{-1}$ is a neighborhood of 0 the map $(S + H)^{-1}$ is in particular a single-valued localization of $(F + H)^{-1}$ around 0 for $\bar{x}$. By assumption g is $\mathcal{C}^2$ on $\operatorname{int}(\operatorname{dom} g) \ni (\bar{x}, \bar{y})$. Invoking [DR09, Exercise 1D.10] we obtain that

$$\widehat{\operatorname{lip}}_y(G; (\bar{x}, \bar{y})) = \limsup_{(x,y) \to (\bar{x}, \bar{y})} |\nabla_y G(x, y)| = \lim_{(x,y) \to (\bar{x}, \bar{y})} |\nabla^2_{yx} g(x, y)| = |\nabla^2_{yx} g(\bar{x}, \bar{y})| < \infty,$$

as well as

$$\widehat{\operatorname{lip}}_x(e; (\bar{x}, \bar{y})) = \limsup_{(x,y) \to (\bar{x}, \bar{y})} |\nabla_x e(x, y)| = \lim_{(x,y) \to (\bar{x}, \bar{y})} |\nabla^2_{xx} g(x, y) - \nabla^2_{xx} g(\bar{x}, \bar{y})| = 0.$$

This implies that H is a strict estimator of G wrt x uniformly in y at $(\bar{x}, \bar{y}) \in \operatorname{int}(\operatorname{dom} G)$ with constant $\mu := 0$. Invoking Theorem 1.14, we assert that $y \mapsto \{x \in \mathbb{R}^m : 0 \in G(x, y) + F(x)\}$ has a single-valued, Lipschitz localization around $\bar{y}$ for $\bar{x}$ which is denoted by M. It remains to show that $\emptyset \neq \operatorname{arg\,min}_{x \in \mathbb{R}^m} f(x) + g(x, y) \subset M(y)$ and $M = \{x \in \mathbb{R}^m : 0 \in T(x) + \nabla_x g(x, y)\}$, where T is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$.

Since $h(x, y, u)$ is level-bounded in x locally uniformly in (y, u), in view of Theorem 1.13 we know that for any sequence $y^\nu \to \bar{y}$ with $\inf_{x \in \mathbb{R}^m} h(x, y^\nu, 0) < \infty$ there is $x^\nu \in \operatorname{arg\,min}_{x \in \mathbb{R}^m} h(x, y^\nu, 0) \neq \emptyset$ with

$$x^\nu \to \bar{x} = \operatorname*{arg\,min}_{x \in \mathbb{R}^m} h(x, \bar{v}, 0), \qquad \inf_{x \in \mathbb{R}^m} h(x, y^\nu, 0) \to \inf_{x \in \mathbb{R}^m} h(x, \bar{y}, 0).$$

From applying Fermat's rule Lemma 1.3 to the minimization problem above we know that $-\nabla_x g(x^\nu, y^\nu) \in \partial f(x^\nu)$ and for ν sufficiently large we have that $\|\nabla_x g(x^\nu, y^\nu) - \nabla_x g(\bar{x}, \bar{y})\| \leq \varepsilon$ due to the continuity of $\nabla_x g$ on a neighborhood of $(\bar{x}, \bar{y})$. In addition, we have $f(x^\nu) \to f(\bar{x})$ as $\inf_{x \in \mathbb{R}^m} h(x, y^\nu, 0) = f(x^\nu) + g(x^\nu, y^\nu) \to \inf_{x \in \mathbb{R}^m} h(x, \bar{y}, 0) =$

$f(\bar{x}) + g(\bar{x}, \bar{y})$. Overall this means that for any y sufficiently near $\bar{y}$ we have:

$$0 \in T(x) + \nabla_x g(x, y),$$

for some x near $\bar{x}$, where T is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$. This shows that

$$\emptyset \neq \operatorname*{arg\,min}_{x \in \mathbb{R}^m} f(x) + g(x, y) \subset \{x \in \mathbb{R}^m : 0 \in T(x) + \nabla_x g(x, y)\}, \tag{2.11}$$

where ∂f is replaced by T. Now let y sufficiently near $\bar{y}$ and take some $x \in \{x \in \mathbb{R}^m : 0 \in T(x) + \nabla_x g(x, y)\}$. For ε in the definition of T sufficiently small and y sufficiently near $\bar{y}$, since M is a localization of $y \mapsto \{x \in \mathbb{R}^m : 0 \in G(x, y) + F(x)\}$ around $\bar{y}$ for $\bar{x}$, we have $x \in M(y)$ and

$$\emptyset \neq \operatorname*{arg\,min}_{x \in \mathbb{R}^m} f(x) + g(x, y) \subset \{x \in \mathbb{R}^m : 0 \in T(x) + \nabla_x g(x, y)\} \subset M(y), \tag{2.12}$$

holds with equality since M is single-valued. □

The next step in our proof of local Lipschitz continuity of the proximal mapping of f under prox-regularity of f at $(\bar{x}, \bar{v})$ is to validate that $\bar{x} = P_\lambda^\phi f(\bar{y}_\lambda)$ for $\bar{y}_\lambda := \bar{x} + \nabla\phi^*(\lambda\bar{v})$ for some $\lambda > 0$. In other words this means that the anisotropic proximal mapping of f at the perturbed point $\bar{x} + \nabla\phi^*(\lambda\bar{v})$ is $\bar{x}$.

For that purpose we prove the following lemma:

Lemma 2.10. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and anisotropically prox-bounded relative to the anisotropic prox-potential $\phi \in \Gamma_0(\mathbb{R}^m)$ with threshold $\lambda_f > 0$. In addition assume that ϕ is very strictly convex. Assume that f is finite at $\bar{x}$ and let $\bar{v} \in \partial f(\bar{x})$ be a proximal subgradient of f at $\bar{x}$. Then the following inequality holds for all $x \in \mathbb{R}^m$ with $x \neq \bar{x}$, $\lambda < \min\{1/r, \lambda_f^{-1}\}$ sufficiently small and $\bar{y}_\lambda := \bar{x} + \nabla\phi^*(\lambda\bar{v})$*

$$f(x) > f(\bar{x}) + \langle \bar{v}, x - \bar{x} \rangle - \frac{1}{\lambda} D_\phi(\bar{y}_\lambda - x, \bar{y}_\lambda - \bar{x}). \tag{2.13}$$

Proof. Since $\bar{v}$ is a proximal subgradient of f at $\bar{x}$ we know there exist $r > 0$ and $\varepsilon > 0$ such that the subgradient inequality holds:

$$f(x) > f(\bar{x}) + \langle \bar{v}, x - \bar{x} \rangle - \frac{r}{2}\|x - \bar{x}\|^2, \tag{2.14}$$

for $\|x - \bar{x}\| < \varepsilon$ with $x \neq \bar{x}$. Furthermore for $\lambda^\nu > 0$ with $\lambda^\nu \to 0$ since $\nabla\phi^*$ is continuous and since $\nabla\phi(0) = 0$, we have that $\nabla\phi^*(\lambda^\nu\bar{v}) \to 0$ and therefore $\bar{y}_{\lambda^\nu} := \bar{x} + \nabla\phi^*(\lambda^\nu\bar{v}) \to \bar{x}$. Since ϕ is very strictly convex we have in view of Lemma 1.20(iii)

$$\begin{aligned} \frac{\mu}{2}\|x - \bar{x}\|^2 &\leq D_\phi(\bar{y}_\lambda - x, \bar{y}_\lambda - \bar{x}) \\ &= D_{\phi(\bar{y}_\lambda - \cdot)}(x, \bar{x}) \\ &= D_\phi(\bar{y}_\lambda - x, \nabla\phi^*(\lambda\bar{v})) \\ &= \phi(\bar{y}_\lambda - x) + \phi^*(\lambda\bar{v}) - \lambda\langle \bar{v}, \bar{y}_\lambda - x \rangle \\ &= \phi(\bar{y}_\lambda - x) - \phi(\bar{y}_\lambda - \bar{x}) + \langle \nabla\phi(\bar{y}_\lambda - \bar{x}), x - \bar{x} \rangle, \end{aligned} \tag{2.15}$$

for some $\mu > 0$ and $\|x - \bar{x}\| < \varepsilon$ and $\bar{y}_\lambda$ near 0. Inequalities (2.14) and (2.15) together yield (2.13), which holds for any $x \neq \bar{x}$ with $\|x - \bar{x}\| < \varepsilon$ and any $\lambda \leq \mu/r$ sufficiently small. To show the assertion we prove that (2.13) also holds for any x with $\|x-\bar{x}\| \geq \varepsilon$ for $\lambda < \min\{\mu/r, \lambda_f\}$ sufficiently small. By anisotropic prox-boundedness it holds for some $\lambda' \in (0, \lambda_f)$ with $\lambda < \lambda'$ and $\bar{y}_\lambda \in \operatorname{dom} f + \operatorname{dom}\phi = \operatorname{dom} e^\phi_{\lambda'} f$ that $+\infty > e^\phi_{\lambda'} f(\bar{y}_\lambda) > -\infty$. We have

$$f(x) \geq e^\phi_{\lambda'} f(\bar{y}_\lambda) - \frac{1}{\lambda'}\phi(\bar{y}_\lambda - x), \tag{2.16}$$

for all $x \in \mathbb{R}^m$ showing, that the desired Inequality (2.13) is implied by

$$e^\phi_{\lambda'} f(\bar{y}_\lambda) - \frac{1}{\lambda'}\phi(\bar{y}_\lambda - x) > f(\bar{x}) + \langle \bar{v}, x - \bar{x}\rangle - \frac{1}{\lambda} D_{\phi(\bar{y}_\lambda - \cdot)}(x, \bar{x}),$$

which is equivalent to

$$(\lambda^{-1} - 1/\lambda') D_{\phi(\bar{y}_\lambda - \cdot)}(x, \bar{x}) > f(\bar{x}) - e^\phi_{\lambda'} f(\bar{y}_\lambda) + (1 - \lambda/\lambda')\langle \bar{v}, x - \bar{x}\rangle + \frac{1}{\lambda'}\phi(\nabla\phi^*(\lambda\bar{v})),$$

and (using Cauchy–Schwarz) implied by

$$(\lambda^{-1} - 1/\lambda')\frac{D_{\phi(\bar{y}_\lambda - \cdot)}(x, \bar{x})}{\|x - \bar{x}\|} > \frac{f(\bar{x}) - e^\phi_{\lambda'} f(\bar{y}_\lambda) + \frac{1}{\lambda'}\phi(\nabla\phi^*(\lambda\bar{v}))}{\|x - \bar{x}\|} + (1 - \lambda/\lambda')\|\bar{v}\|. \tag{2.17}$$

Due to the super-coercivity of ϕ there is $\gamma > 0$ so that Inequality (2.17) holds for any x with $\|x - \bar{x}\| > \gamma$ sufficiently large or for any $\bar{y}_\lambda - x \notin \operatorname{dom}\phi$ and any $\lambda > 0$ sufficiently small. To make (2.17) also hold for $\bar{y}_\lambda - x \in \operatorname{dom}\phi$ with $\varepsilon \leq \|x - \bar{x}\| \leq \gamma$ we show that $D_\phi(\bar{y}_\lambda - x, \bar{y}_\lambda - \bar{x})$ is uniformly bounded from below by a positive constant $\delta > 0$. To this end first observe that for such x we have $D_\phi(\bar{y}_\lambda - x, \bar{y}_\lambda - \bar{x}) > 0$. Now suppose we can find sequences $\varepsilon \leq \|x^\nu - \bar{x}\| \leq \gamma$ and $\lambda^\nu \in [0, \lambda_f]$ with $\bar{y}_{\lambda^\nu} - x^\nu \in \operatorname{dom}\phi$ such that $D_\phi(\bar{y}_{\lambda^\nu} - x^\nu, \bar{y}_{\lambda^\nu} - \bar{x}) \to 0$, for $\nu \to \infty$. Due to compactness we can potentially go to a subsequence and assume $x^\nu \to x^*$ with $\varepsilon \leq \|x^* - \bar{x}\| \leq \gamma$ and $\lambda^\nu \to \lambda^* \in [0, \lambda_f]$ and $\bar{y}_{\lambda^\nu} - \bar{x} = \nabla\phi^*(\lambda^\nu\bar{v}) \in \operatorname{int}(\operatorname{dom}\phi)$. Since ϕ is Legendre, in view of [BB97, Theorem 3.9(iii)] this means $x^* = \bar{x}$, a contradiction. Therefore, $D_\phi(\bar{y}_\lambda - x, \bar{y}_\lambda - \bar{x}) \geq \delta > 0$, for $\bar{y}_\lambda - x \in \operatorname{dom}\phi$ with $\varepsilon \leq \|x - \bar{x}\| \leq \gamma$ and $\lambda \in [0, \lambda_f]$. As a consequence, the desired result is implied by the following inequality

$$(\lambda^{-1} - 1/\lambda')\frac{\delta}{\gamma} > \frac{f(\bar{x}) - e^\phi_{\lambda'} f(\bar{y}_\lambda) + \frac{1}{\lambda'}\phi(\nabla\phi^*(\lambda\bar{v}))}{\varepsilon} + (1 - \lambda/\lambda')\|\bar{v}\|,$$

which holds for any $\lambda > 0$ sufficiently small, since in view of the continuity of the envelope function, see Lemma 2.7, we have $e^\phi_{\lambda'} f(\bar{y}_{\lambda^\nu}) \to e^\phi_{\lambda'} f(\bar{x})$ for $\lambda^\nu \to 0$. □

We are now able to prove the Lipschitz continuity of $P^\phi_\lambda f = (I + \nabla\phi^* \circ \lambda T)^{-1}$ and associated Moreau envelope along with expressions for the anisotropic resolvent and Yosida regularization of T:

Theorem 2.11. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and anisotropically prox-bounded relative to the anisotropic prox-potential $\phi \in \Gamma_0(\mathbb{R}^m)$ with threshold $\lambda_f > 0$. In addition assume that ϕ is very strictly convex. Let f be finite and prox-regular at $\bar{x}$ for $\bar{v} \in \partial f(\bar{x})$. Then, for all $\lambda \in (0, \lambda_f)$ sufficiently small there is a neighborhood of $\bar{y}_\lambda := \bar{x} + \nabla\phi^*(\lambda\bar{v})$ on which:*

(i) $P_\lambda^\phi f$ *is a singled-valued, Lipschitz map such that* $\bar x = P_\lambda^\phi f(\bar y_\lambda)$ *and*

$$P_\lambda^\phi f = (I + \nabla\phi^* \circ \lambda T)^{-1}, \tag{2.18}$$

(ii) $e_\lambda^\phi f$ *is Lipschitz differentiable with*

$$\nabla e_\lambda^\phi f = \lambda^{-1}\nabla\phi \circ (I - P_\lambda^\phi f) = \big(\nabla\phi^* \circ \lambda I + T^{-1}\big)^{-1}, \tag{2.19}$$

where T *is an* f*-attentive localization of* ∂f *at* $\bar x$ *for* $\bar v$*. Indeed, this localization can be chosen so that the set*

$$U_\lambda := \operatorname{rge}(I + \nabla\phi^* \circ \lambda T),$$

serves for all $\lambda > 0$ *sufficiently small as a neighborhood of* $\bar y_\lambda$ *on which these properties hold.*

Proof. Let f be prox-regular at $\bar x$ for $\bar v$ with constants $r \geq 0$ and $\varepsilon > 0$. In particular this means that $\bar v \in \partial f(\bar x)$ is a proximal subgradient of f at $\bar x$. In view of Lemma 2.10, since f is proper lsc and anisotropically prox-bounded with threshold λ_f we have for $\bar y_\lambda = \bar x + \nabla\phi^*(\lambda \bar v)$:

$$f(x) > f(\bar x) + \langle \bar v, x - \bar x\rangle - \frac{1}{\lambda} D_\phi(\bar y_\lambda - x, \bar y_\lambda - \bar x),$$

for all $x \in \mathbb{R}^m$ with $x \neq \bar x$ and $\lambda < \min\{1/r, \lambda_f\}$ sufficiently small. Expanding the Bregman distance

$$\frac{1}{\lambda} D_\phi(\bar y_\lambda - x, \bar y_\lambda - \bar x) = \frac{1}{\lambda}\phi(\bar y_\lambda - x) - \frac{1}{\lambda}\phi(\bar y_\lambda - \bar x) + \frac{1}{\lambda}\langle \nabla\phi(\bar y_\lambda - \bar x), x - \bar x\rangle \tag{2.20}$$

we obtain since $\nabla\phi(\bar y_\lambda - \bar x) = \lambda \bar v$

$$f(x) + \frac{1}{\lambda}\phi(\bar y_\lambda - x) > f(\bar x) + \frac{1}{\lambda}\phi(\bar y_\lambda - \bar x).$$

We define $g(x, y) := \lambda^{-1}\phi(y - x)$. Then the inequality means that

$$P_\lambda^\phi f(\bar y_\lambda) = \operatorname*{arg\,min}_{x \in \mathbb{R}^m} f(x) + g(x, \bar y_\lambda) = \{\bar x\}$$

is single-valued for any $\lambda < \min\{r^{-1}, \lambda_f\}$ sufficiently small. In addition, for any $\lambda > 0$ sufficiently small we know that $\|\bar y_\lambda - \bar x\| = \|\nabla\phi^*(\lambda \bar v)\| \leq \varepsilon$.

Since ϕ is very strictly convex, there is $\theta > 0$ so that for any λ sufficiently small $\nabla^2\phi(\bar v_\lambda - \bar x) \succeq \theta$. This means $\nabla^2_{xx} g(\bar x, \bar y_\lambda) \succeq \sigma I$ with $\sigma = \theta/\lambda > r$.

In view of Lemma 2.6 we have that for any $\lambda \in (0, \lambda_f)$, the function $h : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$, defined via

$$h(x, y, u) := f(x) + g(x, y) - \langle u, x\rangle,$$

is level-bounded in x locally uniformly in (y, u).

Also note the relation $\bar v = \lambda^{-1}\nabla\phi(\bar y_\lambda - \bar x) = -\nabla_x g(\bar x, \bar y_\lambda)$. This verifies all the assumptions in Proposition 2.9 from which we deduce that on a neighborhood of $\bar y_\lambda$ we have $P_\lambda^\phi f$ is a single-valued, Lipschitz map such that $\bar x = P_\lambda^\phi f(\bar y_\lambda)$ and for y near $\bar y_\lambda$

$$P_\lambda^\phi f(y) = \left(T - \frac{1}{\lambda}\nabla\phi(y - \cdot)\right)^{-1}(0) = (I + \nabla\phi^* \circ \lambda T)^{-1}(y), \tag{2.21}$$

where T is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$. Now take y with $\|y - \bar{y}_\lambda\| \leq \varepsilon$ with $\varepsilon > 0$ sufficiently small. Then, there is x near $\bar{x}$ so that $y \in (I + \nabla\phi^* \circ \lambda T)(x)$, i.e., $y \in \operatorname{rge}(I + \nabla\phi^* \circ \lambda T)$. This means $\operatorname{rge}(I + \nabla\phi^* \circ \lambda T)$ is a neighborhood of $\bar{y}_\lambda$. For any $y = x + \nabla\phi^*(\lambda v)$, with (x, v) in $\operatorname{gph} T$, we have $(I + \nabla\phi^* \circ \lambda T)(x) \ni y$, and since $(I + \nabla\phi^* \circ \lambda T)^{-1}$ is single-valued and Lipschitz we have $x = (I + \nabla\phi^* \circ \lambda T)^{-1}(y)$. Therefore $\operatorname{rge}(I + \nabla\phi^* \circ \lambda T)$ is a neighborhood of $\bar{y}_\lambda$ on which the claimed properties hold.

(ii) Choose $y, y' \in U_\lambda$ and $x = P_\lambda^\phi f(y)$, $x' = P_\lambda^\phi f(y')$. Then, by Fermat's rule Lemma 1.3 it holds $v \in \partial f(x)$ such that $v = \lambda^{-1}\nabla\phi(y - x)$. Furthermore, by assumption the subgradient inequality (2.5) holds true at $(x, v) \in \operatorname{gph} T$. This means in particular that v is a proximal subgradient of f at x and in particular $v \in \widehat{\partial} f(x)$. Thus (and using the differentiability of ϕ on $\operatorname{dom}\phi$) one can derive

$$\begin{aligned} e_\lambda^\phi f(y') - e_\lambda^\phi f(y) &= f(x') - f(x) + \frac{1}{\lambda}\phi(y' - x') - \frac{1}{\lambda}\phi(y - x) \\ &\geq \langle v, x' - x\rangle + o(\|x' - x\|) + \frac{1}{\lambda}\langle \nabla\phi(y - x), (y' - y) - (x' - x)\rangle \\ &\quad + o(\|(y' - y) - (x' - x)\|). \end{aligned} \tag{2.22}$$

Using the conclusion from (i) that $y \mapsto x$ is a Lipschitz map on U_λ there is some α such that

$$\|x' - x\| \leq \alpha\|y' - y\|.$$

This shows that $o(\|(y' - y) - (x' - x)\|) + o(\|x' - x\|) = o(\|y' - y\|)$ and we get from (2.22) that

$$e_\lambda^\phi f(y') - e_\lambda^\phi f(y) - \frac{1}{\lambda}\langle \nabla\phi(y - x), y' - y\rangle \geq o(\|y' - y\|). \tag{2.23}$$

On the other hand, we have

$$e_\lambda^\phi f(y') = \inf_{x'' \in \mathbb{R}^m} f(x'') + \frac{1}{\lambda}\phi(y' - x'') \leq f(x) + \frac{1}{\lambda}\phi(y' - x). \tag{2.24}$$

Due to the differentiability of ϕ, we have

$$\phi(y' - x) = \phi(y - x) + \langle \nabla\phi(y - x), y' - y\rangle + o(\|y' - y\|). \tag{2.25}$$

This yields

$$\begin{aligned} e_\lambda^\phi f(y') - e_\lambda^\phi f(y) &\leq f(x) + \frac{1}{\lambda}\phi(y' - x) - f(x) - \frac{1}{\lambda}\phi(y - x) \\ &= \frac{1}{\lambda}\langle \nabla\phi(y - x), y' - y\rangle + o(\|y' - y\|). \end{aligned} \tag{2.26}$$

Combining (2.23) and (2.26), we conclude that $e_\lambda^\phi f$ is differentiable at $y \in U_\lambda$ and $\lambda^{-1}\nabla\phi(y - x) = \nabla e_\lambda^\phi f(y)$.

We have that

$$\nabla e_\lambda^\phi f = \lambda^{-1}\nabla\phi \circ \big(I - (I + \nabla\phi^* \circ \lambda T)^{-1}\big).$$

In view of the inverse-resolvent identity [RW98, Lemma 12.14] we have that

$$I - (I + \nabla\phi^* \circ \lambda T)^{-1} = \big(I + (\nabla\phi^* \circ \lambda T)^{-1}\big)^{-1}. \tag{2.27}$$

Note the equivalences:

$$\begin{aligned} & z \in \big(I + (\nabla\phi^* \circ \lambda T)^{-1}\big)^{-1}(y) \\ \iff \quad & y - z \in (\nabla\phi^* \circ \lambda T)^{-1}(z) \\ \iff \quad & \lambda^{-1}\nabla\phi(z) \in T(y-z) \\ \iff \quad & y - z \in T^{-1}(\lambda^{-1}\nabla\phi(z)). \end{aligned}$$

We introduce the substitution $v = \lambda^{-1}\nabla\phi(z) \iff z = \nabla\phi^*(\lambda v)$:

$$\begin{aligned} & y - z \in T^{-1}(\lambda^{-1}\nabla\phi(z)) \\ \iff \quad & y \in \big(\nabla\phi^* \circ \lambda I + T^{-1}\big)(v) \\ \iff \quad & x \in \big(\nabla\phi^* \circ \lambda I + T^{-1}\big)^{-1}(y) \\ \iff \quad & \lambda^{-1}\nabla\phi(z) \in \big(\nabla\phi^* \circ \lambda I + T^{-1}\big)^{-1}(y). \end{aligned}$$

Overall we obtain

$$\big(I + (\nabla\phi^* \circ \lambda T)^{-1}\big)^{-1} = \nabla\phi^* \circ \lambda\big(\nabla\phi^* \circ \lambda I + T^{-1}\big)^{-1},$$

and therefore

$$\nabla e_\lambda^\phi f = \big(\nabla\phi^* \circ \lambda I + T^{-1}\big)^{-1},$$

as claimed. Furthermore, since ϕ is very strictly convex, $\nabla\phi$ is locally Lipschitz and in particular Lipschitz on every convex compact subset of $\operatorname{dom}\phi$. Since $(I + \nabla\phi^* \circ \lambda T)^{-1}$ is Lipschitz on its domain as well we obtain that $\nabla e_\lambda^\phi f$ is Lipschitz relative to U_λ as claimed. □

If f is hypoconvex and ϕ strongly convex the previous result can be globalized:

Corollary 2.12. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and hypoconvex with constant $r \geq 0$ and let $\phi \in \Gamma_0(\mathbb{R}^m)$ be an anisotropic prox-potential and let $\phi \in \mathcal{C}^2$ be strongly convex such that $\nabla^2\phi(x) \succeq \theta I > 0$, for any $x \in \operatorname{dom}\phi$. Then, for any $0 < \lambda < \theta/r$ it holds that $P_\lambda^\phi f$ is single-valued and locally Lipschitz on $\operatorname{dom} f + \operatorname{dom}\phi$ and $e_\lambda^\phi f$ is differentiable on $\operatorname{dom} f + \operatorname{dom}\phi$ with $\nabla e_\lambda^\phi f$ locally Lipschitz. Furthermore the following identities hold:*

$$P_\lambda^\phi f = (I + \nabla\phi^* \circ \lambda\partial f)^{-1}$$

as well as

$$\nabla e_\lambda^\phi f = \big(\nabla\phi^* \circ \lambda I + (\partial f)^{-1}\big)^{-1}.$$

Proof. Since f is hypoconvex with constant $r \geq 0$ we know that f is in particular prox-bounded with some threshold $\lambda_f \geq 1/r$. Since ϕ is strongly convex, in view of Lemma 2.5, f is also anisotropically prox-bounded with threshold $\theta\lambda_f$. Since f is proper lsc and hypoconvex it is in particular prox-regular on $\operatorname{dom} f$ with uniform constants r and $\varepsilon = +\infty$. In view of the proof of Theorem 2.11 we obtain the claimed formulas, where T is replaced by ∂f, wich are valid on $U_\lambda = \operatorname{rge}(I + \nabla\phi^* \circ \lambda\partial f)$. Let $y \in U_\lambda$. I.e., there exists $x \in \mathbb{R}^m$ so that $y = x + \nabla\phi^*(\lambda v)$, for some $v \in \partial f(x)$ which means that $0 \in \partial f(x) - \lambda^{-1}\nabla\phi(z)$ for $x + z = y$ implying that $y \in \operatorname{dom} f + \operatorname{dom}\phi$ for $x \in \operatorname{dom} f$ and $z \in \operatorname{dom}\phi$. In view of Lemma 2.7 for any $y \in \operatorname{dom} f + \operatorname{dom}\phi$ there is $x \in P_\lambda^\phi f(y)$ such that $0 \in \partial f(x) - \lambda^{-1}\nabla\phi(y - x)$. This implies that $y \in \operatorname{rge}(I + \nabla\phi^* \circ \lambda\partial f)$ and therefore $U_\lambda = \operatorname{dom} f + \operatorname{dom}\phi$. □

2.2.3. Anisotropic prox-regularity: Beyond local strong convexity

Classical prox-regularity, intuitively, involves quadratic lower bounds with uniform curvature which exist relative to the domain of a graphical localization of the subdifferential at a certain reference point. In the previous sections, we therefore chose ϕ to be locally strongly convex and assumed the constant of local strong convexity of ϕ to match the curvature in the definition of prox-regularity. It is tempting to replace these quadratic lower bounds with lower bounds generated by the Legendre function ϕ that is used in the anisotropic proximal mapping. We conjecture, that without local strong convexity of ϕ the scaling $r\phi$ used in the previous section is not ideal, in particular in the context of Lemma 2.10. Instead the epi-scaling $(r \star \phi)(x) = r\phi(x/r)$, for $r > 0$, should be used. For simplicity, for the remainder of this section we will instead choose the scaling λ^{-1} to be fixed. For notational convenience we hide the scaling within ϕ. This leads us to the following generalization of prox-regularity which yields a sharp sufficient condition for the single-valuedness and existence of a resolvent expression of the anisotropic proximal mapping beyond local strong convexity:

Definition 2.13 (anisotropic prox-regularity). *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be an anisotropic prox-potential. A function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is anisotropically prox-regular relative to ϕ at $\bar{x}$ for $\bar{v}$ if f is finite and locally lsc at $\bar{x}$ with $\bar{v} \in \partial f(\bar{x})$, and there exists $\varepsilon > 0$ such that for all $\|x' - \bar{x}\| < \varepsilon$*

$$f(x') \geq f(x) - \phi(x - x' + \nabla\phi^*(v)) + \phi(\nabla\phi^*(v)), \tag{2.28}$$

whenever $\|x - \bar{x}\| < \varepsilon$, $f(x) - f(\bar{x}) < \varepsilon$, $v \in \partial f(x)$, $\|v - \bar{v}\| < \varepsilon$. When this holds for all $\bar{v} \in \partial f(\bar{x})$, f is said to be anisotropically prox-regular at $\bar{x}$.

The anisotropic subgradient inequality (2.28) generalizes the subgradient inequality (2.5) from classical prox-regularity when $\phi = (1/2)\|\cdot\|^2$ and one expands the square in $\phi(x - x' + \nabla\phi^*(v))$. Indeed, when the inequality holds strictly at $x \in \mathbb{R}^m$ for all $x' \neq x$ for some $v \in \partial f(x)$ it implies that the anisotropic proximal mapping $P_1^\phi f$ of f at the perturbed point $y := x + \nabla\phi^*(v)$ equals $x = P_1^\phi f(y)$.

In the next proposition we show that for $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ proper lsc and anisotropically prox-bounded, anisotropic prox-regularity is equivalent to the single-valuedness of the anisotropic proximal mapping and the existence of a resolvent expression which involves a graphical localization of ∂f.

Theorem 2.14. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and anisotropically prox-bounded relative to the anisotropic prox-potential $\phi \in \Gamma_0(\mathbb{R}^m)$ with threshold $\lambda_f > 1$. Let f be finite at $\bar{x}$ and $\bar{v} \in \partial f(\bar{x})$ such that $\{\bar{x}\} = P_1^\phi f(\bar{y})$. Then the following statements are equivalent.*

(i) *f is anisotropically prox-regular relative to ϕ at $\bar{x}$ for $\bar{v}$ such that the subgradient inequality* (2.28) *holds strictly for $x' \neq x$ with x', x near $\bar{x}$.*

(ii) *For all $(x, y), (x', y') \in \operatorname{gph}(I + \nabla\phi^* \circ T)$ with $x' \neq x$ the following anisotropic strict monotonicity inequality holds true:*

$$0 < \phi(y - x') - \phi(y - x) + \phi(y' - x) - \phi(y' - x'). \tag{2.29}$$

(iii) *$P_1^\phi f(y)$ is single-valued on a neighborhood of $\bar{y} = \bar{x} + \nabla\phi^*(\bar{v})$ such that for any y in that neighborhood*

$$P_1^\phi f(y) = (I + \nabla\phi^* \circ T)^{-1}(y),$$

where T is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$.

Proof. (i) $\implies$ (ii): Choose $(x,y),(x',y') \in \operatorname{gph}(I+\nabla\phi^* \circ T)$ with $x' \neq x$. This means there exist $v \in T(x)$ and $v' \in T(x')$ such that $y = x + \nabla\phi^*(v)$ and $y' = x' + \nabla\phi^*(v')$.

By assumption, we have

$$f(x') > f(x) - \phi(y-x') + \phi(y-x),$$

and

$$f(x) > f(x') - \phi(y'-x) + \phi(y'-x').$$

Summing the two we obtain:

$$0 < \phi(y-x') - \phi(y-x) + \phi(y'-x) - \phi(y'-x'),$$

as desired.

(ii) $\implies$ (iii): Let f be anisotropically prox-regular at $\bar{x}$ for $\bar{v}$ with $\varepsilon > 0$ such that the subgradient inequality (2.28) holds strictly for $x' \neq x$ with x', x near $\bar{x}$.

By assumption $P_1^\phi f(\bar{y}) = \bar{x}$ is single-valued.

In view of Lemma 2.7 we know that for any sequence $y^\nu \to \bar{y}$ for ν sufficiently large $y^\nu \in \operatorname{dom} e_1^\phi f$ there is $x^\nu \in P_1^\phi f(y^\nu) \neq \emptyset$ with $x^\nu \to \bar{x}$ and $e_1^\phi f(y^\nu) \to e_1^\phi f(\bar{y})$. From applying Fermat's rule Lemma 1.3 to the minimization problem above we know that $\nabla\phi(y^\nu - x^\nu) \in \partial f(x^\nu)$ and for ν sufficiently large we have that $\|\nabla\phi(y^\nu - x^\nu) - \bar{v}\| \leq \varepsilon$ due to the continuity of ϕ on a neighborhood of $\bar{y} - \bar{x}$. In addition, we have $f(x^\nu) \to f(\bar{x})$ as $e_1^\phi f(y^\nu) = f(x^\nu) + \phi(y^\nu - x^\nu) \to e_1^\phi f(\bar{y}) = f(\bar{x}) + \phi(\bar{y} - \bar{x})$. Overall this means that for any y sufficiently near $\bar{y}$ we have for any $x \in P_1^\phi f(y)$:

$$0 \in T(x) - \nabla\phi(y-x),$$

where T is an f-attentive localization of ∂f at $\bar{x}$ for $\bar{v}$. Therefore we have

$$\emptyset \neq P_1^\phi f(y) \subset \big(T - \nabla\phi(y - \cdot)\big)^{-1}(0) = (I + \nabla\phi^* \circ T)^{-1}(y). \tag{2.30}$$

Suppose that $(I+\nabla\phi^* \circ T)^{-1}(y)$ is not at most a singleton. This means there exists $x \in (I+\nabla\phi^* \circ T)^{-1}(y)$ and $x' \in (I+\nabla\phi^* \circ T)^{-1}(y)$ with $x' \neq x$. Thus there exist $v \in T(x)$ and $v' \in T(x')$ such that $y = x + \nabla\phi^*(v)$ and $y = x' + \nabla\phi^*(v')$. Applying the anisotropic monotonicity inequality (2.29) we obtain since $x' \neq x$

$$0 < \phi(y-x') - \phi(y-x) + \phi(y-x) - \phi(y-x') = 0,$$

a contradiction. This means $(I+\nabla\phi^* \circ T)^{-1}(y)$ is at most a singleton and therefore the inclusion above holds with equality

$$\emptyset \neq P_1^\phi f(y) = \big(T + \nabla\phi(y - \cdot)\big)^{-1}(0) = (I + \nabla\phi^* \circ T)^{-1}(y),$$

and $(I+\nabla\phi^* \circ T)^{-1}$ is single-valued.

(iii) $\implies$ (i): Let the conditions in (iii) hold. Let $(x,v) \in \operatorname{gph} T$ and define $y = x + \nabla\phi^*(v)$ where we choose $\|x - \bar{x}\| \leq \varepsilon$ and $\|v - \bar{v}\| \leq \varepsilon$ with $\varepsilon > 0$ sufficiently small such that y sufficiently near $\bar{y}$. By assumption we have for y that $P_1^\phi f(y) = (I+\nabla\phi^* \circ T)^{-1}(y) = x$. This means that

$$f(x') + \phi(y-x') > f(x) + \phi(y-x),$$

for all $x' \neq x$. Substituting the relation $y = x + \nabla\phi^*(v)$ this means

$$f(x') > f(x) - \phi(x + \nabla\phi^*(v) - x') + \phi(\nabla\phi^*(v)),$$

which holds in particular for any $\|x' - \bar{x}\| \leq \varepsilon$ and the conclusion follows. □

We shall also prove the local $\mathcal{C}^1$ property of the anisotropic Moreau envelope along with an expression of its gradient in terms of the anisotropic Yosida regularization of T under anisotropic prox-regularity:

Theorem 2.15. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and anisotropically prox-bounded relative to the anisotropic prox-potential $\phi \in \Gamma_0(\mathbb{R}^m)$ with threshold $\lambda_f > 1$. In the situation of Theorem 2.14 assume that one of the three equivalent conditions holds true. Then e_1^ϕ is continuously differentiable on a neighborhood of $\bar{y} = \bar{x} + \nabla\phi^*(\bar{v})$ with*

$$\nabla e_1^\phi f = \nabla\phi \circ (I - P_1^\phi f) = \big(\nabla\phi^* + T^{-1}\big)^{-1}. \tag{2.31}$$

Since the negative anisotropic Moreau envelope can be understood as a pointwise maximum $f(x) = \max_{t\in T} f_t(x)$ over a collection of $\mathcal{C}^1$-functions f_t, we will invoke [RW98, Theorem 10.31] which shows that the limiting subdifferential of such a function can be written in terms the convex hull of the gradients of the active $\mathcal{C}^1$-pieces: $\partial f(x) = \operatorname{con}\{\nabla f_t(x) : t \in T(x)\}$ for $T(x) = \{t \in T : f_t(x) = f(x)\}$. This is completely analogous to the subdifferential of a convex function which contains the slopes of all lower supporting affine functions. Subsmoothness is defined according to [RW98, Definition 10.29]:

Definition 2.16 (subsmooth functions)**.** *A function $f : O \to \mathbb{R}$, where O is an open set in $\mathbb{R}^m$, is said to be lower-$\mathcal{C}^1$ on O, if on some neighborhood V of each $\bar{x} \in O$ there is a representation*

$$f(x) = \max_{t\in T} f_t(x), \tag{2.32}$$

in which the functions f_t are of class $\mathcal{C}^1$ on V and the index set T is a compact space such that $f_t(x)$ and $\nabla f_t(x)$ depend continuously not just on $x \in V$ but jointly on $(t, x) \in T \times V$. More generally, f is lower-$\mathcal{C}^k$ on O if such a local representation can be arranged in which the functions f_t are of class $\mathcal{C}^k$, with $f_t(x)$ and all its partial derivatives through order k depending continuously not just on x but on (t, x). Similarly, upper-$\mathcal{C}^k$ functions are defined in terms of a minimum in place of a maximum. Thus, f is upper-$\mathcal{C}^k$ when $-f$ is lower-$\mathcal{C}^k$.

We prove the subsmoothness property of the anisotropic Moreau envelope along with a representation in terms of the anisotropic proximal mapping. In contrast to the other statements in this section we treat the scaling λ explicitly:

Lemma 2.17. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and anisotropically prox-bounded relative to the anisotropic prox-potential $\phi \in \Gamma_0(\mathbb{R}^m)$ with threshold $\lambda_f > 0$. Then for any $\lambda \in (0, \lambda_f)$ the envelope function $-e_\lambda^\phi f$ is lower-$\mathcal{C}^1$ with representations*

$$-e_\lambda^\phi f = \max_{x\in Z} h(x, \cdot), \qquad P_\lambda^\phi f = \operatorname*{arg\,max}_{x\in Z} h(x, \cdot),$$

on $V \subset \operatorname{dom} f + \operatorname{dom}\phi$ being a neighborhood of $\bar{y} \in \operatorname{dom} f + \operatorname{dom}\phi$, $h(x, y) := -f(x) - \lambda^{-1}\phi(y - x)$ and $P_\lambda^\phi f(y) \subset Z$, with both $h(x, y)$, $\nabla_y h(x, y) = -(1/\lambda)\nabla\phi(y - x)$, depending continuously on $(x, y) \in Z \times V$.

Proof. Let $\lambda \in (0, \lambda_f)$ and $\bar{y} \in \operatorname{dom} f + \operatorname{dom} \phi$. Let $V \subset \operatorname{dom} f + \operatorname{dom} \phi$ be a compact neighborhood of $\bar{y}$. Choose $Z := \{x \in P_\lambda^\phi f(y) : y \in V\}$. We have $\arg\max_{x \in Z} h(x, y) = P_\lambda^\phi f(y)$ and due to Lemma 2.7(i) we have

$$-e_\lambda^\phi f(y) = \max_{x \in Z} h(x, y).$$

Furthermore Z is closed. Otherwise there is a sequence $x^\nu \to x$ with $x^\nu \in P_\lambda^\phi f(y^\nu)$ for some $y^\nu \in V$ with $x \notin Z$. Taking a convergent subsequence $y^{\nu_j} \to y \in V$ (which exists as V is bounded) we know from Lemma 2.7(iii) that $x^{\nu_j} \to x \in P_\lambda^\phi f(y) \subset Z$, a contradiction. Z is also bounded, hence compact. Otherwise there is a sequence $\|x^\nu\| \to \infty$ with $x^\nu \in P_\lambda^\phi f(y^\nu)$ for some $y^\nu \in V$. However, going to a convergent subsequence $y^{\nu_j} \to y \in V$ we have for $x^{\nu_j} \in P_\lambda^\phi f(y^{\nu_j})$ using Lemma 2.7(iii) that $\{x^{\nu_j}\}_{\nu \in \mathbb{N}}$ is bounded, a contradiction.

Next we show that f becomes continuous over the compact space Z. Equivalently this means h is continuous over $Z \times V$. We assume the contrary: Suppose that f is not continuous over Z. This means there is a sequence $\{x^\nu\}_{\nu \in \mathbb{N}} \subset Z$ with $x^\nu \to x^* \in Z$ such that $f(x^\nu) \not\to f(x^*)$. For any $x^\nu \in Z$ there exists $y^\nu \in V$ such that $x^\nu \in P_\lambda^\phi f(y^\nu)$ and $f(x^\nu) \leq f(x^\nu) + \lambda^{-1}\phi(y^\nu - x^\nu) = e_\lambda^\phi f(y^\nu) \leq \gamma$, for some $\infty > \gamma$, since $e_\lambda^\phi f$ is continuous and V is compact. Since f is proper, lsc it is also uniformly bounded from below over Z: $-\infty < \delta \leq f(x^\nu)$. This means we can find a subsequence indexed by ν_j such that $f(x^{\nu_j}) \to f^* \geq f(x^*) + \varepsilon$, with $\varepsilon > 0$. Taking another subsequence if necessary we can ensure that $y^{\nu_j} \to y^*$ and $x^* \in P_\lambda^\phi f(y^*)$. By continuity of the envelope function, cf. Lemma 2.7(ii), we then have $e_\lambda^\phi f(y^{\nu_j}) = f(x^{\nu_j}) + \lambda^{-1}\phi(y^{\nu_j} - x^{\nu_j}) \to e_\lambda^\phi f(y^*) = f(x^*) + \lambda^{-1}\phi(y^* - x^*)$. Hence, along that subsequence $f(x^{\nu_j}) \to f(x^*)$, a contradiction.

Since $h(x, \cdot)$ is $\mathcal{C}^1$ both $h(x, y)$ and $\nabla_y h(x, y) = -(1/\lambda)\nabla\phi(y - x)$, depend continuously jointly on $(x, y) \in Z \times V$. Hence $-e_\lambda^\phi f$ is lower-$\mathcal{C}^1$. □

We are now ready to prove Theorem 2.15 invoking [RW98, Theorem 10.31] and the single-valuedness of the anisotropic proximal mapping from Theorem 2.14:

Proof of Theorem 2.15. In view of Lemma 2.17 and invoking [RW98, Theorem 10.31] we obtain that

$$\begin{aligned} \partial(-e_1^\phi f)(y) &= \operatorname{con}\left\{\nabla_y h(x, y) : x \in P_1^\phi f(y)\right\} \\ &= -\nabla\phi(y - \operatorname{con} P_1^\phi f(y)). \end{aligned} \tag{2.33}$$

Due to the assumptions we can invoke Theorem 2.14 and assert that $P_1^\phi f$ is singled-valued and continuous at y near $\bar{y} = \bar{x} + \nabla\phi^*(\bar{v})$. This means that $\partial(-e_1^\phi f)$ is single-valued and continuous around $\bar{y}$. Through [RW98, Corollary 9.19] we obtain that $-e_1^\phi$ is $\mathcal{C}^1$ around $\bar{y}$ with

$$\nabla\phi(y - P_1^\phi f(y)) = \nabla e_1^\phi f(y).$$

The identity for the anisotropic Yosida regularization follows along the same lines as in the proof of Theorem 2.11. □

It is tempting to study the case when anisotropic prox-regularity holds globally and strictly for all $x \in \mathbb{R}^m$ and all $y \in \partial f(x)$ with a uniform constant $\varepsilon = \infty$. According to the theorem above this is equivalent to the global single-valuedness and existence of a resolvent expression of the anisotropic proximal mapping.

Actually, in the quadratic case this condition is equivalent to the hypoconvexity of the classical proximal mapping. In [RW98] such functions are also called λ-*proximal* [RW98, Example 1.44]. In Section 2.3 we will consider the Bregman proximal mapping and obtain that the corresponding globalized Bregmanian subgradient inequality is equivalent to relative hypoconvexity, see Proposition 2.45. The characterization of the class of functions for which the anisotropic subgradient inequality holds in the aforementioned global sense is interesting but future work: We conjecture, that there is a close connection to generalized conjugacy in close analogy to the *proximal transform*, see Definition 3.3 in Chapter 3, originally due to [RW98, Example 11.64]. Related questions are also addressed in [CJT17].

2.2.4. A nonconvex feasibility problem with anisotropic proximal mapping

In this section we showcase a simple toy example for using the anisotropic proximal mapping in an algorithm. We consider the problem of computing a matrix $Y \in \mathbb{R}^{n\times m}$ that has rank r and satisfies a set of linear equality constraints. Mathematically, such a problem can be cast as a feasibility problem where one seeks to find a matrix $Y \in \mathcal{C} \cap \mathcal{D} \subset \mathbb{R}^{n\times m}$ for $\mathcal{C} := \{Y \in \mathbb{R}^{n\times m} : A(Y) = B\}$ for a linear map $A : \mathbb{R}^{n\times m} \to \mathbb{R}^d$ and $\mathcal{D} := \{Y \in \mathbb{R}^{n\times m} : \operatorname{rank} Y = r\}$. Such an experiment has been considered previously for instance in [LP16; Och18]. The problem is traditionally formulated as an optimization problem where the objective $e_\lambda f + g$ is the sum of the quadratic Euclidean distance function $e_\lambda f(Y) = \inf_{X\in\mathcal{C}} \|X - Y\|^2$ to the set $\mathcal{C}$ for $f := \iota_\mathcal{C}$ and the indicator function $g = \iota_\mathcal{D}$ of the set $\mathcal{D}$. Projected gradient descent then results in the standard alternating

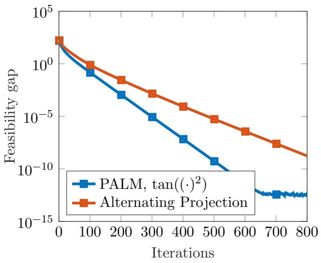


Figure 2.1.: Feasibility problem. Comparison of the alternating projection method and PALM for anisotropic ϕ as defined in (2.34). Achieved precision vs. required iterations is plotted in log-scale for the y-axis.

Table 2.1.: Feasibility problem. Precision vs. required iterations. The values are averaged over 200 randomly generated matrices. It can be seen that PALM with anisotropic ϕ performs consistently better than alternating projection. The dash "–" indicates that the algorithms did not reach the desired precision after a maximum number of 800 iterations.

	Iterations				
Precision	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}
PALM, $\phi = \tan((\cdot)^2)$	119	193	268	344	421
Alternating Projection	244	403	569	–	–

projection method:

$$Y^{t+1} = \mathrm{proj}_{\mathcal{C}}(\mathrm{proj}_{\mathcal{D}}(Y^t)).$$

Instead, we reformulate the problem as an optimization problem of the form

$$\text{mimize}\left\{ e^{\phi}_{\lambda} f(Y) + g(Y) : Y \in \mathbb{R}^{n\times m} \right\}.$$

We replace the indicator function of the set $\mathcal{D}$ with its anisotropic Moreau envelope for the separable anisotropic prox-potential

$$\phi(X) := \sum_{i=1}^{n}\sum_{j=1}^{m} \eta \tan \frac{X_{ij}^2}{\eta^2}, \tag{2.34}$$

where we define $\tan(x^2) = +\infty$, if $x \notin (-\sqrt{\pi/2}, \sqrt{\pi/2})$. Here, η is a scaling parameter chosen sufficiently large, so that $Y^0 \in \operatorname{dom}\phi$. Then we can equivalently rewrite the problem as

$$\min_{X,Y\in\mathbb{R}^{n\times m}} \iota_{\mathcal{D}}(X) + \phi(Y - X) + \iota_{\mathcal{C}}(Y). \tag{2.35}$$

We solve the problem via *proximal alternating linearized minimization* (PALM) [BST14] where in each iteration we replace the coupling term $\phi(Y - X)$ with a proximal linearization:

$$X^{t+1} := \operatorname*{arg\,min}_{X\in\mathbb{R}^{n\times m}} \iota_{\mathcal{D}}(X) - \langle \nabla\phi(Y^t - X^t), X^t - Y^t\rangle + \frac{1}{2\sigma}\|X - X^t\|^2 \tag{2.36}$$

$$Y^{t+1} := \operatorname*{arg\,min}_{Y\in\mathbb{R}^{n\times m}} \iota_{\mathcal{C}}(Y) + \langle \nabla\phi(Y^t - X^{t+1}), X^t - Y^t\rangle + \frac{1}{2\tau}\|Y - Y^t\|^2. \tag{2.37}$$

In each iteration τ, σ are chosen sufficiently small (via a line search), to guarantee a sufficient descent and, at the same time that $Y^t - X^{t+1}, Y^{t+1} - X^{t+1} \in \operatorname{dom}\phi$. This can be guaranteed (locally) via the continuity of the Euclidean projections under prox-regularity and the openness of $\operatorname{dom}\phi$. For a numerical evaluation we consider the setting in [Och18], i.e., we randomly generate matrices $A \in \mathbb{R}^{d\times nm}$, $R_1 \in \mathbb{R}^{n\times r}$ $R_2 \in \mathbb{R}^{r\times m}$ and define $Y^0 := R_1 R_2$, so that it has rank r and $B := AY^0$. This ensures that $Y^0 \in \mathcal{C}\cap\mathcal{D} \neq \emptyset$. We manually choose $m = 110$, $n = 100$, $r = 4$, $d = 150$. The entries of the matrices are normal distributed and normalized to $[-1, 1]$. We compare the alternating projection method with PALM using the anisotropic prox-potential (2.34). The results are shown in Table 2.1 and for a representative example in Figure 2.1. We believe that in some situations the nonlinear preconditioner $\nabla\phi^*$ helps to improve the condition of the problem. Regretfully, we currently do not have a systematic and theoretically justified strategy to obtain good choices for ϕ which provably lead to better performance than plain quadratics. Indeed, the surprisingly good performance of PALM for the tan potential is rather due to a combination with a fine-tuned line search which accepts very large steps in our situation. A theoretical justification which for instance involves or expands upon linear regularity of sets is an interesting direction for future research.

2.3. Bregman–Moreau envelope and proximal mapping

2.3.1. Definition and continuity properties

In this section we consider the Bregman proximal mapping and associated Bregman envelope function. Like the anisotropic proximal mapping the Bregman proximal mapping is a generalization of the Euclidean proximal mapping. In the convex setting [CR13] has revealed an interesting connection between the two via a generalization of Moreau's identity [Mor62; Mor65]. The Bregman proximal mapping is obtained by replacing the quadratic penalty $\|x-y\|^2$ in the Euclidean proximal mapping with a Bregman distance $D(x,y)$, see Definition 1.19. In contrast to the anisotropic proximal mapping, considered in Section 2.2, which yields a formulation based on inf-convolution, the Bregmanian formulation is based on inf-projection:

$$\inf_{x\in\mathbb{R}^m} f(x) + D(x,y).$$

Since Bregman distances are asymmetric in general, in accordance with [BCN06], we will, depending on the order of arguments in D, distinguish a left (x first) and a right (y first) Bregman proximal mapping. In contrast to the left Bregman proximal mapping which, for convex f, is a convex minimization problem, the right Bregman proximal mapping results in a generally nonconvex minimization problem even for convex f, see Example 2.62.

The Bregman proximal mapping was studied extensively in the convex setting, see, e.g., [BCN06; BDL18] and more recently in the nonconvex setting under relative hypoconvexity [KS12]. Relative hypoconvexity of a function f, i.e., $\lambda f+\phi$ is convex for some $\lambda>0$ sufficiently small, is a generalization of Euclidean hypoconvexity, i.e., $\lambda f+\|\cdot\|^2$ is convex. Like its Euclidean counterpart it is a sufficient condition for the single-valuedness of the proximal mapping and the existence of a resolvent identity. However, not every function is (relatively) hypoconvex. For instance a function that is hypoconvex must have a convex domain. Expanding upon existing related work we introduce the concept of relative prox-regularity which also holds for functions which are not relatively hypoconvex. Like Euclidean prox-regularity it is a local condition and yields a sufficient condition for the local single-valuedness of the Bregman proximal mapping and the existence of a Bregman resolvent expression which holds locally in terms of a graphical localization of the limiting subdifferential.

More formally, we define the left Bregman–Moreau envelope and proximal mapping with step-size parameter $\lambda>0$ in accordance with [KS12]. In addition we allow f to be possibly improper.

Definition 2.18 (left Bregman–Moreau envelope and proximal mapping)**.** *Let $\phi\in\Gamma_0(\mathbb{R}^m)$ be Legendre and $f\colon\mathbb{R}^m\to\overline{\mathbb{R}}$ be extended real-valued (and possibly improper). For some $\lambda>0$ and $y\in\mathbb{R}^m$ we define the left Bregman–Moreau envelope (in short: left envelope) of f at y as*

$$\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f(y) = \inf_{x\in\mathbb{R}^m} f(x) \mathbin{\dot{+}} \frac{1}{\lambda} D_\phi(x,y), \tag{2.38}$$

and the associated left Bregman proximal mapping (in short: left prox) of f at y as

$$\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f(y) = \operatorname*{arg\,min}_{x\in\mathbb{R}^m} f(x) \mathbin{\dot{+}} \frac{1}{\lambda} D_\phi(x,y). \tag{2.39}$$

From the definition, it is clear that $\operatorname{dom}(\overleftarrow{\operatorname{prox}}_{\lambda}^{\phi} f) \subset \operatorname{int}(\operatorname{dom}\phi)$ and $\operatorname{dom}(\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f) \subset \operatorname{int}(\operatorname{dom}\phi)$. The arithmetic of extended real values, see Section 1.2.2, ensures well-definedness for improper functions: In particular we have

$$\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f(y) = +\infty, \qquad \overleftarrow{\operatorname{prox}}_{\lambda}^{\phi} f(y) = \emptyset,$$

whenever $y \notin \operatorname{int}(\operatorname{dom}\phi)$ even in case there is $x \in \operatorname{dom}\phi$ with $f(x) = -\infty$.

The set $\overleftarrow{\operatorname{prox}}_{\lambda}^{\phi} f(y)$ is possibly empty in the nonconvex setting. The following prox-boundedness condition, which we adapt from [KS12, Definition 2.3], comes in handy to guarantee that the Bregman proximal mapping is nonempty. In addition, this condition guarantees that the envelope function is continuous relative to $\operatorname{int}(\operatorname{dom}\phi)$ and the proximal mapping is outer semicontinuous, see Definition 1.11, relative to $\operatorname{int}(\operatorname{dom}\phi)$.

Definition 2.19 (relative prox-boundedness)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and supercoercive and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be extended real-valued. We say f is prox-bounded relative to ϕ if there exists $\lambda > 0$ such that $\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f(y) > -\infty$ for some $y \in \operatorname{int}(\operatorname{dom}\phi)$. The supremum of the set of all such λ is the threshold λ_f of the prox-boundedness, i.e.*

$$\lambda_f = \sup\bigl\{\lambda > 0 : \exists\, y \in \operatorname{int}(\operatorname{dom}\phi) : \overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f(y) > -\infty\bigr\}.$$

We complement the results of [KS12] by stating equivalent characterizations of relative prox-boundedness. To this end, we first prove the following lemma, whose proof is analogous to [RW98, Exercise 1.14].

Lemma 2.20. *Let $\phi\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, lsc and coercive with $\operatorname{dom}\phi = \mathbb{R}^m$ and let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper and lsc. Then we have the identity*

$$\liminf_{\|x\|\to\infty} \frac{f(x)}{\phi(x)} = \sup\left\{\gamma \in \mathbb{R} \,:\, \exists\, \beta \in \mathbb{R} \text{ with } f(x) \geq \gamma\phi(x) + \beta \text{ for all } x \in \mathbb{R}^m\right\}.$$

Proof. Note that since ϕ is coercive, we have $\phi(x) \to \infty$, whenever $\|x\| \to \infty$. Let $\bar\gamma := \liminf_{\|x\|\to\infty} f(x)/\phi(x)$ and

$$\gamma \in \Gamma := \left\{\gamma \in \mathbb{R} \,:\, \exists\, \beta \in \mathbb{R} \text{ with } f(x) \geq \gamma\phi(x) + \beta \text{ for all } x \in \mathbb{R}^m\right\}.$$

This means that there exists β such that $f(x) \geq \gamma\phi(x) + \beta$, for all $x \in \mathbb{R}^m$. Dividing by $\phi(x) > 0$ which holds for $\|x\| > t$, for some $t > 0$, and taking the lim inf on both sides yields $\liminf_{\|x\|\to\infty} f(x)/\phi(x) \geq \gamma + 0$, meaning that $\bar\gamma \geq \gamma$. Now let $\gamma \in \mathbb{R}$ with $\gamma < \bar\gamma$. Suppose that for any compact level set $\emptyset \neq \operatorname{lev}_{\leq r}\phi := \{x \in \mathbb{R}^m : \phi(x) \leq r\}$ with $r > 0$ there exists $x \in (\operatorname{lev}_{\leq r}\phi)^c \neq \emptyset$ in the complement of $\operatorname{lev}_{\leq r}\phi$, which is nonempty due to the coercivity of ϕ, with $f(x) < \gamma\phi(x)$. In particular this means that there is a sequence $x^\nu \in \mathbb{R}^m$ with $\phi(x^\nu) \to \infty$ and $f(x^\nu)/\phi(x^\nu) < \gamma$. Taking the lim inf on both sides implies due to the coercivity of ϕ that

$$\bar\gamma = \lim_{r\to\infty}\left(\inf_{r<\|x\|} \frac{f(x)}{\phi(x)}\right) \leq \limsup_{\nu\to\infty} \frac{f(x^\nu)}{\phi(x^\nu)} \leq \gamma < \bar\gamma,$$

which is a contradiction. This means that there is $r > 0$ such that for any $x \in (\operatorname{lev}_{\leq r}\phi)^c \neq \emptyset$ we have $f(x) \geq \gamma\phi(x)$. By assumption $h := f - \gamma\phi$ is proper lsc. In view of [RW98, Corollary 1.10], h is uniformly bounded from below on $\operatorname{lev}_{\leq r}\phi$, showing that for some $\beta \in \mathbb{R}$ sufficiently small, it holds $f(x) \geq \gamma\phi(x) + \beta$ for any $x \in \operatorname{lev}_{\leq r}\phi$. Overall we have $f(x) \geq \gamma\phi(x) + \beta$, for all $x \in \mathbb{R}^m$ and we have $\gamma \in \Gamma$. This shows that we can find a sequence $\{\gamma^\nu\}_{\nu\in\mathbb{N}} \subset \Gamma$ with $\gamma^\nu < \bar{\gamma}$ and $\gamma^\nu \to \bar{\gamma}$ showing that $\sup \Gamma \geq \bar{\gamma}$. Overall we have $\bar{\gamma} = \sup \Gamma$. □

The following proposition adapts [RW98, Exercise 1.24] to a Bregmanian setting where we also account for possibly imporper functions.

Proposition 2.21 (characterization of relative prox-boundedness)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and super-coercive and let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be extended real-valued. Then, the following properties are equivalent:*

(i) *f is prox-bounded relative to ϕ.*

(ii) *for some $r > 0$ the function $f \dot{+} r\phi$ is uniformly bounded from below on $\mathbb{R}^m$.*

If, furthermore, $\operatorname{dom}\phi = \mathbb{R}^m$ *and f is proper and lsc, the above properties are equivalent to*

$$\liminf_{\|x\|\to\infty} \frac{f(x)}{\phi(x)} > -\infty. \tag{2.40}$$

Proof. (i) $\implies$ (ii): Let f be prox-bounded relative to ϕ with threshold $\lambda_f > 0$. This means there is $\lambda \in (0, \lambda_f)$ and $y \in \operatorname{int}(\operatorname{dom}\phi)$ such that

$$\overleftarrow{\operatorname{env}}{}_\lambda^\phi f(y) = \inf_{x\in\mathbb{R}^m} f(x) \dot{+} \frac{1}{\lambda}\phi(x) - \frac{1}{\lambda}\phi(y) - \frac{1}{\lambda}\langle \nabla\phi(y), x-y\rangle > -\infty.$$

This implies that there is $\beta > -\infty$ and we have for $r > 1/\lambda$

$$f(x) \dot{+} r\phi(x) \geq \beta + (r - \lambda^{-1})\phi(x) - \frac{1}{\lambda}\langle \nabla\phi(y), x\rangle + \frac{1}{\lambda}\phi(y) + \frac{1}{\lambda}\langle \nabla\phi(y), y\rangle.$$

Since $\phi \in \Gamma_0(\mathbb{R}^m)$ is super-coercive we know that $(r - \lambda^{-1})\phi(x) - \langle \lambda^{-1}\nabla\phi(y), x\rangle$ is uniformly bounded from below and the assertion follows.

(ii) $\implies$ (i): Let $r > 0$. Then there exists $\beta \in \mathbb{R}$ such that $f(x) \dot{+} r\phi(x) \geq \beta$ for any $x \in \mathbb{R}^m$. Adding $-r\phi(\nabla\phi^*(0))$ to both sides of the inequality yields

$$f(x) \dot{+} rD_\phi(x, \nabla\phi^*(0)) \geq \beta - r\phi(\nabla\phi^*(0)),$$

for all $x \in \mathbb{R}^m$ and the assertion follows for $y := \nabla\phi^*(0)$ and $\lambda := 1/r$.

To show the remaining statement, assume that $\operatorname{dom}\phi = \mathbb{R}^m$ and f is proper and lsc and let (2.40) hold. In view of Lemma 2.20, we have that

$$\sup\{\gamma \in \mathbb{R} \,:\, \exists\, \beta \in \mathbb{R} \text{ with } f(x) \geq \gamma\phi(x) + \beta \text{ for all } x \in \mathbb{R}^m\} > -\infty.$$

Then there exists a finite $+\infty > \gamma > -\infty$ such that $f(x) \geq \gamma\phi(x) + \beta$ holds for some $\beta \in \mathbb{R}$ and any $x \in \mathbb{R}^m$. For $r > \max\{0, -\gamma\}$, we have that $r + \gamma \geq 0$ and

$$f + r\phi \geq (r + \gamma)\phi + \beta > -\infty,$$

since $\phi \in \Gamma_0(\mathbb{R}^m)$ is coercive and therefore bounded from below on $\mathbb{R}^m$, meaning we have (ii).

Assume (ii) holds. By assumption there is some $\beta \in \mathbb{R}^m$ such that for any $x \in \mathbb{R}^m$ we have:

$$f(x) > -r\phi(x) + \beta.$$

Let $\Gamma := \{\gamma \in \mathbb{R} : \exists \beta \in \mathbb{R} \text{ with } f(x) \geq \gamma\phi(x) + \beta \text{ for all } x \in \mathbb{R}^m\}$. Then $-r \in \Gamma$ and in view of Lemma 2.20, we have $\liminf_{\|x\|\to\infty} f(x)/\phi(x) = \sup \Gamma > -\infty$. □

Prox-boundedness also allows us to extract a continuity property for both the Bregman proximal mapping and the Bregman–Moreau envelope. The following result summarizes important properties of the left envelope adapted from [KS12].

Lemma 2.22 (continuity properties of the left prox and envelope)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and super-coercive and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, lsc and relatively prox-bounded with threshold λ_f and let $\lambda \in (0, \lambda_f)$. Assume that* $\operatorname{dom}\phi \cap \operatorname{dom} f \neq \emptyset$*. Then $\overleftarrow{\operatorname{prox}}{}_\lambda^\phi f$ and $\overleftarrow{\operatorname{env}}{}_\lambda^\phi f$ have the following properties:*

(i) $\overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(y) \neq \emptyset$ *is compact for all $y \in \operatorname{int}(\operatorname{dom}\phi)$ and the envelope $\overleftarrow{\operatorname{env}}{}_\lambda^\phi f$ is proper.*

(ii) *The envelope $\overleftarrow{\operatorname{env}}{}_\lambda^\phi f$ is continuous on* $\operatorname{int}(\operatorname{dom}\phi)$*.*

(iii) *For any sequence $y^\nu \to y^* \in \operatorname{int}(\operatorname{dom}\phi)$ and $x^\nu \in \overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(y^\nu)$ we have $\{x^\nu\}_{\nu\in\mathbb{N}}$ is bounded and all its cluster points x^* lie in $\overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(y^*)$.*

For completeness we shall provide a proof generalizing the proof of [Bau+09, Theorem 4.3] for the Bregman projection to Bregman proximal mapping case:

Proof of Lemma 2.22. Fix $\bar{y} \in \operatorname{int}(\operatorname{dom}\phi)$ and $\varepsilon > 0$ such that $B_\varepsilon(\bar{y}) := \{y \in \mathbb{R}^m : \|y - \bar{y}\| \leq \varepsilon\} \subset \operatorname{int}(\operatorname{dom}\phi)$. Choose $\lambda \in (0, \lambda_f)$. Consider the function $h : \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$ defined by

$$h(x, y) = f(x) + \frac{1}{\lambda} D_\phi(x, y) + \iota_{B_\varepsilon(\bar{y})}(y).$$

Observe that $\operatorname{dom} h = (\operatorname{dom} f \cap \operatorname{dom}\phi) \times B_\varepsilon(\bar{y})$ and h is proper lsc. For every $y \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$ we have:

$$\{x \in \mathbb{R}^m : h(x, y) \leq \alpha\} = \begin{cases} \{x \in \mathbb{R}^m : \lambda^{-1} D_\phi(x, y) + f(x) \leq \alpha\}, & \text{if } y \in B_\varepsilon(\bar{y}), \\ \emptyset, & \text{otherwise.} \end{cases} \tag{2.41}$$

Next we prove that h is level-bounded in x locally uniformly in y in order to apply Theorem 1.13: To this end suppose that h is not level-bounded in x locally uniformly in y. This means there is $y^* \in B_\varepsilon(\bar{y})$ and $\alpha \in \mathbb{R}$ and there exist sequences $y^\nu \to y^*$ and x^ν with $\|x^\nu\| \to \infty$ such that $h(x^\nu, y^\nu) \leq \alpha$. On the other hand since f is prox-bounded relative to ϕ, there exists $y \in \operatorname{int}(\operatorname{dom}\phi)$ such that for $\lambda' \in (0, \lambda_f)$ with $\lambda' > \lambda$ we have

$$f(x^\nu) + \frac{1}{\lambda'} D_\phi(x^\nu, y) \geq \inf_{x\in\mathbb{R}^m} f(x) + \frac{1}{\lambda'} D_\phi(x, y) \geq \beta > -\infty. \tag{2.42}$$

Summing yields:

$$\frac{1}{\lambda} D_\phi(x^\nu, y^\nu) + \iota_{B_\varepsilon(\bar{y})}(y^\nu) - \frac{1}{\lambda'} D_\phi(x^\nu, y) \leq \alpha - \beta.$$

We know that $\iota_{B_\varepsilon(\bar{y})}(y^\nu) = 0$ since $h(x^\nu, y^\nu) \leq \alpha$ and $\operatorname{dom} h = (\operatorname{dom} f \cap \operatorname{dom}\phi) \times B_\varepsilon(\bar{y})$ and expanding the Bregman distance yields:

$$\begin{aligned} &\left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right) \phi(x^\nu) - \langle x^\nu, \lambda^{-1}\nabla\phi(y^\nu) - (1/\lambda')\nabla\phi(y)\rangle \\ &\qquad \leq \alpha - \beta + \frac{1}{\lambda}\phi(y^\nu) - \frac{1}{\lambda'}\phi(y) - \frac{1}{\lambda}\langle\nabla\phi(y^\nu), y^\nu\rangle + \frac{1}{\lambda'}\langle\nabla\phi(y), y\rangle. \end{aligned}$$

Since $\nabla\phi$ is continuous relative to $\operatorname{int}(\operatorname{dom}\phi)$ and $y^\nu \to y^* \in \operatorname{int}(\operatorname{dom}\phi)$ and ϕ is super-coercive we have $(1/\lambda - 1/\lambda')\,\phi(x^\nu) - \langle \lambda^{-1}\nabla\phi(y^\nu) - (1/\lambda')\nabla\phi(y), x^\nu\rangle \to \infty$, a contradiction. This means that $h(x,y)$ is level-bounded in x locally uniformly in y.

Define a function p at $y \in \mathbb{R}^m$ by

$$p(y) := \inf_{x\in\mathbb{R}^m} h(x,y) = \begin{cases} \overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f(y), & \text{if } y \in B_\varepsilon(\bar{y}),\\ +\infty, & \text{otherwise.}\end{cases}$$

Then $p = \overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f + \iota_{B_\varepsilon(\bar{y})}$ and

$$P(y) := \operatorname*{arg\,min}_{x\in\mathbb{R}^m}\ h(x,y) = \begin{cases} \overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y), & \text{if } y \in B_\varepsilon(\bar{y}),\\ \emptyset, & \text{otherwise.}\end{cases}$$

Now Theorem 1.13(i) implies that $\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y)$ is nonempty and compact whenever $y \in B_\varepsilon(\bar{y})$. Let $\bar{x} \in \overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(\bar{y})$. As

$$h(\bar{x},\cdot) = f(\bar{x}) + \frac{1}{\lambda} D_\phi(\bar{x},\cdot),$$

is continuous at $\bar{y}$, in view of Theorem 1.13(iii), the function p is continuous at $\bar{y}$ and therefore $\overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f$ is continuous at $\bar{y}$. Since $\bar{y} \in \operatorname{int}(\operatorname{dom}\phi)$ is arbitrary, we have Items (i) and (ii). Item (iii) follows directly from Theorem 1.13(ii). □

Note that in general the left Bregman–Moreau envelope is not lsc relative to $\mathbb{R}^m$, cf. [BCN06, Remark 3.6].

In Bregman proximal algorithms well-definedness is crucial, i.e., the output of one Bregman proximal step must be compatible with the input of the next iteration. Usually, this can be achieved by the property

$$\operatorname{rge}\big(\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f\big) \subset \operatorname{int}(\operatorname{dom}\phi),$$

which, however, requires a *constraint qualification* (CQ):

Lemma 2.23. *Let $\lambda > 0$, $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc. Assume that* $\operatorname{dom}\phi \cap \operatorname{dom} f \neq \emptyset$ *and that the following constraint qualification holds:*

$$\partial^\infty f(x) \cap -N_{\operatorname{dom}\phi}(x) = \{0\}, \tag{2.43}$$

for any $x \in \operatorname{dom} f \cap \operatorname{dom}\phi$. Then we have that

$$\begin{aligned}\operatorname{rge}\big(\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f\big) &\subset \operatorname{rge}\big((\partial(\phi+\lambda f))^{-1} \circ \nabla\phi\big) && (2.44)\\ &\subset \operatorname{rge}\big((\partial(\phi+\lambda f))^{-1}\big) \subset \operatorname{int}(\operatorname{dom}\phi). && (2.45)\end{aligned}$$

Proof. In case $y \notin \operatorname{int}(\operatorname{dom}\phi)$ we have $\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f = \emptyset$ and therefore for the first inclusion only vectors y contained in $\operatorname{int}(\operatorname{dom}\phi)$ matter: Fix $y \in \operatorname{int}(\operatorname{dom}\phi)$. By the definition of the left prox it is clear that $\operatorname{rge}(\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f) \subset \operatorname{dom} f \cap \operatorname{dom}\phi$. For $x \in \operatorname{dom} f \cap \operatorname{dom}\phi$, using Lemma 1.2 and the smoothness of the affine function $\phi(y) + \langle \cdot - y, \nabla\phi(y)\rangle$, we observe that

$$\partial(f + D_\phi(\cdot,y))(x) = \partial\Big(f + \frac{1}{\lambda}\phi\Big)(x) - \frac{1}{\lambda}\nabla\phi(y).$$

Therefore, invoking Fermat's rule Lemma 1.3, the first inclusion follows:

$$\begin{aligned} x \in \overleftarrow{\operatorname{prox}}_\lambda^\phi f(y) &\implies 0 \in \partial(\phi + \lambda f)(x) - \nabla\phi(y) \\ &\implies x \in \big(\partial(\phi + \lambda f)^{-1} \circ \nabla\phi\big)(y). \end{aligned}$$

The second inclusion is clear. For the third inclusion note that

$$\operatorname{rge}\big((\partial(\phi + \lambda f))^{-1}\big) = \operatorname{dom} \partial(\phi + \lambda f).$$

Let $x \in \operatorname{dom} \partial(\phi + \lambda f)$. This means in particular $x \in \operatorname{dom} f \cap \operatorname{dom} \phi$ and there exists $v \in \mathbb{R}^m$ such that $v \in \partial(\phi + \lambda f)(x)$. In view of condition (2.43), we invoke Lemma 1.2 and [RW98, Proposition 8.12] to obtain

$$v \in \partial(\phi + \lambda f)(x) \subset \partial\phi(x) + \lambda\partial f(x).$$

This shows that the subset relation is preserved under the dom-operation: $\operatorname{dom} \partial(\phi+\lambda f) \subset \operatorname{dom} \partial\phi \cap \operatorname{dom} \partial f$. In addition, since ϕ is essentially smooth we know from Lemma 1.16 that $\operatorname{dom} \partial\phi = \operatorname{int}(\operatorname{dom} \phi)$. This yields

$$\operatorname{dom} \partial\phi \cap \operatorname{dom} \partial f \subset \operatorname{dom} \partial\phi = \operatorname{int}(\operatorname{dom} \phi)$$

and overall $\operatorname{rge}((\partial(\phi + \lambda f))^{-1}) \subset \operatorname{int}(\operatorname{dom} \phi)$. □

The CQ (2.43) ensures that the sum-rule Lemma 1.2 holds with inclusion: $\partial(\phi + f) \subset \partial\phi + \partial f$. The condition is valid if, for instance, f is smooth, cf. [RW98, Exercise 8.8] or $\operatorname{dom} \phi$ is open or simply $\operatorname{dom} f \subset \operatorname{int}(\operatorname{dom} \phi)$. In the convex setting, the conclusion also follows when $\operatorname{int}(\operatorname{dom} f) \cap \operatorname{int}(\operatorname{dom} \phi) \neq \emptyset$.

As remarked previously, due to asymmetry, we distinguish a left and a right Bregman proximal mapping. We adopt the definition from [BDL18] for the convex setting:

Definition 2.24 (right Bregman–Moreau envelope and proximal mapping)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be extended real-valued (and possibly improper). For some $\lambda > 0$ and $y \in \mathbb{R}^m$ we define the right Bregman–Moreau envelope (in short: right envelope) of f at y as*

$$\overrightarrow{\operatorname{env}}_\lambda^\phi f(y) = \inf_{x\in\mathbb{R}^m} f(x) \dotplus \frac{1}{\lambda} D_\phi(y, x), \tag{2.46}$$

and the associated right Bregman proximal mapping (in short: right prox) of f at y as

$$\overrightarrow{\operatorname{prox}}_\lambda^\phi f(y) = \operatorname*{arg\,min}_{x\in\mathbb{R}^m} f(x) \dotplus \frac{1}{\lambda} D_\phi(y, x). \tag{2.47}$$

Like for the left Bregman proximal mapping, we seek to find a sufficient condition which guarantees nonemptyness and outer semicontinuity Definition 1.11 of the right Bregman proximal mapping. We therefore formulate a relative right prox-boundness condition:

Definition 2.25 (relative right prox-boundedness)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ with $\operatorname{dom} \phi = \mathbb{R}^m$ be Legendre and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be extended real-valued. We say f is right prox-bounded relative to ϕ if there exists $\lambda > 0$ such that $\overrightarrow{\operatorname{env}}_\lambda^\phi f(y) > -\infty$ for some $y \in \mathbb{R}^m$. The supremum of the set of all such λ is the threshold λ_f of the right prox-boundedness, i.e.*

$$\lambda_f = \sup\big\{\lambda > 0 : \exists\, y \in \mathbb{R}^m : \overrightarrow{\operatorname{env}}_\lambda^\phi f(y) > -\infty\big\}.$$

The above condition arises naturally through a transformation of the right proximal mapping into a left proximal mapping via the identity $D_\phi(y,x) = D_{\phi^*}(\nabla\phi(x), \nabla\phi(y))$, for $x, y \in \operatorname{int}(\operatorname{dom}\phi)$, see Proposition 1.20(ii), and applying left prox-boundedness. This substitution is motivated from [Bau+09; BMW11] who studied the nonconvex right Bregman projection, through the left Bregman projection. This leads us to formulate the following lemma:

Lemma 2.26. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be extended real-valued (and possibly improper). Define $g : \mathbb{R}^m \to \overline{\mathbb{R}}$:*

$$g(z) := \begin{cases} f(\nabla\phi^*(z)), & \text{if } z \in \operatorname{int}\operatorname{dom}\phi^* \\ +\infty & \text{otherwise.} \end{cases}$$

Let $y \in \operatorname{int}(\operatorname{dom}\phi)$. Then we have for the right Bregman envelope

$$\overrightarrow{\operatorname{env}}{}^{\phi}_{\lambda} f(y) = \overleftarrow{\operatorname{env}}{}^{\phi^*}_{\lambda} g(\nabla\phi(y)), \tag{2.48}$$

and the associated right Bregman-prox

$$\overrightarrow{\operatorname{prox}}{}^{\phi}_{\lambda} f(y) = \nabla\phi^*\big(\overleftarrow{\operatorname{prox}}{}^{\phi^*}_{\lambda} g(\nabla\phi(y))\big). \tag{2.49}$$

Proof. Let $y \in \operatorname{int}(\operatorname{dom}\phi)$. In view of Lemma 1.20(ii), we may introduce a substitution $z = \nabla\phi(x)$, for $x \in \operatorname{int}(\operatorname{dom}\phi)$ and rewrite:

$$\begin{aligned}
\overrightarrow{\operatorname{env}}{}^{\phi}_{\lambda} f(y) &= \inf_{x\in\mathbb{R}^m} f(x) \dot{+} \frac{1}{\lambda} D_\phi(y,x) \\
&= \inf_{x\in\operatorname{int}(\operatorname{dom}\phi)} f(x) \dot{+} \frac{1}{\lambda} D_\phi(y,x) \\
&= \inf_{x\in\operatorname{int}(\operatorname{dom}\phi)} f(x) \dot{+} \frac{1}{\lambda} D_{\phi^*}(\nabla\phi(x), \nabla\phi(y)) \\
&= \inf_{z\in\operatorname{int}(\operatorname{dom}\phi^*)} f(\nabla\phi^*(z)) \dot{+} \frac{1}{\lambda} D_{\phi^*}(z, \nabla\phi(y)) \\
&= \overleftarrow{\operatorname{env}}{}^{\phi^*}_{\lambda} g(\nabla\phi(y)).
\end{aligned}$$

By the same argument, we also have:

$$\begin{aligned}
& & x &\in \overrightarrow{\operatorname{prox}}{}^{\phi}_{\lambda} f(y) \\
&\iff & x &\in \operatorname*{arg\,min}_{x\in\operatorname{int}(\operatorname{dom}\phi)} f(x) \dot{+} \frac{1}{\lambda} D_{\phi^*}(\nabla\phi(x), \nabla\phi(y)) \\
&\iff & \nabla\phi(x) &\in \operatorname*{arg\,min}_{z\in\operatorname{int}(\operatorname{dom}\phi^*)} f(\nabla\phi^*(z)) \dot{+} \frac{1}{\lambda} D_{\phi^*}(z, \nabla\phi(y)) \\
&\iff & x &\in \nabla\phi^*\big(\overleftarrow{\operatorname{prox}}{}^{\phi^*}_{\lambda} g(\nabla\phi(y))\big).
\end{aligned}$$

□

The above relations reveal that prox-boundedness of g relative to ϕ^* is equivalent to $\overrightarrow{\operatorname{env}}{}^{\phi}_{\lambda} f(y) > -\infty$ for some $y \in \mathbb{R}^m$ and $\lambda > 0$. If g is also lsc, we can use the continuity properties already proved for the left Bregman proximal mapping to deduce the continuity properties of the right prox assuming right prox-boundedness and super-coercivity of ϕ^*, or equivalently, in view of Lemma 1.16(iv), $\operatorname{dom}\phi = \mathbb{R}^m$. However, g is not necessarily lsc unless $\operatorname{dom}\phi^* = \mathbb{R}^m$. Therefore we will assume that, in addition, ϕ is super-coercive which implies that $\operatorname{dom}\phi^* = \mathbb{R}^m$.

Alternatively to right prox-boundedness and super-coercivity of ϕ, we can assume that f is coercive:

Lemma 2.27. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre with $\operatorname{dom}\phi = \mathbb{R}^m$ and let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, lsc and coercive. Then $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ defined by*

$$g(z) := \begin{cases} f(\nabla\phi^*(z)), & \text{if } z \in \operatorname{int}\operatorname{dom}\phi^* \\ +\infty & \text{otherwise,} \end{cases}$$

is proper and lsc and left prox-bounded with threshold $\lambda_f = +\infty$.

Proof. Since $\operatorname{rge}\nabla\phi^* = \mathbb{R}^m$ and f is proper g is proper. Clearly, $f \circ \nabla\phi^*$ is lsc relative to $\operatorname{int}\operatorname{dom}\phi^*$. Now take $\bar{x} \in \operatorname{bdry}\operatorname{dom}\phi^*$ and consider $\operatorname{int}(\operatorname{dom}\phi^*) \ni x^\nu \to \bar{x}$. Then $\|\nabla\phi^*(x^\nu)\| \to \infty$ since ϕ^* is essentially smooth. Therefore $g(x^\nu) = f(\nabla\phi^*(x^\nu)) \to \infty = g(\bar{x})$ since f is coercive. Therefore g is lsc. In addition for any $\lambda > 0$ and some $y \in \mathbb{R}^m$, since $D_\phi(y, x) \geq 0$, the function $x \mapsto f(x) + (1/\lambda)D_\phi(y, x)$ is proper lsc and coercive in x and therefore bounded from below. This means that f is right prox-bounded relative to ϕ with threshold $+\infty$. □

Lemma 2.28 (continuity properties of the right prox and envelope)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre with $\operatorname{dom}\phi = \mathbb{R}^m$ and let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, lsc. In addition, let one of the following two conditions hold true:*

(a) *either f is relatively right prox-bounded with threshold λ_f and ϕ is super-coercive,*

(b) *or f is coercive (and thus relatively right prox-bounded with threshold $\lambda_f = +\infty$).*

Let $\lambda \in (0, \lambda_f)$. Then $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ defined by

$$g(z) := \begin{cases} f(\nabla\phi^*(z)), & \text{if } z \in \operatorname{int}\operatorname{dom}\phi^* \\ +\infty & \text{otherwise,} \end{cases}$$

is proper lsc and left prox-bounded relative to ϕ^* with threshold λ_f and thus $\overrightarrow{\operatorname{prox}}_\lambda^\phi f$ and $\overrightarrow{\operatorname{env}}_\lambda^\phi f$ have the following properties:

(i) *$\overrightarrow{\operatorname{prox}}_\lambda^\phi f(y) \neq \emptyset$ is compact for all $y \in \mathbb{R}^m$ and the envelope $\overrightarrow{\operatorname{env}}_\lambda^\phi f$ is proper.*

(ii) *The envelope $\overrightarrow{\operatorname{env}}_\lambda^\phi f$ is continuous.*

(iii) *For any sequence $y^\nu \to y^*$ and $x^\nu \in \overrightarrow{\operatorname{prox}}_\lambda^\phi f(y^\nu)$ we have $\{x^\nu\}_{\nu\in\mathbb{N}}$ is bounded and all its cluster points x^* lie in $\overrightarrow{\operatorname{prox}}_\lambda^\phi f(y^*)$.*

Proof. Since $\operatorname{rge}\nabla\phi^* = \mathbb{R}^m$ and f is proper g is proper. Due to Lemma 2.26 we have the identities $\overrightarrow{\operatorname{env}}_\lambda^\phi f(y) = \overleftarrow{\operatorname{env}}_\lambda^{\phi^*} g(\nabla\phi(y))$ and $\overrightarrow{\operatorname{prox}}_\lambda^\phi f(y) = \nabla\phi^*(\overleftarrow{\operatorname{prox}}_\lambda^{\phi^*} g(\nabla\phi(y)))$ for any $y \in \mathbb{R}^m$.

First assume condition (a) holds true. Due to the identity $\overrightarrow{\operatorname{env}}_\lambda^\phi f(y) = \overleftarrow{\operatorname{env}}_\lambda^{\phi^*} g(\nabla\phi(y))$ for any $y \in \mathbb{R}^m$, we know that g is left prox-bounded relative to ϕ^* with the same threshold. In addition, $g = f \circ \nabla\phi^*$ is lsc since $\operatorname{dom}\nabla\phi^* = \mathbb{R}^m$ and f is lsc as desired.

Now assume that instead condition (b) holds true. Invoking Lemma 2.27 we know that g is proper, lsc and left prox-bounded with threshold $\lambda_f = +\infty$.

The second result then follows by Lemma 2.22 applied to g, the super-coercivity of ϕ^* and the continuity of both $\nabla\phi$ and $\nabla\phi^*$, cf. Lemma 1.16, as well as the identities form Lemma 2.26. □

Remark 2.29. Note that coercivity of f eventually guarantees that the function $h(x,y) = f(x) + 1/\lambda D_\phi(y,x)$ in the right Bregman proximal mapping is level-bounded in x locally uniformly in y. Thus the continuity properties of the right Bregman proximal mapping under coercivity of f follow alternatively from Theorem 1.13 applied to h.

2.3.2. Relatively proximal subgradients

We generalize the definition of proximal subgradients [RW98, Definition 8.45] to a Bregmanian setting: A classical proximal subgradient is a regular subgradient for which the error term $o(\|x-\bar{x}\|)$ can be specialized to a negative quadratic: $o(\|x-\bar{x}\|) = -(r/2)\|x-\bar{x}\|^2$. Analogously, a relatively proximal subgradient is a regular subgradient where the error term $o(\|x-\bar{x}\|)$ specializes to a Bregman distance $-rD_\phi(x,\bar{x})$.

Definition 2.30 (relatively proximal subgradients and normals)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. A vector $v \in \mathbb{R}^m$ is called a proximal subgradient of a function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ relative to ϕ at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$, a point where $f(\bar{x})$ is finite, if there exist $r > 0$ and $\varepsilon > 0$ such that for all $x \in \mathbb{R}^m$ with $\|x-\bar{x}\| \le \varepsilon$ we have*

$$f(x) \ge f(\bar{x}) + \langle v, x-\bar{x}\rangle - rD_\phi(x,\bar{x}). \tag{2.50}$$

If $f = \iota_C$ specializes to an indicator function of a set C we shall refer to v as a relatively proximal normal to C.

We shall point out the following relation to classical proximal subgradients and normals [RW98, Definition 8.45], i.e. when $\phi = (1/2)\|\cdot\|^2$.

Proposition 2.31. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $\mathcal{C}^2$ on $\operatorname{int}(\operatorname{dom}\phi)$. Let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be finite at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$. Then the following implication holds: If $v \in \mathbb{R}^m$ is a relatively proximal subgradient of f at $\bar{x}$, then v is a proximal subgradient of f at $\bar{x}$. The converse is true if, furthermore, $\nabla^2\phi(x)$ is positive definite for any $x \in \operatorname{int}(\operatorname{dom}\phi)$, i.e. ϕ is very strictly convex.*

Proof. This is a direct consequence of Lemma 1.20(iii). □

As the following example shows, there exist functions that do not have a proximal subgradient everywhere, but do have a relatively proximal subgradient.

Example 2.32. Choose $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = -|x|^{3/2}$. Then f does not have a classical proximal subgradient at $\bar{x} = 0$. However, obviously $\nabla f(\bar{x}) = 0$ is a proximal subgradient of f at 0 relative to $\phi \in \Gamma_0(\mathbb{R})$ with $\phi(x) = |x|^{3/2}$ Legendre.

Proof. Suppose that $\nabla f(\bar{x}) = 0$ is a proximal subgradient of f at $\bar{x} = 0$. This means there is $\varepsilon > 0$ and $\infty > r$ sufficiently large such that for any $x \in \mathbb{R}$ with $|x| < \varepsilon$ the proximal subgradient inequality holds:

$$-|x|^{3/2} \ge -r|x|^2.$$

Equivalently this means

$$|x|^3 \le r^2|x|^4.$$

This implies for any $x \in \mathbb{R}$ with $0 < |x| < \varepsilon$:

$$\frac{1}{r^2} \le |x|,$$

which is a contradiction, since r is finite. □

Lemma 2.33 (globalization of proximal subgradient inequality)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and super-coercive and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc, relatively prox-bounded with threshold λ_f and finite at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$. Let $\bar{v}$ be a relatively proximal subgradient of f at $\bar{x}$. Then, if $r > 0$ is sufficiently large the subgradient inequality* (2.50) *holds globally for all $x \in \mathbb{R}^m$.*

Proof. Since $\bar{v}$ is a relatively proximal subgradient of f at $\bar{x}$ we know that there exists $r' > 0$ and $\varepsilon > 0$ such that for any $r \geq r'$ we have

$$f(x) \geq f(\bar{x}) + \langle \bar{v}, x - \bar{x}\rangle - rD_\phi(x, \bar{x}), \tag{2.51}$$

whenever $\|x - \bar{x}\| < \varepsilon$ and ε is sufficiently small such that $x \in \operatorname{int}(\operatorname{dom}\phi)$. We prove the assertion by showing that the inequality also holds for any $x \in \mathbb{R}^m$ with $\|x-\bar{x}\| \geq \varepsilon$, when r is chosen to be sufficiently large: Let $\lambda \in (0, \lambda_f)$. Since $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ is prox-bounded and proper, lsc and $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$ we know from Lemma 2.22 that $+\infty > \overleftarrow{\operatorname{env}}_\lambda^\phi f(\bar{x}) > -\infty$ since f is prox-bounded and $f(\bar{x})$ is finite. Then we have

$$f(x) \geq \overleftarrow{\operatorname{env}}_\lambda^\phi f(\bar{x}) - \frac{1}{\lambda} D_\phi(x, \bar{x}), \tag{2.52}$$

for all $x \in \mathbb{R}^m$. Combining (2.51) and (2.52) shows that (2.50) holds with constant $r \geq \max\{r', 1/\lambda\}$, when

$$\overleftarrow{\operatorname{env}}_\lambda^\phi f(\bar{x}) - \frac{1}{\lambda} D_\phi(x, \bar{x}) \geq f(\bar{x}) + \langle \bar{v}, x - \bar{x}\rangle - rD_\phi(x, \bar{x})$$

is satisfied, which is implied (using the Cauchy–Schwarz inequality) by

$$\frac{f(\bar{x}) - \overleftarrow{\operatorname{env}}_\lambda^\phi f(\bar{x})}{\|x - \bar{x}\|} + \|\bar{v}\| \leq \left(r - \frac{1}{\lambda}\right) \frac{D_\phi(x, \bar{x})}{\|x - \bar{x}\|}. \tag{2.53}$$

Using super-coercivity of ϕ, the inequality happens to be true for $r \geq \max\{r', 1/\lambda\}$ and all x with $\|x-\bar{x}\| \geq \mu$ for some $\mu > \varepsilon$. It remains to verify (2.53) for $\mu > \|x-\bar{x}\| \geq \varepsilon > 0$ for some r. However, since for such x, using strict convexity of ϕ, obviously, $D_\phi(x, \bar{x})$ is bounded away from 0, we can find some r sufficiently large such that (2.53) also holds for x with $\mu > \|x - \bar{x}\| \geq \varepsilon$. □

The following lemma shows that analogous to the classical prox, the left Bregman prox and envelope of a tilted function $f - \langle \cdot, v\rangle$ can be written as the Bregman prox and envelope of f at a transformed point, respectively.

Lemma 2.34 (effects of tilt transformation)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc. Let $y \in \operatorname{int}(\operatorname{dom}\phi)$ and $v \in \mathbb{R}^m$. Denote by $z := \nabla\phi^*(\nabla\phi(y) + \lambda v)$ and by $f_0 := f - \langle \cdot, v\rangle$. Then we have the following identities for the prox*

$$\overleftarrow{\operatorname{prox}}_\lambda^\phi f_0(y) = \overleftarrow{\operatorname{prox}}_\lambda^\phi f(z),$$

and the envelope function

$$\overleftarrow{\operatorname{env}}_\lambda^\phi f_0(y) = \overleftarrow{\operatorname{env}}_\lambda^\phi f(z) + \frac{1}{\lambda} D_\phi(z, y) - \langle v, z\rangle.$$

Proof. For $z = \nabla\phi^*(\nabla\phi(y) + \lambda v)$ the identities follow from the following calculation:

$$\begin{aligned} D_\phi(x,z) &= \phi(x) - \phi(z) - \langle \nabla\phi(y) + \lambda v, x - z\rangle \\ &= -\lambda\langle v, x\rangle + D_\phi(x,y) + \phi(y) - \langle \nabla\phi(y), y\rangle + \langle \nabla\phi(y) + \lambda v, z\rangle - \phi(z). \end{aligned}$$

Scaling the equality with $1/\lambda$ and adding $f(x)$ and reordering yields

$$\begin{aligned} f_0(x) + \frac{1}{\lambda}D_\phi(x,y) &= f(x) + \frac{1}{\lambda}\big(D_\phi(x,z) - \phi(y) \\ &\qquad + \langle \nabla\phi(y), y\rangle - \langle \nabla\phi(y) + \lambda v, z\rangle + \phi(z)\big) \\ &= f(x) + \frac{1}{\lambda}D_\phi(x,z) + \frac{1}{\lambda}D_\phi(z,y) - \langle v, z\rangle. \qquad \square \end{aligned}$$

Based on the globalized subgradient inequality from Lemma 2.33 we shall characterize relatively proximal subgradients via the Bregman proximal map. This property is used frequently in the course of this section to assert the single-valuedness of the Bregman proximal mapping.

Proposition 2.35. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and finite at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$. Then the following conditions are equivalent for some $\lambda > 0$:*

(i) *The following inclusion holds:*

$$\bar{x} \in \overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(\nabla\phi^*(\nabla\phi(\bar{x}) + \lambda v)). \tag{2.54}$$

(ii) *The subgradient inequality* (2.50) *holds globally for all $x \in \mathbb{R}^m$ and $r := 1/\lambda$,*

where strict inequality holds for all $x \neq \bar{x}$ by decreasing λ. Equivalently this means the inclusion (2.54) *holds with equality. If, furthermore, ϕ is super-coercive and f relatively prox-bounded with threshold λ_f then the above conditions hold for some $\lambda < \lambda_f$ sufficiently small, if and only if $v \in \mathbb{R}^m$ is a relatively proximal subgradient of f at $\bar{x}$.*

Proof. Let $\lambda > 0$ and let the subgradient inequality (2.50) hold globally for all $x \in \mathbb{R}^m$ and $r := 1/\lambda$. This means:

$$f(x) - \langle v, x\rangle \geq f(\bar{x}) - \langle v, \bar{x}\rangle - \frac{1}{\lambda}D_\phi(x,\bar{x}). \tag{2.55}$$

Define $f_0 := f - \langle v, \cdot\rangle$. Then, by reordering the terms the above is equivalent to:

$$f_0(x) + \frac{1}{\lambda}D_\phi(x,\bar{x}) \geq f_0(\bar{x}) + \frac{1}{\lambda}D_\phi(\bar{x},\bar{x}), \tag{2.56}$$

which holds if and only if $\bar{x} \in \overleftarrow{\operatorname{prox}}{}_\lambda^\phi f_0(\bar{x})$, which, in view of Lemma 2.34, is equivalent to $\bar{x} \in \overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(\nabla\phi^*(\nabla\phi(\bar{x}) + \lambda v))$. Then, clearly, strict inequality holds for all $x \neq \bar{x}$ by decreasing λ, which equivalently means the inclusion (2.54) holds with equality.

Let v be a relatively proximal subgradient of f at $\bar{x}$ with constants $\varepsilon > 0$ and $r > 0$. Then we may invoke Lemma 2.33 to make the subgradient inequality in (2.50) hold globally, for all $x \in \mathbb{R}^m$, $r := 1/\lambda$ and $\lambda > 0$ sufficiently small. Conversely, when the subgradient inequality (2.50) holds globally for $r := 1/\lambda$, this means in particular that v is a relatively proximal subgradient. □

Invoking the above globalization lemma we shall highlight a close relation between relatively proximal subgradients and a *smooth variational description of regular subgradients* due to Mordukhovich [Mor18, Theorem 1.27]: The following proposition shows, that any regular subgradient of a relatively prox-bounded function can be expressed in terms of the Bregman proximal mapping in the same way this is true for proximal subgradients and the classical proximal mapping:

Proposition 2.36 (regular subgradients are relatively proximal subgradients)**.** *Let the function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded relative to $\chi \in \Gamma_0(\mathbb{R}^m)$ Legendre and super-coercive. Let $v \in \widehat{\partial} f(\bar x)$ be a regular subgradient of f at $\bar x \in \operatorname{int}(\operatorname{dom}\chi)$, a point where $f(\bar x)$ is finite. Then there exists $\phi \in \Gamma_0(\mathbb{R}^m)$ Legendre and super-coercive with $\operatorname{dom}\phi = \operatorname{dom}\chi$ such that v is also a proximal subgradient relative to ϕ of f at $\bar x$ and in particular it holds for any $\lambda \leq 1$:*

$$\bar x = \overleftarrow{\operatorname{prox}}{}^{\phi}_{\lambda} f(\nabla\phi^*(\nabla\phi(\bar x) + \lambda v)).$$

Proof. Let $v \in \mathbb{R}^m$ be a regular subgradient of f at $\bar x$. Since f is prox-bounded relative to $\chi \in \Gamma_0(\mathbb{R}^m)$ Legendre and super-coercive, in view of Proposition 2.21(ii), there is $r > 0$ so that $f + r\chi$ is uniformly bounded from below. In view of [Mor18, Theorem 1.27] there exists a convex and smooth function $\psi : \mathbb{R}^m \to \mathbb{R}$ with $-\nabla\psi(\bar x) = v + r\nabla\chi(\bar x)$ so that $f + r\chi + \psi$ attains its global minimum at $\bar x$. This means that for all $x \neq \bar x$ we have

$$(f + r\chi + \psi)(x) > (f + r\chi + \psi)(\bar x) + \langle v, x - \bar x\rangle + \langle \nabla\psi(\bar x) + r\nabla\chi(\bar x), x - \bar x\rangle.$$

Let $\phi := r\chi + \psi \in \Gamma_0(\mathbb{R}^m)$. Then we have $\operatorname{dom}\phi = \operatorname{dom}\chi$. Now consider a sequence $\operatorname{int}(\operatorname{dom}\phi) \ni x^\nu \to x \in \operatorname{bdry}\operatorname{dom}\phi$. We have $\|\nabla\phi(x^\nu)\| = \|r\nabla\chi(x^\nu) + \nabla\psi(x^\nu)\| \to \infty$, since $\psi \in \mathcal{C}^1$ and therefore $\nabla\psi(x^\nu) \to \nabla\psi(x) \in \mathbb{R}^m$. Therefore ϕ is essentially smooth as well. Since ϕ is also essentially strictly convex we have that ϕ is Legendre. In addition ϕ is super-coercive. Rewriting the inequality in terms of the Bregman distance yields for any $0 < \lambda \leq 1$:

$$f(x) > f(\bar x) + \langle v, x - \bar x\rangle - D_\phi(x, \bar x) \geq f(\bar x) + \langle v, x - \bar x\rangle - \frac{1}{\lambda} D_\phi(x, \bar x),$$

which is equivalent to

$$\bar x = \overleftarrow{\operatorname{prox}}{}^{\phi}_{\lambda} f(\nabla\phi^*(\nabla\phi(\bar x) + \lambda v)). \qquad \square$$

An important class of prox-bounded functions f are indicator functions $f = \iota_C$ of a possibly nonconvex set C. Indeed, the threshold of prox-boundedness for such f is $\lambda_f = \infty$. Invoking the above lemma we obtain that v is a relatively proximal normal to C at $\bar x$ if and only if we can perturb $\bar x$ along v in the Bregmanian sense as $y := \nabla\phi^*(\nabla\phi(\bar x) + \lambda v)$ so that by Bregman-projecting the perturbed point y back on C (i.e. computing the left prox of f at y), we recover $\bar x$. Indeed, for $\phi = \|\cdot\|^2$ we obtain the classical definition of proximal normals:

$$N_C^P(\bar x) := \{ r(y - \bar x) : \bar x \in \operatorname{proj}_C(y), r \geq 0, y \in \mathbb{R}^m \},$$

where proj_C denotes the classical Euclidean projection onto the set C. More generally, invoking the above proposition we obtain the following alternative expression for the regular normal cone $\widehat{N}_C(\bar x)$ of the set C:

$$\widehat{N}_C(\bar x) = \left\{ r(\nabla\phi(y) - \nabla\phi(\bar x)) : \bar x \in \overleftarrow{\operatorname{proj}}{}^{\phi}_{C}(y), r \geq 0, y \in \mathbb{R}^m \right\},$$

for some appropriate choice $\phi \in \Gamma(\mathbb{R}^m)$ Legendre. This is illustrated in Figure 2.2.

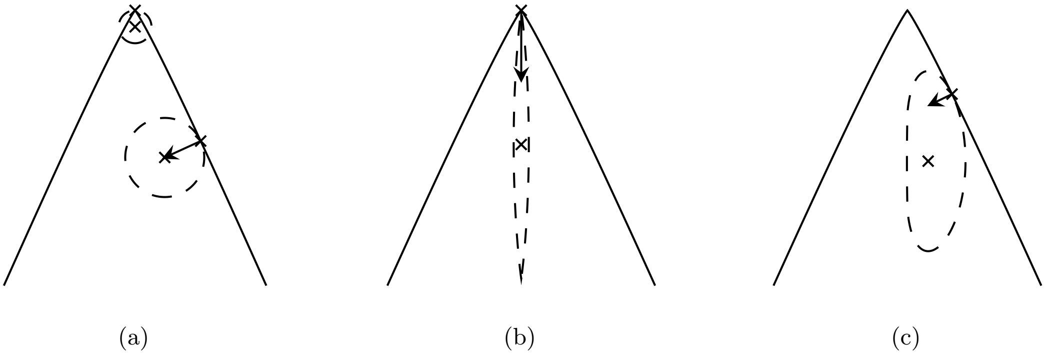


Figure 2.2.: Illustration of Bregman proximal normals and Bregman projections by means of Example 2.56. The set $C = \operatorname{epi} h$, given as the epigraph of $h(x) = 2x^2 - 3|x|^{1.1}$, is indicated as the areas above the solid lines, which correspond to the graph of h around 0. The arrows indicate the (relatively) proximal normals v of C at $\bar{x}$. The dashed lines correspond to the Euclidean (a) resp. Bregman (b),(c) distance balls around the points $\nabla\phi^*(\nabla\phi(\bar{x}) + \lambda v)$ generated by $\phi(x) = \|x\|^2$, see (a) resp. $\phi(x_1, x_2) = x_1^2 + |x_1|^{1.1} + x_2^2$, see (b),(c). Their "radii" are chosen such that their upper surfaces touch the epigraph of h only at $\bar{x}$, which is possible everywhere in the Bregman case (b),(c) if $\lambda > 0$ is sufficiently small. While v at the point $x \neq 0$ is both, a relatively proximal normal of C at $\bar{x}$, see (c) and a classical proximal normal of C at $\bar{x}$, see (a), the situation is different at the point $\bar{x} = 0$: While v is a relatively proximal normal of C at $\bar{x}$, see (b), it is not a classical proximal normal, see (a), since for any $\lambda > 0$ there is no Euclidean ball around $\bar{x} + \lambda v$, below C that touches C only at 0. Also compare to Example 2.32. However, in view of the subgradient inequality (2.50), $\tilde{v} := v + (1/\lambda)\nabla\phi(\bar{x})$ is a classical proximal subgradient of $\tilde{f} := (1/\lambda)\phi + \delta_C$ at $\bar{x}$, cf. also Remark 2.42.

2.3.3. Single-valuedness of the Bregman proximal mapping under relative prox-regularity

We now define relative prox-regularity, generalizing [RW98, Definition 13.27] to a Bregmanian setting: We fix a reference point $(\bar{x}, \bar{v})$, where $f(\bar{x})$ is finite and $\bar{v} \in \partial f(\bar{x})$ is a relatively proximal subgradient, and require the subgradient inequality (2.50) to hold uniformly on an f-attentive neighborhood of $(\bar{x}, \bar{v})$:

Definition 2.37 (relative prox-regularity)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. We say a function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ is relatively prox-regular at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$ for $\bar{v} \in \mathbb{R}^m$ if f is finite and locally lsc at $\bar{x}$ with $\bar{v} \in \partial f(\bar{x})$ and there exist $\varepsilon > 0$ and $r \geq 0$ such that for all $\|x' - \bar{x}\| < \varepsilon$, $\|x - \bar{x}\| < \varepsilon$, with ε sufficiently small such that $x, x' \in \operatorname{int}(\operatorname{dom}\phi)$, it holds that:*

$$f(x') \geq f(x) + \langle v, x' - x\rangle - rD_\phi(x', x), \tag{2.57}$$

whenever $f(x) - f(\bar{x}) < \varepsilon$, $v \in \partial f(x)$, $\|v - \bar{v}\| < \varepsilon$. When this property holds for all $\bar{v} \in \partial f(\bar{x})$, f is said to be relatively prox-regular at $\bar{x}$.

For examples of relatively prox-regular functions we refer to Section 2.3.5 below. We first clarify a relation between the new relative prox-regularity property and classical prox-regularity.

Proposition 2.38. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $\mathcal{C}^2$ on $\operatorname{int}(\operatorname{dom}\phi)$. Let $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be extended real-valued and $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$. Then the following implication holds: If f is relatively prox-regular at $\bar{x}$ for $\bar{v}$, then f is also prox-regular at $\bar{x}$ for $\bar{v}$. The converse is true if, furthermore, $\nabla^2\phi(x)$ is positive definite for any $x \in \operatorname{int}(\operatorname{dom}\phi)$, i.e. ϕ is very strictly convex.*

Proof. This is a direct consequence of Lemma 1.20(iii). □

The relative prox-regularity property of a tilted function is preserved as the following lemma shows:

Lemma 2.39 (invariance under tilt transformation)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre, $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be extended real-valued and $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$. Then f is relatively prox-regular at $\bar{x}$ for $\bar{v} \in \partial f(\bar{x})$ if and only if $f_0 := f - \langle v, \cdot\rangle$ is relatively prox-regular at $\bar{x}$ for $0 \in \partial f_0(\bar{x})$.*

Proof. This is clear from the definition of relative prox-regularity. □

The following theorem is analogous to [PR96, Theorem 3.2] or [RW98, Theorem 13.36] for classical prox-regularity. More precisely, for a function f and a reference point $(\bar{x}, \bar{v}) \in \operatorname{gph}\partial f$ we provide an equivalent characterization of relative prox-regularity in terms of the relative hypomonotonicity of an f-attentive graphical localization T of the subdifferential ∂f at $(\bar{x}, \bar{v})$, cf. Definition 1.6. Under a constraint qualification for the sum-rule for the subdifferential of $f + r\nabla\phi$, such a statement can be seen as a localized analogue to the equivalence between the relative hypoconvexity of f, i.e. $f + r\phi$ is proper lsc and convex on $\operatorname{int}(\operatorname{dom}\phi)$ for some $r \geq 0$, and the relative hypomonotonicity of ∂f. An important difference to note is that, in the following statement, we also require $\bar{v}$ to be a relatively proximal subgradient of f at $\bar{x}$.

Theorem 2.40. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and super-coercive and the function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc, prox-bounded with threshold λ_f and finite at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$. Then the following conditions are equivalent:*

(i) *f is relatively prox-regular at $\bar{x}$ for $\bar{v}$.*

(ii) *$\bar{v} \in \partial f(\bar{x})$ is a relatively proximal subgradient and ∂f has an f-attentive ε-localization $T : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ around $(\bar{x}, \bar{v})$ such that $T + r\nabla\phi$ is monotone for some $r > 0$.*

(iii) *For $\bar{v} \in \partial f(\bar{x})$ and $\lambda < \lambda_f$ sufficiently small it holds that $\overleftarrow{\operatorname{prox}}{}_\lambda^\phi f$ is a singled-valued map near the point $\bar{y} := \nabla\phi^*(\nabla\phi(\bar{x}) + \lambda\bar{v})$ such that $\{\bar{x}\} = \overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(\bar{y})$ and*

$$\overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(y) = \big((\nabla\phi + \lambda T)^{-1} \circ \nabla\phi\big)(y), \tag{2.58}$$

for some f-attentive ε-localization $T : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ of ∂f around $(\bar{x}, \bar{v})$ and y near $\bar{y}$.

Proof. (i) $\implies$ (ii): Let f be relatively prox-regular at $\bar{x}$ for $\bar{v} \in \partial f(\bar{x})$. This means that there exist constants $\varepsilon > 0$ and $r > 0$ such that the subgradient inequality (2.57) holds for $x', x \in \mathbb{R}^m$ with $\|x' - \bar{x}\| < \varepsilon$, $\|x - \bar{x}\| < \varepsilon$ and $v \in \partial f(x)$, $v' \in \partial f(x')$, $\|v' - \bar{v}\| < \varepsilon$, $\|v - \bar{v}\| < \varepsilon$. In particular this implies that $\bar{v}$ is a relatively proximal subgradient at $\bar{x}$ and we have:

$$f(x') \geq f(x) + \langle v, x' - x\rangle - rD_\phi(x', x), \quad f(x) \geq f(x') + \langle v', x - x'\rangle - rD_\phi(x, x').$$

Adding these inequalities yields:

$$\begin{aligned}0 &\geq \langle v, x' - x\rangle + \langle v', x - x'\rangle - r\big(D_\phi(x', x) + D_\phi(x, x')\big)\\ &= -\langle v - v', x - x'\rangle - r\,\langle \nabla\phi(x) - \nabla\phi(x'), x - x'\rangle\,.\end{aligned}$$

This shows that the corresponding map $T + r\nabla\phi$ is monotone, where T is the f-attentive ε-localization of ∂f at $(\bar{x}, \bar{v})$.

(ii) $\implies$ (iii): By assumption, $\bar{v}$ is a relatively proximal subgradient. Then we may invoke Proposition 2.35 to obtain that $\{\bar{x}\} = \overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(\bar{y})$ is a singleton for $\lambda < \min\{\lambda_f, 1/r\}$ being sufficiently small. Due to the prox-boundedness, we can invoke Lemma 2.22 to assert that $\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y) \neq \emptyset$ for any $y \in \operatorname{int}(\operatorname{dom}\phi)$. Furthermore, for any sequence $x^\nu \in \overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y^\nu)$, $y^\nu \to \bar{y}$ we have $\{x^\nu\}_{\nu\in\mathbb{N}}$ is bounded and all its cluster points lie in $\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(\bar{y}) = \{\bar{x}\}$, meaning $x^\nu \to \bar{x}$ and $\overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f(y^\nu) \to \overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f(\bar{y})$. In addition we have $f(x^\nu) \to f(\bar{x})$ as

$$\overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f(y^\nu) = f(x^\nu) + \frac{1}{\lambda} D_\phi(x^\nu, y^\nu) \to \overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f(\bar{y}) = f(\bar{x}) + \frac{1}{\lambda} D_\phi(\bar{x}, \bar{y}).$$

Overall this shows that for any y, which is sufficiently near to $\bar{y}$, we have $x \in \overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y)$, $\|x - \bar{x}\| < \varepsilon$, $|f(x) - f(\bar{x})| < \varepsilon$ and $\|v - \bar{v}\| < \varepsilon$, for $v := (1/\lambda)\nabla\phi(y) - (1/\lambda)\nabla\phi(x)$ due to the continuity of $\nabla\phi$ (cf. Lemma 1.16). By applying Fermat's rule Lemma 1.3 to $\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y)$ we obtain

$$0 \in \partial f(x) + \frac{1}{\lambda}(\nabla\phi(x) - \nabla\phi(y)),$$

or equivalently

$$0 \in T(x) + r(\nabla\phi(x) - \nabla\phi(y)),$$

where $\partial f(x)$ is replaced by $T(x)$ due to the arguments above. This means

$$\begin{aligned}\emptyset \neq \overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y) &\subset \left\{x \in \mathbb{R}^m : 0 \in T(x) + \frac{1}{\lambda}(\nabla\phi(x) - \nabla\phi(y))\right\}\\ &= \big((\nabla\phi + \lambda T)^{-1} \circ \nabla\phi\big)(y),\end{aligned}$$

is at most a singleton due to the strict monotonicity of $T + (1/\lambda)\nabla\phi$ for $\lambda < 1/r$, cf. Lemma 1.10. This implies $\{x\} = \overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y)$ is a singleton for y near $\bar{y}$.

(iii) $\implies$ (i): Let T be some f-attentive ε-localization of ∂f at $\bar{x}$ for $\bar{v}$, which has the properties in (iii). Let $x \in \mathbb{R}^m$ with $\|x - \bar{x}\| < \varepsilon$, $f(x) < f(\bar{x}) + \varepsilon$ and $v \in \partial f(x)$, $\|v - \bar{v}\| < \varepsilon$. We have $v \in T(x)$ and for $\varepsilon > 0$ sufficiently small $x \in \operatorname{int}(\operatorname{dom}\phi)$ and $y := \nabla\phi^*(\nabla\phi(x) + \lambda v)$ near $\nabla\phi^*(\nabla\phi(\bar{x}) + \lambda\bar{v}))$, due to the continuity of $\nabla\phi^*$ guaranteed by Lemma 1.16. Then for such y we have that $x \in ((\nabla\phi + \lambda T)^{-1} \circ \nabla\phi)(y)$ and by assumption

$$\overleftarrow{\mathrm{prox}}^{\phi}_{\lambda} f(y) = \big((\nabla\phi + \lambda T)^{-1} \circ \nabla\phi\big)(y) \ni x.$$

Invoking Proposition 2.35 we obtain the subgradient inequality (2.57) for $r := 1/\lambda$, which holds even globally, cf. Lemma 2.33. We may conclude f is relatively prox-regular at $\bar{x}$ for $\bar{v}$. □

Remark 2.41. We would highlight that items (i) and (ii) in the above theorem only depend on the local structure of the epigraph of f near $(\bar{x}, f(\bar{x}))$, while in contrast, (iii) depends on its global structure. This means that (i) resp. (ii) hold for f if and only if they hold for $\tilde{f} := f + \iota_C$ for $C := \{x \in \mathbb{R}^m : \|x - \bar{x}\| \leq \varepsilon, f(x) \leq f(\bar{x}) + \varepsilon\}$, where $\tilde{f}$ is always proper lsc and prox-bounded whenever f is locally lsc at $\bar{x}$, a point where f is finite. This shows that the equivalence between (i) and (ii) holds even if we relax the globally lsc assumption towards locally lsc at $\bar{x}$ and entirely drop the prox-boundedness assumption. In that sense (iii) can be seen as an auxiliary statement applied to $\tilde{f}$ to show the direction (ii) implies (i), which is also used as a strategy in the proof of [RW98, Theorem 13.36].

Remark 2.42. When ϕ is strongly convex on compact convex subsets $K \subset \operatorname{int}(\operatorname{dom}\phi)$ (which is implied by very strict convexity, see Lemma 1.20(iii), but holds more generally, for, e.g. $\phi(x) = (1/p)|x|^p$, $p \in (1,2)$), the direction (ii) implies (i) follows alternatively from [PR96, Theorem 3.2] or [RW98, Theorem 13.36]. To this end, let $\bar{v}$ be a relatively proximal subgradient of f at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$ and let T be a relatively hypomonotone, f-attentive ε-localization of ∂f at $\bar{x}$ for $\bar{v}$. This means that there is $r > 0$ such that $\tilde{v} := \bar{v} + r\nabla\phi(\bar{x})$ is a classical proximal subgradient of $\tilde{f} := f + r\phi$ at $\bar{x}$ and $\widetilde{T} := T + r\nabla\phi$ is monotone. Furthermore $\widetilde{T}$ is an $\tilde{f}$-attentive graphical localization of $\partial\tilde{f}$ at $(\bar{x}, \tilde{v})$. Invoking [RW98, Theorem 13.36] this means that $\tilde{f}$ is classically prox-regular at $\bar{x}$ for $\tilde{v}$. Due to the strong convexity of ϕ on compact convex subsets $K \subset \operatorname{int}(\operatorname{dom}\phi)$ we can bound the negative quadratic term $-(1/2)\|x' - x\|^2$ in the classical subgradient inequality (locally) by a Bregman distance $-\theta D_\phi(x', x)$. Rewriting the estimate gives us the result. In the general case we provide a generalization by means of the above theorem.

Corollary 2.43. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and super-coercive. Let the function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc, prox-bounded with threshold λ_f, and relatively prox-regular at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$ for $\bar{v}$. Assume that ϕ is very strictly convex. Then for $\lambda < \lambda_f$ sufficiently small, $\overleftarrow{\operatorname{prox}}{}^\phi_\lambda f$ is a Lipschitz map on a neighborhood of $\bar{y} := \nabla\phi^*(\nabla\phi(\bar{x}) + \lambda\bar{v})$.*

Proof. Since f is relatively prox-regular at $\bar{x}$ for $\bar{v} \in \partial f(\bar{x})$ due to Theorem 2.40 there exists $r > 0$ such that $T + r\nabla\phi$ is monotone. This means for $(x', v'), (x, v) \in \operatorname{gph} T$ we have:

$$\left\langle v' - v, x' - x \right\rangle + r\left\langle \nabla\phi(x') - \nabla\phi(x), x' - x \right\rangle \geq 0.$$

Let $x \in \overleftarrow{\operatorname{prox}}{}^\phi_\lambda f(y)$ and $x' \in \overleftarrow{\operatorname{prox}}{}^\phi_\lambda f(y')$. Due to Theorem 2.40 we know that $(1/\lambda)(\nabla\phi(y) - \nabla\phi(x)) \in T(x)$ and $(1/\lambda)(\nabla\phi(y') - \nabla\phi(x')) \in T(x')$. This means we have

$$\frac{1}{\lambda}\left\langle \nabla\phi(y') - \nabla\phi(y), x' - x \right\rangle \geq \left(\frac{1}{\lambda} - r\right)\left\langle \nabla\phi(x') - \nabla\phi(x), x' - x \right\rangle.$$

Since ϕ is very strictly convex we may invoke Lemma 1.20(iii) to assert that there are constants Θ and θ such that for any $x, x' \in \operatorname{int}(\operatorname{dom}\phi)$ near $\bar{x}$:

$$\begin{aligned} \|\nabla\phi(x') - \nabla\phi(x)\| &\leq \Theta\|x - x'\|, \\ \langle \nabla\phi(x') - \nabla\phi(x), x' - x \rangle &\geq \theta\|x - x'\|^2. \end{aligned}$$

This yields

$$\left\langle \nabla\phi(y') - \nabla\phi(y), x' - x \right\rangle \geq (1 - \lambda r)\,\theta\|x - x'\|^2,$$

and via Cauchy–Schwarz

$$\begin{aligned} \left\langle \nabla\phi(y') - \nabla\phi(y), x' - x \right\rangle &\leq \|\nabla\phi(y') - \nabla\phi(y)\| \cdot \|x' - x\| \\ &\leq \Theta\|y' - y\| \cdot \|x' - x\|, \end{aligned}$$

and overall

$$\|x - x'\| \leq \frac{\Theta}{\theta(1 - \lambda r)} \|y - y'\|. \qquad \square$$

2.3.4. From local to global: Relative hypoconvexity of functions

Next we define relative hypoconvexity of functions. This setting was studied extensively in [KS12] (see [CKS12] for infinite dimensions):

Definition 2.44 (relatively hypoconvex functions)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. Then we say a function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is relatively hypoconvex if there is some $r \geq 0$ such that $f + r\phi$ is convex.*

The following proposition is an extension of [KS12, Theorem 4.2] and [KS12, Theorem 4.3] and provides equivalent characterizations of relative hypoconvexity. In particular, if a constraint qualification holds, relative hypoconvexity can be regarded as uniform relative prox-regularity, i.e., $\varepsilon = \infty$ and the constant r can be chosen to be uniform in the definition of relative prox-regularity:

Proposition 2.45 (characterizations of relative hypoconvexity)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and super-coercive. Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded relative to ϕ with threshold λ_f such that* $\operatorname{dom} f \cap \operatorname{dom} \phi \neq \emptyset$*. Let $\lambda \in (0, \lambda_f)$. Then the following conditions are equivalent:*

(i) *f is relatively hypoconvex with constant $r = 1/\lambda$, i.e., $f + r\phi$ is convex and therefore $f + r\phi \in \Gamma_0(\mathbb{R}^m)$.*

(ii) *$\overleftarrow{\mathrm{prox}}_\lambda^\phi f \circ \nabla \phi^*$ is maximal monotone,*

(iii) *$\overleftarrow{\mathrm{prox}}_\lambda^\phi f \circ \nabla \phi^* = \partial(\phi + \lambda f)^*$,*

(iv) *$\partial(\lambda f + \phi)$ is monotone,*

while $\lambda f + \phi$ is essentially strictly convex if and only if $\partial(\lambda f + \phi)$ is strictly monotone if and only if $\overleftarrow{\mathrm{prox}}_\lambda^\phi f$ is single-valued on $\operatorname{int}(\operatorname{dom} \phi)$*. If, in addition, the following constraint qualification holds:*

$$\partial^\infty f(x) \cap -N_{\operatorname{dom} \phi}(x) = \{0\}, \tag{2.59}$$

for any $x \in \operatorname{dom} f \cap \operatorname{dom} \phi$, conditions (i)*,* (ii)*,* (iii) *and* (iv) *are equivalent to the following statement: The globalized subgradient inequality holds true at any $x \in \operatorname{int}(\operatorname{dom} \phi) \cap \operatorname{dom} f$:*

$$f(x') \geq f(x) + \langle v, x' - x \rangle - \frac{1}{\lambda} D_\phi(x', x),$$

for any $v \in \partial f(x)$ and any $x' \in \mathbb{R}^m$. Furthermore we have $\overleftarrow{\mathrm{prox}}_\lambda^\phi f = (\nabla \phi + \lambda \partial f)^{-1} \circ \nabla \phi$.

Proof. The equivalence between (i), (ii) and (iii) holds due to [KS12, Theorem 4.2]. The equivalence between (i) and (iv) follows due to [RW98, Theorem 12.17] and since $\operatorname{dom} f \cap \operatorname{dom} \phi \neq \emptyset$ we have $f + r\phi \in \Gamma_0(\mathbb{R}^m)$. Furthermore we have $\lambda f + \phi$ is essentially strictly convex if and only if $\partial(\lambda f + \phi)$ is strictly monotone by [RW98, Theorem 12.17]. We have $\overleftarrow{\mathrm{prox}}_\lambda^\phi f$ is single-valued on $\operatorname{int}(\operatorname{dom} \phi)$ if and only if $\lambda f + \phi$ is essentially strictly convex by [KS12, Theorem 4.3].

Let the constraint qualification above hold and assume (i) holds. Since f is proper lsc and $\phi \in \Gamma_0(\mathbb{R}^m)$ and $\operatorname{dom} f \cap \operatorname{dom} \phi \neq \emptyset$, we have $f + r\phi$ is proper lsc and convex. Let $x \in \operatorname{int}(\operatorname{dom} \phi) \cap \operatorname{dom} f$ and $v \in \partial f(x)$. Due to [RW98, Exercise 8.8(c)] we have $v + r\nabla\phi(x) \in \partial(f + r\phi)(x)$ and since $f + r\phi$ is proper lsc and convex this means that for all $x' \in \mathbb{R}^m$:

$$f(x') + r\phi(x') \geq f(x) + r\phi(x) + \langle v + r\nabla\phi(x), x' - x\rangle.$$

Reordering yields the subgradient inequality as claimed:

$$f(x') \geq f(x) + \langle v, x' - x\rangle - rD_\phi(x', x), \quad \forall x' \in \mathbb{R}^m.$$

Now let $x, x' \in \operatorname{dom} \partial(f + r\nabla\phi) \subset \operatorname{dom} \partial f + \operatorname{dom} \partial\phi \subset \operatorname{int} \operatorname{dom} \phi \cap \operatorname{dom} f$, where the inclusions hold due to the constraint qualification and $\operatorname{dom} \partial\phi = \operatorname{int}(\operatorname{dom} \phi)$. Then the subgradient inequality above yields for any $v \in \partial f(x)$ and $v' \in \partial f(x')$:

$$f(x') \geq f(x) + \langle v, x' - x\rangle - rD_\phi(x', x), \qquad f(x) \geq f(x') + \langle v', x - x'\rangle - rD_\phi(x, x').$$

Summing the two estimates yields

$$\langle v' + r\nabla\phi(x') - v - r\nabla\phi(x), x' - x\rangle \geq 0,$$

Hence, $\partial f + r\nabla\phi$ is monotone. This means we have (iv).

Via the observation $\partial(\lambda f + \phi)^* = (\partial(\lambda f + \phi))^{-1} = (\lambda\partial f + \nabla\phi)^{-1}$ the last statement follows. □

Specialized to a quadratic setting $\phi = (1/2)\|\cdot\|^2$, if f is r-hypoconvex, $P_\lambda f$ and $\nabla e_\lambda f$ are in addition Lipschitz continuous with uniform constants, if $\lambda < 1/r$:

Proposition 2.46. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and hypoconvex with constant $r > 0$. Then for any $\lambda \in (0, 1/r)$ we have for the proximal mapping and the associated Moreau envelope of f:*

(i) *The proximal mapping $P_\lambda f$ is Lipschitz continuous with constant $1/(1 - r\lambda)$ and in particular single-valued.*

(ii) *The Moreau envelope $e_\lambda f$ of f is differentiable with Lipschitz continuous gradient and the following gradient formula holds for any $y \in \mathbb{R}^m$:*

$$\nabla e_\lambda f(y) = \frac{1}{\lambda}\big(y - P_\lambda f(y)\big).$$

Proof. (i) First note that f is prox-bounded with some threshold $\lambda_f > 1/r$. In view of [CWP20, Lemma 3.1(a)] we have $\operatorname{con} P_\lambda f = \partial(\lambda f + (1/2)\|\cdot\|^2)^*$. By assumption $\lambda f + (1/2)\|\cdot\|^2$ is $(1 - r\lambda)$-strongly convex, proper lsc. In view of [RW98, Proposition 12.60] $(\lambda f + (1/2)\|\cdot\|^2)^*$ is differentiable with $(1/(1 - r\lambda))$-Lipschitz continuous gradient and thus

$$P_\lambda f = \nabla\big(\lambda f + (1/2)\|\cdot\|^2\big)^*,$$

is single-valued and in particular Lipschitz with constant $1/(1 - r\lambda)$ as claimed.

(ii) Expanding the square we obtain a well known expression of the negative Moreau envelope in terms of the convex conjugate, see, e.g., [RW98, Example 11.26]:

$$\begin{aligned}
-e_\lambda f(y) &= \sup_{x\in\mathbb{R}^m} -\frac{1}{2\lambda}\|x-y\|^2 - f(x)\\
&= \frac{1}{\lambda}\left(\sup_{x\in\mathbb{R}^m} \langle x,y\rangle - (1/2)\|x\|^2 - \lambda f(x)\right) - \frac{1}{2\lambda}\|y\|^2\\
&= \frac{1}{\lambda}\big(\lambda f + (1/2)\|\cdot\|^2\big)^*(y) - \frac{1}{2\lambda}\|y\|^2. \qquad (2.60)
\end{aligned}$$

Using the previous result, $e_\lambda f$ is differentiable with Lipschitz continuous gradient with modulus

$$\frac{1}{\lambda(1-r\lambda)} + \frac{1}{\lambda}$$

and for any $y \in \mathbb{R}^m$ we have:

$$\nabla e_\lambda f(y) = \frac{1}{\lambda}\Big(y - \nabla\big(\lambda f + (1/2)\|\cdot\|^2\big)^*(y)\Big). \qquad \square$$

2.3.5. Relatively amenable functions

An important source for examples of prox-regular functions is strong amenability [RW98, Definition 10.23], i.e. functions f that can locally be represented as a composition of a convex function with a smooth function and a certain constraint qualification. In the following we generalize this concept to the Bregmanian setting. To this end, the recently introduced generalization of Lipschitz differentiable functions to relatively smooth functions [BBT17; Bol+18; LFN18] (called smooth adaptable in [BBT17; Bol+18]) is used. We state a slightly modified version, where we introduce an additional open subset $V \subseteq \operatorname{int}(\operatorname{dom}\phi)$ of $\operatorname{int}(\operatorname{dom}\phi)$ and require the property to hold only on V instead of $\operatorname{int}(\operatorname{dom}\phi)$.

Definition 2.47 (relatively smooth function)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. A function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ that is $\mathcal{C}^1$ on an open subset $V \subseteq \operatorname{int}(\operatorname{dom}\phi)$ is called smooth relative to ϕ on V, if there exists $L \geq 0$ such that both $L\phi - f$ and $L\phi + f$ are convex on V.*

The following lemma, which we adopted from Lemma [Bol+18, Lemma 2.1], is a generalization of the classical full descent lemma to the relatively L-smooth case:

Lemma 2.48 (full extended descent lemma)**.** *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre. Then, a function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ that is $\mathcal{C}^1$ on an open subset $V \subseteq \operatorname{int}(\operatorname{dom}\phi)$ is smooth relative to ϕ on V with constant $L \geq 0$ if and only if the following holds for all $x, y \in V$*

$$|f(x) - f(y) - \langle \nabla f(y), x-y\rangle| \leq L D_\phi(x,y).$$

Definition 2.49. *A function $F\colon \mathbb{R}^m \to \mathbb{R}^n$ that is $\mathcal{C}^1$ on an open subset $V \subseteq \operatorname{int}(\operatorname{dom}\phi)$ is called L-smooth relative to ϕ on V with $L \geq 0$ if each coordinate function F_i is L-smooth relative to ϕ on V.*

We extend [RW98, Definition 10.23] to a setting where the inner smooth map is relatively smooth. Note that this property is required to hold only on a local neighborhood of a reference point. The first part recapitulates the definition of an amenable function from [RW98, Definition 10.23(a)], while a relatively amenable function generalizes the notion of strong amenability [RW98, Definition 10.23(b)].

Definition 2.50 (relatively amenable functions). *A function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ is amenable at $\bar{x}$, a point where $f(\bar{x})$ is finite, if there is an open neighborhood $V \subset \mathbb{R}^m$ of $\bar{x}$ on which f can be represented in the form $f = g \circ F$ for a $\mathcal{C}^1$ mapping $F\colon V \to \mathbb{R}^n$ and a proper, lsc, convex function $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ such that, in terms of $D = \operatorname{cl}(\operatorname{dom} g)$,*

$$\textit{the only } y \in N_D(F(\bar{x})) \textit{ with } \nabla F(\bar{x})^* y = 0 \textit{ is } y = 0. \tag{2.61}$$

If the mapping F is L-smooth relative to $\phi \in \Gamma_0(\mathbb{R}^m)$ on $V \subseteq \operatorname{int}(\operatorname{dom}\phi)$ it is called relatively amenable at $\bar{x} \in V$ relative to ϕ.

Clearly, the constraint qualification (2.61) is satisfied whenever $F(\bar{x}) \in \operatorname{int}(\operatorname{dom} g)$.

In the following proposition, we show that relatively amenable functions are indeed relatively prox-regular, which is completely analogous to the classical setting of strong amenability and prox-regularity [RW98, Proposition 13.32]. Actually, this also generalizes the classical Euclidean setting with $\phi = (1/2)\|\cdot\|^2$ to requiring the inner functions to be only $\mathcal{C}^1$ with a locally Lipschitz continuous gradient instead of $\mathcal{C}^2$.

Proposition 2.51. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be relatively amenable at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$ relative to ϕ. Then f is subdifferentially continuous and prox-regular relative to ϕ at $\bar{x}$.*

Proof. Since f is relatively amenable at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi)$ relative to ϕ, there exists an open neighborhood $V \subseteq \operatorname{int}(\operatorname{dom}\phi)$ of $\bar{x}$ on which $f = g \circ F$ for a proper lsc convex function $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and a $\mathcal{C}^1$ map $F\colon V \to \mathbb{R}^n$ that is L-smooth relative to ϕ on V. Clearly, f is lsc relative to V and therefore in particular locally lsc at $\bar{x}$. Note that the constraint qualification (2.61) holds not only at $\bar{x}$ but also on V, by possibly narrowing V. Otherwise there exists a sequence $x^\nu \to \bar{x}$ and $0 \neq y^\nu \in N_D(F(x^\nu))$ with $\nabla F(x^\nu)^* y^\nu = 0$ where we may assume $\|y^\nu\| = 1$ by normalizing. Taking a convergent subsequence of $\{y^\nu\}_{\nu\in\mathbb{N}}$ we have at the limit point y that $\nabla F(\bar{x})^* y = 0$ and $\|y\| = 1$, which is a contradiction.

In view of the chain rule from [RW98, Theorem 10.6], we have for all $x \in V$ that $\partial f(x) = \nabla F(x)^* \partial g(F(x))$. This means for $x \in V$, it holds that for any $v \in \partial f(x)$ there is some $u \in \partial g(F(x))$ such that $v = \nabla F(x)^* u$. Fix $\bar{v} \in \partial f(\bar{x})$. We want to show that there exist $\varepsilon > 0$ and $\eta > 0$ such that for any x with $\|x - \bar{x}\| < \varepsilon$ and $u \in \partial g(F(x))$ with the property $\|v - \bar{v}\| = \|\nabla F(x)^* u - \bar{v}\| < \varepsilon$ we have that $\|u\| < \eta$. Since g is convex it is locally Lipschitz on $\operatorname{int}(\operatorname{dom} g)$. This means whenever $F(\bar{x}) \in \operatorname{int}(\operatorname{dom} g)$ there is $\varepsilon > 0$ sufficiently small such that due to continuity of F we have $F(x) \in \operatorname{int}(\operatorname{dom} g)$ near $F(\bar{x})$ and there is some finite $\eta > 0$ such that $\|u\| < \eta$ for any $u \in \partial g(F(x))$. Now assume $F(\bar{x}) \in \operatorname{bdry}(\operatorname{dom} g)$ and suppose that there exist sequences $x^\nu \to \bar{x}$ and $u^\nu \in \partial g(F(x^\nu))$ with $\nabla F(x^\nu)^* u^\nu \to \bar{v}$ and $\|u^\nu\| \to \infty$. For a decomposition $u^\nu = u_0^\nu + u_i^\nu$ with $u_0^\nu \in \ker \nabla F(x^\nu)^*$ and $u_i^\nu \in \operatorname{rge} \nabla F(x^\nu)$, this yields $\|u_0^\nu\| \to \infty$. Through [RW98, Proposition 8.12] u^ν is in particular a regular subgradient of g at $F(x^\nu)$. Obviously, by possibly going to a subsequence, $u^\nu/\|u^\nu\|$ converges to a point on the unit circle, which, by definition, belongs to the horizon subgradient and, by [RW98, Proposition 8.12], to $N_{\operatorname{cl}(\operatorname{dom} g)}(F(\bar{x}))$. Moreover, this point lies in $\ker \nabla F(\bar{x})^*$, since $u_i^\nu/\|u^\nu\| \to 0$. This is a contradiction to the constraint qualification. Let $\|x - \bar{x}\| < \varepsilon$, $\|x' - \bar{x}\| < \varepsilon$ and $\nabla F(x)^* u = v \in \partial f(x)$ with $\|v - \bar{v}\| < \varepsilon$ for some $u \in \partial g(F(x))$. Due to the argument above we have $\|u\| \leq \eta$ and therefore also $\|u\|_1 \leq \gamma$ for some $\gamma > 0$. Then, since F is component-wise relatively L-smooth, thanks to Lemma 2.48, we can make the following computation. We have for some $r \geq \gamma L$:

$$\begin{aligned} f(x') - f(x) &= g(F(x')) - g(F(x)) \\ &\geq \langle u, F(x') - F(x)\rangle \\ &\geq \langle u, \nabla F(x)(x'-x)\rangle - \sum_{i=1}^{n} |u_i| L D_\phi(x', x) \\ &\geq \langle u, \nabla F(x)(x'-x)\rangle - \gamma L D_\phi(x', x) \\ &\geq \langle \nabla F(x)^* u, x'-x\rangle - r D_\phi(x', x) \\ &= \langle v, x'-x\rangle - r D_\phi(x', x), \end{aligned} \tag{2.62}$$

which shows that f is prox-regular at $\bar{x}$ for $\bar{v}$ relative to ϕ.

Subdifferential continuity of f at $\bar{x}$ for $\bar{v}$ follows from the same arguments as in the proof of [RW98, Proposition 13.32]. □

Remark 2.52. Note that the estimate in the proof also holds when each component function F_i is L-smooth relative to a potentially different ϕ_i.

Corollary 2.53. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ such that $f = g \circ F$ for $g : \mathbb{R}^n \to \mathbb{R}$ proper lsc and globally Lipschitz and $F : \mathbb{R}^m \to \mathbb{R}^n$ is L-smooth relative to ϕ on $V = \operatorname{int}(\operatorname{dom}\phi)$, then f is r-relatively hypoconvex, i.e., $f + r\phi$ is proper lsc and convex on $\operatorname{int}(\operatorname{dom}\phi)$ for some $r > 0$ sufficiently large.*

Proof. The assertion follows from the fact that due to global Lipschitz continuity of g subgradients $u \in \partial g(x)$ are uniformly bounded, i.e., there is $\infty > \eta > 0$ such that $\|u\| \leq \eta$ for all $(x, u) \in \operatorname{gph} \partial g$ and Inequality (2.62) in the proof of Proposition 2.51. □

Amenable functions whose representation $g \circ F$ involves a change of coordinates F have rich properties. Thanks to calculus of prox-regularity [PR10, Theorem 3.1] even for a prox-regular (outer) function g, and $F \in \mathcal{C}^2$ the composition is also prox-regular if a constraint qualification holds at the reference point.

For completeness we provide a proof for the special case ∇F is nonsingular at the reference point. Unlike [PR10, Theorem 3.1] we assume that $F \in \mathcal{C}^1$ is locally Lipschitz, which is implied by $F \in \mathcal{C}^2$.

Proposition 2.54. *Let $f \colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be finite at $\bar{x}$. Let $V \subset \mathbb{R}^m$ be an open neighborhood of $\bar{x}$ on which f can be represented in the form $f = g \circ F$ for a $\mathcal{C}^1$ mapping $F \colon V \to \mathbb{R}^m$ and a function $g \colon \mathbb{R}^m \to \overline{\mathbb{R}}$. If g is prox-regular at $F(\bar{x})$ for $\bar{u} \in \partial g(F(\bar{x}))$, $\nabla F(\bar{x})$ is nonsingular and ∇F is Lipschitz on V, then f is prox-regular at $\bar{x}$ for $\bar{v} = \nabla F(\bar{x})^* \bar{u}$.*

Proof. Since g is prox-regular at $F(\bar{x})$ for $\bar{u} \in \partial g(F(\bar{x}))$ and due to the continuity of F, there exists a constant $r' > 0$ such that for any $\varepsilon' > 0$ sufficiently small we have:

$$g(F(x')) \geq g(F(x)) + \langle u, F(x') - F(x)\rangle - \frac{r'}{2}\|F(x') - F(x)\|^2$$

for $\|F(\bar{x}) - F(x')\| < \varepsilon'$, $\|F(\bar{x}) - F(x)\| < \varepsilon'$, $\|\bar{u} - u\| < \varepsilon'$ with $u \in \partial g(F(x))$ and $|g(F(x)) - g(F(\bar{x}))| < \varepsilon'$. Since F is locally Lipschitz, there exists r'' such that

$$\frac{r'}{2}\|F(x') - F(x)\|^2 \leq \frac{r''}{2}\|x' - x\|^2.$$

For the inner product, we use the fact that the component functions F_i satisfy

$$F_i(x') - F_i(x) = \int_0^1 \langle \nabla F_i(x + t(x' - x)), x' - x \rangle \, \mathrm{d}t$$

and since ∇F is Lipschitz we have for some $L > 0$

$$\begin{aligned} \langle u, F(x') - F(x) \rangle &= \langle u, \nabla F(x)(x' - x) \rangle \\ &\quad + \int_0^1 \big\langle u, \big(\nabla F(x + t(x' - x)) - \nabla F(x)\big)(x' - x) \big\rangle \, \mathrm{d}t \\ &\leq \langle \nabla F(x)^* u, x' - x \rangle \\ &\quad + \|u\| \int_0^1 \|\nabla F(x + t(x' - x)) - \nabla F(x)\| \cdot \|x' - x\| \, \mathrm{d}t \\ &\leq \langle \nabla F(x)^* u, x' - x \rangle + \|u\| \cdot \|x' - x\|^2 \frac{L}{2}. \end{aligned}$$

Combining the inequalities we obtain, since u is bounded around $\bar{u}$, that for some $r > 0$ we have

$$g(F(x')) \geq g(F(x)) + \langle \nabla F(x)^* u, x' - x \rangle - \frac{r}{2}\|x' - x\|^2, \tag{2.63}$$

whenever $\|F(\bar{x}) - F(x')\| < \varepsilon'$, $\|F(\bar{x}) - F(x)\| < \varepsilon'$, $\|\bar{u} - u\| < \varepsilon'$ with $u \in \partial g(F(x))$ and $|g(F(x)) - g(F(\bar{x}))| < \varepsilon'$.

As F is $\mathcal{C}^1$ with $\nabla F(\bar{x})$ nonsingular, in view of the inverse function theorem, we know that F is invertible on a small neighborhood of $\bar{x}$.

In view of [RW98, Exercise 10.7], the chain rule holds on a neighborhood of $\bar{x}$, i.e. we have $\nabla F(x)^* \partial g(F(x)) = \partial f(x)$ when x near $\bar{x}$. This means for such x and any $v \in \partial f(x)$ there exists $u \in \partial g(F(x))$ such that $v = \nabla F(x)^* u$.

Let $v \in \partial f(x)$ near $\bar{v}$ and x near $\bar{x}$. Let $u \in \partial g(F(x))$ such that $v = \nabla F(x)^* u$. In view of the Lipschitz continuity of ∇F, we can make the following computation:

$$\begin{aligned} \sigma_{\min}(\nabla F(x))\|u - \bar{u}\| &\leq \|\nabla F(x)^* u - \nabla F(x)^* \bar{u}\| \\ &\leq \|\nabla F(x)^* u - \nabla F(\bar{x})^* \bar{u}\| + \|\nabla F(\bar{x})^* \bar{u} - \nabla F(x)^* \bar{u}\| \\ &\leq \|v - \bar{v}\| + \|\bar{u}\| L \|\bar{x} - x\|. \end{aligned}$$

Since $\nabla F(x)$ is invertible we know that the smallest singular value $\sigma_{\min}(\nabla F(x))$ of $\nabla F(x)$ is positive. Since the ordered singular value map as well as ∇F are continuous we have a uniform bound $\sigma_{\min}(\nabla F(x)) \geq \delta > 0$ for $\|x - \bar{x}\| \leq \varepsilon$. Then dividing the inequality by $\sigma_{\min}(\nabla F(x))$ shows that for $v \in \partial f(x)$ near $\bar{v}$ and x near $\bar{x}$ we guarantee u near $\bar{u}$.

Overall, this means we can find $\varepsilon > 0$ sufficiently small, such that whenever $\|\bar{x} - x\| < \varepsilon$, $\|\bar{x} - x'\| < \varepsilon$ and $\|\bar{v} - v\| < \varepsilon$, $v \in \partial f(x)$ and $|f(x) - f(\bar{x})| < \varepsilon$, we guarantee via the continuity of F that $\|F(\bar{x}) - F(x')\| < \varepsilon'$, $\|F(\bar{x}) - F(x)\| < \varepsilon'$, $|g(F(x)) - g(F(\bar{x}))| < \varepsilon'$ and $\|\bar{u} - u\| < \varepsilon'$. Then, in view of (2.63), we have:

$$f(x') \geq f(x) + \langle v, x' - x \rangle - \frac{r}{2}\|x' - x\|^2.$$

Since g is in particular finite and locally lsc at $F(\bar{x})$, f is finite and locally lsc at $\bar{x}$. We may conclude that f is prox-regular at $\bar{x}$ for $\bar{v}$. □

A particularly interesting choice for F in context of the right Bregman proximal mapping is $F = \nabla\phi^*$, for a Legendre function ϕ:

Corollary 2.55. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be finite at $\bar{x}$. Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre, very strictly convex and $\nabla^2\phi$ locally Lipschitz at $\bar{x}$. Then f is prox-regular at $\bar{x}$ for $\bar{v} \in \partial f(\bar{x})$ if and only if $f \circ \nabla\phi^*$ is prox-regular at $\nabla\phi(\bar{x})$ for $\bar{u} = \nabla^2\phi^*(\nabla\phi(\bar{x}))\bar{v} \in \partial(f \circ \nabla\phi^*)(\nabla\phi(\bar{x}))$.*

Proof. Since ϕ is very strictly convex, we know that $\nabla^2\phi(x)$ is positive definite for $x \in \operatorname{int}(\operatorname{dom}\phi)$ and therefore nonsingular. In view of Lemma 1.18 we know that $\nabla^2\phi^*(x) = (\nabla^2\phi(\nabla\phi^*(x)))^{-1}$, which is locally Lipschitz as the composition of the inverse matrix map, $\nabla^2\phi$ and $\nabla\phi^*$, all of which are locally Lipschitz. The conclusion then follows from applying Proposition 2.54 to $f \circ \nabla\phi^*$ resp. $f = (f \circ \nabla\phi^*) \circ \nabla\phi$. □

In particular, combining Lemma 2.26, Theorem 2.40 and the Corollary 2.55 above we may guarantee the local single-valuedness of the right Bregman proximal mapping $\overrightarrow{\operatorname{prox}}_{\lambda}^{\phi} f$ of f under prox-regularity of f and very strict convexity of ϕ.

The class of relatively amenable functions is a wide source of examples for relatively prox-regular functions:

Example 2.56. Choose $f\colon \mathbb{R}^2 \to \overline{\mathbb{R}}$ with $f(x_1, x_2) = g(F(x_1, x_2))$ for $g\colon \mathbb{R} \to \overline{\mathbb{R}}$ with $g := \iota_{\mathbb{R}_{\leq 0}}$ and $F\colon \mathbb{R}^2 \to \mathbb{R}$ with $F(x_1, x_2) = 2x_1^2 - 3|x_1|^{1.1} - x_2$. Choose $\phi\colon \mathbb{R}^2 \to \mathbb{R}$ with $\phi(x_1, x_2) = x_1^2 + |x_1|^{1.1} + x_2^2$. Then, clearly, F is L-smooth relative to ϕ for $L = 3$. Since $\nabla F(0) = (0, -1)$ is full rank, f is relatively amenable at 0 and in view of Proposition 2.51, relatively prox-regular at 0. Note that f is the indicator function of the epigraph of the nonconvex function $h(x) = 2x^2 - 3|x|^{1.1}$ and therefore neither hypoconvex relative to ϕ nor classically prox-regular at 0.

The above example is illustrated in Figure 2.2.

2.3.6. Smoothness of the Bregman–Moreau envelope

So far we know that relative prox-regularity provides us with a sufficient condition for the local single-valuedness of the left and right Bregman proximal mapping. This in turn allows us to guarantee that the Bregman envelope functions are locally $\mathcal{C}^1$ providing an explicit formula for their gradients, which involves the corresponding Bregman proximal mappings. The formulas for both the left and the right envelope have been proven previously in the convex setting [BCN06, Proposition 3.12] and for the left envelope in a more general relatively hypoconvex setting [KS12, Corollary 3.1].

We provide an interesting additional (global) regularity property of the Bregman envelope function, which comes in handy for proving that the envelope function is $\mathcal{C}^1$: Both the left envelope $\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^*$ and the right envelope $\overrightarrow{\operatorname{env}}_{\lambda}^{\phi} f$ have the one-sided smoothness property relative to ϕ^* resp. ϕ, and therefore yield promising candidates for optimization with Bregman proximal gradient methods [BBT17; Bol+18].

Proposition 2.57. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper. Then it holds that:*

(i) *If $\operatorname{dom} f \cap \operatorname{dom}\phi$ is nonempty, f is prox-bounded relative to ϕ with threshold λ_f and ϕ is super-coercive, then for any $\lambda \in (0, \lambda_f)$,*

$$\frac{1}{\lambda}\phi^* - \overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^* = \left(f + \frac{1}{\lambda}\phi\right)^* \left(\frac{\cdot}{\lambda}\right)$$

is proper, lsc and convex.

(ii) *If f is right prox-bounded relative to ϕ with threshold λ_f and* $\operatorname{dom}\phi = \mathbb{R}^m$, *then for any $\lambda \in (0, \lambda_f)$,*

$$\frac{1}{\lambda}\phi - \overrightarrow{\operatorname{env}}_{\lambda}^{\phi} f = \left(g + \frac{1}{\lambda}\phi^*\right)^*\left(\frac{\cdot}{\lambda}\right),$$

with $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ defined by

$$g(z) := \begin{cases} f(\nabla\phi^*(z)), & \text{if } z \in \operatorname{int}\operatorname{dom}\phi^* \\ +\infty & \text{otherwise,} \end{cases}$$

is proper, lsc and convex.

Proof. Let f be prox-bounded relative to ϕ super-coercive with threshold $\lambda_f > 0$ and let $\lambda \in (0, \lambda_f)$. From [KS12, Theorem 2.4] we obtain that we have for all $y \in \mathbb{R}^m$:

$$\frac{1}{\lambda}\phi^*(y) - \left(f + \frac{1}{\lambda}\phi\right)^*\left(\frac{y}{\lambda}\right) = \left(\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^*\right)(y).$$

Then, thanks to Lemma 1.16(iv), $\operatorname{dom}(\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^*) = \mathbb{R}^m$. Furthermore, in view of Proposition 2.21(ii), $f + (1/\lambda)\phi$ is bounded below and proper. Then, clearly, $\operatorname{con}(f + (1/\lambda)\phi)$ is proper, and in view of [RW98, Theorem 11.1], $(f + (1/\lambda)\phi)^*(\cdot/\lambda)$ is proper, lsc and convex. Since in view of Lemma 1.16(iv) also $\operatorname{dom}\phi^* = \mathbb{R}^m$ we can reorder the terms and obtain that

$$\frac{1}{\lambda}\phi^*(y) - \left(\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^*\right)(y) = \left(f + \frac{1}{\lambda}\phi\right)^*\left(\frac{y}{\lambda}\right),$$

and the assertion follows.

Part (ii) follows from a similar argument invoking [BDL18, Proposition 2.4(ii)] and the observation that right prox-boundedness of f relative to ϕ implies prox-boundedness of g relative to ϕ^*. □

Interestingly, the above result shows, that even if f is not lsc, the negative left and right envelopes are always proper and lsc if f is left resp. right prox-bounded and ϕ is super-coercive resp. $\operatorname{dom}\phi = \mathbb{R}^m$.

The one-sided relative L-smoothness property from Proposition 2.57(i) yields an additive decomposition of the negative left envelope function $-\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^*$ into a finite convex part and a smooth part $-\lambda^{-1}\phi^* \in \mathcal{C}^1$. As a consequence, in view of [RW98, Theorem 10.33] and [RW98, Exercise 10.35](a), the negative envelope is in particular subsmooth, see Definition 2.16.

To prove the desired smoothness property of the envelope function, along with a gradient formula, we need a stronger constructive result where the pointwise max representation in the definition of subsmoothness is explicitly given in terms of the proximal mapping:

Lemma 2.58. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc. If* $\operatorname{dom} f \cap \operatorname{dom}\phi$ *is nonempty, f is prox-bounded relative to ϕ with threshold λ_f and ϕ is super-coercive, then for any $\lambda \in (0, \lambda_f)$ the envelope function $-\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^*$ is lower-$\mathcal{C}^1$ with representations*

$$-\overleftarrow{\operatorname{env}}_{\lambda}^{\phi} f \circ \nabla\phi^* = \max_{x \in Z} h(x, \cdot), \qquad \overleftarrow{\operatorname{prox}}_{\lambda}^{\phi} f \circ \nabla\phi^* = \operatorname*{arg\,max}_{x \in Z} h(x, \cdot),$$

on V being a neighborhood of $\bar{y} \in \mathbb{R}^m$, $h(x, y) := -f(x) - \frac{1}{\lambda}(\phi(x) + \phi^*(y) - \langle y, x\rangle)$,

$\overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y)) \subset Z$, with both $h(x,y)$ and $\nabla_y h(x,y) = -(1/\lambda)(\nabla\phi^*(y) - x)$, depending continuously on $(x,y) \in Z \times V$.

Proof. Let $\lambda \in (0, \lambda_f)$ and $\bar{y} \in \mathbb{R}^m$. Let $V \subset \mathbb{R}^m$ be a compact neighborhood of $\bar{y}$. Choose $Z := \{x \in \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y)) : y \in V\}$. We have $\arg\max_{x\in Z} h(x,y) = \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y))$ and due to Lemma 2.22(i) we have

$$-\overleftarrow{\mathrm{env}}_\lambda^\phi f(\nabla\phi^*(y)) = \max_{x\in Z} h(x,y).$$

Furthermore Z is closed. Otherwise there is a sequence $x^\nu \to x$ with $x^\nu \in \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y^\nu))$ for some $y^\nu \in V$ with $x \notin Z$. Taking a convergent subsequence $y^{\nu_j} \to y \in V$ (exists as V is bounded) we know from Lemma 2.22(iii) and the continuity of $\nabla\phi^*$ that $x^{\nu_j} \to x \in \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y)) \subset Z$, a contradiction. Z is also bounded, hence compact. Otherwise there is a sequence $\|x^\nu\| \to \infty$ with $x^\nu \in \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y^\nu))$ for some $y^\nu \in V$. However, going to a convergent subsequence $y^{\nu_j} \to y \in V$ we have for $x^{\nu_j} \in \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y^{\nu_j}))$ using Lemma 2.22(iii) that $\{x^{\nu_j}\}_{\nu\in\mathbb{N}}$ is bounded, a contradiction.

Next we show that f becomes continuous over the compact space Z. Equivalently this means h is continuous over $Z \times V$. We assume the contrary: Suppose that f is not continuous over Z. This means there is a sequence $\{x^\nu\}_{\nu\in\mathbb{N}} \subset Z$ with $x^\nu \to x^* \in Z$ such that $f(x^\nu) \not\to f(x^*)$. For any $x^\nu \in Z$ there exists $y^\nu \in V$ such that $x^\nu \in \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y^\nu))$ and $f(x^\nu) \leq f(x^\nu) + \lambda^{-1} D_\phi(x^\nu, \nabla\phi^*(y^\nu)) = \overleftarrow{\mathrm{env}}_\lambda^\phi f(\nabla\phi^*(y^\nu)) \leq \gamma$, for some $\infty > \gamma$, since $\overleftarrow{\mathrm{env}}_\lambda^\phi f \circ \nabla\phi^*$ is continuous and V is compact. Since f is proper, lsc it is also uniformly bounded from below over Z: $-\infty < \delta \leq f(x^\nu)$. This means we can find a subsequence indexed by ν_j such that $f(x^{\nu_j}) \to f^* \geq f(x^*) + \varepsilon$, with $\varepsilon > 0$. Taking another subsequence if necessary we can ensure that $y^{\nu_j} \to y^*$ and $x^* \in \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y^*))$. By continuity of the envelope function, cf. Lemma 2.22(ii), we then have $\overleftarrow{\mathrm{env}}_\lambda^\phi f(\nabla\phi^*(y^{\nu_j})) = f(x^{\nu_j}) + \lambda^{-1} D_\phi(x^{\nu_j}, \nabla\phi^*(y^{\nu_j})) \to \overleftarrow{\mathrm{env}}_\lambda^\phi f(\nabla\phi^*(y^*)) = f(x^*) + \lambda^{-1} D_\phi(x^*, \nabla\phi^*(y^*))$. Hence, along that subsequence $f(x^{\nu_j}) \to f(x^*)$, a contradiction.

Since $h(x,\cdot)$ is $\mathcal{C}^1$ as ϕ^* is $\mathcal{C}^1$ on $\operatorname{dom}\phi^* = \mathbb{R}^m$ both $h(x,y)$ and $\nabla_y h(x,y) = -(1/\lambda)(\nabla\phi^*(y) - x)$, depend continuously jointly on $(x,y) \in Z \times V$. Hence $-\overleftarrow{\mathrm{env}}_\lambda^\phi f \circ \nabla\phi^*$ is lower-$\mathcal{C}^1$. □

The following proposition provides us with an explicit formula for the gradient of the composition $\overleftarrow{\mathrm{env}}_\lambda^\phi f \circ \nabla\phi^*$. The gradient formulas of both the left and right envelope are direct consequences of this underlying formula.

Proposition 2.59. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ be Legendre and super-coercive and the function $f\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded with threshold λ_f. Let f be relatively prox-regular at $\bar{x} \in \operatorname{int}(\operatorname{dom}\phi) \cap \operatorname{dom} f$ for $\bar{v} \in \partial f(\bar{x})$.*

If $\lambda \in (0, \lambda_f)$ is sufficiently small, we have that $\overleftarrow{\mathrm{env}}_\lambda^\phi f \circ \nabla\phi^*$ is $\mathcal{C}^1$ around

$$\bar{y} := \nabla\phi(\bar{x}) + \lambda\bar{v},$$

with

$$\nabla\big(\overleftarrow{\mathrm{env}}_\lambda^\phi f \circ \nabla\phi^*\big)(y) = \frac{1}{\lambda}\big(\nabla\phi^*(y) - \overleftarrow{\mathrm{prox}}_\lambda^\phi f(\nabla\phi^*(y))\big), \tag{2.64}$$

and y sufficiently close to $\bar{y}$. If, furthermore, ϕ is very strictly convex, then $\nabla(\overleftarrow{\mathrm{env}}_\lambda^\phi f \circ \nabla\phi^*)$ is Lipschitz continuous on a neighborhood of $\bar{y}$.

Proof. In view of Lemma 2.58 and invoking [RW98, Theorem 10.31] we obtain that

$$\begin{aligned}\partial\big(-\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*\big)(y) &= \mathrm{con}\,\big\{\nabla_y h(x,y) : x\in \overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f(\nabla\phi^*(y))\big\}\\ &= -\frac{1}{\lambda}\Big(\nabla\phi^*(y)-\mathrm{con}\big(\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f(\nabla\phi^*(y))\big)\Big).\end{aligned} \tag{2.65}$$

Due to the assumptions we can invoke Theorem 2.40(iii) and assert that for $\lambda\in(0,\lambda_f)$ sufficiently small $\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f\circ\nabla\phi^*$ is singled-valued and continuous at y near $\bar y=\nabla\phi(\bar x)+\lambda\bar v$. In view of Equation (2.65), this means that $\partial(-\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*)$ is single-valued and continuous around $\bar y$. Through [RW98, Corollary 9.19] we obtain that $-\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*$ is $\mathcal{C}^1$ around $\bar y$ with

$$\frac{1}{\lambda}\big(\nabla\phi^*(y)-\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f(\nabla\phi^*(y))\big)=\nabla\big(\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*\big)(y).$$

If, furthermore, ϕ is very strictly convex, we know due to Corollary 2.43 that $\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f$ is locally Lipschitz at $\nabla\phi^*(\bar y)$. Then $\nabla(\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*)$ is locally Lipschitz at $\bar y$ as a composition resp. sum of locally Lipschitz maps. □

Corollary 2.60. *Let $\phi\in\Gamma_0(\mathbb{R}^m)$ be Legendre, super-coercive and $\mathcal{C}^2$ on* $\mathrm{int}(\mathrm{dom}\,\phi)$ *and $f\colon\mathbb{R}^m\to\overline{\mathbb{R}}$ be proper lsc and prox-bounded with threshold λ_f. Let f be relatively prox-regular at $\bar x\in\mathrm{int}(\mathrm{dom}\,\phi)\cap\mathrm{dom}\,f$ for $\bar v\in\partial f(\bar x)$. If $\lambda\in(0,\lambda_f)$ is sufficiently small we have that $\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f$ is $\mathcal{C}^1$ around*

$$\bar y:=\nabla\phi^*(\nabla\phi(\bar x)+\lambda\bar v),$$

with

$$\nabla\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f(y)=\frac{1}{\lambda}\nabla^2\phi(y)\big(y-\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f(y)\big), \tag{2.66}$$

and y sufficiently close to $\bar y$. If, furthermore, ϕ is very strictly convex and $\nabla^2\phi$ is locally Lipschitz on $\mathrm{int}(\mathrm{dom}\,\phi)$*, then $\nabla\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f$ is Lipschitz continuous on a neighborhood of $\bar y$.*

Proof. The result follows from the identity $\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f=(\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*)\circ\nabla\phi$ via the chain rule and Proposition 2.59: Then we have for y near $\bar y$ that

$$\begin{aligned}\nabla\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f(y) &= \nabla\big((\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*)\circ\nabla\phi\big)(y)\\ &= \nabla^2\phi(y)\cdot\nabla\big(\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f\circ\nabla\phi^*\big)(\nabla\phi(y))\\ &= \frac{1}{\lambda}\nabla^2\phi(y)\big(\nabla\phi^*(\nabla\phi(y))-\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f(\nabla\phi^*(\nabla\phi(y)))\big)\\ &= \frac{1}{\lambda}\nabla^2\phi(y)\big(y-\overleftarrow{\mathrm{prox}}_{\lambda}^{\phi} f(y)\big).\end{aligned}$$

If, furthermore, ϕ is very strictly convex and $\nabla^2\phi$ is locally Lipschitz, clearly, $\nabla\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f$ is locally Lipschitz at $\bar y$ as it is given as the product of two locally Lipschitz maps. □

In view of Lemma 2.26, the right Bregman envelope involves the expression $\overleftarrow{\mathrm{env}}_{\lambda}^{\phi^*}(f\circ\nabla\phi^*)\circ\nabla\phi$. This allows us to invoke the proposition above to derive a gradient formula for the right envelope.

Corollary 2.61. *Let $\phi\in\Gamma_0(\mathbb{R}^m)$ be Legendre with* $\mathrm{dom}\,\phi=\mathbb{R}^m$ *and the function $f\colon\mathbb{R}^m\to\overline{\mathbb{R}}$ be proper lsc. In addition, let one of the following two conditions hold true:*

(a) *either, f is relatively right prox-bounded with threshold λ_f and ϕ is super-coercive,*

(b) *or f is coercive (and thus relatively right prox-bounded with threshold $\lambda_f = +\infty$).*

Define $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ by

$$g(z) := \begin{cases} f(\nabla\phi^*(z)), & \textit{if } z \in \operatorname{int}\operatorname{dom}\phi^* \\ +\infty & \textit{otherwise.} \end{cases}$$

For $\bar{x} \in \operatorname{dom} f$ let g be prox-regular relative to ϕ^* at $\nabla\phi(\bar{x})$ for $\bar{v} \in \partial g(\nabla\phi(\bar{x}))$. If $\lambda \in (0, \lambda_f)$ is sufficiently small we have that $\overrightarrow{\operatorname{env}}_\lambda^\phi f$ is $\mathcal{C}^1$ around

$$\bar{y} = \bar{x} + \lambda\bar{v}$$

with

$$\begin{aligned} \nabla\,\overrightarrow{\operatorname{env}}_\lambda^\phi f(y) &= \frac{1}{\lambda}\big(\nabla\phi(y) - \nabla\phi(\,\overrightarrow{\operatorname{prox}}_\lambda^\phi f(y))\big) \\ &= \frac{1}{\lambda}\big(\nabla\phi(y) - \overleftarrow{\operatorname{prox}}_\lambda^{\phi^*} g(\nabla\phi(y))\big), \end{aligned} \tag{2.67}$$

and y sufficiently close to $\bar{y}$. If, furthermore, ϕ is very strictly convex, $\nabla\,\overrightarrow{\operatorname{env}}_\lambda^\phi f$ is Lipschitz continuous on a neighborhood of $\bar{y}$.

Proof. In view of Lemma 1.16, ϕ^* is super-coercive. Then the result follows from the identities

$$\overrightarrow{\operatorname{env}}_\lambda^\phi f = \overleftarrow{\operatorname{env}}_\lambda^{\phi^*} g \circ \nabla\phi$$

and

$$\overrightarrow{\operatorname{prox}}_\lambda^\phi f = \nabla\phi^* \circ \overleftarrow{\operatorname{prox}}_\lambda^{\phi^*} g \circ \nabla\phi,$$

cf. Lemma 2.26 as well as Lemma 2.28 and Proposition 2.59 applied to $\overleftarrow{\operatorname{env}}_\lambda^{\phi^*} g \circ \nabla\phi$. □

Note that when ϕ is very strictly convex and in addition $\nabla^2\phi$ is Lipschitz at $\bar{x}$, in view of Corollary 2.55 and Proposition 2.38, the relative prox-regularity assumption on $f \circ \nabla\phi^*$ is equivalent to classical prox-regularity of f at $\bar{x}$ for $\nabla^2\phi(\bar{x})\bar{v} \in \partial f(\bar{x})$.

The gradient formula of the right Bregman envelope provides us with an explicit sufficient condition for the local $\mathcal{C}^1$ property of the right Bregman distance function of a convex set, which, in view of [Bau+09, Theorem 6.6] and [Bau+09, Example 7.5], does not hold globally in general. To illustrate this we revisit and extend [Bau+09, Example 7.5].

Example 2.62. Define the Legendre function $\phi\colon \mathbb{R}^2 \to \mathbb{R}$ as

$$\phi(x_1, x_2) := \exp(x_1) + \exp(x_2).$$

Then the convex conjugate is given as

$$\phi^*(y_1, y_2) = \begin{cases} y_1\log(y_1) - y_1 + y_2\log(y_2) - y_2 & \text{if } y_1 \geq 0, y_2 \geq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Define $f := \iota_C$ for

$$C := \{(x, 2x) : x \in [0, 1]\},$$

and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ by

$$g(z) := \begin{cases} f(\nabla\phi^*(z)), & \text{if } z \in \operatorname{int}\operatorname{dom}\phi^* \\ +\infty & \text{otherwise.} \end{cases}$$

Then $g = \iota_{\nabla\phi(C)}$ for $\nabla\phi(C) = \{(\exp(x), \exp(2x)) : x \in [0,1]\}$ which is obviously a compact and nonconvex set. Let $\bar{x} \in C$. Since ϕ^* is $\mathcal{C}^3$ on $\operatorname{int}(\operatorname{dom}\phi^*)$ and very strictly convex and therefore $\nabla^2\phi^*$ is full rank at $\nabla\phi(\bar{x})$ and $g(\nabla\phi(\bar{x})) = f(\bar{x})$ is finite we have that g is strongly amenable at $\nabla\phi(\bar{x})$. In view of [RW98, Proposition 13.32] g is also prox-regular at $\nabla\phi(\bar{x})$. Let $\bar{v} \in \partial g(\nabla\phi(\bar{x}))$, i.e. $\bar{v}$ is a limiting normal of $\nabla\phi(C)$ at $\nabla\phi(\bar{x})$ and, in view of the prox-regularity of g at $\nabla\phi(\bar{x})$, even a proximal normal of $\nabla\phi(C)$ at $\nabla\phi(\bar{x})$. In view of Proposition 2.38, g is also prox-regular relative to ϕ^* at $\nabla\phi(\bar{x})$. In addition we know that f is proper, lsc and coercive. Then we can invoke Corollary 2.61 to assert, that for $\lambda > 0$ being sufficiently small, the right Bregman distance function $\overrightarrow{\operatorname{env}}_\lambda^\phi f = (1/\lambda)\,\overrightarrow{\operatorname{env}}_1^\phi \iota_C$ is $\mathcal{C}^1$ around the point $\bar{y} = \bar{x} + \lambda\bar{v}$ and the proximal mapping is single-valued, even locally Lipschitz on that neighborhood.

In view of Corollary 2.55, the local $\mathcal{C}^1$ property of the right Bregman distance function $\overrightarrow{\operatorname{env}}_1^\phi \iota_C$ even holds for nonconvex C with ι_C prox-regular.

2.3.7. Example of a simple Bregman proximal mapping

We present an analytically solvable Bregman proximal mapping for the relatively prox-regular function $(1/p)|x|^p$ for $p \in (0,1)$. While the function is also prox-regular, the classical proximal mapping cannot be solved analytically, except for $p = 1/2$. For each $p \in (0,1)$, we define a Legendre function ϕ relative to which $(1/p)|x|^p$ is prox-regular and the left Bregman proximal mapping can be solved easily. This example is potentially interesting for applications that involve optimization with sparsity regularization, as for example in compressed sensing.

Example 2.63. Let $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = (1/p)|x|^p$ with $p \in (0,1)$ and choose $\phi(x) = (1/q)|x|^q$ with $q > 1$. For some $y \in \mathbb{R}^m$ we seek a closed form solution of the left Bregman proximal mapping

$$\overleftarrow{\operatorname{prox}}_\lambda^\phi f(y) = \operatorname*{arg\,min}_{x\in\mathbb{R}} \frac{1}{p}|x|^p + \frac{1}{\lambda}D_\phi(x,y) = \operatorname*{arg\,min}_{x\in\mathbb{R}} \frac{1}{p}|x|^p + \frac{1}{q\lambda}|x|^q - cx,$$

for $c := (1/\lambda)\operatorname{sign}(y)|y|^{q-1}$. Let $\bar{x} \in \overleftarrow{\operatorname{prox}}_\lambda^\phi f(y)$. Note that

$$\partial f(x) = \begin{cases} \operatorname{sign}(x)|x|^{p-1} & \text{if } x \neq 0, \\ \mathbb{R}, & \text{otherwise.} \end{cases}$$

and $\phi'(x) = \operatorname{sign}(x)|x|^{q-1}$. Then, the first order necessary optimality condition is given as follows:

$$0 \in \partial f(\bar{x}) + \frac{1}{\lambda}\phi'(\bar{x}) - c = \begin{cases} \bar{x}^{p-1} + \frac{1}{\lambda}\bar{x}^{q-1} - c & \text{if } \bar{x} > 0, \\ \mathbb{R} & \text{if } \bar{x} = 0, \\ -\bar{x}^{p-1} - \frac{1}{\lambda}\bar{x}^{q-1} - c, & \text{otherwise.} \end{cases}$$

This shows that the left Bregman proximal mapping can be evaluated by checking the three conditions individually and combining the minimum objective solutions. Note that the first and the last condition are exclusive while the first two and the last two conditions can potentially be satisfied simultaneously. Indeed, the Bregman proximal mapping of the given f can be multivalued. Assume that $\bar{x} > 0$ as the other case follows analogously. I.e. we seek a point $\bar{x} > 0$ that satisfies

$$\bar{x}^{p-1}\left(1 + \frac{1}{\lambda}\bar{x}^{q-p} - c\bar{x}^{1-p}\right) = 0.$$

Let $\alpha \in \{2, 3, 4, \ldots\}$ and choose q according to the following condition:

$$\frac{q-p}{1-p} = \alpha,$$

which is equivalent to $q = \alpha + (1-\alpha)p$. Now, the substitution

$$\bar{x}^{1-p} = u \iff \bar{x} = u^{1/(1-p)}$$

leads to the following root-finding problem

$$u^{-1}\left(1 + \frac{1}{\lambda}u^{\frac{q-p}{1-p}} - cu\right) = 0 \qquad \iff \qquad 1 + \frac{1}{\lambda}u^{\alpha} - cu = 0,$$

which can be solved analytically (at least) for $\alpha \in \{2, 3, 4\}$. Verification that f is relatively prox-regular is yet to be performed. Let $\bar{x} > 0$. We can choose $\varepsilon > 0$ such that the ε-ball around $\bar{x}$ lies in $\mathbb{R}_{>0}$. Then we find $r > 0$ sufficiently large such that for all $x \in \mathbb{R}$ with $|x - \bar{x}| < \varepsilon$ the second order derivative of $f + r\phi$ at x, given as $(f + r\phi)''(x) = (1/(p-1))x^{p-2} + r(1/(q-1))x^{q-2} \geq 0$, is nonnegative, which is asserted for

$$r \geq \frac{q-1}{1-p} \cdot \inf\left\{x^{(p-2)/(q-2)} : |x - \bar{x}| < \varepsilon\right\} > 0,$$

since $(q-1)/(1-p) > 0$. This implies that $f + r\phi$ is convex on the open ε-ball around $\bar{x}$ and therefore f is relatively prox-regular at $\bar{x}$. The case $\bar{x} < 0$ follows by symmetry. Now, we choose $\bar{x} = 0$ and fix $\bar{v} \in \partial f(0) = \mathbb{R}$. Since $\lim_{x \to 0, x \neq 0} |f'(x)| \to \infty$ we can find ε sufficiently small such that the graph of the ε-localization T of ∂f around $(\bar{x}, \bar{v})$ degenerates to

$$\operatorname{gph} T = \{(\bar{x}, v) : |v - \bar{v}| < \varepsilon\}.$$

Relative prox-regularity of f at $\bar{x} = 0$ for $\bar{v}$ is then asserted by verifying the subgradient inequality (2.57) for all $(x, v) \in \operatorname{gph} T$. Indeed, we can find $\varepsilon > 0$, such that for all such $|v - \bar{v}| < \varepsilon$ we have $f(x') \geq vx'$ for all $|x' - \bar{x}| < \varepsilon$, which shows that f is relatively prox-regular also at 0.

2.4. Beyond prox-regularity: Upper-$\mathcal{C}^1$ functions

In this section, we consider a class of functions that can be written (locally) in terms of a pointwise maximum/minimum of a *finite* collection of smooth functions. Such functions are differentiable almost everywhere. In addition this is a particular special case of subsmoothness, where the index set T is finite and therefore compact under the discrete topology. In the definition of subsmoothness, Definition 2.16, one distinguishes upper-$\mathcal{C}^1$ and lower-$\mathcal{C}^1$ functions depending on pointwise maximization/minimization. However, while pointwise maxima are in particular prox-regular [RW98, Proposition 13.33], the situation is different for pointwise minima: At points at which multiple functions are *active* in the sense that multiple graphs intersect, the resulting function is possibly not

Clarke regular and in particular not prox-regular and the proximal mapping is typically multivalued. A goal of this section therefore is to identify sufficient conditions under which the subgradient formula of the Moreau envelope holds in terms of the limiting subdifferential. In particular each element of the proximal mapping (if it is multivalued) shall give rise to a limiting subgradient of the Moreau envelope. We show that such a condition can be formulated in terms of the *linear independence constraint qualification* of the hypograph of the function.

As a convention and in accordance with Definition 2.16 we treat pointwise minima in terms of a negative pointwise maximum:

$$\min_{t\in T} f_t(x) = -\max_{t\in T} -f_t(x). \tag{2.68}$$

Here, the index set T is finite and $f_t : (O \subset \mathbb{R}^m) \to \mathbb{R}$ with O open is $\mathcal{C}^1$. According to [RW98, Exercise 10.27(c)] pointwise maxima are semidifferentiable, see [RW98, Definition 7.20]. According to [RW98, Theorem 10.31], the limiting subdifferential $\partial(-f)$ of a pointwise minimum or equivalently, a negative pointwise maximum, enjoys a one-sided inclusion:

$$-\partial(-f)(x) \subset \{\nabla f_t(x) : t \in T(x)\}, \qquad T(x) = \operatorname*{arg\,max}_{t\in T} f_t(x), \tag{2.69}$$

where $T(x)$ is the active set. Indeed, the inclusion is strict as the following simple example reveals:

Example 2.64. Let $f(x) = \max_{t\in T} f_t(x)$ with $f_1(x) = -(x+1)^2, f_2(x) = -(x-1)^2$, and $f_3(x) = -x^2 - 1$ and $T = \{1,2,3\}$. Then $T(0) = \{1,2,3\}$, and therefore $\{\nabla f_t(0) : t \in T(0)\} = \{1,-1,0\}$ whereas $-\partial(-f)(0) = \{\nabla f_1(0), \nabla f_2(0)\} = \{-1,1\}$. However, since f is locally Lipschitz, we have $\{\nabla f_t(0) : t \in T(0)\} \subset -\partial_C(-f)(0) = -\operatorname{con}\partial(-f)(0) = [-1,1]$, where $\partial_C f$ is the Clarke subdifferential.

If, in addition, the gradient normals of the pieces f_t are linearly independent, the inclusion in (2.69) holds with equality as shown in the next proposition. This provides a refinement of [RW98, Theorem 10.31] under a certain linear independence regularity condition:

Definition 2.65 (linear independence constraint qualification (LICQ))**.** *Let $g_t : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable for each $t \in T$ and T is finite. Let $\Omega := \{x \in \mathbb{R}^n : g_t(x) \le 0\}$. Then we say Ω satisfies the* linear independence constraint qualification (LICQ) *at $\bar{x} \in \Omega$ if the active gradients $\{\nabla g_t(\bar{x}) : t \in T(\bar{x})\}$ with $T(\bar{x}) = \{t \in T : g_t(\bar{x}) = 0\}$ are linearly independent.*

Proposition 2.66. *Let $f(x) = \max_{t\in T} f_t(x)$ and $f_t : \mathbb{R}^m \to \mathbb{R}$ be continuously differentiable. Let $\bar{x} \in \mathbb{R}^m$. Assume that* $\operatorname{epi} f = \{(x,y) : f_t(x) - y \le 0, \forall t \in T\}$ *satisfies the* LICQ *at $(\bar{x}, f(\bar{x}))$. Then the Inclusion* (2.69) *holds with equality:*

$$-\partial(-f)(x) = \{\nabla f_t(x) : t \in T(x)\}.$$

Proof. To show the desired result we construct for any $t \in T(\bar{x})$ a direction $v_t \in \mathbb{R}^m$ such that for any $\lambda > 0$ sufficiently small we have

$$T(\bar{x} + \lambda v_t) = \{t\}. \tag{2.70}$$

Then, due to the continuity of the individual pieces f_t, there is a sufficiently small neighborhood of $\bar{x} + \lambda v_t$ so that $T(\bar{x})$ is a singleton on this neighborhood and we can conclude that f is continuously differentiable at $\bar{x}+\lambda v_t$ with $\nabla f(\bar{x}+\lambda v_t) = \nabla f_t(\bar{x}+\lambda v_t)$, which proves that $\nabla f_t(\bar{x}) \in \partial f(\bar{x})$. Suppose such a neighborhood of $\bar{x}+\lambda v_t$ does not exist. Then there exist sequences $x^\nu \to \bar{x}+\lambda v_t$ and $t^\nu \in T(x^\nu)$ with $t^\nu \neq t$ and by going to a subsequence if necessary we can assume $t^\nu = t^*$ constant, such that $f(x^\nu) = f_{t^*}(x^\nu)$ and, due to Equality (2.70), $f(\bar{x}+\lambda v_t) = f_t(\bar{x}+\lambda v_t) < f_{t^*}(\bar{x}+\lambda v_t)$, which yields a contradiction since f is given as a pointwise minimum over finitely many continuous functions and therefore continuous.

To prove the condition (2.70), we define the matrix $U \in \mathbb{R}^{(m+1)\times|T(\bar{x})|}$ such that

$$U := \begin{pmatrix} | & & | \\ \nabla f_1(\bar{x}) & \dots & \nabla f_{|T(\bar{x})|}(\bar{x}) \\ | & & | \\ -1 & \dots & -1 \end{pmatrix}. \tag{2.71}$$

For each t, let $v_t \in \mathbb{R}^m$, $y_t \in \mathbb{R}$, $\alpha_t \in \mathbb{R}^{|T(\bar{x})|}$ such that $(v_t, y_t)^\top = U\alpha_t$. Now choose α_t as follows. Let $e_t \in \mathbb{R}^{|T(\bar{x})|}$ be the t^{th} unit vector. Since $\operatorname{epi} f$ satisfies the LICQ at $(\bar{x}, f(\bar{x}))$, the matrix U has full column-rank. Therefore the following linear system $U^\top U\alpha_t = -e_t$ has a unique solution $\alpha_t = (U^\top U)^{-1}e_t$. This implies that

$$-1 = \langle \nabla f_t(\bar{x}), v_t\rangle - y_t < \langle \nabla f_{t'}(\bar{x}), v_t\rangle - y_t = 0,$$

for all $t' \in T(\bar{x}) \setminus \{t\}$. For $\lambda > 0$ sufficiently small this means that

$$\langle \nabla f_t(\bar{x}), v_t\rangle + \frac{o(\lambda)}{\lambda} < \langle \nabla f_{t'}(\bar{x}), v_t\rangle + \frac{o(\lambda)}{\lambda} \tag{2.72}$$

and hence

$$f_t(\bar{x}+\lambda v_t) < f_{t'}(\bar{x}+\lambda v_t), \tag{2.73}$$

for all $t' \in T(\bar{x}) \setminus \{t\}$. Furthermore, for $\lambda > 0$ sufficiently small it holds that $T(\bar{x}+\lambda v_t) \subset T(\bar{x})$, due to the continuity of f. Thus, we verify that the active set $T(\bar{x}+v_t)$ is a singleton with $T(\bar{x}+\lambda v_t) = \{t\}$ for $\lambda \to 0^+$ as desired, and the conclusion follows. □

Indeed, the functions in Example 2.64 violate the LICQ of $\operatorname{epi} f$ at $(0,-1)$.

With this result at hand, we can prove that the gradient formula for the Moreau envelope of a piecewise convex function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, defined by

$$f(x) = \min_{t\in T} f_t(x),$$

with T finite and f_t proper, lsc and convex holds in terms of the limiting subdifferential. To this end note that $e_\lambda f(x) = \min_{t\in T} e_\lambda f_t(x)$ and $P_\lambda f(x) = \{P_\lambda f_t(x) : t \in \arg\min_{t\in T} e_\lambda f_t(x)\}$.

Corollary 2.67. *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, defined by*

$$f(x) = \min_{t\in T} f_t(x),$$

with T finite and f_t proper, lsc and convex. Let $\bar{x} \in \mathbb{R}^m$. Assume that $\operatorname{epi} -e_\lambda f$ *satisfies the LICQ at $(\bar{x}, e_\lambda f(\bar{x}))$. Then the following subgradient formula holds for the Moreau envelope $e_\lambda f$ of f:*

$$\partial e_\lambda f(\bar{x}) = \left\{\frac{1}{\lambda}(\bar{x} - P_\lambda f_t(\bar{x})) : t \in \arg\min_{t\in T} e_\lambda f_t(\bar{x})\right\} = \frac{1}{\lambda}(\bar{x} - P_\lambda f(\bar{x})). \tag{2.74}$$

Note that the functions in Example 2.64 are the negative Moreau envelopes $e_1 f_i$ of the functions $f_1 = \iota_{\{-1\}}, f_2 = \iota_{\{1\}}$ and $f_3 = \iota_{\{0\}} + 1$. For $f := \min_{t \in \{1,2,3\}} f_t = f_1 + f_2 + f_3$ it holds $P_1 f(0) = \{-1, 0, 1\}$, but not every element in $P_1 f(0)$ gives rise to a limiting subgradient of $e_1 f = \min_{t \in \{1,2,3\}} e_1 f_t$ at 0. This also shows that the inclusion in the limiting subgradient formula for the Moreau envelope [RW98, Theorem 10.31] is strict in general. An exception are indicator functions of closed sets where the inclusion always holds with equality [RW98, Example 8.53] even for the Bregman–Moreau envelope, [Bau+09, Theorem 5.4].

Chapter 3.

Generalized conjugate functions, proximal transform and a nonconvex proximal average

3.1. Generalized conjugate functions

This chapter surveys existing results from [RW98] on generalized conjugacy and the proximal transform and on the nonconvex proximal average due to [CWP20]. Leveraging the proximal transform we point out a duality relation between Proximal Point and gradient descent. This is based on [LOC21]. In the previous chapter we have considered lower envelopes which are obtained by inf-convolution, or, more generally, inf-projection wrt a discrepancy measure. Under suitable assumptions, these lower envelopes inherit the smoothness of the discrepancy measure, at least locally. In this chapter we consider a complementary approach to construct lower envelopes based on proximal hulls, or, more generally, generalized conjugacy. An important special case is the convex envelope of a function, i.e., the largest lsc convex function below the input function, which, in contrast to the Moreau envelope of a convex function, is nonsmooth in general. In convex conjugacy, an extended real-valued function is convex and lsc if and only if it can be written in terms of a pointwise supremum over a collection of affine functions. Given a convex lsc function, this collection of lower supporting affine functions constitutes the dual representation of the convex function. In addition, the slope of a lower supporting hyperplane at a point is called a subgradient of the convex function.

Likewise, in nonconvex conjugacy, we seek for a representation of a nonconvex function in terms of a pointwise supremum over a family of lower nonconvex, possibly concave, functions. Indeed, this is the key idea behind generalized conjugate functions [RW98, Chapter 11L*]:

Definition 3.1 (generalized conjugate functions)**.** *Let X and Y be nonempty sets. Let $\Phi : X \times Y \to \overline{\mathbb{R}}$ be any function. Let $f : X \to \overline{\mathbb{R}}$. Then the Φ-conjugate of f on Y at $y \in Y$ is defined by*

$$f^{\Phi}(y) := \sup_{x \in X} \Phi(x, y) \underset{\cdot}{-} f(x), \tag{3.1}$$

and the Φ-biconjugate of f back on X at $x \in X$ is given by

$$f^{\Phi\Phi}(x) := \sup_{y \in Y} \Phi(x, y) \underset{\cdot}{-} f^{\Phi}(y). \tag{3.2}$$

We say that f is a Φ-envelope on X if f can be written in terms of a pointwise supremum of a collection of elementary functions $x \mapsto \Phi(x, y) \underset{\cdot}{-} \alpha$, where $(\alpha, y) \in \overline{\mathbb{R}} \times Y$ is the *parameter element.*

Let $g : Y \to \overline{\mathbb{R}}$. Then, the Φ-conjugate of g on X at $x \in X$ is defined by

$$g^{\Phi}(x) := \sup_{y \in Y} \Phi(x, y) \underset{\cdot}{-} g(y), \tag{3.3}$$

and the Φ-biconjugate of g back on Y at $y \in Y$ is given by

$$g^{\Phi\Phi}(y) := \sup_{x \in X} \Phi(x, y) \underset{\cdot}{-} g^{\Phi}(x). \tag{3.4}$$

We say that g is a Φ-envelope on Y if g can be written in terms of a pointwise supremum of a collection of elementary functions $y \mapsto \Phi(x, y) \underset{\cdot}{-} \beta$, where $(\beta, x) \in \overline{\mathbb{R}} \times X$ is the parameter element.

If $X = Y$ and Φ is not symmetric we distinguish a left and a right Φ-conjugate of f.

Clearly, if $X = Y = \mathbb{R}^m$ and $\Phi(x, y) = \langle x, y \rangle$, one recovers the classical convex conjugate.

The following lemma is adapted from [RW98, Exercise 11.63] and connects lower envelopes to certain generalized biconjugates:

Lemma 3.2. *Let $\Phi : \mathbb{R}^m \times \mathbb{R}^n \to \overline{\mathbb{R}}$ be an extended real-valued function. Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$. Then, the left Φ-biconjugate $f^{\Phi\Phi}$ of f is the pointwise supremum of all elementary functions $x \mapsto \Phi(x, y) \underset{\cdot}{-} \beta$ that are majorized by f where $(\beta, y) \in \overline{\mathbb{R}} \times \mathbb{R}^n$ is the parameter element. In addition, if f is a left Φ-envelope, we have the identity $f^{\Phi\Phi} = f$.*

Likewise, for $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, the right Φ-biconjugate $g^{\Phi\Phi}$ is the pointwise supremum of all elementary functions $y \mapsto \Phi(x, y) \underset{\cdot}{-} \alpha$ that are majorized by g where $(\alpha, x) \in \overline{\mathbb{R}} \times \mathbb{R}^m$ is the parameter element. If, in addition, g is a right Φ-envelope, the identity $g^{\Phi\Phi} = g$ holds. In particular, we have for any function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ that $(f^{\Phi\Phi})^{\Phi} = f^{\Phi}$ and any function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ that $(g^{\Phi\Phi})^{\Phi} = g^{\Phi}$.

Proof. Consider the pointwise supremum over functions $q_{y,\beta} = \Phi(\cdot, y) \underset{\cdot}{-} \beta$ that are majorized by f. The property $\Phi(x, y) \underset{\cdot}{-} \beta = q_{y,\beta}(x) \le f(x)$ for all $x \in \mathbb{R}^m$ means in the notation of Φ-conjugates $\beta \ge f^{\Phi}(y)$. Therefore, we can actually take the pointwise supremum $\sup_{y \in \mathbb{R}^m} q_y(x)$ over functions $q_y(x) := \Phi(x, y) \underset{\cdot}{-} f^{\Phi}(y)$ and we get

$$\sup_{y \in \mathbb{R}^m} q_y(x) = f^{\Phi\Phi}(x),$$

as claimed. Let f be a left Φ-envelope. I.e., $f = \sup_{(\beta, y) \in W} q_{y,\beta}$, for some $W \subset \overline{\mathbb{R}} \times \mathbb{R}^n$. Since $f^{\Phi\Phi}$ is the pointwise supremum of all elementary functions $x \mapsto \Phi(x, y) \underset{\cdot}{-} \beta$ that are majorized by f we have $f^{\Phi\Phi} = f$. The claims for $g^{\Phi\Phi}$ and the right Φ-envelope are obtained analogously.

By definition, f^{Φ} is a right Φ-envelope and we obtain the claimed formula for f^{Φ}. Likewise, g^{Φ} is a left Φ-envelope and we obtain the claimed formula for g^{Φ}. □

In Chapter 5 we will consider lifting to measures in Lagrangian relaxations and dual discretizations. The approach will involve Φ-conjugates where the dimensions of X and Y do not agree: We will have $X \subset \mathbb{R}^m$ compact, nonempty and Y is the infinite-dimensional space of continuous functions on X. The coupling functional $\Phi(x, \lambda) = \int_X \lambda(x') \, \mathrm{d}\delta_x(x') = f(x)$ first lifts the input x to the Dirac measure δ_x centered at $x \in X$ and then couples the continuous function $\lambda \in \mathcal{C}(X)$ with δ_x via the appropriate dual pairing. We will later on discretize $\mathcal{C}(X)$ in terms of a finite-dimensional subspace $\Lambda \subset \mathcal{C}(X)$. Then the "discretized" Φ-biconjugate is the largest function below f which (up to constant translation) can be written in terms of a pointwise supremum of functions $\lambda \in \Lambda$.

Indeed, when $\Lambda = \{x \mapsto \langle x, y\rangle - r/2\|x\|^2 : y \in \mathbb{R}^m, r \in \mathbb{R}\}$, the coupling functional specializes to $\Phi(x,(y,r)) = \langle x, y\rangle - r/2\|x\|^2$ and one recovers the *basic quadratic transform* [RW98, Example 11.66] which identifies all proper lsc prox-bounded functions as Φ-envelopes.

3.2. Proximal transform

In what follows we restrict $x \mapsto \Phi(x,y) \underset{\cdot}{-} \alpha$ to a family of lower concave quadratics whose curvature parameter r is fixed and not adaptive as in the basic quadratic transform. In contrast to the basic quadratic transform such a family of functions is not a subspace since scaling is not permitted. This yields the proximal transform [RW98, Example 11.64] which sets up a connection between Moreau envelopes and generalized conjugate functions:

Definition 3.3 (proximal transform)**.** *For fixed $\lambda > 0$ pair $X = \mathbb{R}^m = Y$ with itself:*

$$\Phi(x,y) = -\frac{1}{2\lambda}\|x-y\|^2.$$

Then for any function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ one has

$$f^\Phi = -e_\lambda f, \qquad f^{\Phi\Phi} = -e_\lambda(-e_\lambda f).$$

We adopt the definition from [RW98, Example 1.44]:

Definition 3.4 (proximal hull)**.** *For a function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ and $\lambda > 0$ the λ-proximal hull of f is the function $h_\lambda : \mathbb{R}^m \to \overline{\mathbb{R}}$ defined as the pointwise supremum of the collection of all the quadratic functions of the elementary form*

$$x \mapsto \alpha - \frac{1}{2\lambda}\|x-y\|^2,$$

where (α, y) is the parameter element, that are majorized by f. We say f is λ-proximal if $h_\lambda f$ agrees with f everywhere.

In view of Lemma 3.2 the Φ-biconjugate $-e_\lambda(-e_\lambda f)$ is the largest function below f that can be written as a pointwise supremum over concave quadratics with fixed curvature $1/\lambda$, i.e., $-e_\lambda(-e_\lambda f)$ the proximal hull $h_\lambda f$ of f:

$$-e_\lambda(-e_\lambda f) = h_\lambda f.$$

In the language of proximal conjugacy the Φ-envelopes are precisely the λ-proximal functions. The next lemma identifies all lsc $1/\lambda$-hypoconvex functions as Φ-envelopes, in the same way all lsc convex functions are convex envelopes:

Lemma 3.5. *Let $\lambda > 0$. For any function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ we have*

$$-e_\lambda f = \left(f + \frac{1}{2\lambda}\|\cdot\|^2\right)^* \circ \lambda I - \frac{1}{2\lambda}\|\cdot\|^2, \qquad h_\lambda f = \left(f + \frac{1}{2\lambda}\|\cdot\|^2\right)^{**} - \frac{1}{2\lambda}\|\cdot\|^2,$$

and therefore $-e_\lambda f$ is λ-proximal. If, in addition, f is proper, $h_\lambda f = f$ if and only if f

is lsc and

$$f + \frac{1}{2\lambda}\|\cdot\|^2,$$

is convex.

Proof. See [RW98, Example 11.26]. □

In the same way the lower supporting affine functions in convex conjugacy correspond to the subgradients of a convex function, in proximal conjugacy, the lower supporting quadratics represent proximal subgradients: In particular we have for $\bar{x} \in P_\lambda f(y)$ with $y \in \mathbb{R}^m$ that:

$$f(x) \geq f(\bar{x}) + \frac{1}{2\lambda}\|\bar{x} - y\|^2 - \frac{1}{2\lambda}\|x - y\|^2,$$

for all $x \in \mathbb{R}^m$. This shows that the parametrized map $x \mapsto f(\bar{x}) + \frac{1}{2\lambda}\|\bar{x} - y\|^2 - \frac{1}{2\lambda}\|x - y\|^2$ is below f and therefore contained in the family of functions in definition of the proximal hull. Since for $y = \bar{x}$ equality holds at $x = \bar{x}$ we have that $h_\lambda f(\bar{x}) = f(\bar{x})$ and therefore $h_\lambda f$ agrees with f on the range of $P_\lambda f$.

3.3. A nonconvex proximal average

In this section we discuss an interesting application of proximal conjugacy, the so-called *proximal average* [BMR04; BLT08; Bau+08]. Given a collection of convex input functions $f_1, f_2, \ldots, f_N$ and weights $\pi \in \Pi$ with

$$\Pi := \left\{ \pi \in \mathbb{R}^N : \sum_{i=1}^N \pi_i = 1, \pi_i \geq 0 \right\}, \tag{3.5}$$

the proximal average is a recipe for averaging functions and compares favorably to the pointwise arithmetic average $\sum_{i=1}^N \pi_i f_i$ or the epigraphical average $\pi_1 \star f_1 \oplus \pi_2 \star f_2 \oplus \cdots \oplus \pi_N \star f_N$ in some ways [Bau+08]: Desirable properties are its ability to continuously (in an epigraphical sense) transform one function into another [BLT08, Theorem 5.4] and a simple expression of its proximal mapping in terms of the pointwise arithmetic average of its proximal mappings [Bau+08, Theorem 6.7]. [Har09; HP14] are the first to extend the proximal average to the nonconvex setting considering a slightly different construction based on Lasry–Lions envelopes. The construction enjoys powerful stability properties. However, it doesn't recover the proximal average for convex functions as a special case. More recently, as a remedy, [Yu+15; CWP20] consider a proximal average construction based on proximal hulls, which can be seen as a limiting case of the Lasry–Lions envelope construction. The formulation, indeed, strictly generalizes the convex proximal average. Like the convex proximal average it enjoys a certain homotopy property and continuously transforms one proximal hull into another [CWP20, Corollary 6.9]. In addition, for λ-proximal functions, one recovers the pointwise average expression for the proximal mapping [CWP20, Theorem 5.4].

In this section we collect results from related works [Yu+15; CWP20]. In particular, we extend some results for the proximal average for $N = 2$ functions due to [CWP20], to an arbitrary number of functions. This is straightforward using the arguments developed by [CWP20]. For completeness, we will adapt their proofs to the case $N \geq 2$.

We consider the proximal average for $N \geq 2$ functions [Yu+15, Definition 1]:

$$\mathcal{A}_\lambda(f, \pi) = -e_\lambda \left(\sum_{i=1}^{N} -\pi_i e_\lambda f_i \right). \tag{3.6}$$

We collect some key properties of the proximal average from [CWP20], extended from $N = 2$ to $N \geq 2$.

Lemma 3.6. *Let $f_i : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded with threshold λ_{f_i}. Let $\lambda < \min_{1 \leq i \leq N} \lambda_{f_i}$. Then we have*

$$e_\lambda \mathcal{A}_\lambda(f, \pi) = \sum_{i=1}^{N} \pi_i e_\lambda f_i$$

and for the associated proximal mapping for $y \in \mathbb{R}^m$:

$$P_\lambda \mathcal{A}_\lambda(f, \pi)(y) = \sum_{i=1}^{N} \pi_i \operatorname{con} P_\lambda f_i(y).$$

If, in addition, f_i are λ-proximal we have $P_\lambda \mathcal{A}_\lambda(f, \pi)(y) = \sum_{i=1}^{N} \pi_i P_\lambda f_i(y)$.

Proof. For $N = 2$ this is [CWP20, Theorem 5.1(a)] and [CWP20, Theorem 5.4]. Adapting the proof for $N > 2$ observe that due to Lemma 3.5 $\sum_{i=1}^{N} -\pi_i e_\lambda f_i$ is λ-proximal. Invoking the second part of Lemma 3.5 we have

$$e_\lambda \mathcal{A}_\lambda(f, \pi) = -h_\lambda \left(\sum_{i=1}^{N} -\pi_i e_\lambda f_i \right) = \sum_{i=1}^{N} \pi_i e_\lambda f_i.$$

Since $-e_\lambda f_i$ is regular and locally Lipschitz, the sum rule Lemma 1.2 yields

$$\partial(-e_\lambda \mathcal{A}_\lambda(f, \pi))(y) = \sum_{i=1}^{N} \pi_i \partial(-e_\lambda f_i)$$

and therefore using [RW98, Example 10.32]

$$\frac{1}{\lambda}(\operatorname{con} P_\lambda \mathcal{A}_\lambda(f, \pi)(y) - y) = \frac{1}{\lambda} \sum_{i=1}^{N} \pi_i (\operatorname{con} P_\lambda f_i(y) - y).$$

Since $\mathcal{A}_\lambda(f, \pi)$ is λ-proximal using [CWP20, Proposition 2.6] we know that $P_\lambda \mathcal{A}_\lambda(f, \pi)(y)$ is convex and therefore $P_\lambda \mathcal{A}_\lambda(f, \pi)(y) = \operatorname{con} P_\lambda \mathcal{A}_\lambda(f, \pi)(y)$. Reordering then yields

$$P_\lambda \mathcal{A}_\lambda(f, \pi)(y) = \sum_{i=1}^{N} \pi_i \operatorname{con} P_\lambda f_i(y).$$

If, in addition, f_i are λ-proximal, again using [CWP20, Proposition 2.6], we have $\operatorname{con} P_\lambda f_i(y) = P_\lambda f_i(y)$ and therefore $P_\lambda \mathcal{A}_\lambda(f, \pi)(y) = \sum_{i=1}^{N} \pi_i P_\lambda f_i(y)$. □

Proposition 3.7. *Let $f_i : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded with threshold λ_{f_i}. Let $\lambda < \min_{1 \leq i \leq N} \lambda_{f_i}$.*

(i) *$\mathcal{A}_\lambda(f,\pi)$ is proper, lsc. If, in addition, each f_i is r_i-hypoconvex then $\mathcal{A}_\lambda(f,\pi)$ is hypoconvex with parameter $r = \max_{1\leq i\leq N} r_i$.*

(ii) *The following inequalities hold true:*

$$e_\lambda\, \mathcal{A}_\lambda(f,\pi) \leq \mathcal{A}_\lambda(f,\pi) \leq \sum_{i=1}^N \pi_i h_\lambda f_i \leq \sum_{i=1}^N \pi_i f_i.$$

(iii) *If $\bigcap_{i=1}^N \operatorname{arg\,min} f_i \neq \emptyset$ and $\pi \in \operatorname{relint}\Pi$ we have $\operatorname{arg\,min} \sum_{i=1}^N \pi_i f_i = \operatorname{arg\,min} \mathcal{A}_\lambda(f,\pi)$ and $\min \sum_{i=1}^N \pi_i f_i = \min \mathcal{A}_\lambda(f,\pi)$.*

(iv) *If any function f_i is continuously differentiable with Lipschitz continuous gradient and λ-proximal, then for every $\pi \in \operatorname{relint}\Pi$ the proximal average $\mathcal{A}_\lambda(f,\pi)$ is continuously differentiable with Lipschitz continuous gradient as well.*

Proof. (i) Expanding the square we obtain a well-known expression of the negative Moreau envelope in terms of the convex conjugate. See, e.g., [RW98, Example 11.26]:

$$\begin{aligned}
-e_\lambda f(y) &= \sup_{x\in\mathbb{R}^m} -\frac{1}{2\lambda}\|x-y\|^2 - f(x) \\
&= \frac{1}{\lambda}\left(\sup_{x\in\mathbb{R}^m} \langle x,y\rangle - (1/2)\|x\|^2 - \lambda f(x)\right) - \frac{1}{2\lambda}\|y\|^2 \\
&= \frac{1}{\lambda}\big(\lambda f + (1/2)\|\cdot\|^2\big)^*(y) - \frac{1}{2\lambda}\|y\|^2. \qquad (3.7)
\end{aligned}$$

Then we have:

$$\sum_{i=1}^N -\pi_i e_\lambda f_i = \frac{1}{\lambda}\sum_{i=1}^N \pi_i\big(\lambda f_i + (1/2)\|\cdot\|^2\big)^* - \frac{1}{2\lambda}\|\cdot\|^2,$$

where $(\lambda f_i + (1/2)\|\cdot\|^2)^*$ are $1/(1-r_i\lambda)$-Lipschitz differentiable and therefore, in view of [RW98, Proposition 12.60]

$$\frac{1}{\lambda}\left(\sum_{i=1}^N \pi_i\big(\lambda f_i + (1/2)\|\cdot\|^2\big)^*\right)^*$$

is $(1/\lambda - r)$-strongly convex. In view of the definition of the proximal average and expression (3.7) we have

$$\begin{aligned}
\mathcal{A}_\lambda(f,\pi) + \frac{r}{2}\|\cdot\|^2 &= -e_\lambda\left(\sum_{i=1}^N -\pi_i e_\lambda f_i\right) + \frac{r}{2}\|\cdot\|^2 \\
&= \frac{1}{\lambda}\left(\sum_{i=1}^N \pi_i\big(\lambda f_i + (1/2)\|\cdot\|^2\big)^*\right)^* - \frac{1/\lambda - r}{2}\|\cdot\|^2,
\end{aligned}$$

is convex.

(ii) For $N = 2$ this is [CWP20, Theorem 6.4(a)]. For $N > 2$ we adapt the arguments as follows: We have

$$\begin{aligned} e_\lambda \sum_{i=1}^N \pi_i g_i(y) &= \inf_{x\in\mathbb{R}^m} \sum_{i=1}^N \pi_i g_i(x) + \frac{1}{2\lambda}\|x-y\|^2 \\ &\geq \inf_{x_i\in\mathbb{R}^m} \sum_{i=1}^N \pi_i g_i(x_i) + \frac{1}{2\lambda}\|x_i-y\|^2 = \sum_{i=1}^N \pi_i e_\lambda g_i(y). \end{aligned}$$

Thus we have for $g_i := -e_\lambda f_i$:

$$\mathcal{A}_\lambda(f,\pi) = -e_\lambda\left(\sum_{i=1}^N -\pi_i e_\lambda f_i\right) \leq \sum_{i=1}^N -\pi_i e_\lambda(-e_\lambda f_i) = \sum_{i=1}^N \pi_i h_\lambda f_i.$$

The first inequality follows by definition of the Moreau envelope and the last inequality due to the definition of the proximal hull, see Definition 3.4.

(iii) For $N=2$ this is [CWP20, Theorem 7.2]. For $N>2$ observe that for any $x \in \bigcap_{i=1}^N \arg\min f_i \neq \emptyset$ we have $\mathcal{A}_\lambda(f,\pi)(x) \leq \sum_{i=1}^N \pi_i f_i(x) = \sum_{i=1}^N \pi_i e_\lambda f_i(x) = e_\lambda\,\mathcal{A}_\lambda(f,\pi)(x) \leq \mathcal{A}_\lambda(f,\pi)(x)$ and therefore $\mathcal{A}_\lambda(f,\pi)(x) = \sum_{i=1}^N \pi_i f_i(x) = \min\sum_{i=1}^N \pi_i f_i = \min\sum_{i=1}^N \pi_i e_\lambda f_i \leq \min\mathcal{A}_\lambda(f,\pi)$. This implies $\min\sum_{i=1}^N \pi_i f_i = \min\mathcal{A}_\lambda(f,\pi)$.

It is easy to verify that $\arg\min\sum_{i=1}^N \pi_i f_i = \bigcap_{i=1}^N \arg\min f_i$ and $\arg\min\sum_{i=1}^N \pi_i e_\lambda f_i = \bigcap_{i=1}^N \arg\min e_\lambda f_i$. Since $\arg\min g = \arg\min e_\lambda g$ we have

$$\arg\min\sum_{i=1}^N \pi_i f_i = \arg\min\sum_{i=1}^N \pi_i e_\lambda f_i,$$

and

$$\arg\min\mathcal{A}_\lambda(f,\pi) = \arg\min e_\lambda\,\mathcal{A}_\lambda(f,\pi) = \arg\min\sum_{i=1}^N \pi_i e_\lambda f_i.$$

Overall we have $\arg\min\sum_{i=1}^N \pi_i f_i = \arg\min\mathcal{A}_\lambda(f,\pi)$ as claimed.

(iv) For $N=2$ this is [CWP20, Proposition 8.8]. For $N>2$ we adapt the arguments as follows: Define

$$j_\lambda = \frac{1}{2\lambda}\|\cdot\|^2.$$

Let f_1 be continuously differentiable with Lipschitz continuous gradient and λ-proximal. Then $f_1 + j_\lambda$ is proper convex lsc and Lipschitz differentiable. In view of [RW98, Proposition 12.60] $(f_1+j_\lambda)^*$ is strongly convex. We have

$$\begin{aligned} \mathcal{A}_\lambda(f,\pi) &= -e_\lambda\left(\sum_{i=1}^N -\pi_i e_\lambda f_i\right) && (3.8) \\ &= \left(\sum_{i=1}^N \pi_i(f_i+j_\lambda)^*\right)^* - j_\lambda. && (3.9) \end{aligned}$$

Thus we know that $\sum_{i=1}^N \pi_i(f_i+j_\lambda)^*$ is strongly convex and therefore $(\sum_{i=1}^N \pi_i(f_i+j_\lambda)^*)^*$ is Lipschitz differentiable which concludes the proof. □

In Section 4.3.2 we will consider variance bounds for the proximal average: In particular, Lemma 4.14 shows that in the smooth case the gradient of the arithmetic average can be bounded by the gradient of the proximal average under bounded variance of the gradients of the individual functions.

3.4. On the duality between gradient descent and Proximal Point

We conclude this chapter with another application of the proximal transform: It is well known that Proximal Point with sufficiently small λ on a hypoconvex function is equivalent to gradient descent on its Moreau envelope. Proximal conjugacy reveals that also the converse is true: Gradient descent on a function with Lipschitz gradient is equivalent to Proximal Point applied to a certain possibly nonsmooth function obtained via the proximal transform:

Theorem 3.8. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be L-smooth. Let $\lambda \leq 1/L$. Define $d_\lambda := -e_\lambda(-f)$ to be the* inf*-deconvolution of f. Then, d_λ is hypoconvex with some $r < 1/\lambda$ and the following relation holds:*

$$f = e_\lambda d_\lambda. \tag{3.10}$$

In addition we have for all $x \in \mathbb{R}^m$:

$$x - \lambda \nabla f(x) = P_\lambda d_\lambda(x). \tag{3.11}$$

Proof. Let f be L-smooth. It is well known that L-Lipschitz continuity of ∇f implies that

$$\frac{L}{2}\|\cdot\|^2 - f,$$

is convex, and thus, $-f$ is L-hypoconvex. Let $\lambda \leq 1/L$ and consider

$$d_\lambda = -e_\lambda(-f) = \left(\frac{1}{2}\|\cdot\|^2 - \lambda f\right)^* - \frac{1}{2\lambda}\|\cdot\|^2.$$

Using Lemma 3.5 we have $e_\lambda d_\lambda = e_\lambda(-e_\lambda(-f)) = f$. Rewriting the above expression for d_λ we have

$$d_\lambda + \frac{1}{2\lambda}\|\cdot\|^2 = \left(\frac{1}{2}\|\cdot\|^2 - \lambda f\right)^*,$$

where $(1/2)\|\cdot\|^2 - \lambda f$ is convex, proper lsc and differentiable with Lipschitz continuous gradient. Invoking [RW98, Proposition 12.60] we have that $d_\lambda + (1/(2\lambda))\|\cdot\|^2$ is strongly convex and therefore d_λ is hypoconvex with some $r < 1/\lambda$. In particular this means that d_λ is prox-bounded with some threshold $\lambda_f > \lambda$.

Invoking [RW98, Example 10.32] we obtain

$$-\nabla f(x) = \partial(-e_\lambda d_\lambda)(x) = \frac{1}{\lambda}(\operatorname{con} P_\lambda d_\lambda(x) - x),$$

and therefore a gradient step $x - \lambda \nabla f(x) = P_\lambda d_\lambda(x)$ is a Proximal Point step on the deconvolution. □

In the context of the minimization of a finite sum $\sum_{i=1}^N \pi_i f_i$, with $\pi \in \Pi$, for f_i being $\mathcal{C}^1$ with L-Lipschitz gradient the above result reveals a connection between deconvolutions and the proximal average: The proximal average of the deconvolutions $g := (d_\lambda f_i)_{i=1}^N$ of

f_i with weights π is

$$\mathcal{A}_\lambda(g,\pi) = -e_\lambda\left(\sum_{i=1}^N -\pi_i e_\lambda d_\lambda f_i\right) = -e_\lambda\left(\sum_{i=1}^N -\pi_i f_i\right) = d_\lambda \sum_{i=1}^N \pi_i f_i,$$

i.e., the deconvolution of the finite sum and therefore the Moreau envelope of the proximal average of g is

$$e_\lambda\,\mathcal{A}_\lambda(g,\pi) = \sum_{i=1}^N \pi_i f_i.$$

As a consequence, a proximal point step with parameter λ wrt $\mathcal{A}_\lambda(g,\pi)$ is a gradient step with step-size λ on $\sum_{i=1}^N \pi_i f_i$. Using this reformulation, gradient descent on a finite sum can be interpreted in terms of an averaged Proximal Point iteration applied to the deconvolutions g. Introducing identical copies of the variable for each summand, averaged Proximal Point takes the form of a block coordinate descent minimization applied to a penalty objective. Averaged Proximal Point aka block coordinate descent can be implemented in a stochastic fashion such that the objective decreases surely, as considered in the next chapter: The key idea is to update a random sample of blocks in each iteration, while the other blocks are left unchanged. Therefore, the aforementioned equivalence between gradient descent and block coordinate descent can inform the design of stochastic formulations of gradient descent in which a sure descent of the penalty objective function is guaranteed: In Section 4.3 we consider a stochastic averaged proximal point method. The method specializes to the Finito/MISO algorithm [DDC14; Mai15] when applied to the deconvolutions $d_\lambda f_i = -e_\lambda(-f_i)$.

Chapter 4.

Alternating inexact Proximal Point with applications to weakly supervised and federated learning

4.1. Overview

In this chapter we consider alternating Proximal Point and discuss the applications of weakly supervised and federated learning. As we will explain in the course of this chapter these algorithms are well-suited to solve such a class of problems. This chapter is based on [LWC19], [LOC20] and [LOC21]: In particular, Algorithm 1 and Theorem 4.1 are adopted from [LOC20]. The stochastic averaged proximal point method and its application to federated learning in Section 4.3 and in particular Algorithm 6 were considered in a very simplified setup under the anisotropic prox geometry and without an analysis in [LWC19]. Otherwise the results in this chapter are based on [LOC21].

The alternating Proximal Point typically solves a certain relaxation to the original problem. However, leveraging prox-regularity and the gradient formula of the Moreau envelope we will be able to show that the relaxation is useful: In federated learning, for instance, we invoke prox-regularity to show that under bounded variance of the gradients of the Moreau envelopes of the individual risks a stationary point wrt the relaxation is near stationary wrt the original problem and approximate consensus between the clients is attained in the limit, see Corollary 4.20.

4.2. Alternating inexact Bregman Proximal Point with application to weakly supervised learning

4.2.1. On the duality between induction and transduction in optimization for machine learning

First we consider alternating Bregman Proximal Point in a general nonconvex setting under relative prox-regularity. Alternating Bregman minimization was considered before in a convex setting [BCN06]. We are interested in the following coupled optimization problem:

$$\text{minimize}\,\{f(x) + D_\phi(x,y) + g(y) : (x,y) \in \operatorname{dom}\phi \times \operatorname{int}(\operatorname{dom}\phi)\}, \tag{4.1}$$

where $\phi \in \Gamma_0(\mathbb{R}^m)$ is a Legendre function and $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ are extended real-valued functions. The Bregman distance $D_\phi(x,y) \geq 0$ penalizes the discrepancy between x and y and attains the value 0 if and only if $x = y$. Therefore the problem can be seen as a relaxation to the problem

$$\text{minimize}\left\{f(x) + \iota_{\{0\}}(x-y) + g(y) : (x,y) \in \operatorname{dom}\phi \times \operatorname{int}(\operatorname{dom}\phi)\right\}. \tag{4.2}$$

In what follows we elaborate on duality relations for this class of problems: For instance, it was shown in [BCN06] that if, among other properties, D_ϕ is jointly convex, and f, g are convex, a classical convex duality relation holds between the primal problem (4.1) and a Fenchel–Rockafellar dual problem

$$\max_{x^*,y^*\in\mathbb{R}^m} -f^*(x^*) - g^*(y^*) - D_\phi^*(-x^*, -y^*).$$

In the nonconvex setting there still holds a certain double-min duality relation depending on the order of minimization: In some situations a certain hierarchy is imposed on the variables x and y resp. the minimization wrt x or y: Suppose that either we are interested in the solution y, and x is merely an auxiliary or hidden variable in the optimization or, alternatively, x and y have their roles flipped. Then we are interested in the cost function dependent on y (or x) that is obtained when x (or y) is eliminated by inf-projection. This hierarchy of minimization tasks can be expressed in terms of a certain bilevel optimization problem which distinguishes an upper and a lower level minimization problem: We have

$$\begin{aligned} \underset{y\in\mathbb{R}^m}{\text{minimize}} \quad & g(y) + D_\phi(x(y), y) + f(x(y)) \\ \text{subject to} \quad & x(y) = \operatorname*{arg\,min}_{x\in\mathbb{R}^m} f(x) + D_\phi(x,y), \end{aligned} \tag{4.3}$$

if x is the hidden variable or

$$\begin{aligned} \underset{x\in\mathbb{R}^m}{\text{minimize}} \quad & g(y(x)) + D_\phi(x, y(x)) + f(x) \\ \text{subject to} \quad & y(x) = \operatorname*{arg\,min}_{y\in\mathbb{R}^m} g(y) + D_\phi(x,y), \end{aligned} \tag{4.4}$$

if the hierarchy is flipped. The parametric lower level minimization problems $x(y)$ and $y(x)$ are merely the left and right Bregman proximal operators, see Definition 2.18 and Definition 2.24, studied in Section 2.3. Therefore we can equivalently write the bilevel problems in terms of left and right Bregman–Moreau regularized optimization problems: The left Bregman relaxation reads for $\lambda = 1$:

$$\text{minimize}\left\{\overleftarrow{\operatorname{env}}_\lambda^\phi f(y) + g(y) : y \in \mathbb{R}^m\right\}, \tag{4.5}$$

and, accordingly, the right Bregman relaxation to the coupled problem is:

$$\text{minimize}\left\{\overrightarrow{\operatorname{env}}_\lambda^\phi g(x) + f(x) : x \in \mathbb{R}^m\right\}. \tag{4.6}$$

We point out an interesting connection to double-min duality [RW98, Chapter 11L*], where the duality relations take the form of min = min instead of min = max: In the terminology of duality we can associate the left relaxation with a certain primal and the right relaxation with a certain dual problem. Indeed, in view of [RW98, Proposition 1.35] these problems are actually equivalent to the joint problem (4.1). However, in a nonconvex setting, if one applies inexact Gauss–Seidel minimization to the joint problem (4.1), one can only expect to find a stationary point of Problem (4.1). In general, such a stationary point does not translate to a stationary point of the bilevel or Bregman–Moreau regularized problem, where the minimization of the lower-level problem is understood to be solved to global optimality. A sufficient condition for such a translation is relative prox-regularity, studied in Section 2.3, which we will show in Section 4.2.2.

In the entropic/*Kullback–Leibler* (KL) divergence setting, the hierarchy of minimization in model 4.1 is connected to a double-min duality between *induction* and *transduction* in machine learning. Such a connection was observed previously by [Lau+18], who consider a nonsmooth setting with a hinge loss in place of a KL divergence. The formulations find applications in modern weakly and self-supervised learning, parameter learning and transductive learning: Here, the question of whether one solves a weakly supervised (aka inductive) or a transductive learning problem is induced by the hierarchy of upper and lower level problem in Equation 4.1.

To this end let

$$\phi(x) := \sum_{i=1}^{m} x_i \log(x_i) - x_i, \tag{4.7}$$

with $0\log(0) := 0$ be the Boltzmann–Shannon entropy. Then $\operatorname{dom}\phi = \{x \in \mathbb{R}^m : x_i \geq 0\}$ is the nonnegative orthant and D_ϕ is the Kullback–Leibler divergence:

$$D_\phi(x,y) = \mathrm{KL}(x,y) = \begin{cases} \sum_{i=1}^{m} x_i \log(x_i/y_i) - x_i + y_i & \text{if } 0 \leq x_i \leq 1 \text{ and } 0 < y_i < 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Suppose that $y = \mathcal{H}_w(s)$ are the predicted labels in probability vector representation $y_j \in \operatorname{relint}(\Pi)$ for $\Pi = \{\pi \in \mathbb{R}^d : \pi_k \geq 0, \sum_{k=1}^{d} \pi_k = 1\}$ of a finite sequence $s \in (\mathbb{R}^J)^M$ of M training inputs $s_j \in \mathbb{R}^J$. Here, $\mathcal{H}_w$ is a prediction aka hypothesis function parametrized by weights $w \in \mathbb{R}^n$. Note that the assumption $(\mathcal{H}_w(s))_j \in \operatorname{relint}(\Pi)$ is mild, since this can always be enforced by concatenating the output $(\mathcal{H}_w(s))_j$ with a softmax function whose k^{th} compenent is defined by

$$\operatorname{softmax}(y)_k = \frac{\exp(y_k)}{\sum_{k=1}^{d} \exp(y_k)}.$$

The softmax function is the gradient of the logexp-function $\operatorname{logexp}(x) = \log(\sum_{k=1}^{d} \exp(x_k))$ which yields a smooth approximation to the vecmax-function given by $\operatorname{vecmax}(x) = \max\{x_1, x_2, \ldots x_d\}$ [RW98, Example 1.30]. The vector $x \in (\mathbb{R}^d)^M$ is a variable representing the labels corresponding to s. In classical supervised learning the labels t_j are known and therefore the variable x is constrained to coincide with the known sequence of labels t. Then, supervised learning can be written in terms of Problem (4.1) adopting the following choices for f and g: $f(x) = \iota_{\{t\}}(x)$ constrains x to be fixed and coincide with the known labels t in *one-hot* (unit vector) representation and

$$g(y) := \inf\{R(w) : w \in \mathbb{R}^n, \mathcal{H}(w;s) = y\}, \tag{4.8}$$

is the image function of R wrt the nonlinear mapping $w \mapsto \mathcal{H}(w;s)$. It constrains the variable $y = \mathcal{H}_w(s)$ to lie in the range of $\mathcal{H}(\cdot;s)$ and R is a regularizer on the weights w. Substituting these specializations in Problem (4.1) we obtain:

$$\operatorname*{arg\,min}_{w\in\mathbb{R}^n}\ \mathrm{KL}(t, \mathcal{H}(w;s)) + R(w) = \operatorname*{arg\,min}_{w\in\mathbb{R}^n}\ \mathcal{L}(t, \mathcal{H}(w;s)) + R(w),$$

where the KL-divergence reduces to the cross-entropy loss $\mathcal{L}$.

More generally, in weakly and self-supervised learning, the labels t are (partially) unknown. In that case f represents a certain prior on the labels. For instance, in image segmentation f can be a Potts and/or normalized cut [SM00] energy, that favors segmentations that are spatially smooth and/or balanced [Tan+18]. Alternatively, f can model a convex relaxation of a hard balancing constraint that ensures that the assignment of inputs to labels is balanced [PKD15] in the sense that the cardinality of the preimage of each label is bounded from below by some positive constant. In the unsupervised case we choose $f(x) = \sum_{j=1}^{M} \iota_{\Pi}(x_j)$. Then Gauss–Seidel minimization of Problem (4.1) specializes to the *expectation maximization* (EM) algorithm, while for nontrivial choices of f one obtains a certain EM algorithm with posterior regularization [Gan+10]. Since the labels $x_j \in \Pi$ are soft probability vectors rather than one-hot vectors, the prior f, typically, is a convex function or even a linear function [PKD15] over the probability simplex. However, there also exist nonconvex priors f [Tan+18] that take the form of a Rayleigh quotient. Typically, in weakly supervised learning and expectation maximization x is merely a hidden variable and one seeks to find the hypothesis $\mathcal{H}_w$ parametrized by $w \in \mathbb{R}^n$. One therefore minimizes the left Bregman relaxation (4.3) which in that case amounts to

$$\min_{w \in \mathbb{R}^n} \overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f(\mathcal{H}(w; s)) + R(w).$$

This also reveals that the left Bregman relaxation in our case corresponds to the training of the hypothesis $\mathcal{H}_w$ with loss function $\overleftarrow{\mathrm{env}}_{\lambda}^{\phi} f$ and regularizer R.

Now we consider the right Bregman relaxation (4.6) where the minimization in x is the upper level problem. This viewpoint is certainly less common but can be connected to transductive learning [GVV98]: In (weakly) supervised and unsupervised learning (aka inductive learning), one distinguishes a training and a inference phase: At training one is interested in the estimation of a hypothesis $\mathcal{H}_w$, where both, the labeled and/or the unlabeled inputs, are considered training data. The estimated labels of the unlabeled inputs are merely auxiliary variables which are discarded after training. At inference time, the weights w are kept fixed and one applies the hypothesis $\mathcal{H}_w$ to infer the unknown labels of certain test or production time inputs. In contrast, in transductive learning, training and inference is combined: Given a combined set of labeled training and unlabeled production time test inputs one directly infers the labels x of the test inputs. Therefore transductive learning can potentially adapt to the distribution of the production time test data, while inductive learning ignores the information hidden in the test data. In the framework of alternating Bregman minimization transductive learning corresponds to the minimization of the right bilevel problem (4.4). Here, the hypothesis parametrized by w is rather an auxiliary variable. One therefore obtains the right Bergman relaxation (4.6) which corresponds to the dual problem in the terminology of double-min duality. Since transductive learning involves the solution of a potentially expensive optimization problem at inference time, it is more costly at production time than the inductive approach which typically amounts to a simple problem.

However, considering the fact that in supervised learning we aim to minimize the risk, rather than the empirical risk, actually the opposite is true: The training phase in (weakly) supervised learning attempts to solve a more difficult (typically intractable) problem (the minimization of the risk) which is merely used as an intermediate step in a possibly simpler problem, the inference of the labels of some given test data: This is precisely the philosophy in transduction: According to Vapnik, “when solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one" [Vap06] or, more philosophically, "we shall reach the conclusion that Socrates is mortal with a greater approach to certainty if we make our argument purely inductive than if we go by way of 'all men are mortal' and then use deduction" [Rus12, Chapter VII][1].

We highlight that there exists another interesting duality relation between left and right Bregman relaxation in the entropic/EM setting considering the notion of exponential families to parametrize probability distributions, see, e.g. [KKS21]: As we have seen in Lemma 2.26 the right Bregman prox can be transformed into a left Bregman prox (and vice versa) using Lemma 1.20(ii). In the context of exponential families and information geometry the log and exp corresponding to $\nabla\phi$ and $\nabla\phi^*$ precisely constitute a bijection between the *mean* and the *natural* parameters of a probability distribution. In our optimization problem, from an exponential family point of view, we resort to the categorial distribution, where the softmax in the hypothesis function retracts the mean parameters back to the natural parameters.

In the next two subsections we consider two specializations of g: Firstly, in Section 4.2.2, we choose g to be an image function as in Equation (4.8) and secondly, in Section 4.2.3, we choose $g := \iota_C$ to model a subspace constraint $C = \{y \in (\mathbb{R}^m)^N : y_1 = y_2 = \cdots = y_N\}$. In both cases our focus is on the effects of relative prox-regularity which is a sufficient condition to guarantee that inexact Gauss–Seidel minimization converges subsequentially to a stationary point of the Bregman–Moreau regularized problems. In a nonconvex setting, in particular under inexact updates, this cannot be taken for granted.

4.2.2. Alternating inexact Bregman Proximal Point

In this section we consider Model (4.1) and choose g to be an image function as in Equation (4.8):

$$g(y) := \inf\{R(w) : y = A(w), w \in \mathbb{R}^n\},$$

where $A : \mathbb{R}^n \to \mathbb{R}^m$ is a nonlinear $\mathcal{C}^1$-map with range $\operatorname{rge} A \subset \operatorname{int}(\operatorname{dom}\phi)$ and let $\phi \in \Gamma_0(\mathbb{R}^m)$ be a general Legendre function. Then, Problem (4.1) specializes to:

$$\text{minimize}\left\{H_\lambda(x,w) \equiv f(x) + \frac{1}{\lambda}D_\phi(x, A(w)) + R(w) : (x,w) \in \operatorname{dom}\phi \times \mathbb{R}^n\right\}, \quad (4.9)$$

where $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ and $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper lsc functions and $\lambda > 0$.

Via inf-projection with respect to x, the model is equivalent to the left Bregman relaxation:

$$\text{minimize}\left\{\big(\overleftarrow{\operatorname{env}}_\lambda^\phi f \circ A\big)(w) + R(w) : w \in \mathbb{R}^n\right\}. \quad (4.10)$$

The first algorithm we consider, Algorithm 1, is Gauss–Seidel minimization of Problem (4.9), with a proximal regularization of both variables. This is a Bregmanian generalization of proximal alternating minimization [Att+10].

Note that the w-update is in general a difficult problem. We may therefore replace the coupling function $D_\phi(x, A(w))$ with a proximal linearization as in *proximal alternating linearized minimization* (PALM) [BST14]. This is, to some extent, captured in the Bregman proximal term $D_\psi(w, w^t)$ in our formulation:

[1] These quotes are borrowed from the Wikipedia article on transduction in machine learning `https://en.wikipedia.org/wiki/Transduction_(machine_learning)`

Algorithm 1 Bregman Gauss–Seidel minimization with proximal regularization

```
Require: Choose appropriate Legendre functions $\chi \in \Gamma_0(\mathbb{R}^m)$ and $\psi \in \Gamma_0(\mathbb{R}^n)$ with
  $\operatorname{dom}\psi = \mathbb{R}^n$ and $\operatorname{dom}\chi \supset \operatorname{dom}\phi$ and initialize $x^0 \in \operatorname{int}(\operatorname{dom}\phi)$ and $w^0 \in \mathbb{R}^n$.
  for all $t = 1, 2, \dots$ do
      $x^{t+1} := \operatorname*{arg\,min}_{x\in\mathbb{R}^m} \ H_\lambda(x, w^t) + D_\chi(x, x^t)$      (4.11)
      $w^{t+1} := \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \ H_\lambda(x^{t+1}, w) + D_\psi(w, w^t).$      (4.12)
  end for
  return $(x^t, w^t)$.
```

Suppose that $A = \nabla\phi^* \circ B$ for ϕ^* being classically L-smooth and B linear. Thus

$$D_\phi(x^{t+1}, \nabla\phi^*(B(w))) = \phi(x^{t+1}) + \phi^*(B(w)) - \langle w, B^* x^{t+1}\rangle$$

is guaranteed to be $L\|B\|^2$-smooth in w and we may choose

$$\psi(w) := \frac{M}{2\lambda}\|w\|^2 - \frac{1}{\lambda}\phi^*(B(w)),$$

for $M > L\|B\|^2$. Then the w-update (4.12) becomes a classical proximal gradient step on H_λ as in PALM:

$$w^{t+1} = \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \ R(w) + \frac{1}{\lambda}\langle w, B^*(\nabla\phi^*(Bw^t) - x^{t+1})\rangle + \frac{M}{2\lambda}\|w - w^t\|^2.$$

Theorem 4.1. *Let $\phi, \chi \in \Gamma_0(\mathbb{R}^m)$, $\psi \in \Gamma_0(\mathbb{R}^n)$ be Legendre with $\operatorname{dom}\psi = \mathbb{R}^n$ and $\operatorname{dom}\chi \supset \operatorname{dom}\phi$ and let $\phi \in \mathcal{C}^2$ be super-coercive. Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ and $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper lsc and $\operatorname{dom} f \cap \operatorname{dom}\phi$ be nonempty. Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be $\mathcal{C}^1$ with $\operatorname{rge} A \subset \operatorname{int}(\operatorname{dom}\phi)$. Assume further that f is prox-bounded relative to ϕ with some threshold $\lambda_f > \lambda$ and $R + \psi$ is coercive. Assume that the sequence of iterates $\{w^t\}_{t\in\mathbb{N}}$ produced by Algorithm 1 is bounded and $\{x^t\}_{t\in\mathbb{N}} \subset C \subset \operatorname{int}\operatorname{dom}\chi$ where C is closed. Then $\{x^t\}_{t\in\mathbb{N}}$ is bounded as well. Let (x^*, w^*) be a limit point of $\{(x^t, w^t)\}_{t\in\mathbb{N}}$. Then (x^*, w^*) is a stationary point of H_λ, i.e., $0 \in \partial H_\lambda(x^*, w^*)$ which means*

$$0 \in \partial(f + \lambda^{-1}\nabla\phi)(x^*) - \frac{1}{\lambda}\nabla\phi(A(w^*)), \tag{4.13}$$

$$0 \in \partial R(w^*) + \frac{1}{\lambda}\nabla A(w^*)^*\nabla^2\phi(A(w^*))(A(w^*) - x^*). \tag{4.14}$$

If, in addition, the qualification condition $\partial^\infty f(x^*) \cap -N_{\operatorname{dom}\phi}(x^*) = \{0\}$ holds, we have $x^* \in \operatorname{int}(\operatorname{dom}\phi)$ and

$$0 \in \partial f(x^*) + \frac{1}{\lambda}(\nabla\phi(x^*) - \nabla\phi(A(w^*))). \tag{4.15}$$

To prove the above theorem we need the following auxiliary result:

Lemma 4.2. *Let $\chi \in \Gamma_0(\mathbb{R}^m)$ Legendre. Assume that $\{x^t\}_{t\in\mathbb{N}} \subset C \subset \operatorname{int}(\operatorname{dom}\chi)$ with C compact. Then $D_\chi(x^{t+1}, x^t) \to 0$ implies that $\|x^{t+1} - x^t\| \to 0$ as well as $\|\nabla\chi(x^{t+1}) - \nabla\chi(x^t)\| \to 0$.*

Proof. Suppose that $\|x^{t+1} - x^t\| \not\to 0$. Then, since $\{x^t\}_{t\in\mathbb{N}} \subset C$ is bounded $\{\|x^{t+1} - x^t\|\}_{t\in\mathbb{N}}$ is bounded and there are infinitely many t such that $\|x^{t+1} - x^t\|$ is bounded away from 0. Thus there exists a convergent subsequence indexed by j such that $\|x^{t_j+1} - x^{t_j}\| \to \varepsilon > 0$. Since $\{(x^{t_j+1}, x^t_j)\}_{j\in\mathbb{N}} \subset C \times C$ is bounded as well by taking another subsequence if necessary we have $(x^{t_j+1}, x^{t_j}) \to (p^*, x^*)$ with $p^* \neq x^*$. Since C is closed we also have $p^*, x^* \in C \subset \operatorname{int}(\operatorname{dom}\chi)$. Passing $j \to \infty$ we also have due to the joint continuity of $D_\chi : C \times C \to \mathbb{R}$ that $D_\chi(x^{t_j+1}, x^{t_j}) \to D_\chi(p^*, x^*) = 0$. By Lemma 1.20(i), this implies that $p^* = x^*$, a contradiction. We conclude $\|x^{t+1} - x^t\| \to 0$. Now suppose that $\|\nabla\chi(x^{t+1}) - \nabla\chi(x^t)\| \not\to 0$. Since $\{x^t\}_{t\in\mathbb{N}} \subset C$ and $\nabla\chi : C \to \mathbb{R}^m$ is continuous, also $\{\nabla\chi(x^t)\}_{t\in\mathbb{N}} \subset \chi(C)$ is a compact set. Then there exists a convergent subsequence $(\nabla\chi(x^{t_j+1}), \nabla\chi(x^{t_j})) \to (w^*, q^*) \in \chi(C) \times \chi(C)$ with $w^* \neq q^*$. By considering another subsequence if necessary in combination with the previous result we have $x^{t_j+1} \to x^*$ and $x^{t_j} \to x^*$. By continuity we have $w^* = \nabla\chi(x^*)$ and $q^* = \nabla\chi(x^*)$ which contradicts $w^* \neq q^*$. We conclude that $\|\nabla\chi(x^{t+1}) - \nabla\chi(x^t)\| \to 0$. □

Proof of Theorem 4.1. Since $\operatorname{rge} A \subset \operatorname{int}(\operatorname{dom}\phi)$ the function H_λ is proper and lsc. Since f is proper, lsc and prox-bounded relative to ϕ with threshold $\lambda_f > \lambda$ and $\operatorname{dom} f \cap \operatorname{dom}\phi \neq \emptyset$ and $R + \psi$ is proper lsc and coercive and $D_\phi(x, A(\cdot))$ is continuous and nonnegative, the iterates $\{(x^t, w^t)\}_{t\in\mathbb{N}}$ are well-defined.

By the definition of the x-update we have that

$$H_\lambda(x^{t+1}, w^t) + D_\chi(x^{t+1}, x^t) \leq H_\lambda(x^t, w^t)$$

and by the definition of the w-update

$$H_\lambda(x^{t+1}, w^{t+1}) + D_\psi(w^{t+1}, w^t) \leq H_\lambda(x^{t+1}, w^t).$$

Summing the two yields

$$H_\lambda(x^{t+1}, w^{t+1}) + D_\chi(x^{t+1}, x^t) + D_\psi(w^{t+1}, w^t) \leq H_\lambda(x^t, w^t). \tag{4.16}$$

This shows that $H_\lambda(x^t, w^t)$ is monotonically decreasing. Suppose that $\|x^t\| \to \infty$. Since f is prox-bounded with threshold $\lambda_f > \lambda$, there is $\lambda_f > \lambda' > \lambda$ such that $f + (1/\lambda')\phi > \alpha > -\infty$. Then using Cauchy–Schwarz

$$\begin{aligned} H_\lambda(x^t, w^t) \geq \left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right)\phi(x^t) + \alpha& \\ - \frac{1}{\lambda}\phi(A(w^t)) - \frac{1}{\lambda}\|\nabla\phi(A(w^t))\| \cdot \|x^t - A(w^t)\| + R(w^t).& \end{aligned}$$

Since w^t is bounded and both $\nabla\phi \circ A$ and $\phi \circ A$, are continuous and ϕ is super-coercive, $H_\lambda(x^t, w^t) \to \infty$, a contradiction. Thus $\{x^t\}_{t\in\mathbb{N}}$ is bounded. In particular this means $H_\lambda(x^t, w^t) > \beta > -\infty$. We sum the estimate form $t = 0$ to T and obtain that

$$\begin{aligned} -\infty < \beta - H_\lambda(x^0, w^0) &\leq H_\lambda(x^T, w^T) - H_\lambda(x^0, w^0) \\ &= \sum_{t=0}^{T} H_\lambda(x^{t+1}, w^{t+1}) - H_\lambda(x^t, w^t) \\ &\leq -\sum_{t=0}^{T} \big(D_\chi(x^{t+1}, x^t) + D_\psi(w^{t+1}, w^t)\big). \end{aligned}$$

We take $T \to \infty$ and deduce that

$$D_\chi(x^{t+1}, x^t) + D_\psi(w^{t+1}, w^t) \to 0,$$

and therefore $D_\chi(x^{t+1}, x^t) \to 0$ and $D_\psi(w^{t+1}, w^t) \to 0$. Since $\{x^t\}_{t\in\mathbb{N}} \subset C$ is bounded and C is closed we can invoke Lemma 4.2 and obtain that $\|x^{t+1} - x^t\| \to 0$ and $\|w^{t+1} - w^t\| \to 0$ as well as $\|\nabla\chi(x^{t+1}) - \nabla\chi(x^t)\| \to 0$ and $\|\nabla\psi(w^{t+1}) - \nabla\psi(w^t)\| \to 0$. In view of the x- and w-updates Fermat's rule Lemma 1.3 yields:

$$\begin{aligned}0 \in \partial(f(x^{t+1}) + \lambda^{-1}\nabla\phi(x^{t+1})) - \frac{1}{\lambda}\nabla\phi(A(w^{t+1}))\\ + \nabla\chi(x^{t+1}) - \nabla\chi(x^t) + \frac{1}{\lambda}(\nabla\phi(A(w^{t+1})) - \nabla\phi(A(w^t))),\end{aligned}$$

and

$$0 \in \partial R(w^{t+1}) + \frac{1}{\lambda}\nabla A(w^{t+1})^*\nabla^2\phi(A(w^{t+1}))(A(w^{t+1}) - x^{t+1}) + \nabla\psi(w^{t+1}) - \nabla\psi(w^t).$$

Since H_λ is the sum of a proper lsc, separable part $(f + \lambda^{-1}\phi)(x) + R(w)$ and a smooth part $\lambda^{-1}\langle\nabla\phi(A(w)), A(w) - x\rangle - \lambda^{-1}\phi(A(w))$ we can invoke [RW98, Exercise 8.8(c)] and [RW98, Proposition 10.5] which yields:

$$\begin{aligned}\partial H_\lambda(x^{t+1}, w^{t+1}) = \partial(f + \lambda^{-1}\phi)(x^{t+1}) \times \partial R(w^{t+1})\\ + \frac{1}{\lambda}\big(-\nabla\phi(A(w^{t+1})), \nabla A(w^{t+1})^*\nabla^2\phi(A(w^{t+1}))(A(w^{t+1}) - x^{t+1})\big).\end{aligned}$$

and therefore

$$\begin{pmatrix}\nabla\chi(x^t) - \nabla\chi(x^{t+1}) + \lambda^{-1}(\nabla\phi(A(w^t)) - \nabla\phi(A(w^{t+1})))\\ \nabla\psi(w^t) - \nabla\psi(w^{t+1})\end{pmatrix} \in \partial H_\lambda(x^{t+1}, w^{t+1}).$$

Since the iterates are bounded we may consider a convergent subsequence $\{(x^{t_j}, w^{t_j})\}_{j\in\mathbb{N}} \subset \{(x^t, w^t)\}_{t\in\mathbb{N}}$. Let (x^*, w^*) denote the limit point. Next we prove that $H_\lambda(x^{t_j}, w^{t_j}) \to H_\lambda(x^*, w^*)$. Due to the update of x^{t_j+1} we have:

$$\begin{aligned}f(x^{t_j+1}) + \frac{1}{\lambda}D_\phi(x^{t_j+1}, A(w^{t_j})) + D_\chi(x^{t_j+1}, x^{t_j})\\ \le f(x^*) + \frac{1}{\lambda}D_\phi(x^*, A(w^{t_j})) + D_\chi(x^*, x^{t_j}).\end{aligned}$$

Due to the update of w^{t_j+1} we have

$$\begin{aligned}R(w^{t_j+1}) + \frac{1}{\lambda}D_\phi(x^{t_j+1}, A(w^{t_j+1})) + D_\psi(w^{t_j+1}, w^{t_j})\\ \le R(w^*) + \frac{1}{\lambda}D_\phi(x^{t_j+1}, A(w^*)) + D_\psi(w^*, w^{t_j}).\end{aligned}$$

Summing yields:

$$\begin{aligned}H_\lambda(x^{t_j+1}, w^{t_j+1}) \le H_\lambda(x^*, w^*) - \frac{1}{\lambda}D_\phi(x^*, A(w^*)) - \frac{1}{\lambda}D_\phi(x^{t_j+1}, A(w^{t_j}))\\ + \frac{1}{\lambda}D_\phi(x^*, A(w^{t_j})) + \frac{1}{\lambda}D_\phi(x^{t_j+1}, A(w^*))\\ + D_\chi(x^*, x^{t_j}) + D_\psi(w^*, w^{t_j}) - D_\chi(x^{t_j+1}, x^{t_j}) - D_\psi(w^{t_j+1}, w^{t_j}).\end{aligned}$$

Passing $j \to \infty$ yields since $\|x^{t+1} - x^t\| \to 0$ and due to the joint continuity of $D_\phi(\cdot, A(\cdot))$ and D_ψ, D_χ and since $H_\lambda(x^{t_j+1}, w^{t_j+1}) \to H_\lambda^*$, that $H_\lambda^* \leq H_\lambda(x^*, w^*)$. Since H_λ is lsc we also have that $H_\lambda^* \geq H_\lambda(x^*, w^*)$. Hence $H_\lambda^* = H_\lambda(x^*, w^*)$.

In view of the closedness of $\operatorname{gph} \partial H_\lambda$ under the H_λ-attentive topology, we have for $j \to \infty$, since $H_\lambda(x^{t_j}, w^{t_j}) \to H_\lambda(x^*, w^*)$, the continuity of $\nabla\chi, \nabla\psi, \nabla\phi \circ A$ and $\|x^{t+1} - x^t\| \to 0$ and $\|w^{t+1} - w^t\| \to 0$ that:

$$0 \in \partial H_\lambda(x^*, w^*).$$

Assume that in addition we have $\partial^\infty f(x^*) \cap -N_{\operatorname{dom}\phi}(x^*) = \{0\}$. Since $0 \in \partial H_\lambda(x^*, w^*)$ we have that

$$\lambda^{-1}\nabla\phi(A(w^*)) \in \partial(f + \lambda^{-1}\phi)(x^*).$$

Invoking Lemma 1.2 and [RW98, Proposition 8.12] we have

$$\nabla\phi(A(w^*)) \in \partial\phi(x) + \lambda\partial f(x).$$

In particular this means that $x \in \operatorname{dom}(\partial\phi)$ implying via Lemma 1.16 that $x \in \operatorname{int}(\operatorname{dom}\phi)$. We also conclude that the optimality conditions (4.13) and (4.14) hold. □

However, in general a stationary point of H_λ does not translate to a stationary point of the left Bregman relaxation (4.10): Indeed, in general the implication

$$\begin{aligned} &0 \in \partial f(x^*) + \frac{1}{\lambda}\big(\nabla\phi(x^*) - \nabla\phi(A(w^*))\big) \\ &\quad\implies \frac{1}{\lambda}\nabla A(w^*)^*\nabla^2\phi(A(w^*))(A(w^*) - x^*) \in \partial\big(\overleftarrow{\mathrm{env}}^{\phi}_{\lambda} f \circ A\big)(w^*), \end{aligned} \tag{4.17}$$

is false. A sufficient condition for this translation of stationarity is relative prox-regularity as explored next: In addition, we illustrate our findings considering a generalization of Algorithm 1 (if $\psi = \chi = \|\cdot\|^2$), where the proximal subproblems are solved inexactly.

Algorithm 2 inexact Bregman Gauss–Seidel minimization

```
Require: Initialize $x^0 \in \operatorname{int}(\operatorname{dom}\phi)$ and $w^0 \in \mathbb{R}^n$.
  for all $t = 1, 2, \ldots$ do
    solve inexactly such that Assumption 4.3 holds true:
        $x^{t+1} \approx \operatorname*{arg\,min}_{x \in \mathbb{R}^m} H_\lambda(x, w^t),$    (4.18)
        $w^{t+1} \approx \operatorname*{arg\,min}_{w \in \mathbb{R}^n} H_\lambda(x^{t+1}, w).$    (4.19)
  end for
  return $(x^t, w^t)$.
```

The following assumptions constitute [ABS13, H1–H2] and are standard for convergence of Gauss–Seidel minimization in a nonconvex setting. The continuity condition Assumption [ABS13, H3] in our case is substituted by a subdifferential continuity condition at the limit point, see Definition 1.7.

Assumption 4.3 (inexact updates Bregman Gauss–Seidel minimization). We assume there exist constants $\gamma, \tau > 0$ such that for each $t \in \mathbb{N}$:

(i) a sufficient descent in x and w is generated, i.e. we have

$$H_\lambda(x^{t+1}, w^t) \le H_\lambda(x^t, w^t) - \frac{\gamma}{2}\|x^{t+1} - x^t\|^2, \tag{4.20}$$

and for the w-update

$$H_\lambda(x^{t+1}, w^{t+1}) \le H_\lambda(x^{t+1}, w^t) - \frac{\gamma}{2}\|w^{t+1} - w^t\|^2, \tag{4.21}$$

(ii) the x- and w-update satisfies a relative error condition: This means $x^{t+1} \in \operatorname{int}\operatorname{dom}\phi$ and there is

$$u^{t+1} \in \partial_x H_\lambda(x^{t+1}, w^t), \quad \|u^{t+1}\| \le \tau\|x^{t+1} - x^t\|, \tag{4.22}$$

and accordingly the w-update meets

$$v^{t+1} \in \partial_w H_\lambda(x^{t+1}, w^{t+1}), \quad \|v^{t+1}\| \le \tau\|w^{t+1} - w^t\|. \tag{4.23}$$

The relative error and sufficient descent conditions are mild. For instance this can be achieved for w via a single line-search proximal gradient step where the coupling part $D_\phi(x^{t+1}, A(w))$ is linearized in w. The conditions are also valid when the updates are perturbed with a quadratic proximal regularization as in the first algorithm 1 with $\psi = \chi = \|\cdot\|^2$.

Next we discuss that relative prox-regularity yields a certain stability condition of the Bregman proximal mapping which allows us to prove that the algorithm solves the left Bregman relaxation. In particular prox-regularity allows us to show that the above implication (4.17) holds true, and the stationarity gap vanishes along the iterates. However, since prox-regularity is a local condition our results have the taste of a local convergence theorem: I.e., we assume that the iterates are sufficiently close to a certain fixed-point of the proximal mapping at which f is prox-regular and subdifferentially continuous and the chosen λ was sufficiently small.

Theorem 4.4. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ such that $\phi \in \mathcal{C}^2$ on $\operatorname{int}(\operatorname{dom}\phi)$ and $\lambda > 0$. Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded with some threshold $\lambda_f > \lambda$ and $\operatorname{dom}\phi \cap \operatorname{dom} f \neq \emptyset$. Let $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper lsc. In addition let Assumption 4.3 hold true. Assume the iterates $\{w^t\}_{t\in\mathbb{N}}$ generated by Algorithm 2 are bounded. Let $w^{t_j} \to w^*$ be a convergent subsequence of the sequence of iterates. Let f be prox-regular relative to ϕ and subdifferentially continuous at $x^* \in \operatorname{int}(\operatorname{dom}\phi)$ for $v^* := \lambda^{-1}(\nabla\phi(A(w^*)) - \nabla\phi(x^*)) \in \partial f(x^*)$ such that $\{x^*\} = \overleftarrow{\operatorname{prox}}_\lambda^\phi f(A(w^*))$. Let R be subdifferentially continuous at w^* for*

$$q^* := \lambda^{-1}\nabla A(w^*)^*\nabla^2\phi(A(w^*))(A(w^*) - x^*).$$

Let λ sufficiently small. Then $\overleftarrow{\operatorname{env}}_\lambda^\phi f$ is continuously differentiable around $A(w^*)$. Assume that for j sufficiently large the iterates $\{x^{t_j}\}_{j\in\mathbb{N}}$ are contained in a sufficiently small neighborhood around x^*. Then we have

$$\operatorname{dist}\left(0, \partial(\overleftarrow{\operatorname{env}}_\lambda^\phi f \circ A + R)(w^{t_j})\right) \to 0,$$

for $j \to \infty$. In particular this implies that w^* is a stationary point of (4.10)*.*

Proof. Summing the sufficient descent conditions (4.20) and (4.21) from Assumption 4.3(i) we obtain:

$$H_\lambda(x^{t+1}, w^{t+1}) \leq H_\lambda(x^t, w^t) - \frac{\gamma}{2}\|x^{t+1} - x^t\|^2 - \frac{\gamma}{2}\|w^{t+1} - w^t\|^2.$$

Since f is proper lsc and relatively prox-bounded with some threshold $\lambda_f > \lambda$ and A is continuous with $\operatorname{rge} A \subset \operatorname{int}(\operatorname{dom}\phi)$ we have in view of Lemma 2.22(ii) that $\overleftarrow{\operatorname{env}}^{\phi}_{\lambda} f \circ A : \mathbb{R}^n \to \mathbb{R}$ is continuous. Since the iterates w^t are bounded and since R is proper lsc we know that $-\infty < \beta \leq \overleftarrow{\operatorname{env}}^{\phi}_{\lambda} f(A(w^t)) + R(w^t) \leq H_\lambda(w^t, x^t)$ is uniformly bounded from below. Summing the estimate form $t = 0$ to $t = T$ we obtain:

$$\begin{aligned} -\infty < \beta - H_\lambda(x^0, w^0) &\leq H_\lambda(x^{T+1}, w^{T+1}) - H_\lambda(x^0, w^0) \\ &\leq -\frac{\gamma}{2}\sum_{t=0}^{T}\big(\|x^{t+1} - x^t\|^2 + \|w^{t+1} - w^t\|^2\big). \end{aligned}$$

Passing $T \to \infty$ shows that $\sum_{t=0}^{T}(\|x^{t+1} - x^t\|^2 + \|w^{t+1} - w^t\|^2) \to 0$ and in particular $\|x^{t+1} - x^t\| \to 0$ and $\|w^{t+1} - w^t\| \to 0$.

Let $w^{t_j} \to w^*$ be a convergent subsequence of the sequence of iterates such that $x^* = \overleftarrow{\operatorname{prox}}^{\phi}_{\lambda} f(A(w^*)) \in \operatorname{int}(\operatorname{dom}\phi)$ and f is prox-regular relative to ϕ at x^* for $v^* = \lambda^{-1}(\nabla\phi(A(w^*)) - \nabla\phi(x^*))$. The relative error condition from Assumption 4.3(ii) guarantees that there is

$$u^{t_j} \in \partial_x H_\lambda(x^{t_j}, w^{t_j - 1}) = \partial f(x^{t_j}) + \lambda^{-1}(\nabla\phi(x^{t_j}) - \nabla\phi(A(w^{t_j - 1}))),$$

such that $\|u^{t_j}\| \leq \tau\|x^{t_j} - x^{t_j - 1}\|$ and therefore $u^{t_j} \to 0$. Rewriting the above inclusion we have for $z^{t_j} := \nabla\phi^*(\nabla\phi(A(w^{t_j - 1})) + \lambda u^{t_j})$

$$0 \in \partial f(x^{t_j}) + \lambda^{-1}(\nabla\phi(x^{t_j}) - \nabla\phi(z^{t_j})),$$

where $z^{t_j} = \nabla\phi^*(\nabla\phi(A(w^{t_j - 1})) + \lambda u^{t_j}) \to A(w^*)$. Since x^{t_j} is near x^* for j sufficiently large we have $\lambda^{-1}(\nabla\phi(z^{t_j}) - \nabla\phi(x^{t_j}))$ near $v^* \in \partial f(x^*)$ and due to the subdifferential continuity of f at x^* for $\lambda^{-1}(\nabla\phi(A(w^*)) - \nabla\phi(x^*))$ we have $f(x^{t_j})$ near $f(x^*)$. This shows that for j sufficiently large we have:

$$0 \in T(x^{t_j}) + \lambda^{-1}(\nabla\phi(x^{t_j}) - \nabla\phi(z^{t_j})),$$

where T is an f-attentive ε-localization of ∂f at x^* for v_i^* with ε chosen sufficiently small such that in view of Theorem 2.40(iii) we have $x^{t_j} = \overleftarrow{\operatorname{prox}}^{\phi}_{\lambda} f(z^{t_j})$ and by Lemma 2.22(iii), we also have $x^{t_j} \to x^*$. In view of Corollary 2.60 we have $\nabla\overleftarrow{\operatorname{env}}^{\phi}_{\lambda} f$ is continuously differentiable around $A(w^*)$ with

$$\nabla\overleftarrow{\operatorname{env}}^{\phi}_{\lambda} f(z^{t_j}) = \frac{1}{\lambda}\nabla^2\phi(z^{t_j})(z^{t_j} - x^{t_j}) \to \nabla\overleftarrow{\operatorname{env}}^{\phi}_{\lambda} f(A(w^*)).$$

The relative error condition from Assumption 4.3(ii) guarantees that there is:

$$v^{t_j} \in \partial_w H_\lambda(x^{t_j}, w^{t_j}) = \partial R(w^{t_j}) + \lambda^{-1}\nabla A(w^{t_j})^*\nabla^2\phi(A(w^{t_j}))(A(w^{t_j}) - x^{t_j}).$$

Define

$$p^{t_j} := v^{t_j} - \lambda^{-1}\nabla A(w^{t_j})^*\nabla^2\phi(A(w^{t_j}))(A(w^{t_j}) - x^{t_j}) + \nabla A(w^{t_j})^*\nabla\overleftarrow{\operatorname{env}}^{\phi}_{\lambda} f(A(w^{t_j})).$$

Then $p^{t_j} \to 0$ for $j \to \infty$ and we have

$$p^{t_j} \in \partial R(w^{t_j}) + \nabla A(w^{t_j})^* \nabla \overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f(A(w^{t_j}))$$

passing $j \to \infty$ shows that

$$\mathrm{dist}(0, \partial R(w^{t_j}) + \nabla A(w^{t_j})^* \nabla \overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f(A(w^{t_j}))) \to 0.$$

Since in addition R is proper lsc and subdifferentially continuous at w^* for

$$q^* = \lambda^{-1} \nabla A(w^*)^* \nabla^2 \phi(A(w^*))(A(w^*) - x^*)$$

we have $R(w^{t_j}) \to R(w^*)$ and therefore

$$0 \in \partial(R + \overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f \circ A)(w^*) = \partial R(w^*) + \nabla A(w^*)^* \nabla \overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f(A(w^*)). \qquad \square$$

Note that under Assumption 4.3 one can actually derive the convergence of the whole sequence (x^t, w^t), if, in addition, H_λ satisfies the Kurdyka–Łojasiewicz property [Att+10], see [ABS13, Theorem 2.9]. This, however, is beyond the scope of this thesis.

If prox-regularity holds globally with uniform constants, i.e., f is relatively hypoconvex, the previous result can be globalized:

Corollary 4.5. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ such that $\phi \in \mathcal{C}^2$ on $\mathrm{int}(\mathrm{dom}\,\phi)$. Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and relatively hypoconvex with constant r and $\mathrm{dom}\,\phi \cap \mathrm{dom}\, f \neq \emptyset$ such that $\partial^\infty f(x) \cap -N_{\mathrm{dom}\,\phi}(x) = \{0\}$ at all $x \in \mathrm{dom}\, f \cap \mathrm{dom}\,\phi \neq \emptyset$. Let $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper lsc. Let R be continuous over $\mathrm{dom}\, R$. In addition let Assumption 4.3 hold true and assume the iterates $\{w^t\}_{t\in\mathbb{N}}$, generated by Algorithm 2 are bounded. Let $w^{t_j} \to w^*$ be a convergent subsequence of the sequence of iterates. Let $\lambda < 1/r$. Then $\overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f \circ A$ is continuously differentiable and*

$$\mathrm{dist}\left(0, \partial(\overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f \circ A + R)(w^{t_j})\right) \to 0,$$

for $j \to \infty$. In particular this implies that w^* is a stationary point of (4.10).

Finally we discuss a variant of Algorithm 1 where the x-update is solved exactly. As a short computation reveals, when the gradient formula for the envelope $\overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f \circ \nabla \phi^*$ holds (which happens to be true locally whenever f is relatively prox-regular around the limit point and relatively prox-bounded and $\lambda > 0$ is sufficiently small) and $A = \nabla \phi^*$, we can rewrite the algorithm as the following Bregman proximal gradient update:

$$w^{t+1} = \operatorname*{arg\,min}_{w \in \mathbb{R}^m} \; R(w) + \left\langle \nabla\left(\overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f \circ \nabla\phi^*\right)(w^t), w - w^t \right\rangle + D_{\frac{1}{\lambda}\phi^* + \psi}(w, w^t).$$

This illustrates a close relationship between alternating (Bregman) minimization and the (Bregman) proximal gradient method, which is known from the quadratic case. Indeed, in view of Proposition 2.57, for ϕ super-coercive and f relatively prox-bounded with threshold λ_f the function

$$\frac{1}{\lambda}\phi^* - \overleftarrow{\mathrm{env}}{}^{\phi}_{\lambda} f \circ \nabla \phi^*$$

is proper lsc and convex if $\lambda \in (0, \lambda_f)$ and therefore f locally satisfies the one-sided extended descent lemma with constant $1/\lambda$ when f is relatively prox-regular and λ sufficiently small. Overall, this means that existing convergence results from [BBT17; Bol+18] for the Bregman proximal gradient method carry over, at least locally.

4.2.3. Inexact averaged Bregman Proximal Point

Next we consider an inexact averaged Bregman proximal point algorithm

$$x^{t+1} \approx \sum_{i=1}^{N} \pi_i \overleftarrow{\operatorname{prox}}^{\phi}_{\lambda} f_i(x^t),$$

for weights $\pi \in \operatorname{relint} \Pi$, with

$$\Pi := \left\{ \pi \in \mathbb{R}^M : \pi_i \geq 0, \sum_{i=1}^{N} \pi_i = 1 \right\}.$$

This algorithm is of particular interest in the context of federated learning which is covered in great detail in Section 4.3 specializing to a Euclidean and globally hypoconvex setting. In this section, however, our focus is on the more technical aspects of the algorithm in a Bregmanian setting under relative prox-regularity.

In contrast to the previous setting, it will be even more important to guarantee that the iterates x^t stay in the interior $\operatorname{int}(\operatorname{dom} \phi)$ to ensure the iterates are well-defined. Under exact proximal updates a possible strategy is to assume a constraint qualification $\partial^\infty f(x) \cap -N_{\operatorname{dom} \phi}(x) = \{0\}$ at all $x \in \operatorname{dom} f \cap \operatorname{dom} \phi \neq \emptyset$.

Then we are able to apply the product space trick known from the Euclidean setting to reformulate the above iteration in terms of a Gauss–Seidel method as before. To this end we define:

$$C := \left\{ x \in (\mathbb{R}^m)^N : x_1 = \cdots = x_N \right\},$$

and $\Phi : (\mathbb{R}^m)^N \to \overline{\mathbb{R}}$

$$\Phi(x) = \sum_{i=1}^{M} \pi_i \phi(x_i)$$

and write

$$\text{minimize} \left\{ H_\lambda(x, y) \equiv \sum_{i=1}^{M} \pi_i f_i(x_i) + \frac{1}{\lambda} D_\Phi(x, y) + \iota_C(y) : x, y \in (\operatorname{dom} \phi)^N \right\}. \tag{4.24}$$

The complete algorithm is listed in Algorithm 3.

We need the following generalization of [BCN06, Proposition 5.2(ii)] (for non-$\mathcal{C}^2$ ϕ) as part of our analysis which shows that the right Bregman projection of the consensus space C is the arithmetic average:

Lemma 4.6. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ and $\pi \in \Pi$. Let $x \in \operatorname{int}(\operatorname{dom} \phi)^N$. Then $y \in (\mathbb{R}^m)^N$ with $y_1 = y_2 = \cdots = y_N = \sum_{i=1}^{N} \pi_i x_i$ is the unique solution to*

$$y = \operatorname*{arg\,min}_{y \in C} \ D_\Phi(x, y) = \overrightarrow{\operatorname{prox}}^{\Phi}_{1} \iota_C(x),$$

and in particular we have $y \in \operatorname{int}(\operatorname{dom} \phi)^N$.

Proof. We adopt the arguments in [Ban+05, Proposition 1]. By construction $y \in \operatorname{int}(\operatorname{dom} \phi)^N$ since $x \in \operatorname{int}(\operatorname{dom} \phi)^N$ and $\pi \in \Pi$. By definition we have $D_\Phi(x, y') = +\infty$ for $y' \notin \operatorname{int}(\operatorname{dom} \phi)^N$. We make the following computation. For any $y' \in \operatorname{int}(\operatorname{dom} \phi)^N \cap C$

with $y' \neq y$ we have:

$$\begin{aligned} D_\Phi(x,y') - D_\Phi(x,y) &= \sum_{i=1}^N \pi_i (D_\phi(x_i, y_i') - D_\phi(x_i, y_i)) \\ &= \sum_{i=1}^N \pi_i(\phi(y_1) - \phi(y_1') - \langle \nabla\phi(y_1'), x_i - y_1'\rangle + \langle \nabla\phi(y_1), x_i - y_1\rangle) \\ &= \phi(y_1) - \phi(y_1') - \langle \nabla\phi(y_1'), y_1 - y_1'\rangle + \langle \nabla\phi(y_1), y_1 - y_1\rangle \\ &= D_\phi(y_1, y_1'). \end{aligned}$$

Since $y, y' \in C$ and $y \neq y'$ we have $y_1 \neq y_1'$ and therefore $D_\Phi(x,y') > D_\Phi(x,y)$. □

Algorithm 3 inexact averaged Bregman minimization

```
Require: Initialize $x^0, y^0 \in (\mathrm{int}(\mathrm{dom}\,\phi))^N$.
  for all $t = 1, 2, \dots$ do
    for all $i \in \{1, 2, \dots, N\}$ do
      solve inexactly such that Assumption 4.7 holds true:
          $$x_i^{t+1} \approx \operatorname*{arg\,min}_{x \in \mathbb{R}^m} \; f_i(x) + \frac{1}{\lambda} D_\phi(x, y_i^t). \qquad (4.25)$$
    end for
      $$y^{t+1} := \operatorname*{arg\,min}_{y \in (\mathbb{R}^m)^N} \; H_\lambda(x^{t+1}, y) = \sum_{i=1}^N \pi_i x_i^{t+1}. \qquad (4.26)$$
  end for
  return $(x^t, y^t)$.
```

In place of an exact proximal step assume that x is updated inexactly.

Assumption 4.7 (inexact update averaged Bregman Proximal Point). Let $h_i(x,y) := f_i(x) + \lambda^{-1} D_\phi(x,y)$.

(i) There is $\gamma > 0$ such that for each i and each t

$$h_i(x_i^{t+1}, y_i^t) \leq h_i(x_i^t, y_i^t) - \frac{\gamma}{2}\|x_i^{t+1} - x_i^t\|^2,$$

(ii) There is $\tau > 0$ such that for each i and each t there is $u_i^{t+1} \in \mathbb{R}^m$ such that $x_i^{t+1} \in \mathrm{int}\,\mathrm{dom}\,\phi$ and

$$u_i^{t+1} \in \partial_x h_i(x_i^{t+1}, y_i^t), \quad \|u_i^{t+1}\| \leq \tau \|x_i^{t+1} - x_i^t\|.$$

We will see that under relative prox-regularity this solves the left Bregman relaxation:

$$\text{minimize} \left\{ \sum_{i=1}^M \pi_i \overleftarrow{\mathrm{env}}_\lambda^\phi f_i(y) : y \in \mathrm{dom}\,\phi \right\}. \qquad (4.27)$$

Theorem 4.8. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ such that $\phi \in \mathcal{C}^2$ on $\operatorname{int}(\operatorname{dom}\phi)$ and $\pi \in \operatorname{relint}\Pi$ and $\lambda > 0$. Let f_i be proper lsc and prox-bounded with some threshold $\lambda_f > \lambda$. Assume that $\operatorname{dom}\phi \cap \operatorname{dom} f_i \neq \emptyset$ and Assumption 4.7 holds true. Then, the iterates $\{(x^t, y^t)\}_{t\in\mathbb{N}}$, generated by Algorithm 3, are well-defined, i.e., $x^{t+1}, y^{t+1} \in \operatorname{int}(\operatorname{dom}\phi)^N$, with $y_1^{t+1} = y_2^{t+1} = \cdots = y_N^{t+1} = \sum_{i=1}^N \pi_i x_i^{t+1}$. In addition assume that the iterates are bounded. Let $y^{t_j} \to y^*$ be a convergent subsequence of the sequence of iterates such that $y^* \in (\operatorname{int}(\operatorname{dom}\phi))^N$ and f_i is prox-regular relative to ϕ and subdifferentially continuous at x_i^* for $v_i^* := \lambda^{-1}(\nabla\phi(y_i^*) - \nabla\phi(x_i^*))$ such that $\{x_i^*\} = \overleftarrow{\operatorname{prox}}{}_\lambda^\phi f(y_i^*)$. Then, $y_1^* = y_2^* = \cdots = y_N^* = \sum_{i=1}^N \pi_i x_i^* \in \operatorname{int}(\operatorname{dom}\phi)$. Let λ sufficiently small. Then $\overleftarrow{\operatorname{env}}{}_\lambda^\phi f_i$ is continuously differentiable around y_1^*. In addition assume that $x_i^{t_j}$ is contained in a sufficiently small neighborhood around x_i^* for j suffciently large. Then we have*

$$\sum_{i=1}^N \pi_i \nabla \overleftarrow{\operatorname{env}}{}_\lambda^\phi f_i(y_1^{t_j}) \to 0,$$

as $j \to \infty$. In particular this implies that y_1^* is a stationary point of (4.27).

Proof. By assumption we know that

$$H_\lambda(x^{t+1}, y^{t+1}) \leq H_\lambda(x^t, y^t) - \frac{\gamma}{2}\sum_{i=1}^N \pi_i \|x_i^{t+1} - x_i^t\|^2.$$

Since the iterates x_i^t are bounded and since f_i is proper lsc we know that $\sum_{i=1}^N \pi_i f_i(x_i^t)$ is uniformly bounded from below. Since $D_\phi(x^t, y^t) \geq 0$ there is $-\infty < \beta$ such that

$$\beta \leq H_\lambda(x^t, y^t).$$

Summing the estimate form $t = 0$ to $t = T$ we obtain:

$$-\infty < \beta - H_\lambda(x^0, y^0) \leq H_\lambda(x^{T+1}, y^{T+1}) - H_\lambda(x^0, y^0) \leq -\frac{\gamma}{2}\sum_{t=0}^T\sum_{i=1}^N \pi_i \|x_i^{t+1} - x_i^t\|^2.$$

Passing $T \to \infty$ shows that $\sum_{t=0}^T \sum_{i=1}^N \pi_i \|x_i^{t+1} - x_i^t\|^2 \to 0$ and since $\pi \in \operatorname{relint}\Pi$ and therefore $\pi_i > 0$ in particular $\|x_i^{t+1} - x_i^t\| \to 0$. Since by assumption $x^{t+1} \in (\operatorname{int}(\operatorname{dom}\phi))^N$, in view of Lemma 4.6, we have $y^{t+1} \in (\operatorname{int}(\operatorname{dom}\phi))^N$ and $y_1^{t+1} = y_2^{t+1} = \cdots = y_N^{t+1} = \sum_{i=1}^N \pi_i x_i^{t+1}$.

Let $y^{t_j} \to y^*$ be a convergent subsequence of the sequence of iterates such that $y^* \in (\operatorname{int}(\operatorname{dom}\phi))^N$ and for each i let f_i be prox-regular relative to ϕ at

$$x_i^* = \overleftarrow{\operatorname{prox}}{}_\lambda^\phi f_i(y_i^*) \in \operatorname{int}(\operatorname{dom}\phi),$$

for $v_i^* := \lambda^{-1}(\nabla\phi(x_i^*) - \nabla\phi(y_i^*))$. By assumption we have

$$u_i^{t_j+1} \in \partial f_i(x_i^{t_j+1}) + \lambda^{-1}(\nabla\phi(x_i^{t_j+1}) - \nabla\phi(y_i^{t_j}))$$

for $\|u_i^{t_j+1}\| \leq \tau\|x_i^{t_j+1} - x_i^{t_j}\|$ and therefore $u_i^{t_j+1} \to 0$. Rewriting the above inclusion we have for $z_i^{t_j+1} := \nabla\phi^*(\nabla\phi(y_i^{t_j}) + \lambda u_i^{t_j+1})$

$$0 \in \partial f_i(x_i^{t_j+1}) + \lambda^{-1}(\nabla\phi(x_i^{t_j+1}) - \nabla\phi(z_i^{t_j+1})).$$

where $z_i^{t_j+1} \to y_i^*$.

Since $x_i^{t_j}$ and therefore $x_i^{t_j+1}$ is near x_i^* for j sufficiently large we have $\lambda^{-1}(\nabla\phi(z_i^{t_j+1}) - \nabla\phi(x_i^{t_j+1}))$ near $v_i^* \in \partial f(x_i^*)$ and due to the subdifferential continuity of f_i at x_i^* for $v_i^* = \lambda^{-1}(\nabla\phi(y_i^*) - \nabla\phi(x_i^*)$ we have $f_i(x_i^{t_j+1})$ near $f_i(x_i^*)$. This shows that for j sufficiently large we have:

$$0 \in T_i(x_i^{t_j+1}) + \lambda^{-1}(\nabla\phi(x_i^{t_j+1}) - \nabla\phi(z_i^{t_j+1})),$$

where T_i is a f_i-attentive ε-localization of ∂f_i at x_i^* for v_i^* with ε chosen sufficiently small such that in view of Theorem 2.40(iii) we have

$$x_i^{t_j+1} = \overleftarrow{\mathrm{prox}}_\lambda^\phi f_i(z_i^{t_j+1}),$$

and by Lemma 2.22(iii) we also have $x_i^{t_j+1} \to x_i^*$. In view of Corollary 2.60 we have $\nabla \overleftarrow{\mathrm{env}}_\lambda^\phi f_i$ is continuously differentiable around y_i^* with

$$\nabla \overleftarrow{\mathrm{env}}_\lambda^\phi f_i(z_i^{t_j+1}) = \frac{1}{\lambda}\nabla^2\phi(z_i^{t_j+1})(x_i^{t_j+1} - z_i^{t_j+1}) \to \nabla \overleftarrow{\mathrm{env}}_\lambda^\phi f_i(y_i^*).$$

Summing over i yields since $y_1^* = y_2^* = \cdots = y_N^* = \sum_{i=1}^N \pi_i x_i^*$:

$$\begin{aligned}\sum_{i=1}^N \pi_i \nabla \overleftarrow{\mathrm{env}}_\lambda^\phi f_i(z_i^{t_j+1}) = \frac{1}{\lambda}\sum_{i=1}^N \pi_i \nabla^2\phi(z_i^{t_j+1})(x_i^{t_j+1} - z_i^{t_j+1}) &\to \frac{1}{\lambda}\sum_{i=1}^N \pi_i \nabla^2\phi(y_i^*)(x_i^* - y_i^*) \\ &= 0 = \sum_{i=1}^N \pi_i \nabla \overleftarrow{\mathrm{env}}_\lambda^\phi f_i(y_1^*).\end{aligned}$$

Due to continuity we also have that

$$\sum_{i=1}^N \pi_i \nabla \overleftarrow{\mathrm{env}}_\lambda^\phi f_i(y_1^{t_j}) \to 0. \qquad \square$$

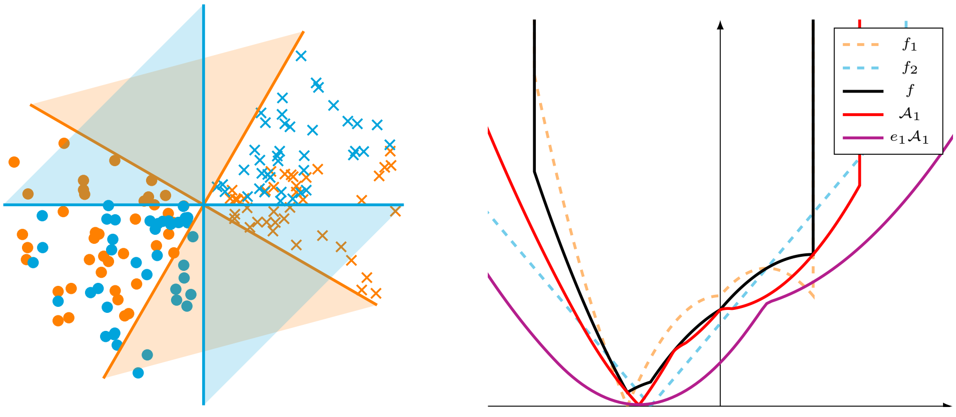
Again, the result can be globalized when f_i satisfies a relative hypoconvexity property in place of prox-regularity and subdifferential continuity.

Corollary 4.9. *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ such that $\phi \in \mathcal{C}^2$ on $\mathrm{int}(\mathrm{dom}\,\phi)$ and $\pi \in \mathrm{relint}\,\Pi$ and $\lambda > 0$. Let f_i be proper lsc. Assume that for all $1 \le i \le N$ we have $\mathrm{dom}\,\phi \cap \mathrm{dom}\, f_i \neq \emptyset$ such that $\partial^\infty f_i(x) \cap -N_{\mathrm{dom}\,\phi}(x) = \{0\}$ at all $x \in \mathrm{dom}\, f_i \cap \mathrm{dom}\,\phi \neq \emptyset$. Let Assumption 4.7 hold true. Then, the iterates $\{(x^t, y^t)\}_{t\in\mathbb{N}}$, generated by Algorithm 3, are well-defined, i.e., $x^{t+1}, y^{t+1} \in \mathrm{int}(\mathrm{dom}\,\phi)^N$, with $y_1^{t+1} = y_2^{t+1} = \cdots = y_N^{t+1} = \sum_{i=1}^N \pi_i x_i^{t+1}$. Assume that f_i is relatively hypoconvex with constant $r_i > 0$. In addition assume that the iterates are bounded. Let $y^{t_j} \to y^*$ be a convergent subsequence of the sequence of iterates such that $y^* \in (\mathrm{int}(\mathrm{dom}\,\phi))^N$. Then, $y_1^* = y_2^* = \cdots = y_N^* = \sum_{i=1}^N \pi_i x_i^*$. Let $r := \max_{1\le i\le N} r_i$ and $0 < \lambda < 1/$. Then $\overleftarrow{\mathrm{env}}_\lambda^\phi f_i$ is continuously differentiable on $\mathrm{int}(\mathrm{dom}\,\phi)$ and we have*

$$\sum_{i=1}^N \pi_i \nabla \overleftarrow{\mathrm{env}}_\lambda^\phi f_i(y_1^{t_j}) \to 0,$$

as $j \to \infty$. In particular this implies that y_1^* is a stationary point of (4.27).

We conclude this section with the remark that one can derive analogous results for the



(a)

(b)

Figure 4.1.: (a) A linearly separable data set with two classes (cross and dot) is partitioned into a cyan and an orange subset. Hyperplanes (with fixed bias) within the orange cone (resp. cyan cone) separate the orange data (resp. cyan data). The full data set is separated by hyperplanes that lie in the intersection of both cones, which reformulates the binary classification problem as a feasibility problem. The same situation arises in modern, highly over-parametrized deep learning applications. (b) Effects of relaxing the finite sum f (black) as a finite sum of Moreau envelopes (purple), which approximately minimizes the original function f, since the minimizers of the individual functions f_1 (orange, dashed) and f_2 (cyan, dashed) are close together. (red) is the proximal average $\mathcal{A}_1$ of (f_1, f_2) with weights π whose Moreau envelope $e_1\mathcal{A}_1$ is the weighted sum of the Moreau envelopes of f_1 and f_2, see Section 3.3

right Bregman envelope starting from the problem

$$\text{minimize}\left\{H_\lambda(x,y) \equiv f(x) + \frac{1}{\lambda}D_\phi(y,x) + g(y) : (x,y) \in \mathbb{R}^m \times \mathbb{R}^m\right\}, \tag{4.28}$$

with $\operatorname{dom}\phi = \mathbb{R}^m$. In this case, we aim to find stationary points of

$$\text{minimize}\left\{\overrightarrow{\text{env}}^{\phi}_{\lambda} f(y) + g(y) : y \in \mathbb{R}^m\right\}, \tag{4.29}$$

via alternating minimization of the upper problem.

4.3. Stochastic averaged Proximal Point with application to federated learning

4.3.1. Feasibility problems and federated learning

In this section we consider a stochastic extension of the Euclidean averaged proximal point method. The resulting algorithm can be interpreted as the Finito/MISO algorithm [DDC14; Mai15] applied to a finite sum of Moreau envelopes. The method is well-suited but not limited to the problem of distributed and federated learning.

In federated learning [KMR15; McM+17; Kai+19] the goal is to train a machine learning model parametrized by a vector $x \in \mathbb{R}^m$ in a collaborative and distributed fashion by a set of N clients. Each client, indexed by i, is associated with a batch

$\mathcal{B}_i \subset \mathbb{R}^n \times \mathbb{R}^d$ of training pairs (source and target) denoted by $(s,t) \in \mathcal{B}_i$ and owns a copy of the parameters $x_i \in \mathbb{R}^m$. Then the clients attempt to minimize an associated (regularized) empirical risk

$$f_i(x_i) = \frac{1}{|\mathcal{B}_i|} \sum_{(s,t)\in\mathcal{B}_i} \ell(\mathcal{H}(s;x_i),t) + R(x_i). \tag{4.30}$$

Here $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a loss function which measures the discrepancy between the target t and the predicted label $\mathcal{H}(s;x)$ of source s. $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$ is a prediction function parametrized by x and $R : \mathbb{R}^m \to \mathbb{R}$ is a regularizer. In addition one enforces (approximate) consensus between the clients and/or an additional server that aggregates the parameters. Under exact consensus this leads to the following optimization problem for weights $\pi \in \operatorname{relint} \Pi$:

$$\min_{x\in(\mathbb{R}^m)^N} \sum_{i=1}^N \pi_i f_i(x_i) + \iota_C(x) = \min_{x\in\mathbb{R}^m} \sum_{i=1}^N \pi_i f_i(x), \tag{4.31}$$

where $C := \big\{x \in (\mathbb{R}^m)^N : x_1 = x_2 = \cdots = x_N\big\}$.

Federated learning involves distributed optimization and therefore inherits its challenges: Typically, the clients are computationally powerful (mobile) devices. The connections between the clients and/or the server, however, may suffer from low bandwidth and/or unstable connections. Therefore a major goal is to reduce communication overhead and allow for broken connections without blocking the joint training progress: Distributed implementations of (stochastic) gradient descent where each client only computes a single gradient which is communicated in each round is therefore not an option. Instead we apply the averaged (Bregman) proximal point iteration from the previous section, where the individual clients perform proximal point steps wrt the associated risks and the server averages the outputs of the proximal mappings after aggregating the results from the clients. In contrast to distributed formulations of (stochastic) gradient descent distributed averaged Proximal Point potentially enables “more progress” within one round if the damping-parameter $\lambda > 0$ in the proximal map is “not too small” and the risks are “similar” in a certain sense. Inexact averaged Proximal Point for federated learning has been considered previously in [Li+20]: There, the proximal term has been interpreted in terms of a re-parametrization of fedAvg [McM+17] which is considered the classical federated learning algorithm. However, [Li+20] do not consider the relations between their method and the proximal point algorithm or Moreau envelopes. As a consequence their formulation is not applicable to nonsmooth problems. In addition, the Proximal Point interpretation offers a unifying view for the formulation and reveals fundamental connections to feasibility problems and the proximal average, see Section 3.3.

We extend the averaged proximal point iteration by means of stochastic sampling of the clients. In federated learning and distributed optimization this is a fundamental feature and allows for randomly interrupted connections between clients and the server. For simplicity we specialize to a Euclidean setting and assume the empirical risks to be prox-regular in a uniform and global sense, i.e., hypoconvex.

As discussed in the previous section the (inexact) averaged proximal point iteration attempts to solve a relaxation of model (4.31), the penalty formulation (4.24). Specialized to a Euclidean setting, i.e., $\phi = 1/2\|\cdot\|^2$, the formulation reads:

$$\min_{x,y\in(\mathbb{R}^m)^N} \left\{ H_\lambda(x,y) \equiv \sum_{i=1}^N \pi_i f_i(x_i) + \frac{\pi_i}{2\lambda}\|x_i - y_i\|^2 + \iota_C(y) \right\}. \tag{4.32}$$

The relaxation of the hard consensus constraint to a soft penalization translates to a certain Moreau smoothing of the original problem; compare to problem (4.27). In the Euclidean setting this smoothed problem reads:

$$\min_{y\in\mathbb{R}^m} \sum_{i=1}^N \pi_i e_\lambda f_i(y). \tag{4.33}$$

This relaxation is particularly useful in a federated learning context: In federated learning we can expect the risks f_i to be similar in a certain sense: In an over-parametrized regime, which is a valid assumption in modern deep learning [Du+18; AZLL19], one can expect the sets of global minimizers $\arg\min f_i$ of the individual risks f_i to have a nonempty intersection:

$$\bigcap_{i=1}^N \arg\min f_i \neq \emptyset. \tag{4.34}$$

In these cases the above relaxation is tight. Indeed, the relaxation is equivalent to the minimization of the Moreau envelope $e_\lambda \mathcal{A}_\lambda(f,\pi)$ of the proximal average $\mathcal{A}_\lambda(f,\pi)$, see Lemma 3.6 and Proposition 3.7(iii) in Section 3.3. In addition this closely resembles the structure of a feasibility problem, see Figure 4.1: Given a collection of closed sets $C_1, \ldots, C_N$, a feasibility problem seeks to

$$\text{find a point } x \text{ in } \bigcap_{i=1}^N C_i\,. \tag{4.35}$$

Via the choice $f_i := \iota_{C_i}$ being indicator functions of the sets C_i, $\pi_i = 1/N$ the feasibility problem can be cast as an instance of Problem (4.33), where the Moreau envelope $e_\lambda \iota_{C_i}$ for $\lambda = 1$ specializes to the squared distance function to the set C_i

$$e_1\iota_{C_i}(y) = \operatorname{dist}^2(y, C_i) := \inf_{x\in C_i} \frac{1}{2}\|x-y\|^2.$$

The averaged proximal point method, in this case, specializes to the averaged projection method which is a classical and successful approach for solving the feasibility problem.

However, as discussed in the previous section, inexact alternating Proximal Point, in general, only converges to a stationary point of the penalty formulation (4.32) but not to a stationary point of the regularized problem (4.33). A sufficient condition for a translation of stationarity is prox-regularity (locally) or hypoconvexity (globally). Prox-regularity and the gradient formula for the Moreau envelope, derived and studied beyond the Euclidean setting in Chapter 2, turn out helpful in the context of federated learning as explored in the next sections.

4.3.2. Variance bounds for the proximal average

In this section, we study and quantify the relation between stationary points of the regularized problem (4.33) and the original problem (4.31) via gradient dissimilarity or gradient variance bounds which goes beyond the qualitative viewpoint from the previous section. These bounds are closely related to the notion of linear regularity of sets for the convex feasibility problem:

Definition 4.10 (linear regularity)**.** *Let $(C_i)_{i=1}^N \subset (\mathbb{R}^m)^N$ be a collection of closed convex sets with nonempty intersection $C := \bigcap_{i=1}^N C_i \neq \emptyset$ and let $\pi \in \operatorname{relint}\Pi$. Let $\kappa > 0$. Then we say the collection $(C_i)_{i=1}^N$ is κ-linearly regular if for any $y \in \mathbb{R}^m$*

$$\operatorname{dist}^2(y, C) \leq \kappa \mathbb{E}_i[\operatorname{dist}^2(y, C_i)],$$

where $\mathbb{E}_i[Y] = \sum_{i=1}^N \pi_i Y_i$ is the expectation of the random variable Y with probabilites $\pi \in \Pi$.

Linear regularity of sets is the standard assumption to guarantee the (linear) convergence of the averaged projections method. In view of [BBL99, Corollary 4.8], if $\bigcap_{i=1}^N C_i$ is bounded, the condition is implied by the standard constraint qualification, the nonemptyness of the intersection of the relative interiors $C = \bigcap_{i=1}^N \operatorname{relint} C_i \neq \emptyset$.

For intuition, the averaged proximal point algorithm converges to a stationary point of (4.33), i.e., $\sum_{i=1}^N \pi_i \nabla e_\lambda f_i(y^t) \to 0$. However, we would rather like to ensure

$$\nabla e_\lambda \left(\sum_{i=1}^N \pi_i f_i\right)(y^t) \to 0,$$

which has the same minimizer as f. Intuitively, this is achieved whenever the gradients $\nabla e_\lambda f_i$ of the Moreau envelopes of the individual functions point in similar directions. Therefore our goal is to quantify this similarity and connect stationary points of the relaxed problem (4.33) with the original problem (4.31).

Indeed, for the feasibility problem linear regularity implies the following condition [NRP19, Theorem 3] which formalizes the above requirement:

Lemma 4.11. *Let $(C_i)_{i=1}^N$ be a collection of closed convex sets and $\pi \in \operatorname{relint}\Pi$. Let $(C_i)_{i=1}^N$ be κ-linearly regular. Then it holds for any $y \in \mathbb{R}^m$ that*

$$\mathbb{E}_i[\|\nabla \operatorname{dist}^2(y, C_i)\|^2] \leq \kappa \left\|\mathbb{E}_i[\nabla \operatorname{dist}^2(y, C_i)]\right\|^2.$$

Proof. For the proof see first part of the proof of [NRP19, Theorem 3] □

More generally, we assume the following condition that we refer to as bounded gradient dissimilarity of the Moreau envelope:

Definition 4.12 (bounded gradient dissimilarity of Moreau envelope)**.** *Let $\varepsilon \geq 0$ and $\kappa > 0$. Then for a collection of functions $(f_i)_{i=1}^N$ the Moreau envelopes $(e_\lambda f_i)_{i=1}^N$ are (ε, κ)-bounded gradient dissimilar if there is $\lambda_0 > 0$ such that for all $\lambda < \lambda_0$ we have*

$$\mathbb{E}_i\big[\|\nabla e_\lambda f_i(y)\|^2\big] \leq \varepsilon + \kappa \big\|\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big]\big\|^2,$$

for all $y \in \mathbb{R}^m$.

While the squared norm of the expected value of the gradients of the Moreau envelopes $\|\mathbb{E}_i[\nabla e_\lambda f_i(y)]\|^2$ is a stationarity measure of the relaxed problem, a small expected value of the squared norms of the gradients of the Moreau envelopes $\mathbb{E}_i[\|\nabla e_\lambda f_i(y)\|^2]$ implies joint near stationarity for all f_i and thus near stationarity with regards to the original problem. Bounded gradient dissimilarity can be reformulated in terms of a variance bound: Let $\mathbb{V}_i[Y] := \mathbb{E}_i[\|Y - \mathbb{E}_i[Y]\|^2] = \mathbb{E}_i[\|Y\|^2] - \|\mathbb{E}_i[Y]\|^2$ denote the variance of a discrete random variable Y with probabilities $\pi_i = \mathbb{P}_i[Y = y_i]$. Then, for $\kappa \geq 1$, the inequality in the definition reads:

$$\mathbb{V}_i[\nabla e_\lambda f_i(y)] \leq \varepsilon + (\kappa - 1)\|\mathbb{E}_i[\nabla e_\lambda f_i(y)]\|^2.$$

To demonstrate that Definition 4.12 is a versatile concept and commonly valid in practice we collect sufficient conditions that imply Definition 4.12:

Firstly we consider functions $(f_i)_{i=1}^N$ that are globally Lipschitz on their domains $(\operatorname{dom} f_i)_{i=1}^N$ which are linearly regular.

Proposition 4.13. *Let $\pi \in \operatorname{relint} \Pi$. Assume that for $i = 1, \ldots, N$ the function $f_i : \mathbb{R}^m \to \overline{\mathbb{R}}$ can be written as $f_i = h_i + \iota_{C_i}$, where h_i is r_i-hypoconvex and globally Lipschitz with constant L on an open set U which contains C_i. Assume that C_i is closed and convex and $(C_i)_{i=1}^N$ is κ-linearly regular. Let $r := \max_{1 \leq i \leq N} r_i$. Then for any $0 < \lambda < 1/r$ the functions $(e_\lambda f_i)_{i=1}^N$ are $((2\kappa + 1)2L^2, 4\kappa)$-bounded gradient dissimilar.*

Proof. Choose $0 < \lambda < 1/r$. Let $\hat{x}_i = P_\lambda f_i(y) = P_\lambda(h_i + \iota_{C_i})(y)$. Since h_i is Lipschitz on the open set U which contains C_i and $\hat{x}_i \in C_i$ we have $\partial^\infty h_i(\hat{x}_i) = \{0\}$. In view of Fermat's rule Lemma 1.3 and Lemma 1.2 we have:

$$0 \in \partial(h_i + \iota_{C_i})(\hat{x}_i) + \frac{1}{\lambda}(\hat{x}_i - y) \subset \partial h_i(\hat{x}_i) + N_{C_i}(\hat{x}_i) + \frac{1}{\lambda}(\hat{x}_i - y).$$

This means in particular that there is $v_i \in \partial h_i(\hat{x}_i)$ such that $0 \in N_{C_i}(\hat{x}_i) + \frac{1}{\lambda}(\hat{x}_i - (y + \lambda v_i))$, implying that

$$\hat{x}_i = \operatorname{proj}_{C_i}(y + \lambda v_i).$$

We have the following estimate:

$$\begin{aligned}\mathbb{E}_i[\|y - \hat{x}_i\|^2] &= \mathbb{E}_i[\|y + \lambda v_i - \hat{x}_i - \lambda v_i\|^2] \\ &\leq 2\mathbb{E}_i[\|y + \lambda v_i - \hat{x}_i\|^2] + 2\lambda^2 \mathbb{E}_i[\|v_i\|^2].\end{aligned}$$

We further bound $\mathbb{E}_i[\|y + \lambda v_i - \hat{x}_i\|^2]$: Exploiting κ-linear regularity via Lemma 4.11, applying triangle and Jensen's inequality we obtain:

$$\begin{aligned}\mathbb{E}_i[\|y + \lambda v_i - \hat{x}_i\|^2] &\leq \kappa \, \|\mathbb{E}_i[y + \lambda v_i - \hat{x}_i]\|^2 \\ &\leq 2\kappa \, \|\mathbb{E}_i[y - \hat{x}_i]\|^2 + 2\kappa\lambda^2 \mathbb{E}_i[\|v_i\|^2] \\ &\leq 2\kappa \, \|\mathbb{E}_i[y - \hat{x}_i]\|^2 + 2\kappa\lambda^2 L^2,\end{aligned}$$

where the last inequality holds since $v_i \in \partial h_i(\hat{x}_i)$, h_i is Lipschitz on U which contains C_i and $\hat{x}_i \in C_i$. We combine both inequalities:

$$\mathbb{E}_i[\|y - \hat{x}_i\|^2] \leq 4\kappa \, \|\mathbb{E}_i[y - \hat{x}_i]\|^2 + (4\kappa + 2)\lambda^2 L^2.$$

We divide the inequality by λ^2 and obtain since h_i is r-hypoconvex and $\lambda < 1/r$ via $\nabla e_\lambda f_i(y) = \lambda^{-1}(y - \hat{x}_i)$ that:

$$\mathbb{E}_i[\|\nabla e_\lambda f_i(y)\|^2] \leq 4\kappa \|\mathbb{E}_i[\nabla e_\lambda f_i(y)]\|^2 + (2\kappa + 1)2L^2. \qquad \square$$

Now, we consider smooth f_i, $i = 1, \ldots, N$ and assume that the functions satisfy a gradient variance bound $\mathbb{V}_i[\nabla f_i] \leq \varepsilon$. For that purpose we prove the following lemma, which relates the gradient of the proximal average $\mathcal{A}_\lambda(f,\pi)$ to the gradient of the (pointwise) arithmetic average $\mathbb{E}_i[f_i]$.

Lemma 4.14. *Let the functions f_i, $i = 1, \dots, N$ be continuously differentiable with L-Lipschitz continuous gradients and assume that $\mathbb{V}_i[\nabla f_i] \leq \varepsilon$. Then, for any $y \in \mathbb{R}^m$ and any $\lambda > 0$ with $1/\lambda > (\sqrt{2}+1)L$ and $\mu := \frac{1}{\lambda} - L$, we have*

$$\left(\frac{1}{2} - \frac{L^2}{\mu^2}\right) \|\mathbb{E}_i[\nabla f_i(y)]\|^2 \leq \frac{L^2\varepsilon}{\mu^2} + \left\|\mathbb{E}_i\left[\nabla e_\lambda f_i(y)\right]\right\|^2 . \tag{4.36}$$

In addition, since $\mathcal{A}_\lambda(f,\pi)$ is continuously differentiable with Lipschitz gradient, we can further bound

$$\left(\frac{\mu^2}{2} - L^2\right) \|\mathbb{E}_i[\nabla f_i(y)]\|^2 \leq L^2\varepsilon + \frac{1}{\lambda^2}\|\nabla \mathcal{A}_\lambda(f,\pi)(y)\|^2 .$$

Proof. Define $\hat{x}_i := P_\lambda f_i(y)$. Further let

$$h_i(x,y) = f_i(x) + \frac{1}{2\lambda}\|x-y\|^2,$$

which is strongly convex (in x) with constant μ. Then we also have that $\hat{x}_i = \arg\min_{x_i \in \mathbb{R}^n} \; h_i(x_i, y)$, and therefore $0 = \nabla_{x_i} h_i(\hat{x}_i, y) = \nabla f_i(\hat{x}_i) + \frac{1}{\lambda}(\hat{x}_i - y)$. Due to μ-strong convexity of $h_i(\cdot, y)$ we have:

$$\frac{\mu}{2}\|y - \hat{x}_i\|^2 \leq \frac{1}{2\mu}\|\nabla_{x_i} h_i(y,y)\|^2 = \frac{1}{2\mu}\|\nabla f_i(y)\|^2. \tag{4.37}$$

Furthermore we have $\nabla e_\lambda f_i(y) = \lambda^{-1}(y - \hat{x}_i) = \nabla f_i(\hat{x}_i)$. Then we have using triangle and Jensen's inequality:

$$\begin{aligned}\frac{1}{2}\left\|\mathbb{E}_i\left[\nabla f_i(y)\right]\right\|^2 &= \frac{1}{2}\left\|\mathbb{E}_i[\nabla f_i(y) - \nabla f_i(\hat{x}_i)] + \mathbb{E}_i\left[\nabla e_\lambda f_i(y)\right]\right\|^2 \\ &\leq \mathbb{E}_i\left[\, \|\nabla f_i(y) - \nabla f_i(\hat{x}_i)\|^2 \,\right] + \left\|\mathbb{E}_i\left[\nabla e_\lambda f_i(y)\right]\right\|^2 .\end{aligned}$$

We have in view of Inequality (4.37):

$$\|\nabla f_i(y) - \nabla f_i(\hat{x}_i)\|^2 \leq L^2\|y - \hat{x}_i\|^2 \leq \frac{L^2}{\mu^2}\|\nabla f_i(y)\|^2.$$

Combining the estimates we obtain:

$$\frac{1}{2}\left\|\mathbb{E}_i\left[\nabla f_i(y)\right]\right\|^2 \leq \frac{L^2}{\mu^2}\mathbb{E}_i\left[\|\nabla f_i(y)\|^2\right] + \|\nabla e_\lambda \mathcal{A}_\lambda(f,\pi)(y)\|^2 .$$

Since $\mathbb{V}_i[\nabla f_i(y)] = \mathbb{E}_i[\|\nabla f_i(y)\|^2] - \|\mathbb{E}_i[\nabla f_i(y)]\|^2 \leq \varepsilon$ we can further bound:

$$\frac{1}{2}\|\mathbb{E}_i[\nabla f_i(y)]\|^2 \leq \frac{L^2\varepsilon}{\mu^2} + \frac{L^2}{\mu^2}\left\|\mathbb{E}_i\left[\nabla f_i(y)\right]\right\|^2 + \left\|\mathbb{E}_i\left[\nabla e_\lambda f_i(y)\right]\right\|^2,$$

which yields

$$\left(\frac{1}{2} - \frac{L^2}{\mu^2}\right) \|\mathbb{E}_i[\nabla f_i(y)]\|^2 \leq \frac{L^2\varepsilon}{\mu^2} + \left\|\mathbb{E}_i\left[\nabla e_\lambda f_i(y)\right]\right\|^2,$$

where the constant $1/2 - L^2/\mu^2 > 0$ for $1/\lambda > (\sqrt{2}+1)L$.

In view of Proposition 3.6 we have:

$$\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big] = \nabla e_\lambda \mathcal{A}_\lambda(f,\pi)(y).$$

Define $\hat{x} := P_\lambda \mathcal{A}_\lambda(f,\pi)(y)$ and

$$h(x,y) = \mathcal{A}_\lambda(f,\pi)(x) + \frac{1}{2\lambda}\|x-y\|^2.$$

which, in view of Proposition 3.7(i), is strongly convex (in the first argument) with constant μ. Then we also have that

$$\hat{x} = \underset{x\in\mathbb{R}^n}{\arg\min}\ h(x,y).$$

By strong convexity of $h(\cdot,y)$ and in view of Proposition 3.7(iv):

$$\begin{aligned}\frac{\mu}{2}\lambda^2\|\nabla e_\lambda \mathcal{A}_\lambda(f,\pi)(y)\|^2 &= \frac{\mu}{2}\,\|y-\hat{x}\|^2\\ &\le \frac{1}{2\mu}\|\nabla_x h(y,y)\|^2\\ &= \frac{1}{2\mu}\|\nabla \mathcal{A}_\lambda(f,\pi)(y)\|^2.\end{aligned}$$ □

Next we invoke the lemma above to obtain the desired result:

Proposition 4.15. *Let the functions f_i, $i=1,\dots,N$ be continuously differentiable with L-Lipschitz continuous gradients and assume that $\mathbb{V}_i[\nabla f_i(x)] \le \varepsilon$ is bounded. Then, for some λ_0 sufficiently small and for any $\lambda < \lambda_0$ the functions $(e_\lambda f_i)_{i=1}^N$ are $(3\varepsilon/2, 3)$-bounded gradient dissimilar.*

Proof. Let $\lambda > 0$ such that $1/\lambda > (\sqrt{2}+1)L$. Adding $\varepsilon/2 - \varepsilon L^2\mu^2$ to both sides of the inequality (4.36) form Lemma 4.14 we obtain:

$$\begin{aligned}\left(\frac{1}{2}-\frac{L^2}{\mu^2}\right)\left(\|\mathbb{E}_i[\nabla f_i(y)]\|^2+\varepsilon\right) &\le \frac{L^2\varepsilon}{\mu^2}+\frac{\varepsilon}{2}-\frac{\varepsilon L^2}{\mu^2}+\big\|\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big]\big\|^2\\ &= \frac{\varepsilon}{2}+\big\|\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big]\big\|^2.\end{aligned}$$

Since $\mathbb{V}_i[\nabla f_i(y)] = \mathbb{E}_i[\|\nabla f_i(y)\|^2] - \|\mathbb{E}_i[\nabla f_i(y)]\|^2 \le \varepsilon$ we can further lower bound the left hand side

$$\left(\frac{1}{2}-\frac{L^2}{\mu^2}\right)\mathbb{E}_i\big[\|\nabla f_i(y)\|^2\big] \le \left(\frac{1}{2}-\frac{L^2}{\mu^2}\right)\left(\|\mathbb{E}_i[\nabla f_i(y)]\|^2+\varepsilon\right)$$

and thus

$$\left(\frac{1}{2}-\frac{L^2}{\mu^2}\right)\mathbb{E}_i\big[\|\nabla f_i(y)\|^2\big] \le \frac{\varepsilon}{2}+\big\|\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big]\big\|^2.$$

Invoking Inequality (4.37) from the proof of the above lemma we can further lower bound the left hand side:

$$\mu^2\lambda^2\mathbb{E}_i\big[\|\nabla e_\lambda f_i(y)\|^2\big] = \mu^2\mathbb{E}_i\big[\|y-\hat{x}_i\|^2\big] \le \mathbb{E}_i\big[\|\nabla f_i(y)\|^2\big].$$

Overall this shows

$$\left(\frac{1}{2}-\frac{L^2}{\mu^2}\right)\mu^2\lambda^2\mathbb{E}_i\big[\|\nabla e_\lambda f_i(y)\|^2\big] \leq \frac{\varepsilon}{2}+\big\|\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big]\big\|^2.$$

Multiplying with 2 and since $\mu = 1/\lambda - L$ we obtain:

$$\left(1-\frac{2L^2}{\mu^2}\right)(1-\lambda L)^2\mathbb{E}_i\big[\|\nabla e_\lambda f_i(y)\|^2\big] \leq \varepsilon+2\big\|\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big]\big\|^2.$$

Obviously for $\lambda \to 0^+$ monotonically decreasing the constant on the left hand side is monotonically increasing and converges to 1. Therefore we can find $\lambda_0 > 0$ so that for any $0 < \lambda < \lambda_0$ we can bound

$$\left(1-\frac{2L^2}{\mu^2}\right)(1-\lambda L)^2 \geq \frac{2}{3}.$$

This shows that

$$\mathbb{E}_i\big[\|\nabla e_\lambda f_i(y)\|^2\big] \leq \frac{3\varepsilon}{2}+3\big\|\mathbb{E}_i\big[\nabla e_\lambda f_i(y)\big]\big\|^2. \qquad \square$$

4.3.3. Stochastic inexact averaged Proximal Point

In this section we consider a stochastic extension of the averaged proximal point iteration, where in each iteration only a randomly chosen subset of proximal mappings is computed. In federated learning, this allows for broken client-server connections which also occur in other distributed optimization settings.

The algorithmic scheme is derived by specializing Algorithm 3 from Section 4.2.3 to the Euclidean setting. For simplicity we eliminate additional blocks $y_2 = \cdots = y_N$ from the product space formulation so that the algorithm can be written as Gauss–Seidel minimization of the following simplified penalty function:

$$\min_{x\in(\mathbb{R}^m)^N, y\in\mathbb{R}^m} \left\{H_\lambda(x,y) \equiv \sum_{i=1}^N \pi_i f_i(x_i) + \frac{\pi_i}{2\lambda}\|x_i - y\|^2\right\}. \tag{4.38}$$

The block coordinate descent formulation of the algorithm suggests the following simple extension to a stochastic setting: In each round we update only a random sample $\mathcal{C}^t \subset \{1,2,\ldots,N\}$ of block coordinates $x_i^{t+1} \approx P_\lambda f_i(y^{t+1})$, $i \in \mathcal{C}^t$ while leaving the other blocks unchanged: $x_i^{t+1} = x_i^t$, $i \notin \mathcal{C}^t$. The complete algorithm is listed in Algorithm 4.

For simplicity, for the remainder of this chapter, we will assume that the functions f_i are prox-regular in a global sense, i.e., r-hypoconvex. Under this assumption the proximal mapping admits a globally single-valued $1/(1-r\lambda)$-Lipschitz proximal mapping, see Proposition 2.46.

Like before we allow the proximal mapping to be evaluated inexactly. The inexact update must satisfy the following conditions:

Assumption 4.16. We define

$$h_i(x,y) := f_i(x) + \frac{1}{2\lambda}\|x-y\|^2.$$

Algorithm 4 stochastic inexact averaged Proximal Point

```
Require: Initialize $\lambda > 0$, $x^0 \in \operatorname{dom} f \subseteq (\mathbb{R}^m)^N$, $y^0 \in \mathbb{R}^m$. Fix probabilities $1 \geq \eta_1, \dots, \eta_N > 0$.
  for all $t = 1, 2, \dots$ do
      $$y^{t+1} = \sum_{i=1}^{N} \pi_i x_i^t \qquad (4.39)$$
    Create a non-empty sample $\mathcal{C}^t \subset \{1, 2, \dots, N\}$ such that $\mathbb{P}[i \in \mathcal{C}^t] := \eta_i > 0$.
    for all $i \in \{1, 2, \dots, N\}$ do
      solve the following inexactly such that Assumption 4.16 holds true:
          $$\hat{x}_i^{t+1} \approx P_\lambda f_i(y^{t+1})$$
          $$x_i^{t+1} = \begin{cases} \hat{x}_i^{t+1}, & \text{if } i \in \mathcal{C}^t, \\ x_i^t, & \text{otherwise.} \end{cases} \qquad (4.40)$$
    end for
  end for
  return $(x^t, y^t)$.
```

(i) There is $\gamma > 0$ such that for each t and $i \in \{1, 2 \dots, N\}$

$$h_i(\hat{x}_i^{t+1}, y^{t+1}) \leq h_i(x_i^t, y^{t+1}) - \frac{\gamma}{2}\|\hat{x}_i^{t+1} - x_i^t\|^2,$$

(ii) There is $\tau > 0$ such that for each t and $i \in \{1, 2 \dots, N\}$

$$\|P_\lambda f_i(y^{t+1}) - x_i^t\| \leq \tau \|\hat{x}_i^{t+1} - x_i^t\|.$$

(i) ensures, that a sufficient descent on h_i is generated, when x_i is updated. The condition is mild. For instance it is satisfied if f_i is smooth and x_i is updated by a single gradient descent step on $h_i(\cdot, y^{t+1})$.

(ii) is an abstraction of a relative error condition of the form: $v^{t+1} \in \partial f_i(\hat{x}_i^{t+1}) + \lambda^{-1}(\hat{x}_i^{t+1} - y^{t+1})$, with $\|v^{t+1}\| \leq \tau \|\hat{x}_i^{t+1} - x_i^t\|$, which can be important in a nonsmooth setting: While relative error implies that $\hat{x}_i^{t+1} = P_\lambda f_i(y^{t+1} + \lambda v^{t+1})$ and therefore via Lipschitz continuity of the proximal mapping also

$$\begin{aligned}\|P_\lambda f_i(y^{t+1}) - x_i^t\| &\leq \|\hat{x}_i^{t+1} - x_i^t\| + \|P_\lambda f_i(y^{t+1}) - \hat{x}_i^{t+1}\| \\ &\leq \|\hat{x}_i^{t+1} - x_i^t\| + \frac{\lambda}{1 - r_i\lambda}\|v^{t+1}\| \leq \left(1 + \frac{\tau\lambda}{1 - r_i\lambda}\right)\|\hat{x}_i^{t+1} - x_i^t\|,\end{aligned}$$

in some settings, e.g., prox-linear updates, Assumption 4.16(ii) still holds while relative error is violated, see [DP19, Section 4] for an explanation and in particular [DP19, Theorem 4.5].

Theorem 4.17. *Let $f_i : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and hypoconvex with constant $r_i \geq 0$. Let $r := \max_{1 \leq i \leq N} r_i$ and choose $\lambda < 1/r$. Let $\pi \in \operatorname{relint} \Pi$. Assume that $\sum_{i=1}^N \pi_i e_\lambda f_i$ is* *bounded from below and* $\mathbb{P}[i \in \mathcal{C}^t] := \eta_i > 0$*. Let Assumption 4.16 hold true. Then we have for the actual iterates* (x^t, y^t)*, generated by Algorithm 4, that* $\|P_\lambda f_i(y^{t+1}) - x_i^t\| \xrightarrow{\text{a.s.}} 0$*, for* $1 \leq i \leq N$ *and* $\|y^{t+1} - y^t\| \xrightarrow{\text{a.s.}} 0$ *and*

$$\sum_{i=1}^{N} \pi_i \nabla e_\lambda f_i(y^t) \xrightarrow{\text{a.s.}} 0,$$

as $t \to \infty$*. Thus every limit point of the sequence of random variables* $\{y^t\}_{t\in\mathbb{N}}$ *is a stationary point of Problem* (4.33) *almost surely.*

Our convergence proof relies on the *Super-Martingale Convergence Theorem* [RS71, Theorem 1]:

Theorem 4.18 (Super-martingale convergence theorem)**.** *Let* $\{w_t\}_{t\in\mathbb{N}}$*,* $\{y_t\}_{t\in\mathbb{N}}$*,* $\{\rho_t\}_{t\in\mathbb{N}}$*, and* $\{\mu_t\}_{t\in\mathbb{N}}$ *be sequences of nonnegative random variables such that*

$$\mathbb{E}\big[w_{t+1} \mid \mathcal{F}^t\big] \leq (1+\rho_t)w_t - y_t + \mu_t, \quad \forall t > 1 \text{ w.p. } 1, \tag{4.41}$$

where $\mathcal{F}^t$ *denotes the collection* $w_1, \dots w_t, y_1 \dots y_t, \rho_1, \dots \rho_t, \mu_1, \dots \mu_t$ *and*

$$\sum_{t=1}^{\infty} \rho_t < \infty, \qquad \sum_{t=1}^{\infty} \mu_t < \infty, \quad \text{w.p. } 1.$$

Then the sequence of random variables $\{w_t\}$ *converges almost surely to a nonnegative random variable, and we have*

$$\sum_{t=1}^{\infty} y_t < \infty \quad \text{w.p. } 1.$$

Proof of Theorem 4.17. We have after the y-update

$$\begin{aligned} H_\lambda(x^t, y^{t+1}) - H_\lambda(x^t, y^t) &\leq \sum_{i=1}^{N} \frac{\pi_i}{2\lambda}\|x_i^t - y^{t+1}\|^2 - \sum_{i=1}^{N} \frac{\pi_i}{2\lambda}\|x_i^t - y^t\|^2 \\ &= \frac{1}{2\lambda}\sum_{i=1}^{N} \pi_i(-2\langle x_i^t, y^{t+1} - y^t\rangle + \|y^{t+1}\|^2 - \|y^t\|^2) \\ &= \frac{1}{2\lambda}(-2\langle y^{t+1}, y^{t+1} - y^t\rangle + \|y^{t+1}\|^2 - \|y^t\|^2) \\ &= -\frac{1}{2\lambda}\|y^{t+1} - y^t\|^2. \end{aligned}$$

We fix $\mathcal{C}^t$ for some iteration index t.

(i) Let $i \in \mathcal{C}^t$. Then we have $x_i^{t+1} = \hat{x}_i^{t+1}$ and by Assumption 4.16(i)

$$h_i(x_i^{t+1}, y^{t+1}) - h_i(x_i^t, y^{t+1}) \leq -\frac{\gamma}{2}\|\hat{x}_i^{t+1} - x_i^t\|^2. \tag{4.42}$$

(ii) Let $i \notin \mathcal{C}^t$: Then we have

$$x_i^{t+1} = x_i^t,$$

and in particular:

$$h_i(x_i^{t+1}, y^{t+1}) - h_i(x_i^t, y^{t+1}) = 0.$$

Taking the weighted sum with weights π_i and rearranging yields:

$$\begin{aligned} H_\lambda(x^{t+1}, y^{t+1}) - H_\lambda(x^t, y^t) &= H_\lambda(x^{t+1}, y^{t+1}) - H_\lambda(x^t, y^{t+1}) + H_\lambda(x^t, y^{t+1}) - H_\lambda(x^t, y^t) \\ &\leq \sum_{i=1}^N \pi_i \left(h_i(x_i^{t+1}, y^{t+1}) - h_i(x_i^t, y^{t+1})\right) - \frac{1}{2\lambda}\|y^{t+1} - y^t\|^2. \end{aligned}$$

We bound $\sum_{i=1}^N \pi_i(h_i(x_i^{t+1}, y^{t+1}) - h_i(x_i^t, y^{t+1}))$

$$\sum_{i=1}^N \pi_i \left(h_i(x_i^{t+1}, y^{t+1}) - h_i(x_i^t, y^{t+1})\right) \leq -\frac{\gamma}{2}\sum_{i\in\mathcal{C}^t} \pi_i\|\hat{x}_i^{t+1} - x_i^t\|^2.$$

We introduce the notation: $\Gamma_{t+1} := H_\lambda(x^{t+1}, y^{t+1})$.

$$\begin{aligned} \Gamma_{t+1} - \Gamma_t &\leq -\frac{\gamma}{2}\sum_{i\in\mathcal{C}^t} \pi_i\|\hat{x}_i^{t+1} - x_i^t\|^2 - \frac{1}{2\lambda}\|y^{t+1} - y^t\|^2 \\ &\leq -\frac{\gamma}{2}\sum_{i\in\mathcal{C}^t} \pi_i\|\hat{x}_i^{t+1} - x_i^t\|^2. \end{aligned}$$

We introduce the notation: $\Delta_i^t := \gamma/2\pi_i\|\hat{x}_i^{t+1} - x_i^t\|^2$. We denote by $\mathcal{F}^t$ the collection $y_1, \ldots y_t, x_1, \ldots, x_t$. We take the expectation $\mathbb{E}_{\mathcal{C}^t}[\,\cdot \mid \mathcal{F}^t]$ on both sides of the inequality. We reorder the summation so that certain terms can be marginalized out. Since y^{t+1} and $\hat{x}_i^{t+1}$ do not depend on $\mathcal{C}^t$ and therefore also Γ_t and Δ_i^t do not dependent on $\mathcal{C}^t$ we obtain:

$$\begin{aligned} \mathbb{E}_{\mathcal{C}^t}\left[\Gamma_{t+1} \,\middle|\, \mathcal{F}^t\right] - \Gamma_t &\leq \sum_{\mathcal{C}\in 2^{\{1,\ldots,N\}}} \mathbb{P}[\mathcal{C}^t = \mathcal{C}] \cdot \sum_{i\in\mathcal{C}^t}(-\Delta_i^t) \\ &= \sum_{i=1}^N(-\Delta_i^t) \sum_{\mathcal{C}\in 2^{\{1,\ldots,N\}}: i\in\mathcal{C}} \mathbb{P}[\mathcal{C}^t = \mathcal{C}] \\ &= -\sum_{i=1}^N \Delta_i^t \cdot \mathbb{P}[\mathcal{C}^t \ni i] = -\sum_{i=1}^N \eta_i\Delta_i^t. \end{aligned}$$

By assumption there is $\Xi > -\infty$ such that

$$-\infty < \Xi \leq \sum_{i=1}^N \pi_i e_\lambda f_i(y^{t+1}) = \inf_{x\in(\mathbb{R}^m)^N} H_\lambda(x, y^{t+1}) \leq H_\lambda(x^{t+1}, y^{t+1}) = \Gamma_{t+1}.$$

Subtracting Ξ from both sides of the inequality and reordering the terms we obtain:

$$\mathbb{E}_{\mathcal{C}^t}\left[\Gamma_{t+1} - \Xi \,\middle|\, \mathcal{F}^t\right] \leq \Gamma_t - \Xi - \sum_{i=1}^N \eta_i\Delta_i^t.$$

We now define $w_t := \Gamma_t - \Xi, \rho_t, \mu_t := 0$ and $y_t := \sum_{i=1}^N \eta_i\Delta_i^t$.

Invoking the Super-martingale Convergence Theorem 4.18 we obtain:

$$\Gamma_t - \Xi \xrightarrow{\text{a.s.}} w^*,$$

where w^* is a nonnegative random variable and

$$y_t = \sum_{i=1}^{N} \eta_i \Delta_i^t < \infty,$$

is summable with probability 1. Since $\eta_i > 0$ this implies that $\Delta_i^t \xrightarrow{\text{a.s.}} 0$ and since $\pi_i > 0$ we have

$$\|\hat{x}_i^{t+1} - x_i^t\| \xrightarrow{\text{a.s.}} 0.$$

Also note that

$$\|y^{t+1} - y^t\| \leq \sum_{i=1}^{N} \pi_i \|x_i^{t+2} - x_i^{t+1}\| = \sum_{j \in \mathcal{C}^{t+1}} \pi_i \|\hat{x}_i^{t+2} - x_i^{t+1}\| \xrightarrow{\text{a.s.}} 0$$

By Assumption 4.16(ii) we have:

$$\|P_\lambda f_i(y^{t+1}) - x_i^t\| \leq \tau \|\hat{x}_i^{t+1} - x_i^t\| \xrightarrow{\text{a.s.}} 0.$$

Then have in view of Jensen's inequality:

$$\begin{aligned} \left\| \sum_{i=1}^{N} \pi_i \nabla e_\lambda f_i(y^{t+1}) \right\| &= \frac{1}{\lambda} \left\| y^{t+1} - \sum_{i=1}^{N} \pi_i P_\lambda f_i(y^{t+1}) \right\| \\ &= \frac{1}{\lambda} \left\| \sum_{i=1}^{N} \pi_i (x_i^t - P_\lambda f_i(y^{t+1})) \right\| \\ &\leq \frac{1}{\lambda} \sum_{i=1}^{N} \pi_i \left\| P_\lambda f_i(y^{t+1}) - x_i^t \right\| \xrightarrow{\text{a.s.}} 0. \end{aligned}$$

Since $\|y^{t+1} - y^t\| \xrightarrow{\text{a.s.}} 0$, by continuity of the gradients of the envelope functions, this implies that every limit point of the sequence $\{y^t\}_{t \in \mathbb{N}}$ is a stationary point of Problem (4.33) almost surely. □

Considering the duality relation between gradient descent and Proximal Point in finite sum minimization from Section 3.4 we obtain an interpretation of the Finito/MISO algorithm [DDC14; Mai15], listed in Algorithm 5, along with a convergence proof in the nonconvex setting under very general sampling strategies:

Corollary 4.19. *Let $f_i : \mathbb{R}^m \to \mathbb{R}$ be Lipschitz differentiable. Assume that $\sum_{i=1}^N \pi_i f_i$ is bounded from below and $\mathbb{P}[i \in \mathcal{C}^t] := \eta_i > 0$. Then we have for the actual iterates $\{y^t\}_{t \in \mathbb{N}}$, generated by Algorithm 5, that*

$$\sum_{i=1}^{N} \pi_i \nabla f_i(y^t) \xrightarrow{\text{a.s.}} 0,$$

as $t \to \infty$. In addition, every limit point of the sequence of random variables $\{y^t\}_{t \in \mathbb{N}}$ is a stationary point of Problem (4.31) *almost surely.*

Proof. This follows by Theorem 3.8 and Theorem 4.17. □

Block coordinate descent interpretations of Finito/MISO, SAGA [DBLJ14] and related algorithms have been observed previously in a nonconvex setting in [Haj+16] for SAGA

Algorithm 5 Finito/MISO algorithm [DDC14; Mai15]

```
Require: Initialize $0 < \lambda < 1/L$, $x^0 \in (\mathbb{R}^m)^N$, $y^0 \in \mathbb{R}^m$. Fix probabilities $1 \geq \eta_1, \dots, \eta_N > 0$.
    for all $t = 1, 2, \dots$ do
        $$y^{t+1} = \sum_{i=1}^{N} \pi_i x_i^t = y^t + \sum_{i \in \mathcal{C}^{t-1}} \pi_i x_i^t - \sum_{i \in \mathcal{C}^{t-1}} \pi_i x_i^{t-1} \qquad (4.43)$$
        Create a non-empty sample $\mathcal{C}^t \subset \{1, 2, \dots, N\}$ such that $\mathbb{P}[i \in \mathcal{C}^t] := \eta_i > 0$.
        $$x_i^{t+1} = \begin{cases} P_\lambda d_\lambda f_i(y^{t+1}) = y^{t+1} - \lambda \nabla f_i(y^{t+1}) & \text{if } i \in \mathcal{C}^t \\ x_i^t, & \text{otherwise.} \end{cases} \qquad (4.44)$$
    end for
    return $y^t$.
```

and in [LTP21] for Finito/MISO. However, in contrast to [LTP21; Haj+16], whose Lyapunov functions are based on the forward-backward envelope [PB13; TSP18] or the closely related augmented Lagrangian, our Lyapunov function is based on the proximal transform.

Invoking our variance bounds for the proximal average from Section 4.3.2, Theorem 4.17 can be refined.

Corollary 4.20. *In the situation of Theorem 4.17 assume that $(e_\lambda f_i)_{i=1}^N$ are (ε, κ)-bounded gradient dissimilar. Then, for any limit point (x^*, y^*) of the sequence of random variables generated by Algorithm 4, we have the following:*

(i) *the feasibility gap $\|x_i^* - y^*\| \leq \lambda\varepsilon/\pi_i$ is bounded by $\lambda\varepsilon/\pi_i$ almost surely,*

(ii) *we have $0 \in \sum_{i=1}^N \pi_i \partial f_i(P_\lambda f_i(y^*))$ almost surely, and*

(iii) *$\|P_\lambda f_i(y^*) - x_i^*\| = 0$ almost surely.*

(iv) *If, in addition, f_i is $\mathcal{C}^1$ with L-Lipschitz continuous gradient, then Items* (i), (ii) *and* (iii) *imply that $\| \sum_{i=1}^N \pi_i \nabla f_i(y^*)\| \leq NL\lambda\varepsilon$ almost surely.*

Proof. By definition of (ε, κ)-bounded gradient dissimilarity we have:

$$\sum_{i=1}^{N} \pi_i \|\nabla e_\lambda f_i(y^t)\|^2 \leq \varepsilon + \kappa \left\| \sum_{i=1}^{N} \pi_i \nabla e_\lambda f_i(y^t) \right\|^2,$$

where $\sum_{i=1}^N \pi_i \nabla e_\lambda f_i(y^t) \to 0$ w.p. 1. Now consider a convergent subsequence of random variables $(x^{t_j}, y^{t_j}) \to (x^*, y^*)$. Due to the continuity of $\nabla e_\lambda f_i$ we have for each individual $i = 1, \dots, N$ almost surely:

$$\|\nabla e_\lambda f_i(y^*)\| \leq \frac{\varepsilon}{\pi_i}.$$

By the gradient formula for $\nabla e_\lambda f_i$ this also means

$$\|P_\lambda f_i(y^*) - y^*\| \leq \lambda \frac{\varepsilon}{\pi_i},$$

for all i almost surely. In view of Theorem 4.17 we have:

$$\|y^{t_j+1} - y^{t_j}\| \xrightarrow{\text{a.s.}} 0.$$

Then since the proximal mapping of f_i is continuous we have

$$P_\lambda f_i(y^{t_j+1}) \to P_\lambda f_i(y^*).$$

Since also $\|P_\lambda f_i(y^{t_j+1}) - x_i^{t_j}\| \xrightarrow{\text{a.s.}} 0$ we have $P_\lambda f_i(y^*) = x_i^*$ almost surely. This implies

$$0 \in \partial f_i(x_i^*) + \frac{1}{\lambda}(x_i^* - y^*),$$

almost surely. We compute the average and reorder the terms:

$$\frac{1}{\lambda}\left(y^* - \sum_{i=1}^N \pi_i x_i^*\right) \in \sum_{i=1}^N \pi_i \partial f_i(x_i^*).$$

We have

$$y^{t_j} - y^{t_j+1} = y^{t_j} - \sum_{i=1}^N \pi_i x_i^{t_j} \xrightarrow{\text{a.s.}} 0 = y^* - \sum_{i=1}^N \pi_i x_i^*,$$

almost surely.

If, in addition, f_i is smooth we have almost surely:

$$\begin{aligned}\left\|\sum_{i=1}^N \pi_i \nabla f_i(y^*)\right\| &= \left\|\sum_{i=1}^N \pi_i \nabla f_i(y^*) + \sum_{i=1}^N \pi_i \nabla f_i(x_i^*) - \sum_{i=1}^N \pi_i \nabla f_i(x_i^*)\right\| \\ &\le \left\|\sum_{i=1}^N \pi_i \nabla f_i(x_i^*)\right\| + \sum_{i=1}^N \pi_i \left\|\nabla f_i(y^*) - \nabla f_i(x_i^*)\right\| \\ &\le \left\|\sum_{i=1}^N \pi_i \nabla f_i(x_i^*)\right\| + \sum_{i=1}^N \pi_i L \left\|y^* - x_i^*\right\|.\end{aligned}$$

And therefore $\|\sum_{i=1}^N \pi_i \nabla f_i(y^*)\| \le \sum_{i=1}^N \pi_i L \|y^* - x_i^*\| \le NL\lambda\varepsilon$ almost surely. □

Corollary 4.20 shows that under bounded gradient dissimilarity of the functions $(e_\lambda f_i)_{i=1}^N$ the deviations between the clients and the server $\|x_i^* - y^*\|$ stay small in the limit and the limit point y^* will get close to a point that is stationary almost surely wrt the original problem (4.31).

We propose the following alternative stochastic variant of the abstract minimization framework. Here, at iteration t, the individual clients perform at least one SGD-step on $H_\lambda(\cdot, y^{t+1})$ to update their copies x_i^{t+1}:

$$x_i^{t+1} = x_i^t - \alpha_t(\nabla f_i(x_i^t; \xi_i^t) + \lambda^{-1}(x_i^t - y^{t+1})),$$

where the step-size α_t and stochastic gradient $\nabla f_i(z_i^t; \xi_i^t)$ satisfy the following assumptions:

Assumption 4.21 (stochastic gradient assumptions).

(i) Let $f_i \in \mathcal{C}^1$ with ∇f_i being L-Lipschitz.

(ii) For all $i = 1, \dots, N$ we have $\mathbb{E}_{\xi_i^t}[\nabla f_i(x_i^t; \xi_i^t)] = \nabla f_i(x_i^t)$.

(iii) There is a nonnegative constant σ so that $\mathbb{V}_{\xi_i^t}[\nabla f_i(x_i^t;\xi_i^t)] \leq \sigma^2$ for all $i = 1, \dots, N$.

(iv) We assume that the step-size α_t is square-summable meaning that: $\sum_{t=1}^{\infty} \alpha_t = +\infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$

The complete algorithm is listed in Algorithm 6:

Algorithm 6 particle stochastic gradient descent

```
Require: Let $t_{\max} \in \mathbb{N}$. Initialize a square-summable sequence $(\alpha_t)_{t=1}^{t_{\max}}$, $\lambda > 0$, $x^0 \in (\mathbb{R}^m)^N$, $y^0 \in \mathbb{R}^m$, possibly $x_i \neq x_j$
  for all $t = 1, 2, \dots, t_{\max}$ do
      $y^{t+1} = \sum_{i=1}^{N} \pi_i x_i^t$    (4.45)
    for all $i \in \{1, 2, \dots, N\}$ do
        $x_i^{t+1} = x_i^t - \alpha_t(\nabla f_i(x_i^t;\xi_i^t) + \lambda^{-1}(x_i^t - y^{t+1}))$    (4.46)
    end for
  end for
  Sample $t^* \in \{1, \dots, t_{\max}\}$ with probability $\mathbb{P}[t = t^*] = \frac{\alpha_{t^*}}{\sum_{t=1}^{t_{\max}} \alpha_t}$
  return $(x^{t^*}, y^{t^*+1})$.
```

The following theorem proves that the proximal gradient of the relaxed model vanishes in mean-square:

Theorem 4.22. *Let f_i be $\mathcal{C}^1$ with L-Lipschitz continuous gradient and hypoconvex with constant $0 \leq r_i \leq L$. Choose $r := \max_{1\leq i\leq N}$ and $\lambda < 1/r$. Assume that $\sum_{i=1}^{N} \pi_i e_\lambda f_i$ is bounded from below. Let Assumption 4.21 hold true. Then we have for the randomly returned iterate $y^{t^*+1} = \sum_{i=1}^{N} \pi_i x_i^{t^*}$ produced by Algorithm 6:*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N} \pi_i \nabla e_\lambda f_i(y^{t^*+1})\right\|^2\right] \to 0,$$

i.e., $\|\sum_{i=1}^{N} \pi_i \nabla e_\lambda f_i(y^{t^*+1})\| \xrightarrow{\text{m.s.}} 0$, as $t_{\max} \to \infty$.

Proof. We have after the y-update

$$\begin{aligned} H_\lambda(x^t, y^{t+1}) - H_\lambda(x^t, y^t) &\leq \sum_{i=1}^{N} \frac{\pi_i}{2\lambda}\|x_i^t - y^{t+1}\|^2 - \sum_{i=1}^{N} \frac{\pi_i}{2\lambda}\|x_i^t - y^t\|^2 \\ &= -\frac{1}{2\lambda}\|y^{t+1} - y^t\|^2. \end{aligned} \tag{4.47}$$

We have the estimate

$$\frac{1}{2}\|y^{t+1} - x_i^{t+1}\|^2 - \frac{1}{2}\|y^{t+1} - x_i^t\|^2 = \frac{1}{2}\|x_i^{t+1} - x_i^t\|^2 - \langle y^{t+1} - x_i^t, x_i^{t+1} - x_i^t\rangle \tag{4.48}$$

Assuming that f_i is L-smooth we have the estimate:

$$f_i(x_i^{t+1}) - f_i(x_i^t) \leq \langle \nabla f_i(x_i^t), x_i^{t+1} - x_i^t \rangle + \frac{L}{2}\|x_i^{t+1} - x_i^t\|^2, \tag{4.49}$$

Recall that

$$h_i(x, y) = f_i(x) + \frac{1}{2\lambda}\|x - y\|^2.$$

Summing the two inequalities above we obtain:

$$h_i(x_i^{t+1}, y^{t+1}) - h_i(x_i^t, y^{t+1}) \leq \langle \nabla_{x_i} h_i(x_i^t, y^{t+1}), x_i^{t+1} - x_i^t \rangle + \frac{L + \lambda^{-1}}{2}\|x_i^{t+1} - x_i^t\|^2.$$

Taking the weighted sum over $i = 1, \ldots, N$ with weights π_i yields using Inequality (4.47):

$$\begin{aligned} &H_\lambda(x^{t+1}, y^{t+1}) - H_\lambda(x^t, y^t) \\ &\quad\leq \langle \nabla_x H_\lambda(x^t, y^{t+1}), x^{t+1} - x^t \rangle + \frac{L + \lambda^{-1}}{2}\sum_{i=1}^N \pi_i \|x_i^{t+1} - x_i^t\|^2 - \frac{1}{2\lambda}\|y^t - y^{t+1}\|^2, \end{aligned} \tag{4.50}$$

where

$$\nabla_x H_\lambda(x^t, y^{t+1}) = (\pi_i(\nabla f_i(x_i^t) + \lambda^{-1}(x_i^t - y^{t+1})))_{i=1}^N.$$

In view of the x_i^{t+1}-update we have

$$x^{t+1} - x^t = -\alpha_t(\nabla f_i(x_i^t; \xi_i^t) + \lambda^{-1}(x_i^t - y^{t+1}))_{i=1}^N =: -\alpha_t G^t. \tag{4.51}$$

Substituting this expression in Inequality (4.50):

$$\begin{aligned} &H_\lambda(x^{t+1}, y^{t+1}) - H_\lambda(x^t, y^t) \\ &\quad\leq -\alpha_t \langle \nabla_x H_\lambda(x^t, y^{t+1}), G^t \rangle + \frac{(L + \lambda^{-1})\alpha_t^2}{2}\sum_{i=1}^N \pi_i \|G_i^t\|^2 - \frac{1}{2\lambda}\|y^t - y^{t+1}\|^2, \end{aligned} \tag{4.52}$$

Taking expectations in this inequality with respect to the distributions of $\xi^t := (\xi_i^t)_{i=1}^N$, and noting that G^t and x^{t+1}, but not x^t or y^{t+1}, depend on ξ^t, we obtain via the unbiasedness of the gradient estimate:

$$\begin{aligned} &\mathbb{E}_{\xi^t}[H_\lambda(x^{t+1}, y^{t+1})] - H_\lambda(x^t, y^t) \\ &\quad\leq -\alpha_t \sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2 + \frac{(L + \lambda^{-1})\alpha_t^2}{2}\sum_{i=1}^N \pi_i \mathbb{E}_{\xi_i^t}[\|G_i^t\|^2] - \frac{1}{2\lambda}\|y^t - y^{t+1}\|^2. \end{aligned}$$

It remains to bound $\mathbb{E}_{\xi_i^t}[\|G_i^t\|^2]$. Since the variance $\mathbb{V}_{\xi_i^t}[\nabla f_i(x_i^t; \xi_i^t)] \leq \sigma^2$ is bounded we obtain the estimate

$$\begin{aligned} \mathbb{E}_{\xi_i^t}[\|G_i^t\|^2] &= \mathbb{V}_{\xi_i^t}[G_i^t] + \|\mathbb{E}_{\xi_i^t}[G_i^t]\|^2 \\ &= \mathbb{V}_{\xi_i^t}[\nabla f_i(x_i^t; \xi_i^t)] + \|\mathbb{E}_{\xi_i^t}[G_i^t]\|^2 \\ &\leq \sigma^2 + \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2. \end{aligned} \tag{4.53}$$

This yields:

$$\begin{aligned}\mathbb{E}_{\xi^t}[H_\lambda(x^{t+1}, y^{t+1})] - H_\lambda(x^t, y^t) \leq -\alpha_t \left(1 - \frac{(L+\lambda^{-1})\alpha_t}{2}\right) \sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2 \\ + \frac{(L+\lambda^{-1})\alpha_t^2\sigma^2}{2} - \frac{1}{2\lambda}\|y^t - y^{t+1}\|^2. \end{aligned} \tag{4.54}$$

We assume that $\alpha_t < 1/(L+\lambda^{-1})$. Then the above can be further bounded

$$\begin{aligned}&\mathbb{E}_{\xi^t}[H_\lambda(x^{t+1}, y^{t+1})] - H_\lambda(x^t, y^t) \\ &\quad \leq -\frac{\alpha_t}{2} \sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2 + \frac{(L+\lambda^{-1})\sigma^2}{2}\alpha_t^2 - \frac{1}{2\lambda}\|y^t - y^{t+1}\|^2. \end{aligned} \tag{4.55}$$

Taking the total expectation we obtain:

$$\begin{aligned}&\mathbb{E}[H_\lambda(x^{t+1}, y^{t+1})] - \mathbb{E}[H_\lambda(x^t, y^t)] \\ &\quad \leq -\mathbb{E}\left[\frac{\alpha_t}{2} \sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2\right] + \frac{(L+\lambda^{-1})\sigma^2}{2}\alpha_t^2. \end{aligned} \tag{4.56}$$

By assumption there is $\Xi > -\infty$ such that

$$-\infty < \Xi \leq \sum_{i=1}^N \pi_i e_\lambda f_i(y^{t+1}) = \inf_{x \in (\mathbb{R}^m)^N} H_\lambda(x, y^{t+1}) \leq H_\lambda(x^{t+1}, y^{t+1}) = \Gamma_{t+1}.$$

Summing we obtain:

$$\begin{aligned}\Xi - H_\lambda(x^0, y^0) &\leq \mathbb{E}[H_\lambda(x^{t_{\max}}, y^{t_{\max}})] - H_\lambda(x^0, y^0) \\ &\leq -\frac{\alpha_t}{2} \sum_{t=1}^{t_{\max}} \mathbb{E}\left[\sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2\right] + \sum_{t=1}^{t_{\max}} \frac{(L+\lambda^{-1})\sigma^2}{2}\alpha_t^2. \end{aligned}$$

We let the algorithm run for $t_{\max}$ iterations and then stop and return with probability $\mathbb{P}[t = t^*] = \alpha_{t^*} / \sum_{t=1}^{t_{\max}} \alpha_t$ one of the previous iterates (x^{t^*}, y^{t^*+1}) with $t^* \in \{1, \dots, t_{\max}\}$. Then we have for the complete expected value of $\sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1})\|^2$

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[\sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1})\|^2\right] &= \frac{1}{2}\sum_{t=1}^{t_{\max}} \mathbb{P}[t = t^*] \cdot \mathbb{E}\left[\sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2\right] \\ &= \frac{1}{\sum_{t=1}^{t_{\max}} \alpha_t} \cdot \sum_{t=1}^{t_{\max}} \frac{\alpha_t}{2} \cdot \mathbb{E}\left[\sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^t, y^{t+1})\|^2\right] \\ &\leq \frac{H_\lambda(x^0, y^0) - \Xi + \sum_{t=1}^{t_{\max}} \frac{(L+\lambda^{-1})\sigma^2}{2}\alpha_t^2}{\sum_{t=1}^{t_{\max}} \alpha_t} \\ &\xrightarrow{t_{\max}\to\infty} 0, \end{aligned} \tag{4.57}$$

which, due to the square-summability of α_t, goes to zero for $t_{\max} \to \infty$.

By definition of $\nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1})$ we have

$$\nabla f_i(x_i^{t^*}) + \lambda^{-1}(x_i^{t^*} - (y^{t^*+1} + \lambda\nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1})) = 0.$$

Due to the hypoconvexity of f_i with constant L and since $\lambda < 1/L$ this means that

$$x_i^{t^*} = P_\lambda f_i(y^{t^*+1} + \lambda \nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1})).$$

In addition we have

$$y^{t^*+1} = \sum_{i=1}^N \pi_i x_i^{t^*}.$$

Then we have, since $P_\lambda f_i$ is $1/(1-L\lambda)$-Lipschitz and due to Jensen's inequality and $\pi_i > 0$:

$$\begin{aligned}
\left\|\sum_{i=1}^N \pi_i \nabla e_\lambda f_i(y^{t^*+1})\right\|^2 &= \left\|\sum_{i=1}^N \pi_i \frac{1}{\lambda}(y^{t^*+1} - P_\lambda f_i(y^{t^*+1}))\right\|^2 \\
&= \frac{1}{\lambda^2}\left\|\sum_{i=1}^N \pi_i (P_\lambda f_i(y^{t^*+1}) - x_i^{t^*})\right\|^2 \\
&\leq \frac{1}{\lambda^2}\sum_{i=1}^N \pi_i \left\|P_\lambda f_i(y^{t^*+1}) - P_\lambda f_i(y^{t^*+1} + \lambda \nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1}))\right\|^2 \\
&\leq \frac{1}{(1-L\lambda)^2}\sum_{i=1}^N \pi_i \|\nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1})\|^2.
\end{aligned}$$

Taking expectations on both sides of the inequality shows via (4.57) that

$$\sum_{i=1}^N \pi_i \nabla e_\lambda f_i(y^{t^*+1}) \xrightarrow{\text{m.s.}} 0,$$

converges to 0 in mean-square. □

A particularly interesting special case is obtained, when all clients are initialized differently $x_i \neq x_j$, but sample from the same training data in each iteration. In that case the algorithm specializes to a Gauss–Seidel variant of the vanilla *Elastic Averaging SGD* (EASGD) [ZCL15] which can be interpreted as a consensus method with interacting particles [RB12; Bor+21]. The following corollary shows, that when $\lambda > 0$ is chosen sufficiently small consensus between the clients is attained in mean-square in the limit:

Corollary 4.23. *In the situation of Theorem 4.22 assume that $(e_\lambda f_i)_{i=1}^N$ is $(0,\kappa)$ gradient dissimilar. Then we have*

$$\left\|\sum_{i=1}^N \pi_i \nabla f_i(y^{t^*+1})\right\| \xrightarrow{\text{m.s.}} 0,$$

and in particular $\|x_i^{t^*} - y^{t^*+1}\| \xrightarrow{\text{m.s.}} 0$, i.e., consensus in mean-square is attained in the limit for $t_{\max} \to \infty$.

Proof. Define $\hat{x}_i^{t^*+1} := P_\lambda f_i(y^{t^*+1})$. The corresponding optimality condition reads:

$$0 = \nabla f_i(\hat{x}_i^{t^*+1}) + \frac{1}{\lambda}(\hat{x}_i^{t^*+1} - y^{t^*+1}).$$

Summing over $i = 1, \dots, N$ yields:

$$-\sum_{i=1}^{N} \pi_i \left(\frac{1}{\lambda}(\hat{x}_i^{t^*+1} - y^{t^*+1}) + \nabla f_i(\hat{x}_i^{t^*+1}) - \nabla f_i(y^{t^*+1}) \right) = \sum_{i=1}^{N} \pi_i \nabla f_i(y^{t^*+1}).$$

We take square norms on both sides and upper bound the right-hand side. Via triangle and Jensen's inequality and L-Lipschitz continuity of ∇f_i we obtain:

$$\left\| \sum_{i=1}^{N} \pi_i \nabla f_i(y^{t^*+1}) \right\|^2 \leq 2(1 + L^2\lambda^2) \sum_{i=1}^{N} \pi_i \frac{1}{\lambda^2} \|\hat{x}_i^{t^*+1} - y^{t^*+1}\|^2$$

By $(0, \kappa)$-bounded gradient dissimilarity of $(e_\lambda f_i)_{i=1}^N$ we have that:

$$\sum_{i=1}^{N} \pi_i \|\nabla e_\lambda f_i(y^{t^*+1})\|^2 \leq \kappa \left\| \sum_{i=1}^{N} \pi_i \nabla e_\lambda f_i(y^{t^*+1}) \right\|^2,$$

and therefore we can further bound:

$$\left\| \sum_{i=1}^{N} \pi_i \nabla f_i(y^{t^*+1}) \right\|^2 \leq 2(1 + L^2\lambda^2)\kappa \left\| \sum_{i=1}^{N} \pi_i \nabla e_\lambda f_i(y^{t^*+1}) \right\|^2.$$

Taking expectations on both sides of the inequality yields via Theorem 4.22 that

$$\sum_{i=1}^{N} \pi_i \nabla f_i(y^{t^*+1}) \xrightarrow{\text{m.s.}} 0,$$

converges to 0 in mean-square. In addition, via $(0, \kappa)$-bounded gradient dissimilarity of $(e_\lambda f_i)_{i=1}^N$, we also have that

$$\sum_{i=1}^{N} \pi_i \mathbb{E}\big[\|\nabla e_\lambda f_i(y^{t^*+1})\|^2\big] \to 0,$$

which means that for all $i = 1, \dots, N$ we have $\nabla e_\lambda f_i(y^{t^*+1}) \xrightarrow{\text{m.s.}} 0$ and therefore $\|\hat{x}_i^{t^*+1} - y^{t^*+1}\| \xrightarrow{\text{m.s.}} 0$. Then we have

$$\begin{aligned} \|x_i^{t^*} - y^{t^*+1}\|^2 &\leq 2\|\hat{x}_i^{t^*+1} - y^{t^*+1}\|^2 + 2\|\hat{x}_i^{t^*+1} - x_i^{t^*}\|^2 \\ &= 2\|\hat{x}_i^{t^*+1} - y^{t^*+1}\|^2 + 2\|P_\lambda f_i(y^{t^*+1}) - P_\lambda f_i(y^{t^*+1} + \lambda \nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1}))\|^2 \\ &\leq 2\|\hat{x}_i^{t^*+1} - y^{t^*+1}\|^2 + \frac{2\lambda^2}{(1 - L\lambda)^2} \|\nabla_{x_i} h_i(x_i^{t^*}, y^{t^*+1})\|^2. \end{aligned}$$

Taking the expectation on both sides yields:

$$\|x_i^{t^*} - y^{t^*+1}\| \xrightarrow{\text{m.s.}} 0,$$

converges to 0 in mean-square. □

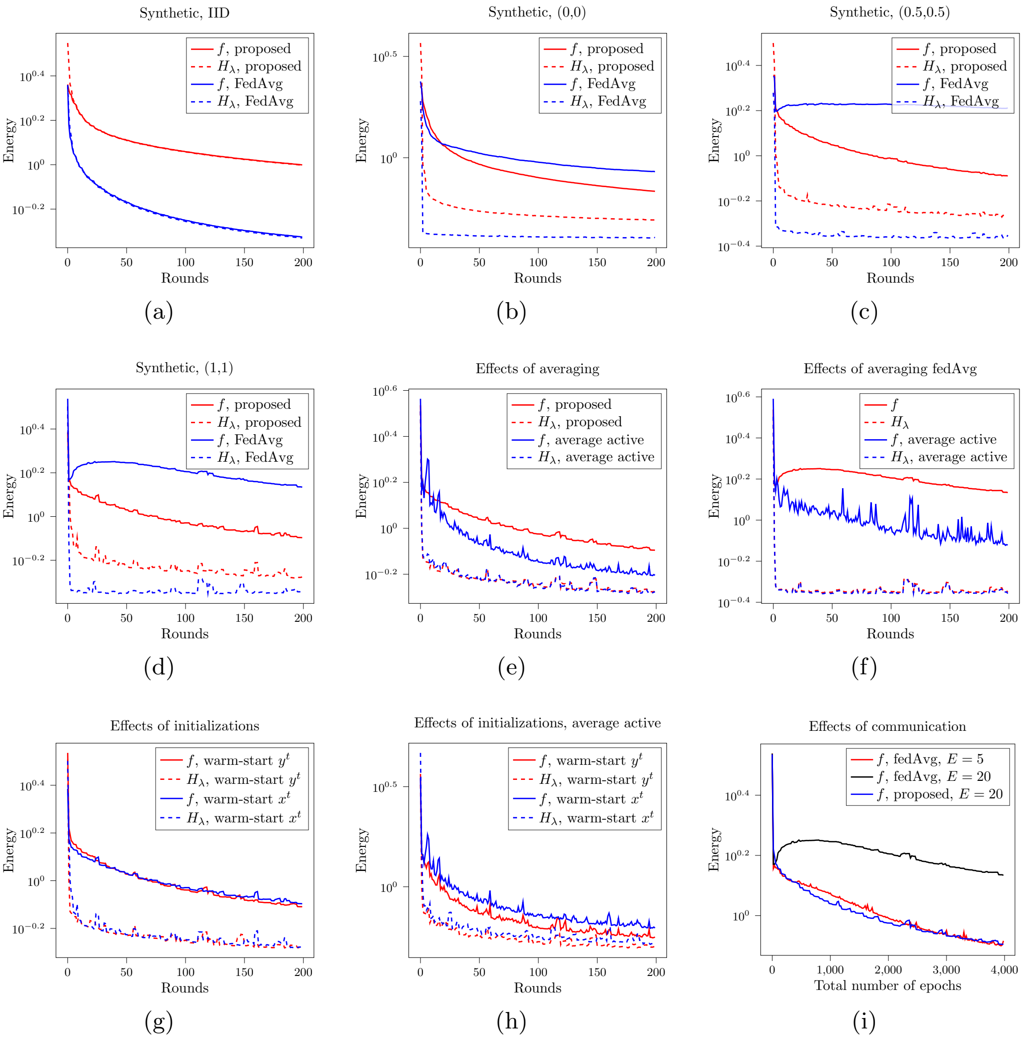


Figure 4.2.: Comparison of fedAvg, [McM+17, Algorithm 1] and our methods Algorithm 4 and Algorithm 6. While in the IID-setting (a) our method is outperformed by fedAvg, in the non-IID cases (b)–(d) our method achieves a superior performance. (e) and (f) show a comparison of the different averaging schemes. (g) and (h) show the effect of warm-starting the clients at y^t rather than x^t under different averaging schemes. In (i) we plot $f(y^t)$ and $H_\lambda(x^t, y^t)$ over the total number of epochs taken by the individual clients for the non-IID dataset Synthetic, $(1, 1)$: While our method with $E = 20$ and fedAvg with $E = 5$ achieve similar performances after the same amount of epochs, our approach requires 4 times less communication. fedAvg with the same amount of communication $E = 20$ performs worse.

4.4. Averaged Proximal Point for federated learning: Numerical results

4.4.1. Logistic regression under heterogeneity

In this experiment we complement the experimental results from [Li+20, Figure 2]: Like [Li+20] we consider a quadratically regularized multinomial logistic regression problem,

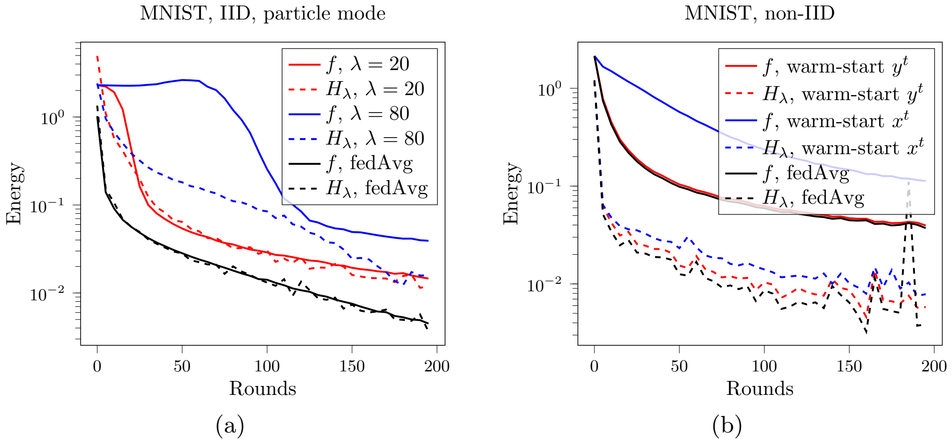


Figure 4.3.: Comparison of fedAvg, [McM+17, Algorithm 1] and our methods Algorithm 4 and Algorithm 6 on the task of distributed training of a convolutional neural network. In (a) we distribute the data in an IID fashion so that each client is assigned $|\mathcal{B}_i| = 3750$ examples. We initialize the clients at random $x_j^0 \neq x_i^0 \neq y^0$ and warm-start the local clients at x^t rather than y^t after each round. As our theory predicts, if λ is chosen sufficiently small the clients attain approximate consensus at y^{200} after 200 rounds by themselves. However, the convergence is overall slower compared to fedAvg. In (b) we distribute the training data in a non-IID fashion where each client is assigned examples of two classes only. We choose $\lambda = 50$ and warm-start the local solvers at y^t after each round. It can be seen that in this setting our method and fedAvg achieve comparable performance.

where $\ell(x; s, t)$ is the multinomial logistic loss and R is a quadratic regularizer weighted by ν (excluding the bias weights): Each client optimizes:

$$f_i(x_i) = \frac{1}{|\mathcal{B}_i|} \sum_{(s,t)\in\mathcal{B}_i} \ell(x_i; s, t) + R(x_i),$$

with $\nu = 10^{-4}$. We define

$$f(x) := \sum_{i=1}^N \pi_i f_i(x).$$

We consider the synthetic federated learning benchmark[2] provided by [Li+20] which comprises 4 datasets with increasingly heterogeneous and unbalanced splits. Each dataset has $M = 9600$ examples (s, t) with 10 classes and $s \in \mathbb{R}^{D-1}$, $D = 61$ and $t \in \{1, 2, \dots, 10\}$ and comes alongs with a split into $N = 30$ subsets $\mathcal{B}_i$. We choose the weights $\pi_i := |\mathcal{B}_i|/M$. For the precise details of the generation of the dataset we refer to [Li+20, Section 5.1]. We apply Algorithm 4 and select 10 (out of $N = 30$) clients per round that approximately solve the proximal subproblem $P_\lambda f_i(y^{t+1})$ in the fashion of Algorithm 6 performing $E = 20$ epochs of mini-batch SGD with constant step-size $\alpha_t = 0.01$ and batch-size 20. We stop the algorithm after 200 rounds. In Figures 4.2a–4.2d we compare the objective values $\sum_{i=1}^N \pi_i f_i(y^t)$ and $H_\lambda(x^t, y^t)$ of our method (initialized randomly with $x_i^0 \neq x_j^0 \neq y^0$) and fedAvg [McM+17, Algorithm 1]. fedAvg happens to be a special case

[2] https://github.com/litian96/FedProx

of our method, where $\lambda = \infty$ and the individual clients are warm-started at y^t after each round. Both methods use the same averaging scheme to obtain the consensus variable y^t. Our method in the smooth setting is closely related to fedProx [Li+20]. However, there exist some key differences: While the same proximal damping terms are used, fedProx invokes a different averaging scheme

$$y^t = \frac{1}{|\mathcal{C}^t|} \sum_{i \in \mathcal{C}^t} x_i^{t+1}, \tag{4.58}$$

which involves the active clients $i \in \mathcal{C}^t$ only. In addition, [Li+20] select the set of active clients according to the probability distribution given by the weights π, while our method selects the clients according to an arbitrary probability distribution η. Indeed, while fedProx can be interpreted in terms of classical SGD applied to the finite sum of Moreau envelopes $\sum_{i=1}^N \pi_i e_\lambda f_i$ our method can be interpreted as the Finito/MISO algorithm [DDC14; Mai15] applied to $\sum_{i=1}^N \pi_i e_\lambda f_i$. Still, our experimental results are in accordance with the experimental findings reported by [Li+20]: While in the IID-setting (Figure 4.2a) our method is outperformed by fedAvg, in the non-IID cases (Figures 4.2b–4.2d) our method achieves a superior performance. Note, that the gap between H_λ and f is a strong indicator for the heterogeneity of the training data and therefore can be used to drive adaptive schemes for the damping parameter λ: In particular in the IID-case (Figure 4.2a) the gap between H_λ and f is small and therefore a proximal damping term is superfluous, and just slows down convergence. Figures 4.2e and 4.2f show a comparison of the different averaging schemes. While our averaging scheme leads to a more robust convergence behavior, averaging the active clients only, results in a better overall performance. Figures 4.2g and 4.2h show the effect of warm-starting the clients at y^t rather than x_i^t. In Figure 4.2i we plot $f(y^t)$ and $H_\lambda(x^t, y^t)$ over the total number of epochs taken by the individual clients for the non-IID dataset Synthetic, $(1, 1)$. It can be seen that our method with $E = 20$ achieves a comparable performance as fedAvg with $E = 5$, however, with reduced communication overhead.

4.4.2. Federated learning for neural networks

We also conduct an experiment using a convolutional neural network. We report results of the distributed training of a nonlinear convolutional neural network classifier on the MNIST dataset using our algorithms resorting to the standard LeNet-5 CNN architecture [LeC+98] given as

$$\begin{aligned}&\mathrm{Conv}_{20,5,1} \to \mathrm{ELU} \to \mathrm{AvgPool}_{2,2} \to \mathrm{Conv}_{50,5,1} \to \mathrm{ELU} \to \mathrm{AvgPool}_{2,2} \to \mathrm{FC}\\ &\to \mathrm{Softmax},\end{aligned}$$

with smooth ELU activation functions and a smooth cross-entropy loss $\ell(x; s, t)$, $s \in \mathbb{R}^{28\times 28}$ and $t \in \{1, \ldots, 10\}$. Each client then optimizes:

$$f_i(x_i) = \frac{1}{|\mathcal{B}_i|} \sum_{(s,t)\in\mathcal{B}_i} \ell(x_i; s, t) + R(x_i),$$

where R is a quadratic regularizer weighted by ν. We choose $\nu = 10^{-6}$ and fix the number of clients $N = 16$ where the individual clients are assigned disjoint subsets $\mathcal{B}_i$ of the $M = 60.000$ training examples and the weights are chosen as $\pi_i = |\mathcal{B}_i|/M$. Within each round we invoke 8 clients chosen at random that perform 150 steps of mini-batch SGD with batch-size 20 and constant step-size $\alpha_t = 0.05$ to solve the proximal subproblem.

The maximum number of rounds is 200. In Figure 4.3a we distribute the data in an IID fashion so that each client is assigned $|\mathcal{B}_i| = 3750$ examples. We initialize the clients at random $x_j^0 \neq x_i^0 \neq y^0$ and warm-start the local clients at x_i^t rather than y^t after each round. In accordance with [Bor+21] we refer to this as the particle mode, as each client aka particle follows its own path to a potentially different limit point. As suggested by our theory, if λ is chosen sufficiently small the clients attain approximate consensus at u^{200} after 200 rounds. However, the convergence is overall slower compared to fedAvg. In Figure 4.3b we distribute the training data in a non-IID fashion where each client is assigned examples of two classes only. We choose $\lambda = 50$ and warm-start the local solvers at y^t after each round. It can be seen that in this setting our method and fedAvg achieve similar performance.

Part II.

Convex relaxation by lifting and dual discretization

Chapter 5.

Lifting and generalized conjugacy in Lagrangian relaxations

5.1. Lagrangian relaxation for Markov Random Fields

The relaxation of the finite sum problem to a finite sum of Moreau envelopes, studied in the previous Section 4.3, yields a lower envelope to the original problem whose minimization is inherently parallelizable via the decomposability of the proximal mapping of the proximal average, see Section 3.3. In addition one obtains a smooth problem whose stationary points are provably near stationary wrt the original problem whenever the variance of the gradients of the Moreau envelopes of the individual functions is small around those points, see Corollary 4.20.

In this chapter a somewhat complementary approach to obtain a lower relaxation is proposed. The approach is based on the Lagrangian relaxation paradigm, and, to remedy duality gaps in nonconvex problems, a certain reformulation over the space of measures or moments is considered. In contrast to the previous approach based on Moreau smoothing, this results in a nonsmooth but convex lower envelope which falls within the regime of highly parallelizable convex optimization tools such as the *primal-dual hybrid-gradient* (PDHG) method [CP11]. In addition we discuss a connection between lower relaxations based on Moreau envelopes and lifting in Lagrangian relaxations via generalized conjugate functions, see Definition 3.1, which are used in Section 3.3 to construct the proximal average.

For the remainder of this chapter we focus on MAP-inference in a pairwise continuous MRF and spatially continuous variational problems with *total variation* (TV) regularization. This chapter is based on [Lau+16] and [Bau+21]. Expanding upon these works a goal of this chapter is to discuss a unifying formulation for both, spatially continuous variational problems and MAP-inference in a continuous MRF.

For an undirected graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V}$ and edges $\mathcal{E}$ MAP-inference in a pairwise continuous MRF amounts to solving the following optimization problem:

$$\min_{x \in (\mathbb{R}^m)^{\mathcal{V}}} \left\{ F(x) \equiv \sum_{u \in \mathcal{V}} f_u(x_u) + \sum_{uv \in \mathcal{E}} f_{uv}(x_u, x_v) \right\}, \tag{5.1}$$

for proper lsc unary functions $f_u : \mathbb{R}^m \to \overline{\mathbb{R}}$ and proper lsc pairwise symmetric functions $f_{uv} : \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$, i.e., $f_{uv}(x_u, x_v) = f_{uv}(x_v, x_u)$. Note that this formulation subsumes the finite sum problem, which is recovered if the graph $\mathcal{G}$ is connected and the pairwise functions are chosen $f_{uv}(x_u, x_v) = \iota_{\{0\}}(x_u - x_v)$.

The classical Lagrangian relaxation is derived as follows: For each $uv \in \mathcal{E}$ one introduces an auxiliary variable x_{uv} and linear constraints $x_{uv} = (x_u, x_v)$. Dualizing these constraints with Lagrange multipliers $\lambda_{uv} \in \mathbb{R}^m \times \mathbb{R}^m$ one obtains a convex optimization problem. This dual problem is amenable to convex optimization tools such as the aforementioned PDHG method or the *alternating direction method of multipliers* (ADMM) [Boy+11] which can exploit the partially separable structure of the objective. However, the approach often suffers form large duality gaps. Indeed, Fenchel–Rockafellar duality (which holds under suitable qualification conditions) applied to the Lagrangian dual problem yields the following primal problem:

$$\min_{x \in (\mathbb{R}^m)^{\mathcal{V}}} \sum_{u \in \mathcal{V}} f_u^{**}(x_u) + \sum_{uv \in \mathcal{E}} f_{uv}^{**}(x_u, x_v), \tag{5.2}$$

where f_u^{**} are the convex biconjugates, i.e., the largest lsc convex functions below f_u. For nonconvex f_u such component-wise convex envelopes, however, are often inaccurate approximations to the original problem and are trivial in some cases.

5.2. Lagrangian relaxation in measure spaces

5.2.1. Every minimization problem is convex

To remedy duality gaps in Lagrangian relaxations we propose to reformulate the original problem in terms of an infinite-dimensional linear program over the space of Radon measures and apply the Lagrangian relaxation after this reformulation. In accordance with Section 1.2.7, we therefore adopt a different notation and write $f_u : X \to \mathbb{R}$ and $f_{uv} : X \times X \to \mathbb{R}$ as finite-valued functions to be defined on $\operatorname{dom} f_u =: X \subset \mathbb{R}^m$. In addition we assume X is nonempty and compact.

The following lemma reveals, that essentially every minimization problem can be equivalently formulated in terms of an infinite-dimensional linear program.

Lemma 5.1. *Let $f : \mathcal{X} \to \mathbb{R}$ be lsc with $\emptyset \neq \mathcal{X} \subset \mathbb{R}^m$ compact. Then we have*

$$\min_{x \in \mathcal{X}} f(x) = \min_{\mu \in \mathcal{P}(\mathcal{X})} \langle \mu, f \rangle, \tag{5.3}$$

and $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$ is a solution to $\min_{x \in \mathcal{X}} f(x)$ if and only if δ_{x^*} is a minimizer of $\min_{\mu \in \mathcal{P}(\mathcal{X})} \langle \mu, f \rangle$.

Proof. The result follows immediately via compactness of $\mathcal{X}$ and the properties of probability measures: By compactness of $\mathcal{X}$ and since f is lsc relative to $\mathcal{X}$ we know x^* exists. Let $f_{\min} = f(x^*)$. By the properties of the Lebesgue integral we have:

$$\int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) \geq \int_{\mathcal{X}} f_{\min}\,\mathrm{d}\mu(x) = f_{\min} \int_{\mathcal{X}} 1\,\mathrm{d}\mu(x) = \min_{x \in \mathcal{X}} f(x),$$

for all $\mu \in \mathcal{P}(\mathcal{X})$ and $\min_{x \in \mathbb{R}^m} f(x) = \langle \delta_{x^*}, f \rangle = f(x^*)$. □

The result shows that convex optimization is not easier than nonconvex optimization. Yet, the formulation turns out helpful to derive more tractable convex reformulations in certain cases: For instance suppose that the objective function belongs to a certain subspace such as the space of polynomials. Thanks to convex algebraic geometry the problem can be reformulated over the space of moments for which there exist tractable characterizations in terms of SDP in some cases. Indeed, this is the starting point in Lasserre's approach [Las01; Las02] for constrained polynomial optimization.

5.2.2. MAP-inference in a pairwise MRF: The local marginal polytope relaxation

Let $\pi_u : X^{\mathcal{V}} \to X$, $\pi_{uv} : X^{\mathcal{E}} \to X^2$ denote the canonical projections onto the u^{th} resp. u^{th} and v^{th} components and $\pi_u \sharp \mu$ is the pushforward of μ w.r.t. π_u defined by: $(\pi_u \sharp \mu)(A) = \mu((\pi_u)^{-1}(A))$ for all $A \subset X$ in the corresponding σ-algebra. Then, in our case, applying the reformulation from Lemma 5.1 directly to the cost function F with $\mathcal{X} = X^{\mathcal{V}}$ we obtain by linearity:

$$\begin{aligned}\min_{\mu \in \mathcal{P}(X^{\mathcal{V}})} \langle \mu, F \rangle &= \min_{\mu \in \mathcal{P}(X^{\mathcal{V}})} \left\langle \mu, \sum_{u \in \mathcal{V}} f_u \circ \pi_u + \sum_{uv \in \mathcal{E}} f_{uv} \circ \pi_{uv} \right\rangle \\ &= \min_{\mu \in \mathcal{P}(X^{\mathcal{V}})} \sum_{u \in \mathcal{V}} \langle \pi_u \sharp \mu, f_u \rangle + \sum_{uv \in \mathcal{E}} \langle \pi_{uv} \sharp \mu, f_{uv} \rangle.\end{aligned}$$

This relaxation is known as the *full marginal polytope relaxation* which is, however, intractable if $\mathcal{V}$ is large as one minimizes over probability measures on the product space $\mathcal{P}(X^{\mathcal{V}})$. Instead, we consider the following linear programming relaxation of (5.1) which is also referred to as the *local marginal polytope relaxation* [Pen+11; FA14; WG14; Ruo15], which is more tractable as the optimization variable lies in the product space of probability measures $\mathcal{P}(X)^{\mathcal{V}}$:

$$\inf \left\{ \sum_{u \in \mathcal{V}} \langle \mu_u, f_u \rangle + \sum_{uv \in \mathcal{E}} \langle \mu_{uv}, f_{uv} \rangle : \mu_u \in \mathcal{P}(X), \mu_{uv} \in \Pi(\mu_u, \mu_v), \forall\, u \in \mathcal{V}, uv \in \mathcal{E} \right\}. \tag{5.4}$$

The constraint set $\Pi(\mu_u, \mu_v)$ consists of all Borel probability measures on X^2 with marginals μ_u and μ_v:

$$\Pi(\mu_u, \mu_v) = \left\{ \mu_{uv} \in \mathcal{P}(X^2) : \pi_u \sharp \mu_{uv} = \mu_u,\ \pi_v \sharp \mu_{uv} = \mu_v \right\}, \tag{5.5}$$

where for $u \in \mathcal{V}$ we denote by $\pi_u : X \times X \to X$ the canonical projection onto the u^{th} component.

Note that for finite state-spaces, i.e., $|X| < \infty$, the set $\mathcal{P}(X)$ can be identified with the standard $(|X| - 1)$-dimensional probability simplex and $\Pi(\mu_u, \mu_v)$ with the set of nonnegative $|X| \times |X|$ matrices whose rows and columns sum up to μ_v and μ_u. In that case, the linear program given in (5.4) is equivalent to the usual finite-dimensional local marginal polytope relaxation for MRFs, which is for example studied in [Wer07]. We further remark that the case of finite X has been extensively studied in the literature, see, e.g., [Kap+13; WJ+08] for recent overviews.

For the more challenging setting of continuous state-spaces, a major difficulty stems from the fact that the linear programming relaxation (5.4) is an infinite-dimensional optimization problem posed in the space of Borel probability measures. Perhaps due to this difficulty, discrete MRF approaches are still routinely applied despite the continuous nature of X. This is typically done by considering a finite sample approximation of X, so that the infinite-dimensional linear program reduces to a finite-dimensional one. This, however, may lead to discretization errors and comes with an exponential overhead in the dimension of X.

Note that for fixed μ_u and μ_v we may absorb the minimization over the variable $\mu_{uv} \in \Pi(\mu_u, \mu_v)$ into an optimal transportation problem [Kan60],

$$\mathrm{OT}_{f_{uv}}(\mu_u, \mu_v) = \inf_{\mu_{uv} \in \Pi(\mu_u, \mu_v)} \langle f_{uv}, \mu_{uv} \rangle, \tag{5.6}$$

with marginals (μ_u, μ_v) and cost f_{uv}. This allows us to rewrite the optimization problem (5.4) more compactly as follows:

$$\inf_{\mu \in \mathcal{P}(X)^{\mathcal{V}}} \left\{ \mathcal{F}(\mu) := \sum_{u \in \mathcal{V}} \langle f_u, \mu_u \rangle + \sum_{uv \in \mathcal{E}} \mathrm{OT}_{f_{uv}}(\mu_u, \mu_v) \right\}. \tag{5.7}$$

Due to the fact that $\Pi(\delta_x, \delta_{x'}) = \{\delta_{(x,x')}\}$ one sees that restricting μ_u (and therefore μ_{uv}) to be Dirac probability measures, the formulation (5.7) reduces to the original problem (5.1). As one instead considers the larger convex set of all probability measures, it is a *relaxation* which lower bounds (5.1), i.e., we have the following important relation:

$$(5.7) \leq (5.1). \tag{5.8}$$

When considering a relaxation, the immediate question arises whether this lower bound is attained, i.e., the relaxation is tight and therefore the above inequality holds with equality. For finite and ordered X the situation is well-understood, see [Wer07]. For infinite $X \subset \mathbb{R}^m$, to our knowledge, the situation is less clear, unless f_u, f_{uv} are convex. An exception is, when f_u, f_{uv} are continuous piecewise linear: Then the continuous formulation is equivalent to a discrete MRF, see [FA14, Theorem 6], and existing results from the discrete case carry over. Despite the lack of theoretical tightness guarantees, in experiments, the infinite-dimensional local marginal polytope relaxation produces accurate lower relaxations, in particular when X is univariate.

Besides a more compact notation, the reformulation in terms of the optimal transportation allows us to invoke results from the well-established optimal transport theory, see, e.g. [Vil08; San15]. For example, it directly follows from the theory that the infimum in (5.7) is attained.

Proposition 5.2. *Let $\emptyset \neq X \subset \mathbb{R}^m$ be compact and $f_u : X \to \mathbb{R}$, $f_{uv} : X^2 \to \mathbb{R}$ be lsc. Then the optimization problem* (5.7) *has an optimal solution.*

Proof. Without loss of generality we can assume the f_u and f_{uv} to be nonnegative. Under the assumptions, it is known that the optimal transportation cost $\mathrm{OT}_{f_{uv}}$ is lsc on $\mathcal{P}(X)^2$ for the weak* convergence of measures, see e.g., [Vil08, Chapter 6] or [CD08, Lemma 5.2]. By [San15, Lemma 1.6], the linear functional $\mu \mapsto \sum_{u \in \mathcal{V}} \int_X f_u(x) \,\mathrm{d}\mu_u(x)$ is also lsc. Existence of a solution directly follows from compactness of $\mathcal{P}(X)^{\mathcal{V}}$. □

The formulation using optimal transport further allows us to study dual formulations of (5.7) in detail, which will eventually be the starting point for our implementation. To further simplify, we assume that the pairwise potentials f_{uv} are chosen to be a lsc metric on X, i.e., $f_{uv}(x, x') = d(x, x')$. Note that this assumption is satisfied in many practical applications, e.g., for discrete total variation type smoothness costs of the form $f_{uv}(x, x') = |x - x'|$. Then, the optimal transportation $\mathrm{OT}_{f_{uv}}(\mu_u, \mu_v)$ in Problem (5.7) is the Wasserstein-1 distance $W_1^d(\mu_u, \mu_v)$ induced by the metric d between μ_u and μ_v. In contrast to the general optimal transportation, which involves two Lagrange multipliers per edge, invoking Kantorovich's duality (see e.g., [San15]), we obtain a more compact dual formulation which involves only a single Lagrange multiplier per edge:

$$W_1^d(\mu_u, \mu_v) = \sup_{\lambda \in \text{Lip}_d(X)} \langle \lambda, \mu_u \rangle - \langle \lambda, \mu_v \rangle = \sigma_{\text{Lip}_d(X)}((\nabla \mu)_{(u,v)}), \tag{5.9}$$

where $\nabla : \mathcal{M}(X)^{\mathcal{V}} \to \mathcal{M}(X)^{\mathcal{E}}$ is a graph gradient operator:

$$(\nabla \mu)_{(u,v)} = \mu_u - \mu_v.$$

The constraint set on the dual variable is the set of 1-Lipschitz continuous functions with respect to the metric d:

$$\text{Lip}_d(X) = \{\lambda \in \mathcal{C}(X) : |\lambda(x) - \lambda(x')| \leq d(x, x')\}. \tag{5.10}$$

Substituting this dual formulation into the relaxation (5.7) and assigning an arbitrary but fixed orientation to the edges in $\mathcal{E}$ an interchange of min and sup yields the following problem:

$$\sup_{\lambda \in \mathcal{C}(X)^{\mathcal{E}}} \left\{ D(\lambda) \equiv - \sum_{u \in \mathcal{V}} \sigma_{\mathcal{P}(X)}(-f_u + (\text{Div}\, \lambda)_u) - \sum_{e \in \mathcal{E}} \iota_{\mathcal{K}}(\lambda_e) \right\}, \tag{5.11}$$

where $\mathcal{K} = \text{Lip}_d(X)$, $\sigma_{\mathcal{P}(X)}(-f_u + (\text{Div}\, \lambda)_u) = \sup_{\mu \in \mathcal{P}(X)} \langle \mu, -f_u + (\text{Div}\, \lambda)_u \rangle$ is the support function of $\mathcal{P}(X)$ at $-f_u + (\text{Div}\, \lambda)_u$ and $\text{Div} : \mathcal{C}(X)^{\mathcal{E}} \to \mathcal{C}(X)^{\mathcal{V}}$ is a graph divergence operator given by:

$$-(\text{Div}\, \lambda)_u = \sum_{v:(u,v) \in \mathcal{E}} \lambda_{(u,v)} - \sum_{v:(v,u) \in \mathcal{E}} \lambda_{(v,u)}. \tag{5.12}$$

Proposition 5.3. *Let $\emptyset \neq X \subset \mathbb{R}^m$ be compact, let $f_u : X \to \mathbb{R}$ be lsc, and $f_{uv} = d : X^2 \to \mathbb{R}$ be lsc and a metric. Then, the following strong duality holds:*

$$(5.7) = (5.11), \tag{5.13}$$

and a maximizer of (5.11) *exists.*

Proof. Define $f : \mathcal{M}(X)^{\mathcal{V}} \to \overline{\mathbb{R}}$

$$f(\mu) := \sum_{u \in \mathcal{V}} \langle f_u, \mu_u \rangle + \iota_{\mathcal{P}(X)}(\mu_u),$$

which is convex, proper and lsc due to [San15, Lemma 1.1.3]. Likewise, define the functional $g : \mathcal{M}(X)^{\mathcal{E}} \to \overline{\mathbb{R}}$

$$g(\nu) := \sum_{e \in \mathcal{E}} \sup_{\lambda_e \in \mathcal{K}} \langle \lambda_e, \nu_e \rangle = \sum_{e \in \mathcal{E}} \sigma_{\mathcal{K}}(\nu_e),$$

which is proper convex lsc, as it is a pointwise supremum over linear functionals. Define $\nabla : \mathcal{M}(X)^{\mathcal{V}} \to \mathcal{M}(X)^{\mathcal{E}}$

$$(\nabla \mu)_{(u,v)} = \mu_u - \mu_v,$$

which is bounded and linear. Then we compute the convex conjugates $f^* : \mathcal{C}(X)^{\mathcal{V}} \to \overline{\mathbb{R}}$ to

$$f^*(\theta) = \sum_{u \in \mathcal{V}} \sigma_{\mathcal{P}(X)}(\theta_u - f_u),$$

and the functional $g^* : \mathcal{C}(X)^{\mathcal{E}} \to \overline{\mathbb{R}}$

$$g^*(\lambda) = \sum_{e \in \mathcal{E}} \iota_{\mathcal{K}}(\lambda_e),$$

and $\nabla^* : \mathcal{C}(X)^{\mathcal{E}} \to \mathcal{C}(X)^{\mathcal{V}} = -\operatorname{Div}$.

Choose $x \in X$ and define $\mu := (\delta_x)^{\mathcal{V}}$. Then $f(\mu) = \sum_{u \in \mathcal{V}} f_u(x) < \infty$. In addition we have $\langle \lambda_e, \delta_x - \delta_x \rangle = 0$ for all $\lambda_e \in \mathcal{C}(X)$ and therefore $g(\nabla \mu) = 0$. Now consider a weakly*-convergent sequence $\mathcal{M}(X)^{\mathcal{E}} \ni \nu^t \overset{*}{\rightharpoonup} \nabla \mu$. This means for all $e \in \mathcal{E}$ and $\lambda_e \in \mathcal{C}(X)$ we have $\langle \lambda_e, \nu_e^t \rangle \to \langle \lambda_e, (\nabla \mu)_e \rangle = 0$. In particular this implies that $g(\nu^t) \to 0$, and therefore g is continuous at $\nabla \mu$. Then we can invoke the Fenchel–Rockafellar duality Theorem 1.33 and obtain that (5.7) = (5.11) and a maximizer of (5.11) exists. □

In what follows, we will now study finite approximations to the dual problem (5.11).

5.3. Dual discretization: A generalized conjugacy perspective

5.3.1. Dual discretization for MRFs

The dual problem (5.11) is formulated over the space of continuous functions $\mathcal{C}(X)$. Therefore, our strategy to obtain a more tractable formulation is to restrict the dual variables $\lambda \in \mathcal{C}(X)$ to a certain subspace $\Lambda \subset \mathcal{C}(X)$ of the space of continuous functions adopting the approach of [FA14]. This eventually leads to a semi-infinite program for which there exist tractable finite formulations in certain cases. In contrast to [FA14] we have drawn a connection to optimal transport levereging Kantorovich's duality to obtain a reduced dual program for the metric pairwise MRF. This eventually leads to a different dual formulation and allows us to derive a tractable implementation beyond subgradient ascent which is left as an open problem in [FA14].

In this section a major goal is to study the primal problem corresponding to the discretized dual problem and draw a connection to the space of moments. In addition we consider hierarchies of dual subspaces

$$\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \mathcal{C}(X),$$

which induce the inequalities

$$\sup_{\lambda \in (\Lambda_1)^{\mathcal{E}}} D(\lambda) \leq \sup_{\lambda \in (\Lambda_2)^{\mathcal{E}}} D(\lambda) \leq \cdots \leq (5.7) = (5.11).$$

In particular we study the convergence properties for a piecewise polynomial hierarchy in $\mathcal{C}(X)$ for the case $X = [a, b]$, $a < b$ is an interval.

Let Λ be spanned by basis functions $\Lambda = \langle \varphi_0, \varphi_1, \dots, \varphi_n \rangle \subset \mathcal{C}(X)$, where $\varphi_0 \equiv 1$ and note that constant components in the dual problem (5.11) do not matter. We identify the basis functions φ_k as component functions of the following mapping $\varphi : X \to \mathbb{R}^n$

$$\varphi(x) = (\varphi_1(x), \dots, \varphi_n(x)). \tag{5.14}$$

The mapping *lifts* the input x to a higher-dimensional space. Therefore we refer to φ as the *lifting map* aka *feature map*. φ also describes a continuous curve which we will refer to as the *moment curve* for a reason that will be discussed below.

Choose $\lambda(x) = \langle p, \varphi(x) \rangle$ for coefficients $p \in \mathbb{R}^n$. Substituting this representation into the support functions of the Lagrangian dual problem (5.11) we obtain in view of

Lemma 5.1:

$$\sigma_{\mathcal{P}(X)}(\lambda - f) = \sup_{\mu \in \mathcal{P}(X)} \langle \lambda - f, \mu \rangle = \max_{x \in X} \langle \varphi(x), p \rangle - f(x). \tag{5.15}$$

We can thus rewrite the discretized dual problem:

$$\sup_{p \in (\mathbb{R}^n)^{\mathcal{E}}} \sum_{u \in \mathcal{V}} \min_{x \in X} \langle -(\operatorname{Div}_\Lambda p)_u, \varphi(x) \rangle + f_u(x) - \sum_{e \in \mathcal{E}} \iota_{\mathcal{K}_\Lambda}(p_e), \tag{5.16}$$

where $\operatorname{Div}_\Lambda : (\mathbb{R}^n)^{\mathcal{E}} \to (\mathbb{R}^n)^{\mathcal{V}}$ with

$$-(\operatorname{Div}_\Lambda p)_u = \sum_{v:(u,v) \in \mathcal{E}} p_{(u,v)} - \sum_{v:(v,u) \in \mathcal{E}} p_{(v,u)},$$

and the constraint set $\mathcal{K}_\Lambda$ is the set of coefficients of Lipschitz functions in Λ:

$$\mathcal{K}_\Lambda := \{ p \in \mathbb{R}^n : \langle p, \varphi(\cdot) \rangle \in \operatorname{Lip}_d(X) \}. \tag{5.17}$$

Observe that the inner minimization problem $\min_{x \in X} \langle -p, \varphi(x) \rangle + f(x)$ closely resembles the structure of a convex conjugate: Indeed, we can write the minimization

$$-\max_{x \in X} \langle p, \varphi(x) \rangle - f(x) = -f_\Lambda^*(p),$$

in terms of the negative convex conjugate of the following extended real-valued function $f_\Lambda : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by:

$$f_\Lambda(y) = \begin{cases} f(x) & \text{if } y = \varphi(x) \text{ for some } x \in X, \\ +\infty & \text{otherwise,} \end{cases} \tag{5.18}$$

which we will refer to as the *lifted* version of f. Thus, we can formulate the discretized dual problem in terms of the lifted convex conjugates f_Λ^* of f_Λ:

$$\sup_{p \in (\mathbb{R}^n)^{\mathcal{E}}} \; - \sum_{u \in \mathcal{V}} (f_u)_\Lambda^*((\operatorname{Div}_\Lambda p)_u) - \sum_{e \in \mathcal{E}} \iota_{\mathcal{K}_\Lambda}(p_e). \tag{5.19}$$

Invoking Fenchel–Rockafellar duality for the dual problem (5.19) one obtains the following discretized primal problem:

$$\min_{y \in (\mathbb{R}^n)^{\mathcal{V}}} \left\{ \mathcal{F}_\Lambda \equiv \sum_{u \in \mathcal{V}} (f_u)_\Lambda^{**}(y_u) + \sum_{e \in \mathcal{E}} \sigma_{\mathcal{K}_\Lambda}\big((\nabla_\Lambda \, y)_e\big) \right\}, \tag{5.20}$$

where ∇_Λ is the adjoint operator of the negative divergence $-\operatorname{Div}_\Lambda$ and $(f_u)_\Lambda^{**}$ are the lifted biconjugates.

A comparison of (5.20) with the convex relaxation (5.2) shows the effect of the dual discretization, where the classical convex biconjugates are replaced with biconjugates of the "lifted" functions $(f_u)_\Lambda$. Indeed, without lifting, i.e., $\varphi(x) = x$, we recover the classical biconjugates and therefore the convex relaxation (5.2).

Note that the choice of a basis φ in the definition of f_Λ is somewhat arbitrary. Indeed, one can get rid of the dependence of a certain basis via a proper treatment of the dual space of Λ. For simplicity, however, we choose a fixed basis and instead remark, that the dual space of Λ is isomorphic to $\mathbb{R}^n$.

Practically, we will approximate the dual variables in terms of piecewise polynomial splines. For intervals, the following proposition shows that either by increasing the number of pieces or the degree of the polynomial the primal-dual gap can be reduced. In our case we approximate the Lipschitz dual variable in terms of a Lipschitz spline. As a consequence existing results such as [FA14, Theorem 2] do not apply. Instead, we use a construction based on Bernstein-polynomials. Then the result follows from [BEP87, Theorem 1].

Proposition 5.4. *Assume that $X = [a, b] \subset \mathbb{R}$, $a < b$, and let the metric d be given by $d(x, y) = |x - y|$. Furthermore, let $\Lambda \subset \mathcal{C}(X)$ be the space spanned by continuous piecewise polynomials on intervals $[t_i, t_{i+1}]$ defined by a regularly spaced grid with nodes given by $t_i = a + (b - a) \cdot (i - 1)/K$, $i = 1, \dots, K + 1$. Then the optimality gap satisfies:*

$$(5.7) - (5.19) = \mathcal{O}(1/(K \cdot \sqrt{\deg})),$$

where deg *is the degree of the polynomial on each piece.*

Proof. We consider the discretized dual problem (5.19) where $\mathcal{K}_\Lambda$ is the set of coefficients corresponding to 1-Lipschitz piecewise polynomials on $[a, b]$ of degree deg with K pieces. Also recall that we have the following relations between the dual and primal problems: $(5.20) = (5.19) \leq (5.11) = (5.7)$.

Now, let us denote a maximizer of (5.11) as $\lambda^* \in \mathrm{Lip}_d(X)^{\mathcal{E}}$. Existence of such a dual maximizer follows by Proposition 5.3. Then, one has for any $\lambda \in \Lambda^{\mathcal{E}}$:

$$\begin{aligned}\min_{x \in X}\ f_u(x) - (\operatorname{Div} \lambda^*)_u(x) &= \min_{x \in X}\ f_u(x) - (\operatorname{Div} \lambda)_u(x) - (\operatorname{Div} \lambda^*)_u(x) + (\operatorname{Div} \lambda)_u(x) \\ &\leq \min_{x \in X}\ f_u(x) - (\operatorname{Div} \lambda)_u(x) + \| - (\operatorname{Div}(\lambda^* - \lambda))_u\|_\infty. \end{aligned} \tag{5.21}$$

This allows us to bound the optimality gap by:

$$\begin{aligned}(5.11) - (5.19) &\leq \sum_{u \in \mathcal{V}} \| - (\operatorname{Div}(\lambda - \lambda^*))_u\|_\infty \\ &\leq \sum_{u \in \mathcal{V}} |d(u)| \cdot \sup_{e \in \mathcal{E}}\ \|\lambda_e - \lambda_e^*\|_\infty \\ &\leq 2|\mathcal{E}| \cdot \sup_{e \in \mathcal{E}} \|\lambda_e - \lambda_e^*\|_\infty, \end{aligned} \tag{5.22}$$

where $d(u)$ denotes the degree of the vertex u.

For a L-Lipschitz function $f : [0, 1] \to \mathbb{R}$ there exists a Bernstein polynomial $p : [0, 1] \to \mathbb{R}$ with $p(0) = f(0)$ and $p(1) = f(1)$ such that $\|p - f\|_\infty \leq \frac{3L}{2} \deg^{-1/2}$ [Car98, Theorem 2.6]. By [BEP87, Theorem 1], this polynomial is L-Lipschitz as well. For each $e \in \mathcal{E}$ we pick the coefficients of the function λ_e such that it approximates the optimal dual variable λ_e^* with such a polynomial individually on each interval $[t_i, t_{i+1}]$. Then one obtains an overall 1-Lipschitz polynomial with the following bound:

$$\|\lambda_e - \lambda_e^*\|_\infty \leq \frac{3(b - a)}{2K\sqrt{\deg}}. \tag{5.23}$$

Inserting this into (5.22) yields:

$$(5.7) - (5.19) \leq 3|\mathcal{E}| \frac{(b - a)}{K\sqrt{\deg}}, \tag{5.24}$$

which gives the stated $\mathcal{O}(1/(K \cdot \sqrt{\deg}))$ rate. □

5.3.2. On the duality between dual discretization and lifting

Next we are going to study the lifted biconjugates f_Λ^{**}. A first question to address is the characterization of the domain $\operatorname{dom} f_\Lambda = \operatorname{con} \varphi(X)$ which is the convex hull of the of the image of φ. For a probability measure $\mu \in \mathcal{P}(X)$ we define the k^{th} moment of μ as $\int_X \varphi_k(x) \,\mathrm{d}\mu(x)$. Then the set of *valid truncated moment sequences* aka the *probability moment space*, $\mathcal{P}_\Lambda$, is the set of all vectors $y \in \mathbb{R}^n$ for which there exists a probability measure $\mu \in \mathcal{P}(X)$ such that the k^{th} component of y is the k^{th} moment of μ:

$$\mathcal{P}_\Lambda = \{y \in \mathbb{R}^n : \exists \mu \in \mathcal{P}(X), y_k = \langle \mu, \varphi_k \rangle\} . \tag{5.25}$$

In other words $\mathcal{P}_\Lambda$ is the set of all "infinite convex combinations" of points $\varphi(x) \in \mathbb{R}^n$ with $x \in X$, i.e., of points that belong to the image of φ. For $x \in X$ the moment vector $\varphi(x) = \int_X \varphi(x') \,\mathrm{d}\delta_x(x')$ of a Dirac measure δ_x sets up a certain correspondence between lifted points $\varphi(x)$ and the Dirac measure δ_x iteself. In this context we call the curve $x \mapsto \varphi(x)$ the moment curve described by Diracs. Then the probability moment space can be written equivalently as the set of all finite convex combinations of moment vectors of Dirac measures, in the same way the unit simplex is the convex hull of the unit vectors as proved in the next proposition.

Proposition 5.5. *Let $\emptyset \neq X$ be compact and $\Lambda = \langle \varphi_0, \ldots, \varphi_n \rangle \subset \mathcal{C}(X)$ with $\varphi_0 \equiv 1$. Then every moment vector of a probability measure is a finite convex combination of moment vectors of Dirac measures, i.e.*

$$\operatorname{con}(\varphi(X)) = \mathcal{P}_\Lambda,$$

and $\mathcal{P}_\Lambda$ is nonempty and compact.

Proof. Note that $\mathcal{P}_\Lambda$ is convex and bounded. It is also closed: To this end consider a sequence $y^t \to y$ with $y^t \in \mathcal{P}_\Lambda$. This means for any t there exists $\mu^t \in \mathcal{P}(X)$ with $y^t_k = \int_X \varphi_k(x) \,\mathrm{d}\mu^t(x)$ for $k = 1, \ldots, n$.

Since X is compact, due to Prokhorov's theorem, see [San15, Section 1.1], there exists a weakly* convergent subsequence $\mu^{t_j} \overset{*}{\rightharpoonup} \mu$ and, hence, $\int_X \varphi_k(x) \,\mathrm{d}\mu^{t_j}(x) = y_k^{t_j} \to \int_X \varphi_k(x) \,\mathrm{d}\mu(x) = y_k$. Therefore $y \in \mathcal{P}_\Lambda$.

Next we show identity of the support functions of the convex sets $\operatorname{con}(\varphi(X))$ and $\mathcal{P}_\Lambda$ as this implies the equality of the sets.

To this end let $f \in \Lambda = \langle \varphi_0, \ldots, \varphi_n \rangle$ with $\varphi_0 = 1$. We write $f(x) = \langle \varphi(x), f \rangle$. Assume that $f_0 = 0$. We have the identities:

$$\sigma_{\varphi(X)}(-f) = -\min_{x \in X} \langle \varphi(x), f \rangle = -\min_{\mu \in \mathcal{P}(X)} \int_X f(x) \,\mathrm{d}\mu(x) = -\min_{y \in \mathcal{P}_\Lambda} \langle y, f \rangle = \sigma_{\mathcal{P}_\Lambda}(-f),$$

where the first equality follows by the definition of the support function, the second equality by Lemma 5.1 and the third equality by the definition of $\mathcal{P}_\Lambda$, the identity

$$\min_{\mu \in \mathcal{P}(X)} \int_X f(x) \,\mathrm{d}\mu(x) = \min_{\mu \in \mathcal{P}(X)} \sum_{k=1}^n f_k \int_X \varphi_k(x) \,\mathrm{d}\mu(x),$$

and the substitution $y_k = \int_X \varphi_k(x) \,\mathrm{d}\mu(x)$. Since for each such f the support functions are equal we have equality of the support functions of $\varphi(X)$ and $\mathcal{P}_\Lambda$. Since X is compact and

φ continuous and the convex hull of a compact set stays compact, cf. [RW98, Corollary 2.30], we can replace $\varphi(X)$ with its convex hull $\operatorname{con}\varphi(X)$ and the conclusion follows. □

The proof of the proposition above also reveals that for $f \in \Lambda$, the lifted biconjugate is a linear function over the probability moment space

$$f_\Lambda^{**}(y) = \langle y, f\rangle + \iota_{\mathcal{P}_\Lambda}(y), \tag{5.26}$$

whose minimization is actually equivalent to minimizing the original function.

Specializing φ_k to the monomial basis and X to a set defined via polynomial inequalities, this is exactly the starting point of the formulation proposed by [Las01; Las02] for constrained polynomial optimization.

More abstractly, a meaningful notion of moments is induced by a lifting map φ which is a homeomorphism between X and $\varphi(X)$.

Definition 5.6 (lifting map)**.** *Let X be compact and nonempty. Then we say the mapping $\varphi : X \to \mathbb{R}^n$ is a lifting map if φ is continuous on X and injective with continuous inverse $\varphi^{-1} : \varphi(X) \to \mathbb{R}^m$.*

It is instructive to discuss possible choices for φ including the ones that correspond to existing discretizations for the continuous MRF such as the discrete approach and the piecewise linear approach. In the latter two cases, the probability moment space is merely the unit simplex. In that sense, the monomial probability moment space can be interpreted as a “nonlinear probability simplex”, which, in contrast to the unit simplex, has infinitely many “vertices”, see Figure 5.1. For the same reason, as we will see in the course of this section, it better suits the continuous nature of our optimization problem.

The discrete sampling-based approach is recovered by the following choice of φ:

Example 5.7. We discretize the interval $X = [a, b]$, $a < b$ and re-define $X := \{t_1, \dots, t_n\}$ with $t_k \in [a, b]$, $t_k < t_{k+1}$. For any $t_k \in X$ let $\varphi(t_k) = e_k$ with $e_k \in \mathbb{R}^n$ being the k^{th} unit vector. As a result φ is the canonical basis and continuous wrt the discrete topology. It spans the space of discrete functions $f : \{t_1, \dots, t_n\} \to \mathbb{R}$ and the probability moment space $\operatorname{con}(\varphi(X)) = \mathcal{P}_\Lambda$ is given by the unit simplex.

The above example can be extended by assigning any points $x \in (t_k, t_{k+1})$ to points on the connecting line between two corresponding Diracs which results in a more continuous formulation:

Example 5.8. Let $X = [a, b]$, $a < b$: Let $t_k < t_{k+1}$ and $t_1 = a$, $t_n = b$ be a sequence of knots that subdivide the interval X into $n - 1$ subintervals $[t_k, t_{k+1}] =: X_k$. We define $\varphi(x) = \alpha e_k + (1 - \alpha)e_{k+1}$, with $e_k \in \mathbb{R}^n$ being the k^{th} unit vector, $\alpha \in [0, 1]$ such that $\alpha t_k + (1 - \alpha)t_{k+1} = x$. The component functions φ_k are the *hat functions* that span the space of continuous piecewise linear functions Λ. φ yields a sparse lifting map. See Section 5.5.5 for the multivariate case. The probability moment space $\operatorname{con}(\varphi(X)) = \mathcal{P}_\Lambda$ is the unit simplex.

For φ_k being chosen as the monomials we obtain the classical notion of moments:

Example 5.9. For $\varphi_0 = 1$, the space of univariate polynomials $\Lambda = \mathbb{R}[x]$ with maximum degree n is spanned by the monomials:

$$\varphi(x) = (x, x^2, \dots, x^n), \tag{5.27}$$

and $\operatorname{con}(\varphi(X)) = \mathcal{P}_\Lambda$ is the monomial probability moment space.

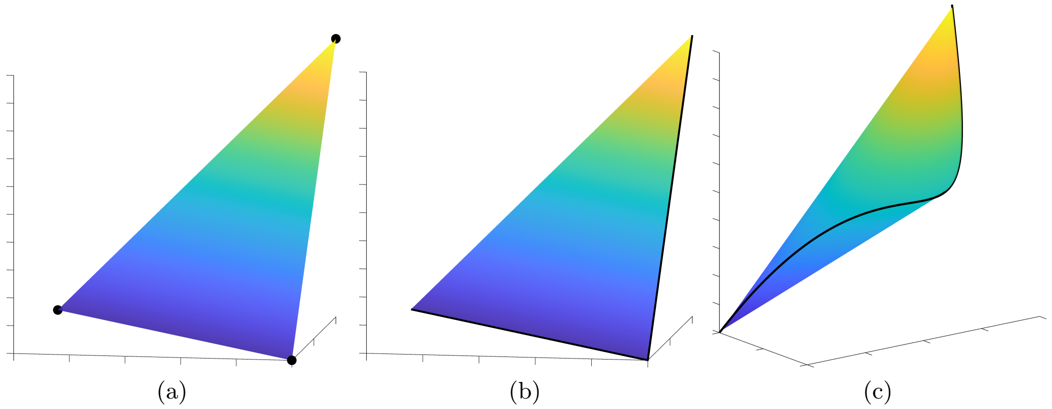


Figure 5.1.: Different finite-dimensional approximations $\mathcal{P}_\Lambda$ of the infinite-dimensional space of probability measures $\mathcal{P}([-1,1])$. Left and middle: 2-dimensional probability simplex and right: Monomial probability moment space $\operatorname{con}\{(x,x^2,x^3) : x \in [-1,1]\}$ of degree 3. The approximations are obtained as the convex hulls of the black curves $x \mapsto \varphi(x)$ for 3 different choices of φ. From left to right, see Example 5.7, Example 5.8 and Example 5.9. In all cases, the black curves themselves correspond to Dirac measures and the convex hulls of the curves correspond to the space of probability measures. In contrast to the simplex that only has a finite number of extreme points, the monomial moment curve comprises a continuum of extreme points so that no Dirac measure on the monomial moment curve can be expressed as a convex combination of other Diracs.

Example 5.10. Identifying $X = \{x \in \mathbb{R}^2 \colon \|x\|_2 = 1\}$ with the complex unit circle $X \cong \{z \in \mathbb{C} \colon |z| = 1\}$ and again assuming $\varphi_0 = 1$, the mapping

$$\varphi(z) = (\operatorname{Re}(z), \operatorname{Im}(z), \dots, \operatorname{Re}(z^n), \operatorname{Im}(z^n)) \in \mathbb{R}^{2n} \tag{5.28}$$

spans the space Λ of real trigonometric polynomials of maximum degree n. Parametrizing elements $z \in X$ via the bijection between $z = e^{i\omega}$ and its angle $\omega \in [0, 2\pi)$ the components of φ are the Fourier basis functions, which define the Carathéodory curve [Car11]. The convex hull $\operatorname{con}(\varphi(X)) = \mathcal{P}$ is the trigonometric probability moment space.

In contrast to the piecewise linear lifting, the (trigonometric) polynomial lifting is extremal in the sense that no Dirac measure on the moment curve can be expressed as a convex combination of other Diracs, which sets up a certain one-to-one correspondece between Diracs δ_x and lifted points $\varphi(x)$. This is illustrated in Figure 5.1 for the monomial case.

More formally, we call a lifting map φ an extremal curve if each $y \in \varphi(X)$ is an extreme point of $\operatorname{con}(\varphi(X))$ defined according to [Roc70, Section 18].

Definition 5.11 (extreme points)**.** *Let C be a convex set and $y \in C$. Then y is called an extreme point of C if there is no way to express y as a convex combination $y = (1-\alpha)x + \alpha z$ of $x, z \in C$ and $0 < \alpha < 1$, except by taking $y = x = z$.*

Definition 5.12 (extremal moment curve)**.** *Let X be compact and nonempty. Then we say the mapping $\varphi : X \to \mathbb{R}^n$ is an extremal moment curve if φ is a lifting map and any point $y \in \varphi(X) \subset \mathbb{R}^n$ is an extreme point of $\operatorname{con} \varphi(X)$.*

Via a change of basis it becomes clear that the definition of extremality is independent of a specific choice of a basis for Λ. Therefore extremality is rather a property of the subspace Λ. This also motivates the following lemma which shows that extremality is inherited along a hierarchy $\Theta \subset \Lambda \subset \mathcal{C}(X)$.

Lemma 5.13 (extremal subspaces). *Let $\emptyset \neq X \subset \mathbb{R}^m$ be compact. Let $\Theta \subset \Lambda \subset \mathcal{C}(X)$ be a hierarchy of finite-dimensional subspaces of the space of continuous functions $\mathcal{C}(X)$. Let $\Theta = \langle \theta_1, \dots, \theta_n \rangle$ such that $\theta : X \to \mathbb{R}^n$ is an extremal curve. Then Λ is spanned by an extremal curve as well.*

Proof. $\{\theta_1, \dots, \theta_n\}$ is a basis of $\Theta \subset \Lambda$ and therefore linearly independent. Since Λ is finite-dimensional in view of the basis extension theorem $\{\theta_1, \dots, \theta_n\}$ can be extended to a basis $\varphi = (\theta_1, \dots, \theta_n, \psi_1, \dots, \psi_k)$ of Λ with vectors $\psi_i \in G$, where $\operatorname{span} G = \Lambda$ and $|G| < \infty$ such that $\operatorname{span} \varphi = \Lambda$.

Now choose $y \in X$ and consider $\varphi(y)$. Let $\alpha \in (0,1)$ and $\varphi(y) = \alpha x + (1-\alpha)z$ for $x, z \in \operatorname{con} \varphi(X) \subset \mathbb{R}^{n+k}$. Due to Carathéodory [RW98, Theorem 2.29] there exist coefficients α_i, $\beta_l > 0$ such that $x = \sum_{i=1}^{n+k+1} \alpha_i \varphi(x_i)$ and $z = \sum_{l=1}^{n+k+1} \beta_l \varphi(z_l)$, $z_l, x_i \in X$ with $\sum_{i=1}^{n+k+1} \alpha_i = 1$, $\sum_{l=1}^{n+k+1} \beta_l = 1$. This implies that $\theta(y) = \alpha \sum_{i=1}^{n+k+1} \alpha_i \theta(x_i) + (1-\alpha) \sum_{l=1}^{n+k+1} \beta_l \theta(z_l)$. Extremality of θ implies that $\theta(y) = \sum_{i=1}^{n+k+1} \alpha_i \theta(x_i) = \sum_{l=1}^{n+k+1} \beta_l \theta(z_l)$ and therefore $\theta(y) = \theta(x_i) = \theta(z_l)$. Since θ is an extremal curve it is injective and therefore $y = z_l = x_i$. This implies that $\varphi(y) = x = z$. □

Lemma 5.14 (extremality of quadratic subspace). *Let $\emptyset \neq X \subset \mathbb{R}^m$ be compact. Let $\varphi(x) = (x_1, x_2, \dots, x_m, \|x\|^2)$. Assume that $\langle \varphi_1, \dots, \varphi_{m+1} \rangle \subset \Lambda \subset \mathcal{C}(X)$. Then Λ is spanned by an extremal curve.*

Proof. Choose $y \in X$. Let $\alpha \in (0,1)$ and $\varphi(y) = \alpha x + (1-\alpha) z$ for $x, z \in \operatorname{con} \varphi(X) \subset \mathbb{R}^{m+1}$. Due to Carathéodory [RW98, Theorem 2.29] there exist coefficients α_i, $\beta_k > 0$ such that $x = \sum_{i=1}^{m+2} \alpha_i \varphi(x_i)$ and $z = \sum_{k=1}^{m+2} \beta_k \varphi(z_k)$, $z_k, x_i \in X$ with $\sum_{i=1}^{m+2} \alpha_i = 1$, $\sum_{k=1}^{m+2} \beta_k = 1$.

Now choose $f(x) = \|x-y\|^2 = \|x\|^2 - 2\langle y, x\rangle + \|y\|^2$. Then we have for $a = (-2y_1, \dots, -2y_m, 1)$ and $a_0 = \|y\|^2$:

$$\begin{aligned} 0 = f(y) = \langle \varphi(y), a \rangle + a_0 &= \left\langle \alpha \sum_{i=1}^{m+2} \alpha_i \varphi(x_i) + (1-\alpha) \sum_{k=1}^{m+2} \beta_k \varphi(z_k), a \right\rangle + a_0 \\ &= \sum_{i=1}^{m+2} \alpha \cdot \alpha_i \cdot \langle \varphi(x_i), a \rangle + \sum_{k=1}^{m+2} (1-\alpha) \cdot \beta_k \cdot \langle \varphi(z_k), a \rangle + a_0 \\ &= \sum_{i=1}^{m+2} \alpha \cdot \alpha_i \cdot f(x_i) + \sum_{k=1}^{m+2} (1-\alpha) \cdot \beta_k \cdot f(z_k) \end{aligned}$$

As $f(x_i) > 0$ for $x_i \neq y$ and $\alpha > 0$ it holds that $x_i \neq y$ implies $\alpha_i = 0$. The same is true for z_k and β_k. Hence $x = \varphi(y) = z$, and therefore $\varphi : X \to \mathbb{R}^{m+1}$ is an extremal curve. In view of Lemma 5.13 Λ is spanned by an extremal curve. □

This shows that in a piecewise polynomial discretization with degree at least 2 the corresponding basis inherits the extreme point property from the extremality of the subspace of quadratic functions.

Extremal curves are key to preserve the cost function when restricted to the set of discretized Diracs $\varphi(x)$:

Theorem 5.15. *Let $X \subset \mathbb{R}^m$ be nonempty and compact and let $f : X \to \mathbb{R}$ be lsc. Furthermore, let $\varphi : X \to \mathbb{R}^n$ be an extremal curve. Then we have*

$$f_\Lambda^{**} \circ \varphi = f \tag{5.29}$$

on X. In addition we have that $\operatorname{con} f_\Lambda = f_\Lambda^{**}$.

Proof. Since f is lsc and X compact f is bounded from below, i.e., there is $\gamma > -\infty$ so that $f(x) \geq \gamma$ for all $x \in X$. We have $\operatorname{dom} f_\Lambda = \varphi(X) \subset \mathbb{R}^n$ and in view of [RW98, Proposition 2.31] it holds for any $y \in \mathbb{R}^n$,

$$\begin{aligned}(\operatorname{con} f_\Lambda)(y) &= \inf\left\{\sum_{i=1}^{n+1} \lambda_i f_\Lambda(y_i) : \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1,\, y = \sum_{i=1}^{n+1} \lambda_i y_i,\, y_i \in \mathbb{R}^n\right\} \\ &= \inf\left\{\sum_{i=1}^{n+1} \lambda_i f(x_i) : \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1,\, y = \sum_{i=1}^{n+1} \lambda_i \varphi(x_i),\, x_i \in X\right\} \geq \gamma.\end{aligned}$$

Let $x \in X$. Since the only possible convex combination of the extreme point $\varphi(x)$ from points $y_i \in \varphi(X)$ is $\varphi(x)$ itself, we have

$$\begin{aligned}(\operatorname{con} f_\Lambda)(\varphi(x)) &= \inf\left\{\sum_{i=1}^{n+1} \lambda_i f_\Lambda(y_i) : \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1,\, \varphi(x) = \sum_{i=1}^{n+1} \lambda_i y_i,\, y_i \in \varphi(X)\right\} \\ &= f_\Lambda(\varphi(x)) = f(x).\end{aligned}$$

This shows that $\operatorname{con} f_\Lambda \circ \varphi = f$. Since $\operatorname{con} f_\Lambda$ is bounded from below $\operatorname{con} f_\Lambda$ is proper.

Let f be lsc. Then f_Λ inherits its lower semicontinuity from f: Assume that $y \in \varphi(X)$. Then there exists $x \in X$ with $y = \varphi(x)$ and we have due to the continuity of φ and φ^{-1}:

$$\begin{aligned}\liminf_{y' \to \varphi(x)} f_\Lambda(y') &:= \lim_{\delta \to 0^+} \inf \{f_\Lambda(y') : y' \in B_\delta(\varphi(x))\} \\ &= \lim_{\delta \to 0^+} \inf \{f_\Lambda(\varphi(x')) : \varphi(x') \in B_\delta(\varphi(x))\} \\ &= \lim_{\varepsilon \to 0^+} \inf \{f(x') : x' \in B_\varepsilon(x)\} \\ &= \liminf_{x' \to x} f(x') \geq f(x) = f_\Lambda(y).\end{aligned}$$

Since X is compact and φ continuous the image $\varphi(X) = \operatorname{dom} f_\Lambda$ is compact as well. Thus f_Λ is super-coercive and bounded from below. Then we can invoke [RW98, Corollary 3.47] and deduce that $\operatorname{con} f_\Lambda$ is proper, lsc and convex. In view of [RW98, Theorem 11.1] we have $\operatorname{con} f_\Lambda = \operatorname{cl} \operatorname{con} f_\Lambda = f_\Lambda^{**}$. □

There is a rich geometric intuition behind this theorem as shown in Figure 5.2.

More practically, for a piecewise polynomial discretization with degree at least 2, due to extremality, the primal discretized energy $\mathcal{F}_\Lambda$ restricted to $\varphi(X)^\mathcal{V}$ agrees with the original energy F. In particular, this implies that an obtained Dirac solution $(\varphi(x_u^*))_{u \in \mathcal{V}}$ of the discretization corresponds to a solution of the original problem in the same way integer solutions of LP relaxations are certificates of optimality for the corresponding ILP.

Proposition 5.16. *Let $\emptyset \neq X \subset \mathbb{R}^m$. Let the metric d be induced by a norm and assume the space of linear functions on X is contained in Λ. Furthermore, assume Λ is spanned* *by an extremal curve $\varphi : X \to \mathbb{R}^n$ and f_u lsc. Then for any $y \in (\mathbb{R}^n)^{\mathcal{V}}$, with $y_u = \varphi(x_u)$, $x_u \in X$ the following identity holds true:*

$$\mathcal{F}_\Lambda(y) = F(x).$$

In particular, whenever y^* is a solution of problem (5.20) *such that $y^* \in (\mathbb{R}^n)^{\mathcal{V}}$, with $y^*_u = \varphi(x^*_u)$, for some $x^*_u \in X$, x^* is a solution of* (5.1).

Proof. Let d be induced by some norm $\|\cdot\|$ and denote its dual norm by $\|\cdot\|_*$. As shown in Theorem 5.15, the unaries preserve the original cost functions at $\varphi(X)$. Hence it remains to show that the for the pairwise costs it holds $\sigma_{\mathcal{K}_\Lambda}((\nabla_\Lambda y)_{(u,v)}) = \|x_u - x_v\|$. By assumption $(\nabla_\Lambda y)_{(u,v)} = \varphi(x_u) - \varphi(x_v)$. We rewrite

$$\sigma_{\mathcal{K}_\Lambda}(\varphi(x_u) - \varphi(x_v)) = \sup_{\lambda \in \Lambda \cap \mathrm{Lip}_d(X)} \lambda(x_u) - \lambda(x_v) \le \|x_u - x_v\|.$$

For any $x_u, x_v \in X$ we have

$$\|x_u - x_v\| = \sup_{p \in \mathbb{R}^m : \|p\|_* \le 1} |\langle x_u - x_v, p\rangle| = \langle x_u - x_v, p^*\rangle,$$

where p^* denotes the maximizer in the supremum, which exists due to the compactness of the unit ball in a finite-dimensional space. Define the linear function $\lambda^* := \langle \cdot, p^*\rangle$ and note that by assumption $\lambda^* \in \Lambda$. In addition we have shown that $\lambda^* \in \mathrm{Lip}_d(X)$ is 1-Lipschitz. This implies $\|x_u - x_v\| = \lambda^*(x_u) - \lambda^*(x_v) \le \sup_{\lambda \in \Lambda \cap \mathrm{Lip}_d(X)} \lambda(x_u) - \lambda(x_v) \le \|x_u - x_v\|$ and hence equality holds. □

5.3.3. A generalized conjugacy perspective

The above results can be obtained from a generalized conjugacy point of view. In particular, the convex conjugate of the lifted function f^*_Λ is comprised by the notion of Φ-conjugacy, see Definition 3.1 in Chapter 3.

For $Y = \mathbb{R}^n$, and $\Phi(x, y) = \langle y, \varphi(x)\rangle$, the convex conjugate f^*_Λ is identical to the Φ-conjugate f^Φ of f, while its biconjugate f^{**}_Λ is the tightest convex extension of f_Λ to $\operatorname{con}\varphi(X)$. The Φ-biconjugate $f^{\Phi\Phi}$ of f at a point $x \in X$ is equal to the classical biconjugate of f_Λ, evaluated at $\varphi(x)$, i.e., $f^{\Phi\Phi} = f^{**}_\Lambda \circ \varphi$ on X, showing that the Φ-biconjugate is a convexely composite function and, therefore, it is nonconvex in general. Actually, Φ-conjugacy also comprises lifting to measures via $\varphi(x) = \delta_x$, Φ the corresponding dual pairing and $Y = \mathcal{C}(X)$.

As a consequence of Lemma 3.2, the considered Φ-conjugacy can be interpreted in terms of under-approximation by functions in Λ. In analogy to the biconjugate f^{**}, which is the pointwise supremum of affine-linear functions majorized by f, the Φ-biconjugate $f^{\Phi\Phi}$ is the pointwise supremum of functions in Λ up to constant translation majorized by f. This point of view also relates $(f^\Phi)^*$ and $f^{\Phi\Phi}$ by each other.

Remark 5.17. The function $(f^\Phi)^* = f^{**}_\Lambda$ is the pointwise supremum of all affine-linear functions $l_{\lambda,\beta} := \langle \cdot, \lambda\rangle - \beta$ for which $q_{\lambda,\beta} := \Phi(\cdot, \lambda) - \beta$ is majorized by f. This can be seen as follows: The Legendre–Fenchel conjugate can be characterized via the identity $(f^\Phi)^*(y) = \sup_{(\lambda,\beta) \in \mathrm{epi}(f^\Phi)} \langle y, \lambda\rangle - \beta$. The observation now follows from the fact that $(\lambda, \beta) \in \mathrm{epi}(f^\Phi)$ if and only if $q_{\lambda,\beta}$ is majorized by f.

The correspondence between f^{**}_Λ and $(f^\Phi)^*$ and between the minorizers $l_{\lambda,\beta}$ and $q_{\lambda,\beta}$ is illustrated in Figure 5.2. Note that this is closely related to the idea of feature maps φ

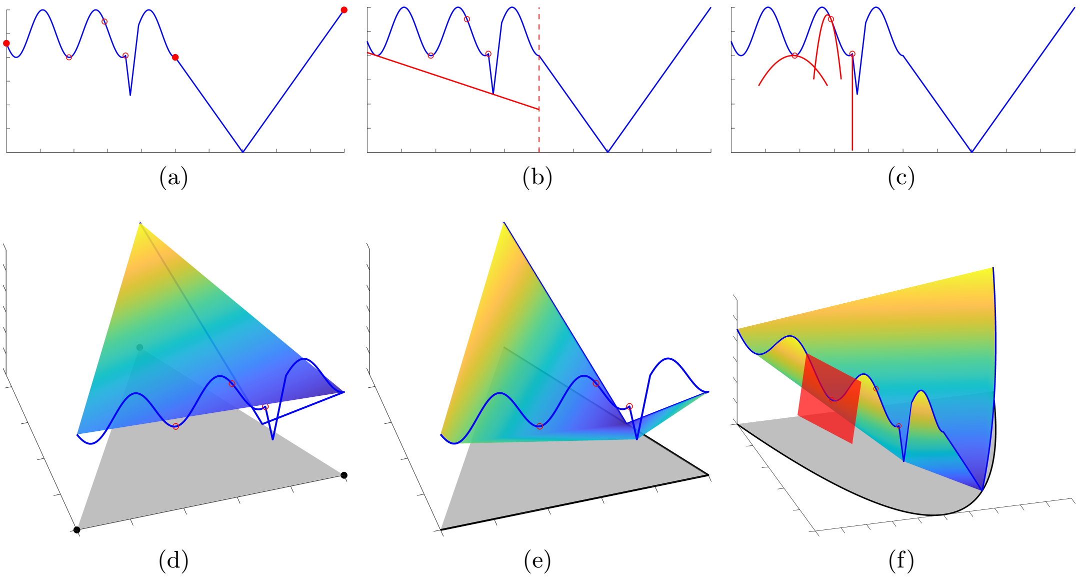


Figure 5.2.: Lifted version f_Λ of the function f for 3 different choices of φ. In the top row the same nonconvex function f (blue curves) is depicted. The colored surfaces in the bottom row correspond to the lifted biconjugates f_Λ^{**}, where the gray shadow areas correspond to their domains $\operatorname{dom} f_\Lambda^{**} = \mathcal{P}_\Lambda$ and the blue curves correspond to the lifted cost f_Λ, see Equation (5.18). The black curves in the bottom row correspond to the moment curve described by Diracs $\varphi(X)$, i.e., the domain of f_Λ. From left to right, see, Example 5.7, Example 5.8 and Example 5.9. The nonlinear supporting dual functions $q_{\lambda,\beta}(x) = \langle \varphi(x), \lambda \rangle - \beta$ (red curves) to f in the top row (middle and right), are transformed into linear supporting hyperplanes $l_{\lambda,\beta}(y) = \langle y, \lambda \rangle - \beta$ (red surfaces) to f_Λ^{**} in the bottom row through the feature map φ. In the language of kernel methods, these functions can be interpreted as the nonlinear decision boundaries that separate individual points on the graph of f from the epigraph of f. Only in the most right case such a separation is possible. As a result the polynomial lifting preserves the nonconvex cost function $f_\Lambda^{**} \circ \varphi = f$ on $\varphi(X)$ whereas the 2-sparse lifting (middle) only leads to a piecewise convex under-approximation $f_\Lambda^{**} \circ \varphi \leq f$.

in linear classifiers.

We call f a Φ-envelope if it can be written in terms of a pointwise supremum over a collection of functions in Λ with constant translation. In view of Lemma 3.2 for such f we have in particular $f = f^{\Phi\Phi}$. Theorem 5.15 therefore identifies all lsc functions $f : X \to \mathbb{R}$ as Φ-envelopes whenever φ is extremal:

Corollary 5.18. *Let $X \subset \mathbb{R}^m$ be nonempty and compact and let $f : X \to \mathbb{R}$ be bounded from below. Furthermore, let $\varphi : X \to \mathbb{R}^n$ be an extremal curve. Then we have*

$$f^{\Phi\Phi} = f_\Lambda^{**} \circ \varphi = \operatorname{cl} f, \tag{5.30}$$

on X.

Proof. Since f is finite-valued and bounded from below on X we have $(\operatorname{cl} f)(x) > -\infty$ for $x \in X$ and therefore $\operatorname{cl} f$ is finite-valued. By [RW98, Exercise 11.63] $f^{\Phi\Phi}$ is the largest

Φ-envelope below f. Since φ is continuous relative to X, $f^{\Phi\Phi} = f_\Lambda^{**} \circ \varphi$ is lsc relative to X. Since $\operatorname{cl} f$ is the largest lsc function below f we have $f^{\Phi\Phi} \leq \operatorname{cl} f$. Since $\operatorname{cl} f \leq f$ we also have $(\operatorname{cl} f)^{\Phi\Phi} \leq f^{\Phi\Phi}$. Invoking Theorem 5.15 we have

$$\operatorname{cl} f = (\operatorname{cl} f)^{\Phi\Phi} \leq f^{\Phi\Phi} \leq \operatorname{cl} f,$$

on X. Therefore $f^{\Phi\Phi} = \operatorname{cl} f$ on X. □

Up to the presence of the compact set X, this result generalizes the basic quadratic transform [RW98, Example 11.66] originally due to [Pol90, Proposition 3.4] (for lsc functions only), which is obtained by choosing $\varphi(x) = (x_1, x_2, \ldots, x_m, \|x\|^2)$.

For this quadratic choice of φ we observe another important relation to the proximal transform from Section 3.2: The above Φ-biconjugate can be seen as a pointwise limiting proximal hull, where one supremizes over the curvature parameter $\lambda > 0$ of the proximal hull h_λ at each point x. We denote by

$$j(x) = \frac{1}{2}\|x\|^2.$$

Then we have

$$f^\Phi(y, r) = \sup_{x \in \mathbb{R}^m} \langle x, y \rangle - rj(x) - f(x) = (f + rj)^*(y).$$

and therefore we can resolve the inner supremum in the Φ-biconjugate to:

$$\begin{aligned} f^{\Phi\Phi}(x) &= \sup_{r \in \mathbb{R}} \sup_{y \in \mathbb{R}^m} \ \langle x, y \rangle - rj(x) - (f + (rj)^*(y) \\ &= \sup_{r \in \mathbb{R}} \ (f + rj)^{**}(x) - rj(x) \\ &= \sup_{\lambda \in \mathbb{R}} \ -e_\lambda(-e_\lambda f)(x) = \sup_{\lambda \in \mathbb{R}} \ h_\lambda f(x). \end{aligned}$$

Also see [RW98, Example 1.44]. In [Bal77, Theorem 1] a similar duality formula is shown for Φ-couplings of a certain "needle-type". Our result is instead based on the extremality condition, which, from a primal point of view, captures an intuitive and sharp (sufficient) condition for the above result for the one-sided linear couplings we consider. For the component functions of φ being the hat basis, see Example 5.8, the class of Φ-envelopes are the piecewise convex functions, see, Figure 5.2 middle.

5.4. Extension to a spatially continuous setting

5.4.1. Lifting to measures for variational problems

Markov random fields are often used to approximate continuous variational problems. For instance, in image processing, the underlying grid graph $\mathcal{G}$ is a discrete approximation to a continuous domain Ω of a corresponding image $I : \Omega \to \mathbb{R}$ which is therefore represented as a function.

This leads us to extend the lifting framework for problems on graphs from the previous section to continuous spatial domains, in the same way the continuous MRF extends the discrete label space in the classical MRF to continuous label spaces.

The continuous variational models that we are interested in take the following form:

$$\inf_{u:\Omega\to X} \left\{ F(u) := \int_\Omega f(x,u(x))\,\mathrm{d}x + TV(u) \right\}. \tag{5.31}$$

Here, the domain $\Omega \subset \mathbb{R}^d$ is nonempty, open and bounded. The range $X \subset (\mathbb{R}^m, \|\cdot\|_2)$ is nonempty and compact. The function $f : \Omega \times X \to \mathbb{R}$ is a possibly nonconvex data fidelity term. For a function $u \in L^1(\Omega;\mathbb{R}^m)$ the functional $TV(u)$ is the *total variation* (TV) of vector-valued functions defined by

$$TV(u) = \sup \left\{ \int_\Omega -\langle \operatorname{Div} p(x), u(x)\rangle\,\mathrm{d}x : p \in \mathcal{C}_c^1(\Omega;\mathbb{R}^d \times \mathbb{R}^m),\ \|p(x)\|_{\mathcal{S}^\infty} \le 1, \forall\, x \in \Omega \right\}, \tag{5.32}$$

where $\mathcal{C}_c^1(\Omega;\mathbb{R}^d \times \mathbb{R}^m)$ is the set of continuously differentiable matrix-valued functions with compact support contained in Ω and $\|\cdot\|_{\mathcal{S}^\infty}$ is the Schatten-∞ norm.

For differentiable functions u the total variation admits a simple primal form

$$TV(u) = \int_\Omega \|\nabla u(x)\|_{\mathcal{S}^1}\,\mathrm{d}x, \tag{5.33}$$

where $\|\cdot\|_{\mathcal{S}^1}$ is the Schatten-1 norm. The TV is a convex functional and favors solutions u that are spatially smooth. At each point $\|\nabla u(x)\|_{\mathcal{S}^1}$ penalizes the sum of the singular values of the Jacobian, which encourages the individual components of u to point in the same direction, see, e.g., [SR96].

The vector-valued TV is actually a special case of the Banach space-valued TV [VL18, Equation 5], which generalizes other existing TV formulations: The framework can be specialized to the total variation of measure-valued functions $\mu : \Omega \to \mathcal{P}(X)$, considered in [Lel+13; Lau+16; VL17] to reformulate the nonconvex functional (5.31) in terms of a convex one:

$$\inf_{\mu:\Omega\to\mathcal{P}(X)} \left\{ \mathcal{F}(\mu) := \int_\Omega \langle f_x, \mu_x\rangle\,\mathrm{d}x + TV_{W_1}(\mu) \right\}. \tag{5.34}$$

Here, $TV_{W_1}(\mu)$ is the Wasserstein-1 total variation [VL18, Equation 3] of measure-valued functions $\mu : \Omega \to \mathcal{P}(X)$ defined by:

$$TV_{W_1}(\mu) = \sup \left\{ \int_\Omega \langle -\operatorname{Div} \lambda(x,\cdot), \mu_x\rangle\,\mathrm{d}x : \lambda \in \mathcal{C}_c^1(\Omega \times X;\mathbb{R}^d),\ \lambda(x,\cdot) \in \operatorname{Lip}(X;\mathbb{R}^d) \right\}, \tag{5.35}$$

where the divergence $\operatorname{Div} \lambda(x,\cdot)$ is taken wrt x, and pointwise in the second argument. Here, $\lambda(x,\cdot) \in \operatorname{Lip}(X;\mathbb{R}^d)$ restricts the Lipschitz constant of $\lambda(x,\cdot)$ to be bounded by 1:

$$\|\lambda(x,y) - \lambda(x,y')\| \le \|y - y'\|, \quad \forall\, y, y' \in X. \tag{5.36}$$

For smooth $p : X \to \mathbb{R}^d$, and if the norms are the 2-norms it is well known that the Lipschitz constraint can be expressed in terms of a bound of the Schatten-∞ norm of the Jacobian:

$$\|\nabla p(x)\|_{\mathcal{S}^\infty} \le 1, \quad \forall\, x \in X. \tag{5.37}$$

For a rigorous treatment of the involved dual and pre-dual spaces, existence of minimizers and well-definedness of the functional we refer to [VL18]: In particular, the existence of a minimizer under the condition that μ is weakly measurable is proved in [VL18, Theorem 1]. The condition ensures in particular that the integrals are well-defined, see [VL18, Lemma 2] in [VL18, Appendix F].

The lifted regularizer TV_{W_1}, can be seen as the natural extension of the Wasserstein-1 distance in the local marginal polytope relaxation (5.7) to a spatially continuous setting.

5.4.2. Discretization with moments and relations to MRFs

To solve the problem on a computer we discretize the domain Ω in terms of a Cartesian grid with points $x_i \in \Omega$. The set of all grid points is denoted by $\mathcal{V}$, $|\mathcal{V}| < \infty$. We define the gradient operator ∇ on a staggered grid using forward differences with von Neumann boundary conditions such that the dual operator $\nabla^* = -\operatorname{Div}$ is the negative divergence with vanishing boundary values.

More generally than existing discretizations for the variational problem (5.34), and in analogy to the previous section, we consider a discretization of the range in terms of moments. To this end we restrict the k^{th} component function of $\lambda(x_i, \cdot) : X \to \mathbb{R}^d$ to a subspace $\lambda_k(x_i, \cdot) \in \Lambda \subset \mathcal{C}(X)$, spanned by basis functions $\Lambda = \langle \varphi_1, \dots \varphi_n \rangle$, e.g. the space of polynomials. Then, at each grid point x_i there exist coefficient matrices $p(x_i) \in \mathbb{R}^{d \times n}$ such that $\lambda(x_i, \cdot) = p(x_i) \cdot \varphi(\cdot)$.

This induces a discretization of the primal variables $\mu(x_i)$ in terms of moments $y \in \mathcal{P}_\Lambda$: The discretized problem (after an interchange of min and sup) amounts to

$$\sup_{\lambda \in (\Lambda^d)^{\mathcal{V}}} \min_{\mu \in \mathcal{P}(X)^{\mathcal{V}}} \sum_{x_i \in \mathcal{V}} \langle f(x_i, \cdot) - (\operatorname{Div} \lambda)(x_i), \mu(x_i) \rangle - \sum_{x_i \in \mathcal{V}} \iota_{\mathcal{K}}(\lambda(x_i, \cdot)), \tag{5.38}$$

where $\mathcal{K} := \operatorname{Lip}(X; \mathbb{R}^d)$ and Div is a discrete divergence operator that maps a function $\lambda : \mathcal{V} \times X \to \mathbb{R}^d$ to $\mathcal{C}(X)^{\mathcal{V}}$. In view of Lemma 5.1 and in analogy to the MRF setting from Section 5.3.1, the formulation can be rewritten:

$$\sup_{p \in (\mathbb{R}^{d \times n})^{\mathcal{V}}} \sum_{x_i \in \mathcal{V}} \min_{z \in X} \langle -(\operatorname{Div}_\Lambda p)(x_i), \varphi(z) \rangle + f(x_i, z) - \sum_{x_i \in \mathcal{V}} \iota_{\mathcal{K}_\Lambda}(p(x_i)), \tag{5.39}$$

which, in terms of the convex conjugates of the lifted versions of $f(x_i, \cdot)$ (5.18), reads:

$$\sup_{p \in (\mathbb{R}^{d \times n})^{\mathcal{V}}} \sum_{x_i \in \mathcal{V}} -f(x_i, \cdot)^*_\Lambda((\operatorname{Div}_\Lambda p)(x_i)) - \sum_{x_i \in \mathcal{V}} \iota_{\mathcal{K}_\Lambda}(p(x_i)). \tag{5.40}$$

Here, $\operatorname{Div}_\Lambda : (\mathbb{R}^{d \times n})^{\mathcal{V}} \to (\mathbb{R}^n)^{\mathcal{V}}$ is a discrete the divergence operator on the coefficients p of λ such that $\langle (\operatorname{Div}_\Lambda p)(x_i), \varphi(\cdot) \rangle = (\operatorname{Div} \lambda)(x_i) \in \Lambda$ when $\lambda(x_i, \cdot) = p(x_i) \cdot \varphi(\cdot)$ for all grid points x_i and the constraint set $\mathcal{K}_\Lambda$ is defined as

$$\mathcal{K}_\Lambda := \left\{ p \in \mathbb{R}^{d \times n} : p \cdot \varphi(\cdot) \in \operatorname{Lip}(X; \mathbb{R}^d) \right\}. \tag{5.41}$$

The corresponding primal problem then amounts to:

$$\min_{y \in (\mathbb{R}^n)^{\mathcal{V}}} \sum_{x_i \in \mathcal{V}} f(x_i, \cdot)^{**}_\Lambda(y(x_i)) + \sum_{x_i \in \mathcal{V}} \sigma_{\mathcal{K}_\Lambda}\big((\nabla_\Lambda\, y)(x_i)\big), \tag{5.42}$$

where $-\operatorname{Div}_\Lambda = \nabla^*_\Lambda$. For further intuition we shall provide an alternative derivation of the Wasserstein-1 TV for moment-valued signals $y : \Omega \to \mathcal{P}_\Lambda$. In contrast to the previous formulation this can be seen as a semidiscrete approach, where $\Omega \subset \mathbb{R}^d$ is a continuous domain but $\mathcal{P}(X)$ is discretized in terms of moments. To this end we adopt the approach in [CCP12] for minimal partitions in a variational framework. For simplicity we restrict $X = [a, b]$, $a < b$ and $y : \Omega \to \mathbb{R}^n$ is a differentiable function. In the unlifted setting, given a differentiable function $u : \Omega \to X$, its total variation is

$$TV(u) = \int_\Omega \|\nabla u(x)\| \, \mathrm{d}x.$$

Consider the lifting map $\varphi : X \to \mathbb{R}^n$ which maps $t \in X$ to the moment curve $\varphi(t) = (t, t^2, \dots, t^n)$ corresponding to the Dirac $\delta_t \in \mathcal{P}(X)$, for $t \in X$. Then we seek to find a new convex functional TV_{W_1} which attains the value $TV_{W_1}(y) = TV(u)$, whenever $y = \varphi \circ u$, i.e., $y(x)$ is a moment vector of a Dirac measure at each point $x \in \Omega$. A tractable approach to obtain such a functional is to consider

$$TV_{W_1}(y) = \int_\Omega \psi^{**}(\nabla y(x)) \, \mathrm{d}x,$$

for an integrand ψ^{**}, which we define as follows: Whenever $y = \varphi \circ u$ for a differentiable function $u : \Omega \to X$, by differentiability of φ, the function y is differentiable as well. We obtain the Jacobian ∇y of $y : \Omega \to \mathbb{R}^n$ by applying the chain rule:

$$\nabla y(x) = \nabla\varphi(u(x)) \otimes \nabla u(x) \in \mathbb{R}^{n \times d},$$

which shows, that for such y, its Jacobian $\nabla y(x)$ is a rank-1 matrix at each $x \in \Omega$. Intuitively, this means that the individual components of y point in the same direction: Then, there exists a direction $\nu \in \mathbb{R}^d$ and $t \in X$ such that $\nabla y(x) = \nabla\varphi(t) \otimes \nu$. This suggests the following choice for ψ:

$$\psi(w) = \begin{cases} \|\nu\| & \text{if } \exists\, t \in X, \nu \in \mathbb{R}^d \text{ such that } w = \nabla\varphi(t) \otimes \nu \\ +\infty & \text{otherwise,} \end{cases} \tag{5.43}$$

which assigns the cost $\|\nu\|$ for Jacobians $\nabla y(x) = \nabla\varphi(t) \otimes \nu$ that are generated by functions of the form $y = \varphi \circ u$, and ∞ otherwise, and therefore $\int_\Omega \psi(\nabla y(x)) \, \mathrm{d}x = TV(u)$, whenever $y = \varphi \circ u$. We are now going to replace $\psi : \mathbb{R}^{n \times d} \to \overline{\mathbb{R}}$ with its biconjugate: The convex conjugate amounts to:

$$\begin{aligned} \psi^*(p) &= \sup_{q \in \mathbb{R}^{n \times d}} \langle q, p \rangle - \psi(q) \\ &= \sup_{t \in X, \nu \in \mathbb{R}^d} \langle \nabla\varphi(t) \otimes \nu, p \rangle - \|\nu\| \\ &= \sup_{t \in X} \sup_{\nu \in \mathbb{R}^d} \langle \nu, p^\top \nabla\varphi(t) \rangle - \|\nu\| \\ &= \begin{cases} 0 & \text{if } \|p^\top \nabla\varphi(t)\|_* \leq 1 \text{ for all } t \in X \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

This is the indicator function of the Lipschitz constraint $p \cdot \varphi(\cdot) \in \mathrm{Lip}(X; \mathbb{R}^d)$ and therefore the biconjugate ψ^{**} is the support function of $\mathcal{K}_\Lambda$ which shows that a polynomial discretization of dual Wasserstein-1 TV is recovered.

Finally we shall discuss the equivalence between the spatially continuous approach and the local marginal polytope relaxation (5.7) for pairwise $f_{uv}(x, x') = |x - x'|$ in the anisotropic case: To this end one specializes the norm on the left hand side of Inequality (5.36) to the ∞-norm such that the constraint separates over the d dimensions. In addition one identifies the Cartesian grid with a grid graph, where the k^{th} component of the gradient operator corresponds to the edges (x_i, x_j) between two adjacent grid points in the k^{th} dimension. The von Neumann boundary conditions ensure that the gradient at the boundary of the grid vanishes. Ignoring these values one ensures that the number of discretized derivatives taken in the d dimensions coincides with the number of edges in the grid graph. Then it easy to see that the formulation (5.20) is recovered. In the isotropic case, however, the discretized continuous approach and the MRF approach are different.

5.5. Derivation of a conic program and implementation

5.5.1. Nonnegativity and moments

After discretization a next step to obtain a practical implementation is to derive finite characterizations of the lifted biconjugates f_Λ^{**} and the constraint set $\mathcal{K}_\Lambda$ in the MRF formulation (5.20) or the discretization of the continuous model (5.42). We will show that the formulations can be rewritten in terms of a semi-infinite conic program which can be implemented using semidefinite programming in the piecewise polynomial case.

Since the MRF formulation is a special case of the variational approach under an anisotropic discretization we limit our discussion to the discretization of the continuous model.

For simplicity, let $f(x_i, \cdot) \in \Lambda$. Expression (5.26) then shows, that the challenging part is to characterize the probability moment space $\mathcal{P}_\Lambda$: The following result shows that up to normalization, $\mathcal{P}_\Lambda$ can be written in terms of the dual cone of the cone of functions in Λ that are nonnegative on X.

Lemma 5.19. *Let $(\mathcal{M}_\Lambda)_+$ be the cone of moments of nonnegative measures defined by*

$$(\mathcal{M}_\Lambda)_+ := \{y \in \mathbb{R}^{n+1} \colon \exists \mu \in \mathcal{M}_+(X), y_k = \langle \mu, \varphi_k \rangle\}, \tag{5.44}$$

and let $\mathcal{N}_\Lambda$ be the cone of the coefficients of the functions in $\Lambda = \langle \varphi_0, \dots, \varphi_n \rangle$ that are nonnegative on X defined as:

$$\mathcal{N}_\Lambda := \{p \in \mathbb{R}^{n+1} : \langle p, \varphi(x) \rangle \geq 0, \ \forall\, x \in X\}. \tag{5.45}$$

Then $(\mathcal{M}_\Lambda)_+$ is equal to $\mathcal{N}_\Lambda^*$, where $\mathcal{N}_\Lambda^*$ denotes the dual cone of $\mathcal{N}_\Lambda$.

If, in addition, $\varphi_0 \equiv 1$ we also have $\mathcal{P}_\Lambda = \{y \in (\mathcal{M}_\Lambda)_+ \colon y_0 = 1\}$.

Proof. Let $y \in (\mathcal{M}_\Lambda)_+$. This means there exists $\mu \in \mathcal{M}_+(X)$ such that $y_k := \int_X \varphi_k(x) \, \mathrm{d}\mu(x)$. Let $p \in \mathcal{N}_\Lambda$. Because of $\langle p, \varphi(\cdot) \rangle \in \Lambda \subset \mathcal{C}(X)$ and $\langle p, \varphi(x) \rangle \geq 0$ for all $x \in X$ and $\mu \in \mathcal{M}_+(X)$ is a nonnegative measure it holds

$$\langle p, y \rangle = \sum_{k=0}^{n} p_k \int_X \varphi_k(x) \, \mathrm{d}\mu(x) = \int_X \langle p, \varphi(x) \rangle \, \mathrm{d}\mu(x) \geq 0,$$

for all $x \in X$. Since $p \in \mathcal{N}_\Lambda$ was an arbitrary choice from $\mathcal{N}_\Lambda$ we have $y \in \mathcal{N}_\Lambda^*$.

Next we show $(\mathcal{M}_\Lambda)_+^* \subseteq \mathcal{N}_\Lambda$ as this implies $\mathcal{N}_\Lambda^* \subseteq (\mathcal{M}_\Lambda)_+^{**} = (\mathcal{M}_\Lambda)_+$, where the last equality holds since $(\mathcal{M}_\Lambda)_+$ is closed and convex. Take $p \in (\mathcal{M}_\Lambda)_+^*$. Let $x \in X$. Now we choose $y \in (\mathcal{M}_\Lambda)_+$ such that $y_k = \langle \delta_x, \varphi_k \rangle = \varphi_k(x)$. Then $p \in (\mathcal{M}_\Lambda)_+^*$ implies that $\langle p, y \rangle \geq 0$. Since the choice $x \in X$ was arbitrary we have $\langle p, \varphi(x) \rangle \geq 0$ for all $x \in X$ and therefore $p \in \mathcal{N}_\Lambda$.

Finally, $\mathcal{P}_\Lambda = \{y \in (\mathcal{M}_\Lambda)_+ \colon y_0 = 1\}$ follows from the fact that $\mu \in \mathcal{M}(X)$ is an element of $\mathcal{P}(X)$ if and only if $\mu \in \mathcal{M}_+(X)$ and $\langle \mu, \varphi_0 \rangle = 1$. □

Thanks to algebraic geometry, there exist finite semidefinite programming characterizations of $\mathcal{N}_\Lambda$ in the univariate piecewise polynomial case.

The constraint set $\mathcal{K}_\Lambda$ corresponds to the Lipschitz constraint of the dual variable $\lambda(x_i, \cdot) \in \mathrm{Lip}(X; \mathbb{R}^d)$. Assume that Λ is closed under differentiation, i.e., φ is differentiable with $\varphi_k' \in \Lambda$. We show, that the Lipschitz constraint can be characterized in terms of nonnegativity of functions in Λ as well. We restrict $X = [a, b]$, $a < b$ to be an interval. Then, the Lipschitz constraint reads:

$$\|\nabla \lambda(x_i, y)\| \leq 1, \; \forall \, y \in [a, b].$$

In the anisotropic case, i.e., $\|\cdot\| = \|\cdot\|_\infty$ the bound separates over the d dimensions and one obtains the constraints $-1 \leq \lambda_k(x_i, \cdot)'(y) \leq 1$ for all $y \in [a, b]$, where $\lambda_k(x_i, \cdot)'$ is the derivative of $\lambda_k(x_i, \cdot)$. Equivalently, this means that the coefficients of the functions $1 + \lambda_k(x_i, \cdot)'$ and $1 - \lambda_k(x_i, \cdot)'$ are in $\mathcal{N}_\Lambda$.

In the isotropic case, i.e., $\|\cdot\| = \|\cdot\|_2$ we restrict Λ to the space of univariate polynomials. Squaring both sides of the inequality one obtains a single nonnegativity constraint of the from $1 - \sum_{k=1}^d \lambda_k(x_i, \cdot)'(y)^2 \geq 0$ for all $y \in [a, b]$.

5.5.2. Semidefinite programming and polynomial duals

As we have seen in the previous section, an important ingredient for a tractable formulation is the efficient characterization of nonnegativity of functions in a finite-dimensional subspace $\Lambda \subset \mathcal{C}(X)$. A promising choice of Λ in that regards is the space of polynomials. Indeed, the characterization of nonnegativity of polynomials is a fundamental problem in *convex algebraic geometry* surveyed in [BPT12]: Let $\mathbb{R}[x_1, \ldots, x_m]$ denote the ring of possibly multivariate polynomials with $p \in \mathbb{R}[x_1, \ldots, x_m]$ then $p = \sum_{\alpha \in I} p_\alpha x^\alpha$ for monomials x^α. Let $\deg p$ denote its degree. A key result from real algebraic geometry is the Positivstellensatz due to [Kri64] and [Ste74] refined in [Sch91] and [Put93]. It characterizes polynomials $p \in \mathbb{R}[x_1, \ldots, x_m]$ that are positive on semi-algebraic sets X, i.e. $p(x) > 0$ for all $x \in X$, where X is defined in terms of polynomial inequalities.

Key to such results is a *certificate of nonnegativity* of the polynomial p that involves *sum-of-squares* (SOS) multipliers q, where q is SOS if $q(x) = \sum_{i=1}^N q_i^2(x)$ for polynomials $q_i \in \mathbb{R}[x_1, \ldots, x_m]$.

Due to Hilbert it is known that in the unconstrained case $X = \mathbb{R}^m$ the set of nonnegative polynomials is equal to the set of SOS-polynomials if and only if p is univariate, p is quadratic, or p is a bivariate quartic polynomial. This yields a SOS characterization for nonnegative polynomials in the unconstrained case $X = \mathbb{R}^m$.

In the constrained case a sharp characterization in terms of SOS can be obtained for univariate polynomials and constraint sets which are intervals $X = [a, b]$: Thanks to [BPT12, Theorem 3.72] originally due to [PR00, Corollary 2.3] we have following result:

Lemma 5.20. *Let $a < b$. Then the univariate polynomial $p \in \mathbb{R}[x]$ is nonnegative on*

$[a, b]$ *if and only if it can be written as*

$$p(x) = \begin{cases} s(x) + (x-a)\cdot(b-x)\cdot t(x) & \text{if } \deg p \text{ is even,} \\ (x-a)\cdot s(x) + (b-x)\cdot t(x) & \text{if } \deg p \text{ is odd,} \end{cases} \tag{5.46}$$

where $s, t \in \mathbb{R}[x]$ *are sum of squares. If* $\deg p = 2n$*, then we have* $\deg s \leq 2n$, $\deg t \leq 2n-2$*, while if* $\deg p = 2n+1$*, then* $\deg s \leq 2n, \deg t \leq 2n$.

The above result can be seen as a refinement of the Positivstellensatz in the univariate case, where X is a closed interval. Remarkably, in contrast to the general case, the above result provides us with explicit upper bounds of the degrees of the SOS multipliers s and t that are important to derive a practical implementation: If such upper bounds are available the SOS constraints can be formulated in terms of semidefinite programming: We adopt [BPT12, Lemma 3.33] and [BPT12, Lemma 3.34]:

Lemma 5.21. *A univariate polynomial* $p \in \mathbb{R}[x]$ *with* $\deg p = 2n$, $n \geq 0$ *is SOS if and only if there exists a positive semidefinite matrix* $Q \in \mathbb{R}^{n+1\times n+1}$ *such that*

$$p_k = \sum_{\substack{0\leq i,j\leq n,\\ i+j=k}} Q_{ij}, \quad \forall\, 0 \leq k \leq 2n. \tag{5.47}$$

Invoking the results above SDP-duality yields the following compact representation of $(\mathcal{M}_\Lambda)_+$:

Lemma 5.22. *Let* $n \geq 0$*. For odd degree* $2n+1$, $y \in (\mathcal{M}_\Lambda)_+$ *if and only if*

$$bM_{0,n}(y) \succeq M_{1,n}(y) \succeq aM_{0,n}(y), \tag{5.48}$$

for Hankel matrices

$$M_{i,n}(y) := \begin{bmatrix} y_i & y_{i+1} & \cdots & y_{i+n} \\ y_{i+1} & y_{i+2} & \cdots & y_{i+n+1} \\ \vdots & & \cdots & \vdots \\ y_{i+n} & & \cdots & y_{i+2n} \end{bmatrix}. \tag{5.49}$$

For even degree $2n$, $y \in (\mathcal{M}_\Lambda)_+$ *if and only if*

$$M_{0,n}(y) \succeq 0, \tag{5.50}$$
$$(a+b)\, M_{1,n-1}(y) - abM_{0,n-1}(y) \succeq M_{2,n-1}(y). \tag{5.51}$$

Proof. Follows by Lemma 5.20 and Lemma 5.21 invoking elementary SDP-duality. □

It is worth noting that this is essentially the same framework adopted by [Las01; Las02] for multivariate constrained polynomial optimization. However, in contrast to [Las01; Las02], our goal is to characterize the lifted biconjugates of $f : X \to \mathbb{R}$ at each grid point whose characterization is equivalent to a univariate polynomial optimization problem. In the MRF setting the relation to the approach in [Las01; Las02] can be made more precise: For polynomial unary terms f_u and pairwise terms f_{uv} the MRF problem (5.1) is a multivariate polynomial formulated over the high-dimensional cube $[a, b]^{\mathcal{V}}$. Indeed, applying the approach of [Las01; Las02] directly to this multivariate polynomial optimization problem corresponds to solving the marginal polytope relaxation which is tight but intractable for large $\mathcal{V}$ as the number of coupling moments explodes for large $|\mathcal{V}|$. In contrast, our framework applies to the local marginal polytope relaxation which is not tight in general but leads to a tractable formulation as it exploits the sparse structure of the optimization problem. For polynomial unaries and pairwise terms the local marginal polytope relaxation is closely related to sparse sum-of-squares approaches [Wak+06; WLT18] and in particular sparse versions of the Positivstellensatz [WLT18].

5.5.3. A First-Order Primal-Dual Algorithm

We are now ready to describe the algorithm for solving the resulting semidefinite program. Again, since the MRF formulation is a special case of the variational approach under an anisotropic discretization we limit our discussion to the implementation of the continuous model. In addition we first consider the case $f(x_i, \cdot) \in \Lambda$ and Λ is the space of univariate polynomials. We propose to use the primal-dual hybrid gradient algorithm [Poc+09; CP11], as it can exploit the partially separable structure of our SDP. The primal-dual algorithm optimizes the problem (5.42) via alternating projected gradient descent/ascent steps applied to the saddle-point formulation of (5.42). The saddle-point problem is obtained by expanding the support function in Problem (5.42) and substituting the expression

$$f(x_i, \cdot)^{**}_{\Lambda}(y(x_i)) = \langle a(x_i), y(x_i) \rangle + \iota_{\mathcal{P}_\Lambda}(y(x_i)),$$

for the linear lifted biconjugates, also see Equation (5.26).

$$\min_{y \in (\mathcal{P}_\Lambda)^{\mathcal{V}}} \max_{p \in (\mathcal{K}_\Lambda)^{\mathcal{V}}} \sum_{x_i \in \mathcal{V}} \langle (\nabla_\Lambda \, y)(x_i), p(x_i) \rangle + \sum_{x_i \in \mathcal{V}} \langle a(x_i), y(x_i) \rangle. \tag{5.52}$$

In each iteration the algorithm performs a projected gradient ascent step in the dual p followed by a projected gradient descent step in the primal variable y. Subsequently it performs an extrapolation step in the primal. For step-sizes $\tau, \sigma > 0$ with $\tau\sigma \|\nabla_\Lambda\| \leq 1$ the update steps read as follows:

$$\begin{cases} p^{t+1} = \operatorname{proj}_{(\mathcal{K}_\Lambda)^{\mathcal{V}}}(p^t + \sigma(\nabla_\Lambda \bar{y}^t + a)) \\ y^{t+1} = \operatorname{proj}_{(\mathcal{P}_\Lambda)^{\mathcal{V}}}(y^t + \tau \operatorname{Div}_\Lambda p^{t+1}) \\ \bar{y}^{t+1} = 2y^{t+1} - y^t. \end{cases} \tag{5.53}$$

The projection operators $\operatorname{proj}_{(\mathcal{P}_\Lambda)^{\mathcal{V}}}$ resp. $\operatorname{proj}_{(\mathcal{K}_\Lambda)^{\mathcal{V}}}$ onto the sets $(\mathcal{P}_\Lambda)^{\mathcal{V}}$ and $(\mathcal{K}_\Lambda)^{\mathcal{V}}$ are separable and can therefore be carried out in parallel on a GPU using the SDP characterizations derived above. For practicality, we introduce additional auxiliary variables and linear constraints to decouple the affine constaints (5.47) resp. Equation (5.49) and the SDP constraints (5.48). The projection operator of the semidefinite cone can then be solved using an eigenvalue decomposition.

5.5.4. Piecewise polynomial duals and nonlinear lifted biconjugates

The polynomial discretization can be extended by means of a continuous piecewise polynomial representation of the dual variables resulting in a possibly more accurate approximation of the dual subspace Λ. Then, both, nonnegativity and Lipschitz continuity can be enforced on each piece X_k individually. Continuity of the piecewise polynomial dual variables can be enforced via linear constraints. The corresponding primal variable y belongs to $y \in (\mathcal{M}_\Lambda)_+ \times (\mathcal{M}_\Lambda)_+ \times \cdots \times (\mathcal{M}_\Lambda)_+$. Then the restriction that y is a moment vector of a probability measure supported on the whole space X yields an additional sum-to-one constraint on the 0^{th} moments $1 = \sum_{k=1}^{K} y_{k,0}$.

Another issue to address is when $f_u \notin \Lambda$ which results in a nonlinear lifted biconjugate over the probability moment space as in Figure 5.2.

The formulation which is derived next addresses both: In particular it allows one to choose Λ independently from f_u which can even be discontinuous, as long as f_u has a certain piecewise polynomial structure. Key to the formulation is to rewrite the inner minimum in the dual formulation 5.39 exploiting a duality between nonnegativity and minimization of functions:

Let $X = [a, b]$, $a < b$ be a compact interval. Let $a = t_1 < t_2 < t_3 < \cdots < t_{K+1} = b$ be a sequence of knots, where $X_k := [t_k, t_{k+1}]$. Let Θ be the space of univariate polynomials with some maximum degree n. Let $f : X \to \mathbb{R}$ be a possibly discontinuous lsc piecewise polynomial function defined by $f(x) = \min_{1 \le k \le K} f_k(x) + \iota_{X_k}(x)$ with $f_k \in \Theta$, i.e., $f_k(x) = \langle \varphi(x), a_k \rangle$ for coefficients $a_k \in \mathbb{R}^{n+1}$, where $\varphi_0 \equiv 1$. First observe the following duality between nonnegativity and minimization of a lsc function:

$$\min_{x \in X} \; f(x) = \max_{q \in \mathbb{R}} \; q - \iota_{\mathcal{N}}(f - q),$$

where we denote by $\mathcal{N}(X) = \{\lambda : X \to \mathbb{R} : \lambda(x) \ge 0, \forall\, x \in X\}$ the cone of nonnegative lsc functions on X. Then we obtain for $Aq = (e_0 q, \dots, e_0 q)$, where $e_0 = (1, 0, \dots, 0) \in \mathbb{R}^{n+1}$ is the 0^{th} unit vector:

$$\begin{aligned} \min_{x \in X} \; f(x) &= \max_{q \in \mathbb{R}} \; q - \iota_{\mathcal{N}(X)}(f - q) && (5.54) \\ &= \max_{q \in \mathbb{R}} \; q - \sum_{k=1}^{K} \iota_{\mathcal{N}_\Theta}(a_k - A_k q). && (5.55) \end{aligned}$$

Fenchel–Rockafellar duality then yields:

$$\begin{aligned} \min_{x \in X} \; f(x) &= \min_{y \in (\mathbb{R}^{n+1})^K} \iota_{\{1\}}(A^* y) + \sum_{k=1}^{K} \iota_{(\mathcal{M}_\Theta)_+}(y_k) + \langle y_k, a_k \rangle \\ &= \min_{\substack{y \in ((\mathcal{M}_\Theta)_+)^K \\ \sum_{k=1}^K y_{k,0} = 1}} \sum_{k=1}^{K} \langle y_k, a_k \rangle. \end{aligned}$$

This formulation can be substituted in the dual problem (5.39) and we obtain

$$\sup_{p \in (\mathbb{R}^{d \times n})^{\mathcal{V}}} \sum_{x_i \in \mathcal{V}} \min_{\substack{y \in ((\mathcal{M}_\Theta)_+)^K \\ \sum_{k=1}^K y_{k,0} = 1}} \sum_{k=1}^{K} \langle y_k, a(x_i)_k - (\operatorname{Div}_\Lambda p)(x_i)_k \rangle - \sum_{x_i \in \mathcal{V}} \iota_{\mathcal{K}_\Lambda}(p(x_i)), \qquad (5.56)$$

Here the dual variables λ are chosen such that $(\operatorname{Div}_\Lambda p)(x_i)$ represents a piecewise polynomial with knots $a = t_1 < t_2 < t_3 < \cdots < t_{K+1} = b$ such that for each piece we have $\langle (\operatorname{Div}_\Lambda p)(x_i)_k, \varphi(\cdot) \rangle \in \Theta$. Note that this does require Λ to be equal the whole space of continuous piecewise polynomials of degree n. Indeed, Λ can be a subspace thereof which covers the case where $f_u \notin \Lambda$.

5.5.5. Numerical optimization for multivariate piecewise linear duals

In this section we consider the multivariate case in the spatially continuous variational setting restricting Λ to the space of piecewise linear functions. We present an alternative derivation of the optimization problem also considering $f(x_i, \cdot) \notin \Lambda$. We assume that $X = [a, b]^m$ is a compact cube in $\mathbb{R}^m$. Then X can be partitioned into a disjoint (up to measure zero) union $X = \bigcup_{\Delta_j \in \mathcal{D}}$ of m-simplices $\Delta_j = \operatorname{con}\{t_{j_0}, t_{j_1}, \dots t_{j_m}\} \in \mathcal{D}$ with $\{t_{j_0}, t_{j_1}, \dots t_{j_m}\}$ affinely independent in $\mathbb{R}^m$ and $|\mathcal{D}| = l$. Also see [RW98, Exercise 2.28] for the technology of simplices. The set of all samples t_{j_k} is denoted by $\mathcal{T}$ where $|\mathcal{T}| = n$. Then we restrict the dual variable λ to a piecewise linear continuous family of functions that are linear on each simplex Δ_j. A natural choice for a basis which spans the space of piecewise linear continuous functions is the *hat basis* $\langle \varphi_1, \dots, \varphi_n \rangle$. Then, the associated lifting map $\varphi : \mathbb{R}^m \to \mathbb{R}^n$ can be written compactly as follows. For any $x \in X$ we have:

$$\varphi(x) = \sum_{k=0}^{m} \pi_k e_{j_k} = E_j \pi, \quad \text{if } x \in \Delta_j \text{ such that } \sum_{k=0}^{m} \pi_k t_{j_k} = x = T_j \pi. \tag{5.57}$$

Here, $\pi \in \Pi := \{\pi \in \mathbb{R}^m : \sum_{k=0}^m \pi_k = 1, \pi_k \geq 0\}$ are the barycentric coordinates of x wrt Δ_j and the matrices $E_j \in \mathbb{R}^{n \times m+1}$ and $T_j \in \mathbb{R}^{m \times m+1}$ are defined by:

$$E_j = (e_{j_0}, e_{j_1}, \dots, e_{j_m}), \qquad T_j = (t_{j_0}, t_{j_1}, \dots, t_{j_m}),$$

where e_{j_k} is the $j_k{}^{\text{th}}$ unit vector. Also see Example 5.8, for the univariate case. In the terminology of approximation of measures in the primal from Section 5.3.2 this yields an interesting sparse embedding $\varphi(x)$ of the Diracs δ_x, where the probability moment space $\mathcal{P}_\Lambda = \operatorname{con} \varphi(X) \subset \mathbb{R}^n$ is the unit simplex.

We proceed to compute the lifted biconjugates f_Λ for the particular choice of φ:

We have:

$$f_\Lambda(y) = \begin{cases} f(T_j \pi) & \text{if } y = E_j \pi \text{ for some } \pi \in \Pi, 1 \leq j \leq l, \\ +\infty & \text{otherwise.} \end{cases}$$

The convex conjugate is then given by:

$$f_\Lambda^*(z) = \max_{1 \leq j \leq l} \sup_{\pi \in \Pi} \langle z, E_j \pi \rangle - f(T_j \pi).$$

Fix j. We introduce a substitution $x = T_j \pi$. Since the columns in T_j are affinely independent we have:

$$\begin{pmatrix} T_j \\ \mathbf{1}^\top \end{pmatrix}^{-1} \begin{pmatrix} x \\ 1 \end{pmatrix} = A_j x + b_j = \pi, \qquad \begin{pmatrix} A_j & b_j \end{pmatrix} = \begin{pmatrix} T_j \\ \mathbf{1}^\top \end{pmatrix}^{-1},$$

where $b_j \in \mathbb{R}^n$ and $A_j \in \mathbb{R}^{n \times m}$. Then we can rewrite the inner supremum in terms of the convex conjugate of $f + \iota_{\Delta_j}$:

$$\sup_{x \in \Delta_j} \langle E_j^\top z, A_j x + b_j \rangle - f(x) = (f + \iota_{\Delta_j})^*(A_j^\top E_j^\top z) + \langle E_j b_j, z \rangle.$$

Overall we obtain:

$$f_\Lambda^*(z) = \max_{1 \leq j \leq l} (f + \iota_{\Delta_j})^*(A_j^\top E_j^\top z) + \langle E_j b_j, z \rangle.$$

We consider the biconjugate which amounts to:

$$f_\Lambda^{**}(y) = \sup_{z \in \mathbb{R}^n} \langle z, y \rangle - \max_{1 \leq j \leq l} (f + \iota_{\Delta_j})^*(A_j^\top E_j^\top z) + \langle E_j b_j, z \rangle.$$

We introduce an additional variable q along with constraints $q \geq (f + \iota_{\Delta_j})^*(A_j^\top E_j^\top z) + \langle E_j b_j, z \rangle$ for all $1 \leq j \leq l$ to absorb the inner discrete maximum.

We will rewrite the nonnegativity constraints in terms of epigraphical constraints: $(A_j^\top E_j^\top z, q - \langle E_j b_j, z \rangle) \in \operatorname{epi}(f + \iota_{\Delta_j})^*$ and obtain:

$$f_\Lambda^{**}(y) = \sup_{\substack{z \in \mathbb{R}^n, \\ q \in \mathbb{R}}} \langle z, y \rangle - q - \sum_{j=1}^{l} \iota_{\operatorname{epi}(f + \iota_{\Delta_j})^*}(A_j^\top E_j^\top z, q - \langle E_j b_j, z \rangle).$$

Like before we propose to use the PDHG algorithm which solves the saddle-point formulation:

$$\min_{y \in (\mathbb{R}^n)^{\mathcal{V}}} \max_{p \in (\mathcal{K}_\Lambda)^{\mathcal{V}}} \sum_{x_i \in \mathcal{V}} f(x_i, \cdot)_\Lambda^{**}(y(x_i)) - \sum_{x_i \in \mathcal{V}} \langle y(x_i), (\operatorname{Div}_\Lambda p)(x_i) \rangle,$$

so that the characterization of f_Λ^{**} given above suffices for a practical implementation, as long as the projections onto $\operatorname{epi}(f + \iota_{\Delta_j})^*$ can be implemented efficiently. In case f is a quadratic we decompose the epigraph in terms of the Minkowski sum of the epigraphs of the individual functions: $\operatorname{epi}(f + \iota_{\Delta_j})^* = \operatorname{epi} f^* + \operatorname{epi} \iota_{\Delta_j}^*$. For the projection onto the epigraph of a multivariate quadratic function we use the method described in [SCC14, Appendix B.2].

Alternatively, we consider f to be polyhedral on each simplex Δ_j. I.e., f attains finite values $f(w)$ at a finite subset of points $w \in \mathcal{W}_j \subset \Delta_j$ and interpolates linearly between them. Then we can write the conjugate:

$$(f + \iota_{\Delta_j})^*(z) = \max_{w \in \mathcal{W}_j} \langle w, z \rangle - f(w).$$

For the projection onto the epigraph of such a function, one solves a quadratic program of the form

$$\min_{x \in \mathbb{R}^m, y \in \mathbb{R}} \frac{1}{2}\|x - c\|^2 + \frac{1}{2}|y - d|^2 \text{ s.t. } \langle w, x \rangle - f(w) \leq y, \forall\, w \in \mathcal{W}_j. \tag{5.58}$$

We implement the primal active-set method described in [NW06, Algorithm 16.3].

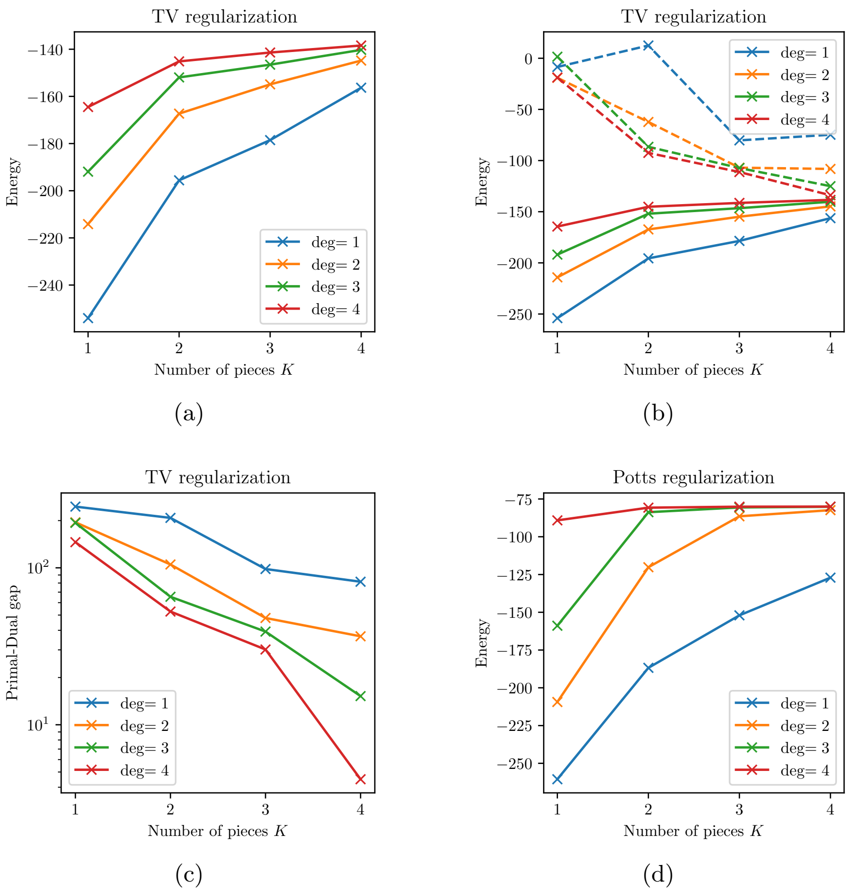
It remains to implement the Lipschitz constraint on the dual variables $\lambda : \mathbb{R}^m \to \mathbb{R}^d$ with $\lambda(x) = p \cdot \varphi(x)$ and $p \in \mathbb{R}^{d \times n}$.

By the choice of the basis $\langle \varphi_1, \ldots, \varphi_n \rangle$ the dual variable λ is a piecewise linear function on X and therefore it suffices to bound the Schatten-∞ norm of the Jacobian $\nabla \lambda$ on each simplex $\Delta_j = \operatorname{con}\{t_{j_0}, t_{j_1}, \ldots t_{j_m}\}$. For $x \in \Delta_j$ we can therefore find $B_j \in \mathbb{R}^{d \times m}$ and $c_j \in \mathbb{R}^d$ such that $\lambda(x) = p \cdot \varphi(x) = B_j x + c_j$. Then the Jacobian $\nabla \lambda(x) = B_j$ is obtained as follows: For $x = t_{j_k}$ we have by definition of the hat basis $\varphi(t_{j_k}) = e_{j_k}$ and therefore for any $1 \leq k \leq m$:

$$\lambda(t_{j_k}) = B_j t_{j_k} + c_j = p_{j_k},$$

where p_{j_k} is the k^{th} column of p. We subtract the m^{th} equation from the other equations and obtain the following system of linear equations:

$$B_j(t_{j_1} - t_{j_m}, t_{j_2} - t_{j_m}, \ldots t_{j_{m-1}} - t_{j_m}) = (p_{j_1} - p_{j_m}, p_{j_2} - p_{j_m}, \ldots, p_{j_{m-1}} - p_{j_m}).$$



(a)

(b)

(c)

(d)

Figure 5.3.: Primal and dual energies for MAP-inference in a continuous MRF with TV regularization using a piecewise polynomial hierarchy of dual variables. (a) shows the dual energy for TV. In (b) the dashed lines correspond to the primal energy at the rounded solution and the solid lines correspond to the dual energy. (c) shows the gap between the nonconvex primal energy at the rounded solution and the dual energy for TV regularization. (d) shows the dual energies for Potts regularization.

Since $\{t_{j_0}, t_{j_1}, \dots t_{j_m}\}$ is affinely independent we obtain:

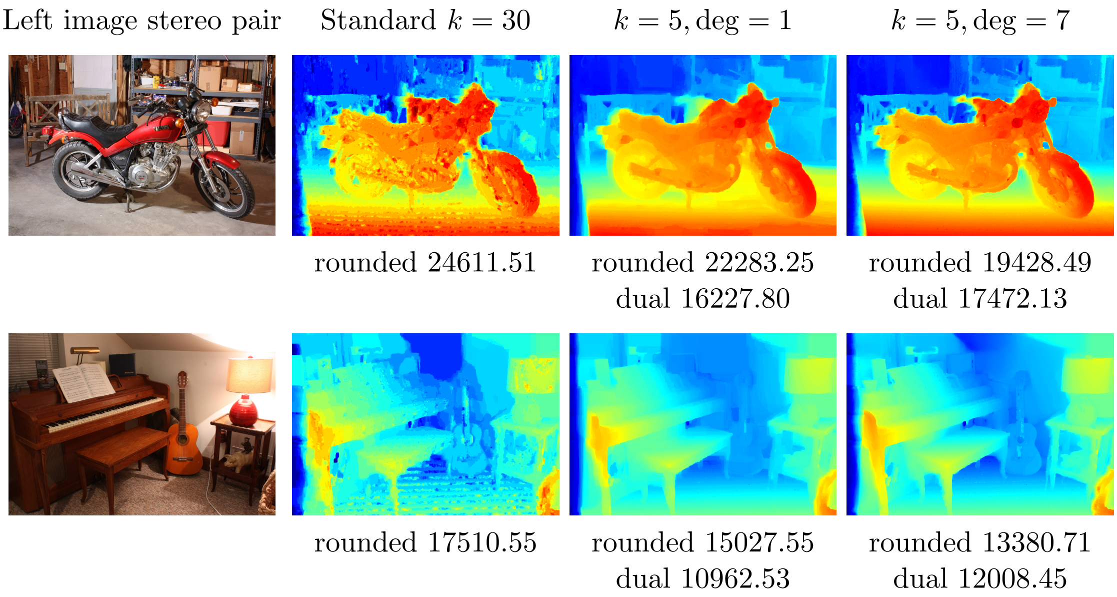
$$B_j = (p_{j_1} - p_{j_m}, p_{j_2} - p_{j_m}, \dots, p_{j_{m-1}} - p_{j_m})(t_{j_1} - t_{j_m}, t_{j_2} - t_{j_m}, \dots t_{j_{m-1}} - t_{j_m})^{-1}.$$

Then the Lipschitz constraint can be formulated in terms of constraints $\|B_j\|_{\mathcal{S}^\infty} \leq 1$ for each simplex $\Delta_j \in \mathcal{D}$. The corresponding projection operators can be solved using a singular value decomposition of B_j.

5.6. Numerical results and applications in computer vision

5.6.1. Empirical convergence for a piecewise polynomial hierarchy

In this first experiment we evaluate the local marginal polytope relaxation of the MRF formulation (5.7) using a piecewise polynomial hierarchy of dual variables. We choose a vertex grid of size 16 × 16. We fix a random polynomial data term of degree 4 at each



Left image stereo pair | Standard $k = 30$ | $k = 5, \deg = 1$ | $k = 5, \deg = 7$

rounded 24611.51 | rounded 22283.25, dual 16227.80 | rounded 19428.49, dual 17472.13

rounded 17510.55 | rounded 15027.55, dual 10962.53 | rounded 13380.71, dual 12008.45

Figure 5.4.: Stereo disparity estimation from a stereo image pair: Left: Standard MRF/OT discretization implemented using a continuous piecewise linear under-approximation for the unaries and piecewise linear duals. middle: piecewise linear duals. right: piecewise polynomial duals. The visual appearence of the solution to the standard MRF/OT discretization shows a strong grid bias. The dual energy gap increases for increasing the degree and/or the number of pieces. Likewise the energy at the rounded solution decreases.

vertex by fitting a random sample of data points. To obtain a high-accuracy solution we solve the primal SDP formulation corresponding to the saddle-point formulation (5.56) specialized to the anisotropic case with MOSEK[1]. For recovering a primal solution at each vertex u we compute the mode wrt the 0^{th} moments to select the best interval denoted by $k^* = \arg\max_{1 \leq k \leq K} (y_u)_{k,0}$. Then we compute the mean of the discretized measure corresponding to $(k^*)^{\text{th}}$ interval as $x_u = (y_u)_{k^*,1}$.

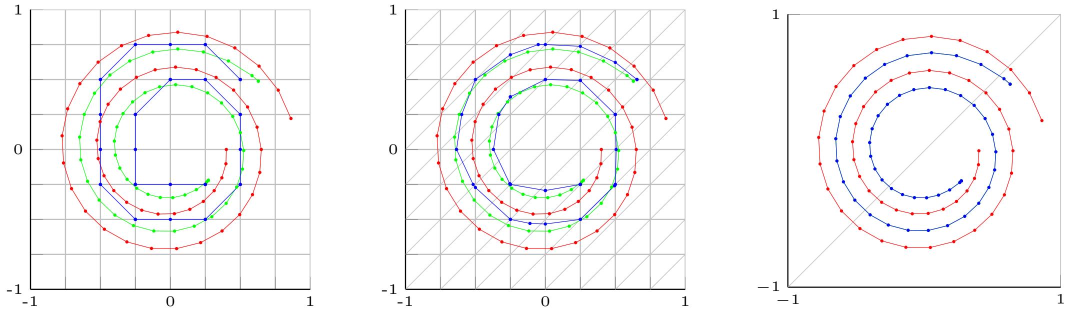
Figure 5.3 visualizes the primal and dual energies for varying degrees and/or number of pieces of the dual variable. While the dual energy strictly increases with higher degrees and/or number of pieces the primal energy is evaluated at the rounded solution and therefore does not strictly decrease in general. While for TV increasing the degree vs. increasing the number of pieces (for $K \cdot \deg$ constant) leads to similar performance, for Potts, in many situations, increasing the degree leads to larger dual energies, e.g., consider $\deg = 4, K = 1$ vs. $\deg = 1, K = 4$, red curve vs. blue curve in Figure 5.3(d). In further experiments, we observed, that this holds in particular when the structure of the dual variables and the unaries match, i.e., $f_u, \lambda_e \in \Lambda$. Note that for Potts, since the dual variables are uniformly bounded on X and the derivative can be unbounded we drop the continuity constraint which leads to a more compact formulation and larger dual energies.

[1] https://www.mosek.com/products/academic-licenses

	Dual energies		
deg	$K = 1$	$K = 3$	$K = 5$
1	14180.08	15733.71	16227.80
2	15052.18	16430.97	16773.75
3	15601.89	16778.87	17055.25
4	15938.92	16998.63	17235.21
5	16191.40	17147.92	17346.49
6	16369.95	17243.47	17422.26
7	16480.22	17308.91	17472.13

	Energies rounded		
deg	$K = 1$	$K = 3$	$K = 5$
1	28982.45	25005.05	22283.26
2	31038.14	22525.12	21049.32
3	28505.41	21500.65	20428.34
4	27255.48	20841.62	20049.04
5	25795.56	20344.17	19764.23
6	24081.51	20032.77	19559.00
7	23142.67	19869.71	19428.50

Table 5.1.: Energies for stereo matching Motorcycle. Left: Dual energies. Right: Primal energies at the rounded solution.



(a) Naive, 81 labels.

(b) [Lel+13], 81 labels.

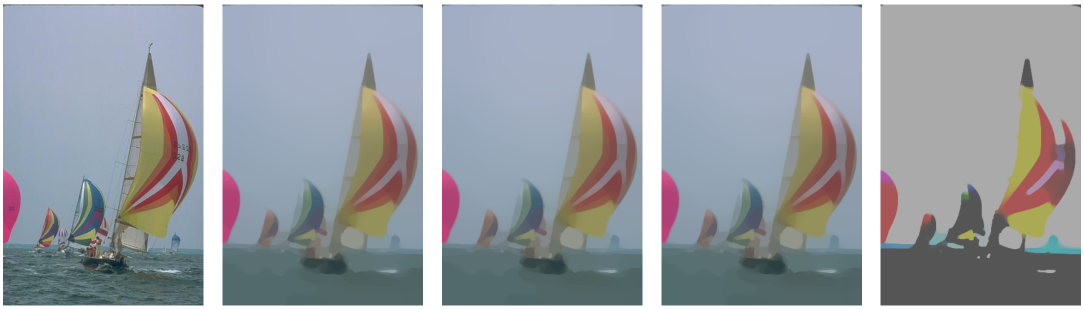
(c) Ours, **4 labels**.

Figure 5.5.: ROF denoising of a vector-valued signal $I : [0,1] \to [-1,1]^2$, discretized on 50 points (shown in red). We compare the proposed approach (c) with the standard discretization for vectorial TV [Lel+13] (a) and (b). The samples t_{j_k} are visualized by the gray grid. In contrast to the standard discretization, the proposed approach does not exhibit any visible grid bias providing fully sublabel-accurate solutions.

5.6.2. Stereo Matching

In this experiment we consider stereo matching using the anisotropic relaxation (5.7). We consider the Motorcycle image from the Middlebury benchmark [Sch+14]. We downsample the image by factor 4. The disparity cost term was first calculated using 135 discrete disparities obtained by shifting the images by the corresponding amount of pixels and comparing the image gradients. Then, the cost dataterm is approximated from below in terms of a continuous piecewise cubic polynomial f_u using 30 pieces at each $u \in \mathcal{V}$. In Figure 5.4 we compare the standard MRF/OT discretization as described in Example 5.7 with our framework using a piecewise polynomial hierarchy of dual variables. The standard MRF/OT discretization is equivalent to a piecewise linear approximation of the data term with piecewise linear duals in our framework. For a fair comparison we use a piecewise linear under-approximation of the continuous piecewise cubic polynomial under-approximation f_u. As the resulting optimization problem is large-scale we solve the formulation (5.20) with PDHG [CP11] as described in Section 5.5.3 using the GPU-based PDHG framework prost[2]. In contrast to the previous experiment which uses a combined

[2] https://github.com/tum-vision/prost



(a) Input image

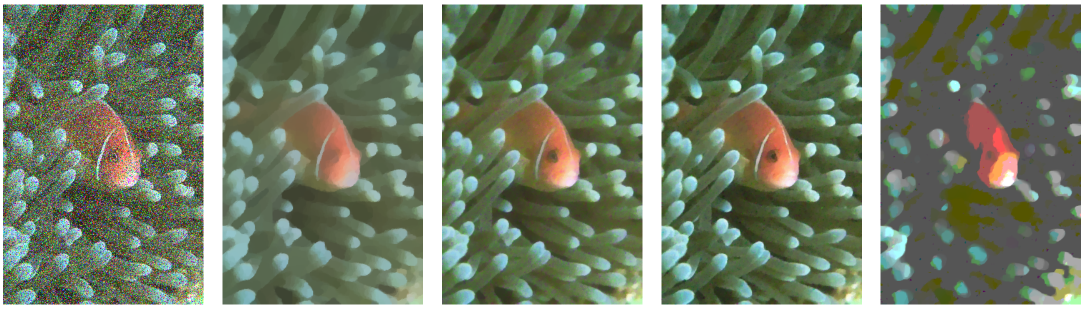
(b) Unlifted Problem, $E = 992.50$

(c) Ours, $|\mathcal{D}| = 1$, $|\mathcal{T}| = 4$, $E = 992.51$

(d) Ours, $|\mathcal{D}| = 6$ $|\mathcal{T}| = 2 \times 2 \times 2$ $E = 993.52$

(e) Baseline, $|\mathcal{D}| = 4 \times 4 \times 4$, $E = 2255.81$

Figure 5.6.: Convex ROF with vectorial TV. Direct optimization and proposed method yield the same result. In contrast to the baseline method [Lel+13] the proposed approach has no discretization artefacts and yields a lower energy. The regularization parameter is chosen as 0.3.



Noisy input

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

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

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

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

Figure 5.7.: ROF with a truncated quadratic dataterm ($\nu = 0.025$). Compared to the standard discretization [Lel+13] the proposed approach yields much better results, already with a very small number of 4 labels. The weight of the total variation was chosen 0.03.

mode and mean rounding procedure we found the plain mean of the discretized measure to produce better results on real data: More explicitly we recover a solution according to $x_u = \sum_{k=1}^{K} t_k(y_u)_{k,0}$ at each vertex $u \in \mathcal{V}$. In Table 5.1 we compare both, dual and nonconvex primal energies, for a larger hierarchy of dual subspaces.

5.6.3. Vectorial ROF Denoising

In the remaining experiments we will evaluate the lifted variational model (5.34) for the isotropic vector-valued setting using a piecewise linear approximation of the dual variable as described in Section 5.5.5. First we will experimentally validate, that our model is exact for convex dataterms: To this end we consider the Rudin-Osher-Fatemi [ROF92] (ROF) model with vectorial TV (5.32). In our model this corresponds to defining $f(x, u(x)) = \frac{1}{2}\|u(x) - I(x)\|^2$, where $I : \Omega \to \mathbb{R}^3$ is a color image. The energy of the solution of the unlifted problem is equal to the energy of the projected solution of our method for $n = 4$ up to machine precision, as can be seen in Figure 5.5 and Figure 5.6.

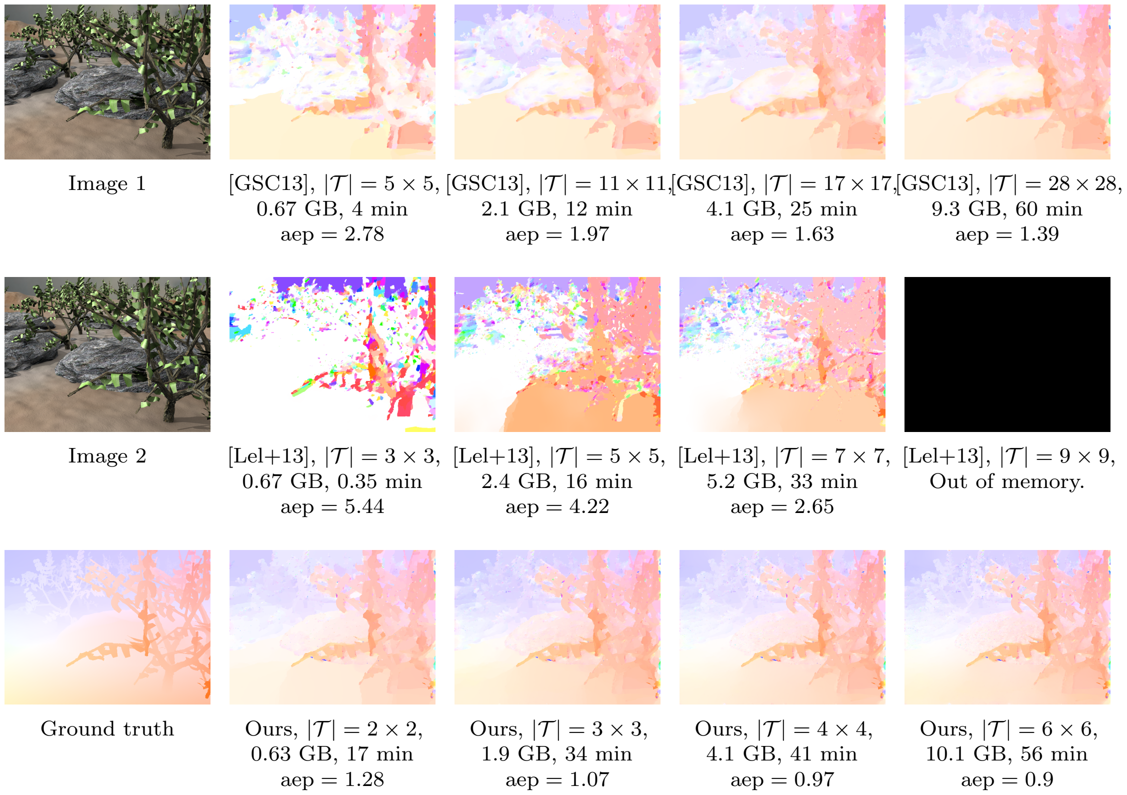


Image 1

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

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

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

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

Image 2

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

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

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

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

Ground truth

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

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

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

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

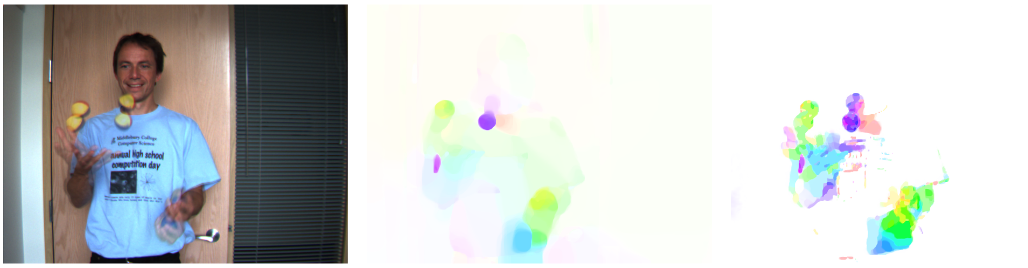
Figure 5.8.: We compute the optical flow using our method, the product space approach [GSC13] and the baseline method [Lel+13] for a varying amount of labels and compare the average endpoint error (aep). The product space method clearly outperforms the baseline, but our approach finds the overall best 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. The run times and memory requirements are to be taken as rough estimates, as they depend on the used stopping criteria and implementation.

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 Figure 5.5 and Figure 5.6, that the baseline method [Lel+13] has a strong label bias. In Figure 5.6 the input image is taken from the Berkeley segmentation database [Mar+01].

5.6.4. Robust color image denoising

In this experiment we consider a robust denoising approach for color images with a truncated quadratic dataterm

$$f(x, u(x)) = \min \left\{ \frac{1}{2} \|u(x) - I(x)\|^2, \nu \right\}, \tag{5.59}$$



(a) Image 1 and 2 (b) Proposed, $|\mathcal{T}| = 2 \times 2$ (c) Baseline, $|\mathcal{T}| = 7 \times 7$

Figure 5.9.: 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+13] in (c). Our method can correctly identify the large motion.

using the lifted variational model (5.34) for vectorial TV: To this end we degrade the input image with both, Gaussian and salt-and-pepper noise. The results are shown in Figure 5.7. It can be seen that increasing the number of samples $|\mathcal{T}|$ 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 larger the number of labels. The baseline method [Lel+13] again produces strong discretization artefacts even for a large number of labels $|\mathcal{T}| = 4 \times 4 \times 4 = 64$. The input image is taken from the Berkeley segmentation database [Mar+01].

5.6.5. Optical flow

In this experiment we compute the optical flow $u : \Omega \to \mathbb{R}^2$ between two input images I_1, I_2 using the lifted isotropic variational model (5.34). The label space $X = [-d, d]^2$ is chosen according to the estimated maximum displacement $d \in \mathbb{R}$ between the images. The dataterm is

$$f(x, u(x)) = \|I_2(x) - I_1(x + u(x))\|. \tag{5.60}$$

We introduce adaptive weights $r(x)$ for the regularizer which are based on the norm of the image gradient $\nabla I_1(x)$:

$$r(x) = s \cdot \begin{cases} 0.1 & \text{if } \|\nabla I_1(x)\| > 0.1, \\ 1 & \text{otherwise.} \end{cases} \tag{5.61}$$

In Figure 5.8 we compare the proposed method to the product space approach [GSC13] and the standard discretization [Lel+13]. For our method, we sample the label space $X = [-15, 15]^2$ on 150 × 150 points t_{j_k} 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 s differently for each method. For the proposed method, we set $s = 0.5$ and for the product space and baseline approach $s = 3$. We can see in Figure 5.8 that our discretization achieves a better average end-point error when compared to existing discretizations. In Figure 5.9 we compare our method on large displacement optical flow to the standard discretization [Lel+13]. The image pairs are taken from the Middlebury optical flow benchmark [Bak+11].

Chapter 6.

Conclusion

In this thesis we have considered lower envelopes for decoupling in additive composite optimization problems. The first approach is based on inf-projection: Here the idea is to relax one or more components in the additive composition by means of their Moreau envelopes. This is a common strategy in feasibility problems and allows one solve the problem in a distributed fashion. An important ingredient is the smoothness of the Moreau envelope and the continuity and single-valuedness of the associated proximal mapping, which holds under prox-regularity. This is for instance leveraged in the analysis of a stochastic inexact averaged proximal point method for federated learning. However, there exist simple even smooth functions that are not prox-regular. As a remedy we have replaced the Euclidean geometry in the proximal mapping and the definition of prox-regularity by an anisotropic or a Bregmanian geometry, both of which are induced by a Legendre function. There, the gradient of the Legendre function appears in the form of a nonlinear preconditioner. Single-valuedness and continuity of these non-Euclidean generalizations of the classical proximal mapping are studied under generalizations of prox-regularity.

As an alternative approach for decoupling in partially separable problems the Lagrangian relaxation paradigm is considered. However, since Lagrangian relaxations for nonconvex problems typically suffer from large duality gaps, reformulations over the space of measures are considered. This can be seen as a certain integer linear programming approach which is common for combinatorial problems such as the Sudoku puzzle. In our case, the focus is on TV-regularized variational problems and MAP-inference in a continuous MRF, where such linear programming formulations over the space of measures are natural. Adopting the framework for Lagrangian relaxations for these infinite programs we derive dual programs via subspace approximations. These are shown to be equivalent to a certain nonlinear lifting to moments in the primal. The approach is studied through a generalized conjugacy perspective which reveals interesting connections to the basic quadratic transform, which shows that dual discretizations, under a certain extremality condition, preserve the original nonconvex problem when one restricts the optimization variable to a certain nonlinear manifold of Diracs. We derive a cone programming formulation using tools from convex algebraic geometry and solve the problem on a GPU using a concretization of a first-order primal-dual algorithm. Experimentally, the approach is applied to stereo matching and optical flow estimation, showing merits over standard discretizations.

Bibliography

[ABS13] H. Attouch, J. Bolte, and B. Svaiter. "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods." In: *Mathematical Programming* 137.1-2 (2013), pp. 91–129. ISSN: 0025-5610. DOI: 10.1007/s10107-011-0484-9.

[AFP00] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems.* Oxford Science Publications. Clarendon Press, 2000. ISBN: 9780198502456. URL: https://books.google.de/books?id=7GUMIh6-5TYC.

[Att+10] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. "Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: An Approach Based on the Kurdyka-Łojasiewicz Inequality." In: *Mathematics of Operations Research* 35 (2010), pp. 438–457.

[Att77] H. Attouch. "Convergence de fonctions convexes, des sous-différentiels et semi-groupes associés." In: *Comptes Rendus de l'Académie des Sciences de Paris* 285 (1977), pp. 539–542.

[Att84] H. Attouch. *Variational Convergence for Functions and Operators.* Pitman Advanced Publishing Program, 1984.

[AZLL19] Z. Allen-Zhu, Y. Li, and Y. Liang. "Learning and generalization in overparameterized neural networks, going beyond two layers." In: *Advances in neural information processing systems.* 2019, pp. 6155–6166.

[Bač+10] M. Bačák, J. M. Borwein, A. Eberhard, and B. Mordukhovich. "Infimal convolutions and Lipschitzian properties of subdifferentials for prox-regular functions in Hilbert spaces." In: *Journal of Convex Analysis* 17 (2010), pp. 732–763.

[Bak+11] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. "A database and evaluation methodology for optical flow." In: *International journal of computer vision* 92.1 (2011), pp. 1–31.

[Bal77] E. J. Balder. "An extension of duality-stability relations to nonconvex optimization problems." In: *SIAM J. Control Optim.* 15.2 (1977), pp. 329–343.

[Ban+05] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. "Clustering with Bregman divergences." In: *Journal of machine learning research* 6.Oct (2005), pp. 1705–1749.

[Bau+08] H. H. Bauschke, R. Goebel, Y. Lucet, and X. Wang. "The proximal average: basic theory." In: *SIAM Journal on Optimization* 19.2 (2008), pp. 766–785.

[Bau+09] H. H. Bauschke, X. Wang, J. Ye, and X. Yuan. "Bregman distances and Chebyshev sets." In: *Journal of Approximation Theory* 159.1 (2009), pp. 3–25.

[Bau+19] H. H. Bauschke, J. Bolte, J. Chen, M. Teboulle, and X. Wang. "On Linear Convergence of Non-Euclidean Gradient Methods without Strong Convexity and Lipschitz Gradient Continuity." In: *Journal of Optimization Theory and Applications* 182.3 (2019), pp. 1068–1087. DOI: 10.1007/s10957-019-01516-9. URL: https://doi.org/10.1007/s10957-019-01516-9.

[Bau+21] H. Bauermeister, E. Laude, T. Möllenhoff, M. Moeller, and D. Cremers. "Lifting the Convex Conjugate in Lagrangian Relaxations: A Tractable Approach for Continuous Markov Random Fields." In: *arXiv preprint arXiv:2107.06028* (2021).

[BB97] H. H. Bauschke and J. M. Borwein. "Legendre functions and the method of random Bregman projections." eng. In: *Journal of Convex Analysis* 4.1 (1997), pp. 27–67. URL: http://eudml.org/doc/227096.

[BBC01] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. "Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces." In: *Communications in Contemporary Mathematics* 3.04 (2001), pp. 615–647.

[BBL99] H. H. Bauschke, J. M. Borwein, and W. Li. "Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization." In: *Mathematical Programming* 86.1 (1999), pp. 135–160.

[BBT17] H. H. Bauschke, J. Bolte, and M. Teboulle. "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications." In: *Mathematics of Operations Research* 42.2 (2017), pp. 330–348.

[BCN06] H. H. Bauschke, P. L. Combettes, and D. Noll. "Joint minimization with alternating Bregman proximity operators." In: *Pacific Journal of Optimization* 2.3 (2006), pp. 401–424.

[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.

[BDL18] H. H. Bauschke, M. Dao, and S. Lindstrom. "Regularizing with Bregman–Moreau Envelopes." In: *SIAM Journal on Optimization* 28.4 (2018), pp. 3208–3228.

[BEP87] B. Brown, D Elliott, and D. Paget. "Lipschitz constants for the Bernstein polynomials of a Lipschitz continuous function." In: *Journal of approximation theory* 49.2 (1987), pp. 196–199.

[BL00] H. H. Bauschke and A. S. Lewis. "Dykstra's algorithm with Bregman projections: A convergence proof." In: *Optimization* 48.4 (2000), pp. 409–427.

[BLT08] H. H. Bauschke, Y. Lucet, and M. Trienis. "How to transform one convex function continuously into another." In: *SIAM review* 50.1 (2008), pp. 115–132.

[BMR04] H. H. Bauschke, E. Matoušková, and S. Reich. "Projection and proximal point methods: convergence results and counterexamples." In: *Nonlinear Analysis: Theory, Methods & Applications* 56.5 (2004), pp. 715–738.

[BMW11] H. H. Bauschke, M. S. Macklem, and X. Wang. “Chebyshev Sets, Klee Sets, and Chebyshev Centers with Respect to Bregman Distances: Recent Results and Open Problems.” In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering.* New York, NY: Springer New York, 2011, pp. 1–21. ISBN: 978-1-4419-9569-8. DOI: `10.1007/978-1-4419-9569-8{\_}1`.

[Bol+18] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. “First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems.” In: *SIAM Journal on Optimization* 28.3 (2018), pp. 2131–2151.

[Bor+21] A. Borovykh, N. Kantas, P. Parpas, and G. Pavliotis. “On stochastic mirror descent with interacting particles: Convergence properties and variance reduction.” In: *Physica D: Nonlinear Phenomena* 418 (2021), p. 132844. ISSN: 0167-2789. DOI: `https://doi.org/10.1016/j.physd.2021.132844`. URL: `https://www.sciencedirect.com/science/article/pii/S0167278921000026`.

[Boy+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers.” In: *Foundations and Trends® in Machine Learning* 3.1 (2011), pp. 1–122.

[BPT12] G. Blekherman, P. A. Parrilo, and R. R. Thomas. *Semidefinite Optimization and Convex Algebraic Geometry.* Philadelphia, PA: Society for Industrial and Applied Mathematics, 2012.

[Bre67] L. M. Bregman. “The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.” In: *USSR Comput. Math. Math. Phys.* 7 (1967), pp. 200–217.

[BST14] J. Bolte, S. Sabach, and M. Teboulle. “Proximal alternating linearized minimization for nonconvex and nonsmooth problems.” In: *Mathematical Programming* 146.1-2 (2014), pp. 459–494.

[Car11] C. Carathéodory. “Über den Variabilitätsbereich der Fourier’schen Konstanten von positiven harmonischen Funktionen.” In: *Rendiconti Del Circolo Matematico di Palermo (1884-1940)* 32.1 (1911), pp. 193–217.

[Car98] N. L. Carothers. *A short course on approximation theory.* 1998.

[CCP12] A. Chambolle, D. Cremers, and T. Pock. “A convex approach to minimal partitions.” In: *SIAM Journal on Imaging Sciences* 5.4 (2012), pp. 1113–1158.

[CD08] P. Clément and W. Desch. “Wasserstein metric and subordination.” In: *Studia Mathematica* 1.189 (2008), pp. 35–52.

[CJT17] A. Cabot, A. Jourani, and L. Thibault. “Envelopes for sets and functions: regularization and generalized conjugacy.” In: *Mathematika* 63.2 (2017), pp. 383–432.

[CKS12] Y. Y. Chen, C. Kan, and W. Song. “The Moreau envelope function and proximal mapping with respect to the Bregman distance in Banach spaces.” In: *Vietnam Journal of Mathematics* 40.2&3 (2012), pp. 181–199.

[CP11] A. Chambolle and T. Pock. “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging.” In: *J. Math. Imaging Vis.* 40 (1 2011), pp. 120–145.

[CR13] P. L. Combettes and N. N. Reyes. “Moreau’s decomposition in Banach spaces.” In: *Mathematical Programming* 139.1-2 (2013), pp. 103–114.

[CWP20] J. Chen, X. Wang, and C. Planiden. “A proximal average for prox-bounded functions.” In: *SIAM Journal on Optimization* 30.2 (2020), pp. 1366–1390.

[DBLJ14] A. Defazio, F. Bach, and S. Lacoste-Julien. “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives.” In: *Advances in neural information processing systems.* 2014, pp. 1646–1654.

[DDC14] A. Defazio, J. Domke, and Caetano. “Finito: A faster, permutable incremental gradient method for big data problems.” In: *Proceedings of the 31st International Conference on Machine Learning.* Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research. Bejing, China: PMLR, 2014, pp. 1125–1133. URL: `http://proceedings.mlr.press/v32/defazio14.html`.

[DP19] D. Drusvyatskiy and C. Paquette. “Efficiency of minimizing compositions of convex functions and smooth maps.” In: *Mathematical Programming* 178.1-2 (2019), pp. 503–558.

[DR09] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings: A View From Variational Analysis.* New York: Springer, 2009.

[Du+18] S. S. Du, X. Zhai, B. Poczos, and A. Singh. “Gradient descent provably optimizes over-parameterized neural networks.” In: *arXiv preprint arXiv:1810.02054* (2018).

[Eck93] J. Eckstein. “Nonlinear Proximal Point Algorithms Using Bregman Functions, with Applications to Convex Programming.” In: *Mathematics of Operations Research* 18.1 (1993), pp. 202–226.

[FA14] A. Fix and S. Agarwal. “Duality and the Continuous Graphical Model.” In: *European Conference on Computer Vision (ECCV).* 2014.

[Fed59] H. Federer. “Curvature measures.” In: *Transactions of the American Mathematical Society* 93.3 (1959), pp. 418–491.

[Gan+10] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar. “Posterior regularization for structured latent variable models.” In: *The Journal of Machine Learning Research* 11 (2010), pp. 2001–2049.

[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.

[GVV98] A. Gammerman, V. Vapnik, and V. Vowk. “Learning by transduction.” In: *UAI.* 1998, pp. 148–156.

[Haj+16] D. Hajinezhad, M. Hong, T. Zhao, and Z. Wang. “NESTT: A nonconvex primal-dual splitting method for distributed and stochastic optimization.” In: *Advances in Neural Information Processing Systems.* 2016, pp. 3215–3223.

[Har09] W. L. Hare. “A proximal average for nonconvex functions: A proximal stability perspective.” In: *SIAM Journal on Optimization* 20.2 (2009), pp. 650–666.

[HP14] W. L. Hare and C Planiden. “The NC-proximal average for multiple functions.” In: *Optimization Letters* 8.3 (2014), pp. 849–860.

[JTZ14] A. Jourani, L. Thibault, and D. Zagrodny. "Differential properties of the Moreau envelope." In: *Journal of Functional Analysis* 266.3 (2014), pp. 1185–1237.

[Kai+19] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. *Advances and Open Problems in Federated Learning.* 2019. arXiv: `1912.04977 [cs.LG]`.

[Kan60] L. V. Kantorovich. "Mathematical methods of organizing and planning production." In: *Management Science* 6.4 (1960), pp. 366–422.

[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.

[KKS21] F. Kunstner, R. Kumar, and M. Schmidt. "Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent." In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics.* Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3295–3303. URL: `http://proceedings.mlr.press/v130/kunstner21a.html`.

[KMR15] J. Konečnỳ, B. McMahan, and D. Ramage. "Federated optimization: Distributed optimization beyond the datacenter." In: *arXiv preprint arXiv:1511.03575* (2015).

[Kri64] J.-L. Krivine. "Anneaux préordonnés." In: *Journal d'analyse mathématique* 12.1 (1964), pp. 307–326.

[KS12] C. Kan and W. Song. "The Moreau envelope function and proximal mapping in the sense of the Bregman distance." In: *Nonlinear Analysis: Theory, Methods & Applications* 75.3 (2012), pp. 1385–1399.

[KZ06] A. J. Kurdila and M. Zabarankin. *Convex functional analysis.* Springer Science & Business Media, 2006.

[Las01] J. B. Lasserre. "Global optimization with polynomials and the problem of moments." In: *SIAM Journal on Optimization* 11.3 (2001), pp. 796–817.

[Las02] J. B. Lasserre. "Semidefinite programming vs. LP relaxations for polynomial programming." In: *Mathematics of Operations Research* 27.2 (2002), pp. 347–360.

[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.* Springer. 2016, pp. 614–627.

[Lau+18] E. Laude, J.-H. Lange, J. Schüpfer, C. Domokos, L. Leal-Taixé, F. R. Schmidt, B. Andres, and D. Cremers. “Discrete-Continuous ADMM for Transductive Inference in Higher-Order MRFs.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[LeC+98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition.” In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[Lel+13] 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.

[Les67] C. Lescarret. “Applications “prox” dans un espace de Banach.” In: *Comptes Rendus de l’Académie des Sciences de Paris Série A* 265 (1967), pp. 676–678.

[LFN18] H. Lu, R. Freund, and Y. Nesterov. “Relatively Smooth Convex Optimization by First-Order Methods and Applications.” In: *SIAM Journal on Optimization* 28.1 (2018), pp. 333–354.

[Li+20] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. “Federated Optimization in Heterogeneous Networks.” In: *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. Ed. by I. S. Dhillon, D. S. Papailiopoulos, and V. Sze. mlsys.org, 2020. URL: `https://proceedings.mlsys.org/book/316.pdf`.

[LOC20] E. Laude, P. Ochs, and D. Cremers. “Bregman proximal mappings and Bregman–Moreau envelopes under relative prox-regularity.” In: *Journal of Optimization Theory and Applications* 184.3 (2020), pp. 724–761.

[LOC21] E. Laude, P. Ochs, and D. Cremers. “On the duality between averaged Proximal Point and gradient descent in finite sum minimization.” In: *Unpublished manuscript* (2021).

[LP16] G. Li and T. K. Pong. “Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems.” In: *Mathematical programming* 159.1-2 (2016), pp. 371–401.

[LTP21] P. Latafat, A. Themelis, and P. Patrinos. “Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems.” In: *Mathematical Programming* (2021), pp. 1–30.

[LWC18] E. Laude, T. Wu, and D. Cremers. “A Nonconvex Proximal Splitting Algorithm under Moreau-Yosida Regularization.” In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, 2018, pp. 491–499. URL: `http://proceedings.mlr.press/v84/laude18a.html`.

[LWC19] E. Laude, T. Wu, and D. Cremers. “Optimization of Inf-Convolution Regularized Nonconvex Composite Problems.” In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 547–556. URL: `http://proceedings.mlr.press/v89/laude19a.html`.

[Mai15] J. Mairal. “Incremental majorization-minimization optimization with application to large-scale machine learning.” In: *SIAM Journal on Optimization* 25.2 (2015), pp. 829–855.

[Mar+01] D. Martin, C. Fowlkes, D. Tal, and J. Malik. “A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics.” In: *Proc. 8th Int’l Conf. Computer Vision*. Vol. 2. 2001, pp. 416–423.

[McM+17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Singh and J. Zhu. Vol. 54. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, 2017, pp. 1273–1282. URL: `http://proceedings.mlr.press/v54/mcmahan17a.html`.

[Mor18] B. S. Mordukhovich. *Variational Analysis and Applications*. Springer, 2018.

[Mor62] J.-J. Moreau. “Fonctions convexes duales et points proximaux dans un espace Hilbertien.” In: *Comptes Rendus de l’Académie des Sciences de Paris Série A* 255 (1962), pp. 2897–2899.

[Mor65] J.-J. Moreau. “Proximité et dualité dans un espace hilbertien.” In: *Bulletin de la Société Mathématique de France* 93.2 (1965), pp. 273–299.

[Mor66] J.-J. Moreau. “Fonctionnelles convexes.” In: *Séminaire Jean Leray* (1966), pp. 1–108.

[MTZ60] C. E. Miller, A. W. Tucker, and R. A. Zemlin. “Integer programming formulation of traveling salesman problems.” In: *Journal of the ACM (JACM)* 7.4 (1960), pp. 326–329.

[NRP19] I. Necoara, P. Richtárik, and A. Patrascu. “Randomized projection methods for convex feasibility: Conditioning and convergence rates.” In: *SIAM Journal on Optimization* 29.4 (2019), pp. 2814–2852.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. 2nd. New York: Springer, 2006.

[Och18] P. Ochs. “Local convergence of the heavy-ball method and iPiano for nonconvex optimization.” In: *Journal of Optimization Theory and Applications* 177.1 (2018), pp. 153–180.

[PB13] P. Patrinos and A. Bemporad. “Proximal Newton methods for convex composite optimization.” In: *52nd IEEE Conference on Decision and Control*. IEEE. 2013, pp. 2358–2363.

[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.

[PKD15] D. Pathak, P. Krahenbuhl, and T. Darrell. “Constrained convolutional neural networks for weakly supervised segmentation.” In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1796–1804.

[Poc+09] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. “An algorithm for minimizing the Mumford-Shah functional.” In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, pp. 1133–1140.

[Pol90] R. A. Poliquin. “Subgradient monotonicity and convex functions.” In: *Nonlinear Analysis: Theory, Methods & Applications* 14.4 (1990), pp. 305–317.

[Pol91] R. A. Poliquin. “Integration of subdifferentials of nonconvex functions.” In: *Nonlinear Analysis: Theory, Methods & Applications* 17.4 (1991), pp. 385–398.

[PR00] V. Powers and B. Reznick. “Polynomials that are positive on an interval.” In: *Transactions of the American Mathematical Society* 352.10 (2000), pp. 4677–4692.

[PR10] R. A. Poliquin and R. T. Rockafellar. “A calculus of prox-regularity.” In: *J. Convex Anal* 17.1 (2010), pp. 203–210.

[PR96] R. A. Poliquin and R. T. Rockafellar. “Prox-regular functions in variational analysis.” In: *Transactions of the American Mathematical Society* 348 (1996), pp. 1805–1838.

[PRT00] R. A. Poliquin, R. T. Rockafellar, and L. Thibault. “Local differentiability of distance functions.” In: *Transactions of the American Mathematical Society* 352.11 (2000), pp. 5231–5249.

[Put93] M. Putinar. “Positive polynomials on compact semi-algebraic sets.” In: *Indiana University Mathematics Journal* 42.3 (1993), pp. 969–984.

[RB12] M. Raginsky and J. Bouvrie. “Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence.” In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE. 2012, pp. 6793–6800.

[Rob80] S. M. Robinson. “Strongly regular generalized equations.” In: *Mathematics of Operations Research* 5 (1980), pp. 43–62.

[Roc67] R. Rockafellar. “Duality and stability in extremum problems involving convex functions.” In: *Pacific Journal of Mathematics* 21.1 (1967), pp. 167–187.

[Roc70] R. T. Rockafellar. *Convex Analysis*. New Jersey: Princeton University Press, 1970.

[ROF92] L. I. Rudin, S. Osher, and E. Fatemi. “Nonlinear total variation based noise removal algorithms.” In: *Physica D: Nonlinear Phenomena* 60.1 (1992), pp. 259–268.

[RS71] H. Robbins and D. Siegmund. “A convergence theorem for non negative almost supermartingales and some applications.” In: *Optimizing methods in statistics*. Elsevier, 1971, pp. 233–257.

[Ruo15] N. Ruozzi. “Exactness of approximate MAP inference in continuous MRFs.” In: *Advances in Neural Information Processing Systems*. 2015.

[Rus12] B. Russell. *The Problems of Philosophy*. Oxford University Press, 1912.

[RW98] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. New York: Springer, 1998.

[San15] F. Santambrogio. *Optimal Transport for Applied Mathematicians*. New York: Birkhäuser, 2015.

[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.

[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.* Springer. 2014, pp. 31–42.

[Sch91] K. Schmüdgen. "The k-moment problem for compact semi-algebraic sets." In: *Mathematische Annalen* 289.1 (1991), pp. 203–206.

[SM00] J. Shi and J. Malik. "Normalized cuts and image segmentation." In: *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000), pp. 888–905.

[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.

[Ste74] G. Stengle. "A Nullstellensatz and a Positivstellensatz in semialgebraic geometry." In: *Mathematische Annalen* 207.2 (1974), pp. 87–97.

[Tan+18] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov. "On regularized losses for weakly-supervised cnn segmentation." In: *Proceedings of the European Conference on Computer Vision (ECCV).* 2018, pp. 507–522.

[Teb92] M. Teboulle. "Entropic Proximal Mappings with Applications to Nonlinear Programming." In: *Mathematics of Operations Research* 17.3 (1992), pp. 670–690.

[TSP18] A. Themelis, L. Stella, and P. Patrinos. "Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms." In: *SIAM Journal on Optimization* 28.3 (2018), pp. 2274–2303.

[Vap06] V. Vapnik. *Estimation of dependences based on empirical data.* Springer Science & Business Media, 2006.

[Vil08] C. Villani. *Optimal Transport: Old and New.* Springer, 2008.

[VL17] T. Vogt and J. Lellmann. "An optimal transport-based restoration method for Q-ball imaging." In: *International Conference on Scale Space and Variational Methods in Computer Vision.* Springer. 2017, pp. 271–282.

[VL18] T. Vogt and J. Lellmann. "Measure-valued variational models with applications to diffusion-weighted imaging." In: *Journal of Mathematical Imaging and Vision* 60.9 (2018), pp. 1482–1502.

[VN50] J. Von Neumann. *Functional operators.* Princeton University Press, 1950.

[Wak+06] H. Waki, S. Kim, M. Kojima, and M. Muramatsu. "Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity." In: *SIAM Journal on Optimization* 17.1 (2006), pp. 218–242.

[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.

[Wex73] D. Wexler. "Prox-mappings associated with a pair of Legendre conjugate functions." In: *Revue française d'automatique informatique recherche opérationnelle* 7.R2 (1973), pp. 39–65.

[WG14] Y. Wald and A. Globerson. “Tightness Results for Local Consistency Relaxations in Continuous MRFs.” In: *UAI*. 2014.

[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.

[WLT18] T. Weisser, J. B. Lasserre, and K.-C. Toh. “Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity.” In: *Mathematical Programming Computation* 10.1 (2018), pp. 1–32.

[Yu+15] Y. Yu, X. Zheng, M. Marchetti-Bowick, and E. Xing. “Minimizing nonconvex non-separable functions.” In: *Artificial Intelligence and Statistics*. 2015, pp. 1107–1115.

[ZCL15] S. Zhang, A. E. Choromanska, and Y. LeCun. “Deep learning with elastic averaging SGD.” In: *Advances in neural information processing systems*. 2015, pp. 685–693.