ℓ_1 Regularized Gradient Temporal-Difference Learning

Dominik Meyer DOMINIK.MEYER@TUM.DE
Hao Shen HAO.SHEN@TUM.DE
Klaus Diepold KLDI@TUM.DE
Institute for Data Processing, Technische Universität München, Germany

Abstract

The family of Gradient Temporal-Difference (GTD) learning algorithms shares a promising property of being stable with both linear function approximation and off-policy training. The success of the GTD family requires a suitable set of features, which are unfortunately not always available in reality. To overcome this difficulty, regularization is often employed as an effective method for feature selection in reinforcement learning. In the present work, we propose and investigate a family of ℓ_1 regularized GTD learning algorithms.

Keywords: Reinforcement Learning (RL), Gradient Temporal-Difference (GTD) learning, linear function approximation, Iterative Soft Thresholding (IST).

1. The Proposed Algorithms

The recently developed family of Gradient Temporal-Difference (GTD) learning algorithms, cf. Sutton et al. (2008, 2009), have demonstrated their promising stability with both linear function approximation and off-policy training. However, the original development does not consider the linear TD learning with regularization. More recently, the work in Painter-Wakefield and Parr (2012) proposes an efficient ℓ_1 regularized TD learning algorithm, namely, the *L1TD* algorithm. The algorithm applies iteratively the soft thresholding operator, which is well known to sparse representation and compressed sensing, cf. Zibulevsky and Elad (2010). Unfortunately, the *L1TD* algorithm only considers the *mean squared Bellman error* (MSBE), which is known to be less robust than the *projected mean squared Bellman error* (MSPBE). In the present work, we study the GTD family with ℓ_1 regularization, and propose a family of Iterative Soft Thresholding (IST) based GTD learning algorithms.

Let us denote by r the reward at the state s, γ a discount factor, θ a parameter vector for linear function approximation, ϕ the feature vector corresponding to the current state s, and ϕ' the feature vector of the state s' transited from s. The GTD learning algorithms minimize either the *norm of the expected TD update* (NEU), cf. Sutton et al. (2008),

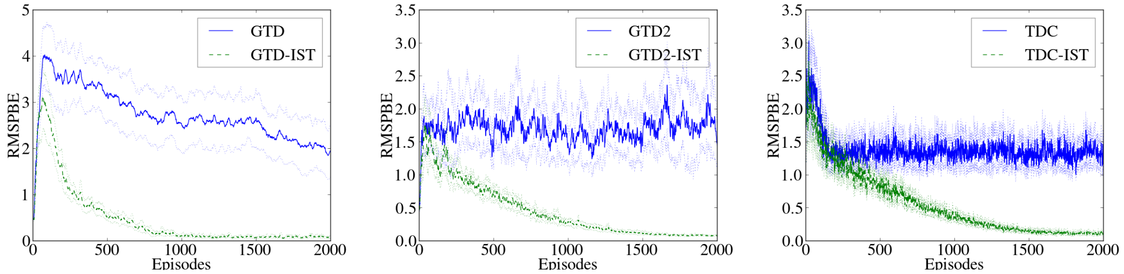
$$J_1(\theta) := \mathrm{NEU}(\theta) = \mathbb{E}[\delta_\theta \phi]^\top \mathbb{E}[\delta_\theta \phi], \tag{1}$$

or the MSPBE, cf. Sutton et al. (2009),

$$J_2(\theta) := \mathrm{MSPBE}(\theta) = \mathbb{E}[\delta_\theta \phi]^\top \mathbb{E}[\phi\phi^\top]^{-1} \mathbb{E}[\delta_\theta \phi], \tag{2}$$

where $\delta_\theta = r + \gamma\theta^\top\phi' - \theta^\top\phi$ is the TD error. Applying an ℓ_1 regularizer to the parameter θ leads to the following objective function that is studied in the present work, i.e.

$$F_i(\theta) := J_i(\theta) + \lambda\|\theta\|_1, \tag{3}$$



(a) GTD $(\alpha = 0.1, \beta = 0.01)$ (b) GTD2 $(\alpha = 0.1, \beta = 0.1)$ (c) TDC $(\alpha = 0.1, \beta = 0.05)$

Figure 1: A comparison of IST based GTD learning family ($\lambda = 0.001$).

where $\lambda > 0$ is a weighting factor and $i \in \{1, 2\}$. Given $x \in \mathbb{R}^m$ and $\nu > 0$, the *soft thresholding operator* applied to x is defined as

$$\Psi_\nu(x) := \operatorname{sgn}(x) \odot \max\{|x| - \nu, 0\}, \tag{4}$$

where $\operatorname{sgn}(\cdot)$ and $\max(\cdot)$ are entry-wise, and $\odot$ is the entry-wise multiplication. Then, minimization of the objective function (3) can be achieved via applying the *soft thresholding operator* iteratively. Straightforwardly, in the form of stochastic gradient descent, we propose a family of IST based GTD learning algorithms as

$$\theta_{t+1} = \Psi_{\alpha_t \lambda} \left(\theta_t + \alpha_t \nabla J_i(\theta_t) \right), \tag{5}$$

where $\alpha_t > 0$, and $\nabla J_i(\theta_t)$ denotes the stochastic gradient update of $J_i(\theta_t)$, or their suitable approximations, cf. Sutton et al. (2008, 2009).

2. Preliminary Results and Outlook

In our experiment, we apply the proposed algorithms to a random walk problem with seven states, where only one action exists and the transition probability of going right or left is equal. A reward of one is only assigned in the rightmost state, which is the terminal state, whereas the rewards are zero everywhere else. The features consist of a binary encoding of the states and ten additional "noisy" features, which are simply Gaussian noise. Figure 1 depicts the learning curves of three GTD learning algorithms, namely, GTD, GTD2, and TDC. It is evident that IST based GTD learning algorithms outperform all their original counterparts. The experimental results demonstrate the effectiveness of IST based GTD learning algorithms. Being aware of advanced developments in the community of sparse representation, we project to employ further state-of-the-art algorithms of sparse representation to reinforcement learning.

3. Acknowledgements

This work has been partially supported by the International Graduate School of Science and Engineering (IGSSE), Technische Universität München, Germany.

References

C. Painter-Wakefield and R. Parr. L_1 regularized linear temporal difference learning. Technical report, Department of Computer Science, Duke University, 2012.

R. S. Sutton, Csaba Szepesvári, and H. R. Maei. A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximations. In *NIPS 21*, pages 1609–1616. The MIT Press, 2008.

R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of ICML 2009*, pages 993–1000, 2009.

M. Zibulevsky and M. Elad. $L1$-$L2$ optimization in signal and image processing. *IEEE Signal Processing Magazine*, 27(3):76–88, 2010.