535520: Optimization Algorithms

Lecture 10 — Mirror Descent

Ping-Chun Hsieh (謝秉均)

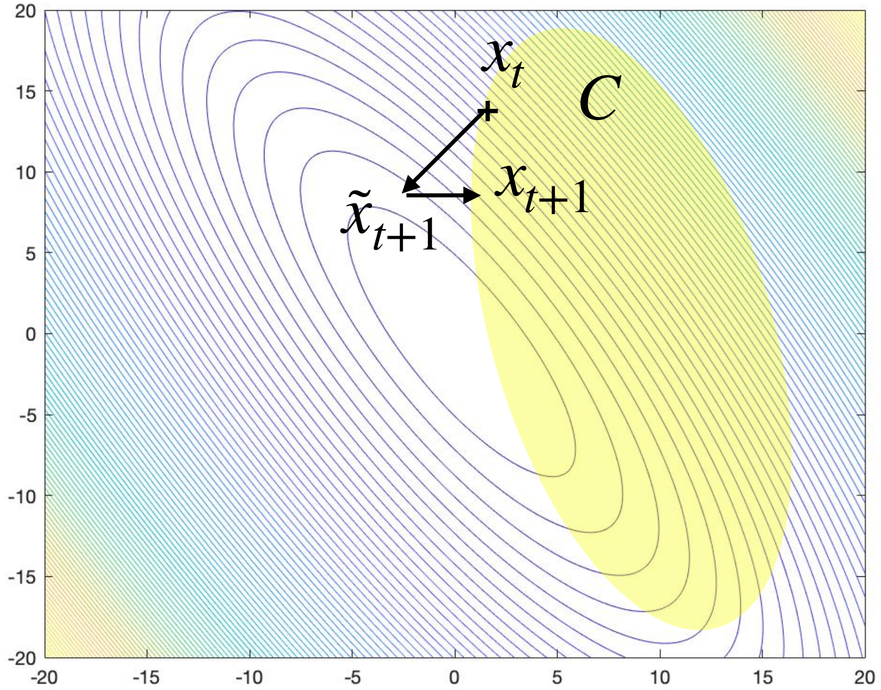
November 18, 2024

This Lecture

1. Mirror Descent

- Reading Material:
  - Amir Beck and Marc Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” Operations Research Letters, 2003.
  - Lecture notes of Prof. Anupam Gupta at CMU (http://www.cs.cmu.edu/~15850/notes/lec19.pdf)
  - Chapter 9 of Amir Beck’s textbook “First-Order Methods in Optimization”
  - Part of the material is adapted from Prof. Yuxin Chen’s lecture notes

Rethinking Projected GD



- Under PGD, the iterates are updated as

$$x_{t+1} = \Pi_C(\underbrace{x_t - \eta_t \nabla f(x_t)}_{=:\tilde{x}_{t+1}})$$

- The PGD update can be rewritten as:

$$x_{t+1} = \arg\min_{x \in C} \|x - (\underbrace{x_t - \eta_t \nabla f(x_t)}_{\tilde{x}_{t+1}})\|^2$$

$$= \arg\min_{x \in C} \left\{ \|x - x_t\|^2 + 2\eta_t (x - x_t)^\top \nabla f(x_t) + \eta_t^2 \|\nabla f(x_t)\|^2 \right\}$$

$$= \arg\min_{x \in C} \left\{ \frac{1}{2\eta_t} \|x - x_t\|^2 + (x - x_t)^\top \nabla f(x_t) \right\}$$

$$= \arg\min_{x \in C} \left\{ \frac{1}{2\eta_t} \|x - x_t\|^2 + (x - x_t)^\top \nabla f(x_t) + f(x_t) \right\}$$

Rethinking Projected GD

- Original viewpoint: $x_{t+1} = \Pi_C(\underbrace{x_t - \eta_t \nabla f(x_t)}_{=:\tilde{x}_{t+1}})$

- "**Proximal**" viewpoint: $x_{t+1} = \arg\min_{x \in C} \{\underbrace{f(x_t) + \nabla f(x_t)^\top (x - x_t)}_{\text{first-order approximation}} + \underbrace{\frac{1}{2\eta_t}\|x - x_t\|^2}_{\text{proximal term}}\}$

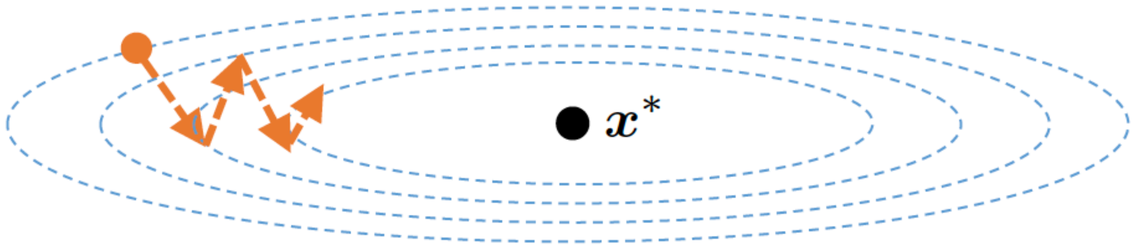
Interpretation: This L_2 proximal term is meant to capture the discrepancy between $f(x)$ and the first-order approximation

Issue: This choice of L_2 proximal term presumes that the local geometry is homogeneous / Euclidean

Question: Can we extend this idea to non-Euclidean geometry?

An Example of Non-Homogeneity

$$\text{minimize}_{x\in\mathbb{R}^2} \quad f(x) := \frac{1}{2}(x-x^*)^\top Q(x-x^*)$$

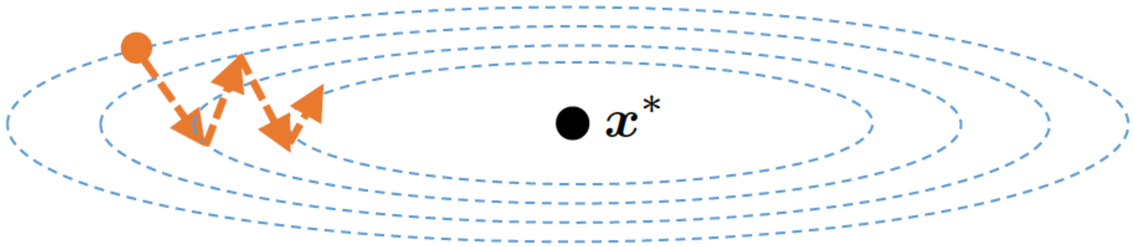


- Suppose $Q = [Q_{11}, 0; 0, Q_{22}]$ is a diagonal matrix with $Q_{11} \gg Q_{22}$

- In this example, the condition number $\kappa \gg 1$ (why?)
- Recall from Lecture 4:
  - Under GD, the convergence rate is $\|x_t - x^*\| \le \left(\frac{\kappa-1}{\kappa+1}\right)^t \|x_0 - x^*\|$
  - Equivalently, the sample complexity is $O(\kappa \log(\frac{1}{\epsilon}))$
- GD converges slowly due to large κ (as GD update does not capture the curvature well)
- Question: How to improve the convergence rate under a large κ?

An Example of Non-Homogeneity: "Scaled" Gradients

$$\text{minimize}_{x \in C} \quad f(x) := \frac{1}{2}(x - x^*)^\top Q (x - x^*)$$



- Suppose $Q = [Q_{11}, 0; 0, Q_{22}]$ is a diagonal matrix with $Q_{11} \gg Q_{22}$

- Idea: Accelerate PGD by *scaling the gradient*

$$x_{t+1} = \Pi_C\left(x_t - \eta_t Q^{-1} \nabla f(x_t)\right) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

- Insight: The above "scaled GD update" is equivalent to

$$x_{t+1} = \arg\min_{x \in C} \left\{ f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2\eta_t} \underline{(x - x_t)^\top Q (x - x_t)} \right\}$$

a scaled L_2 proximal term

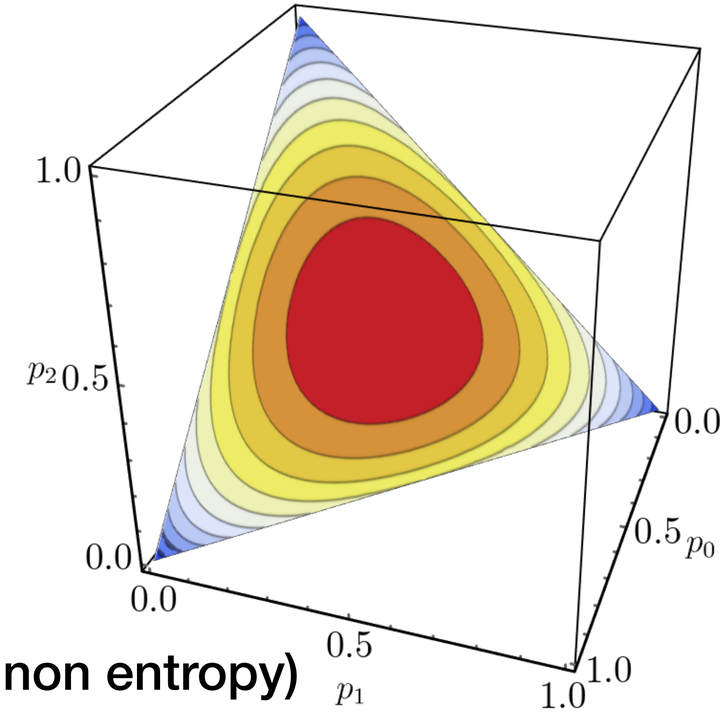
(that better captures the local geometry)

An Example of Non-Euclidean Geometry

$$\text{minimize}_{x \in C} \quad f(x)$$

$$C = \{x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, x_j \geq 0, \forall j\}$$

(probability simplex)



(Contour of Shannon entropy)

- Question: Is it good to measure "the distance of two probability vectors" by L_2 norm?
- Some popular distance measures of probability vectors:
  - Kullback-Leibler (KL) divergence
  - Total variation (TV) divergence
  - Jensen-Shannon (JS) divergence
  - … and more

Mirror Descent (MD)

Key Idea: Generalize the L_2 proximal term to other distance measures!

$$x_{t+1} = \arg\min_{x \in C} \left\{ \underbrace{f(x_t) + \nabla f(x_t)^\top (x - x_t)}_{\text{first-order approximation}} + \frac{1}{\eta_t} \underbrace{D_\phi(x \| x_t)}_{\text{Bregman divergence}} \right\}$$

where $D_\phi(y \| x) := \phi(y) - \phi(x) - \nabla \phi(x)^\top (y - x)$

(with respect to $\phi(\cdot)$ strictly convex and differentiable)

- Remark: Bregman divergence is meant to capture "local geometry" of objective function
- Remark: Bregman divergence is NOT symmetric and hence not a metric
- How about MD in the unconstrained cases?

Bregman Divergence (Formally)

Definition: Given a strictly convex and differentiable function $\phi : X \to \mathbb{R}$, the Bregman divergence $D_\phi(\cdot \| \cdot)$ w.r.t. to ϕ is defined as

$$D_\phi(y\|x) := \phi(y) - \left(\phi(x) + \nabla\phi(x)^\top(y-x)\right)$$

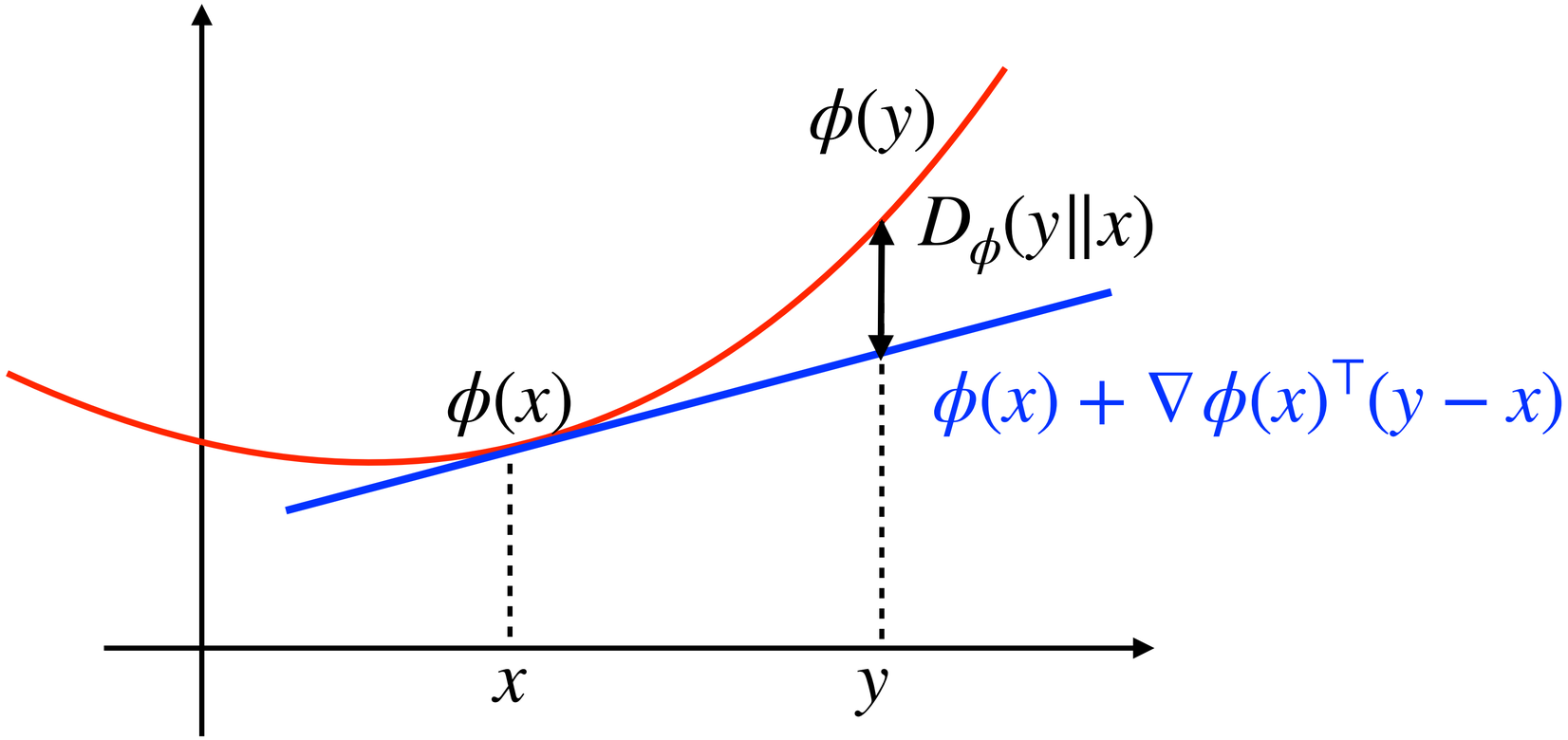
- Intuition: Connection between L_2 norm and Bregman divergence?

$$D_\phi(y\|x) = \phi(y) - \left(\phi(x) + \nabla\phi(x)^\top(y-x)\right) = (y-x)^\top \_\_\_\_\_ (y-x)$$

Bregman divergence is locally a scaled L_2 proximal term

Visualization of Bregman Divergence

$D_\phi(y\|x) = \phi(y) - \phi(x) - \nabla\phi(x)^\top(y-x)$, where $\phi(\cdot)$ is strictly convex



Example #1: Negative Entropy and KL Divergence

Suppose $\phi(x) = \sum_{i=1}^{d} (x_i \log x_i - x_i)$ (unnormalized negative entropy)

$$x \equiv (x_1, \cdots, x_d)$$
$$y \equiv (y_1, \cdots, y_d)$$

- The resulting Bregman divergence is $D_\phi(y\|x) = \sum_{i=1}^{d} y_i \log \frac{y_i}{x_i}$ (KL divergence)

$$D_\phi(y\|x) := \phi(y) - \left(\phi(x) + \nabla\phi(x)^\top (y - x)\right)$$

Example #2: Quadratic Function and Mahalanobis Distance

Suppose $\phi(x) = x^\top A x$ (where A is a positive definite $d \times d$ matrix)

- The resulting Bregman divergence is $D_\phi(y\|x) = (x-y)^\top A(x-y)$

(aka “Mahalanobis distance”)

$$D_\phi(y\|x) := \phi(y) - \Big(\phi(x) + \nabla\phi(x)^\top(y-x)\Big)$$

Example #3: Negative Logarithm and Itakura-Saito Distance

Suppose $\phi(x) = -\log x$

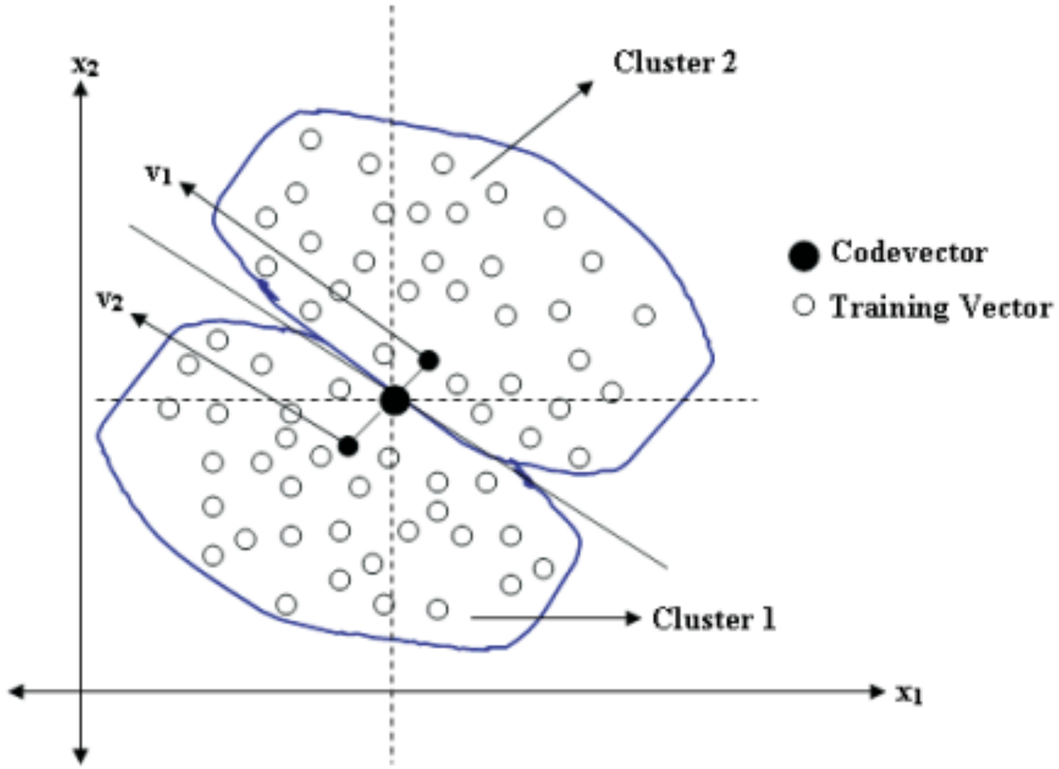
- The resulting Bregman divergence is $D_\phi(y\|x) = \frac{y}{x} - \log\frac{y}{x} - 1$

(aka “Itakura-Saito distance”)

$$D_\phi(y\|x) := \phi(y) - \left(\phi(x) + \nabla\phi(x)^\top(y - x)\right)$$

Itakura-Saito Distance

- Itakura-Saito distance is widely used in vector quantization (Linde, Buzo, Gray, 1980)
  - K-means-like clustering in the frequency domain
  - Especially for speech signal coding



Y. Linde, A. Buzo, and R. Gray, “An algorithm for vector quantizer design,” IEEE Transactions on Communications, 1980

H. B. Kekrea, Prachi Natub, and Tanuja Sarodec, “Color Image Compression Using Vector Quantization and Hybrid Wavelet Transform,” Procedia Computer Science, 2016

Popular Choices of Bregman Divergence

Domain	$\phi(\mathbf{x})$	$d_\phi(\mathbf{x},\mathbf{y})$	**Divergence**
$\mathbb{R}$	x^2	$(x-y)^2$	Squared loss
$\mathbb{R}_+$	$x\log x$	$x\log(\frac{x}{y})-(x-y)$	
$[0,1]$	$x\log x+(1-x)\log(1-x)$	$x\log(\frac{x}{y})+(1-x)\log(\frac{1-x}{1-y})$	Logistic loss [3]
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y}-\log(\frac{x}{y})-1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x-e^y-(x-y)e^y$	
$\mathbb{R}^d$	$\|\mathbf{x}\|^2$	$\|\mathbf{x}-\mathbf{y}\|^2$	Squared Euclidean distance
$\mathbb{R}^d$	$\mathbf{x}^T A\mathbf{x}$	$(\mathbf{x}-\mathbf{y})^T A(\mathbf{x}-\mathbf{y})$	Mahalanobis distance [4]
d-Simplex	$\sum_{j=1}^d x_j\log_2 x_j$	$\sum_{j=1}^d x_j\log_2(\frac{x_j}{y_j})$	KL-divergence
$\mathbb{R}_+^d$	$\sum_{j=1}^d x_j\log x_j$	$\sum_{j=1}^d x_j\log(\frac{x_j}{y_j})-\sum_{j=1}^d(x_j-y_j)$	Generalized I-divergence

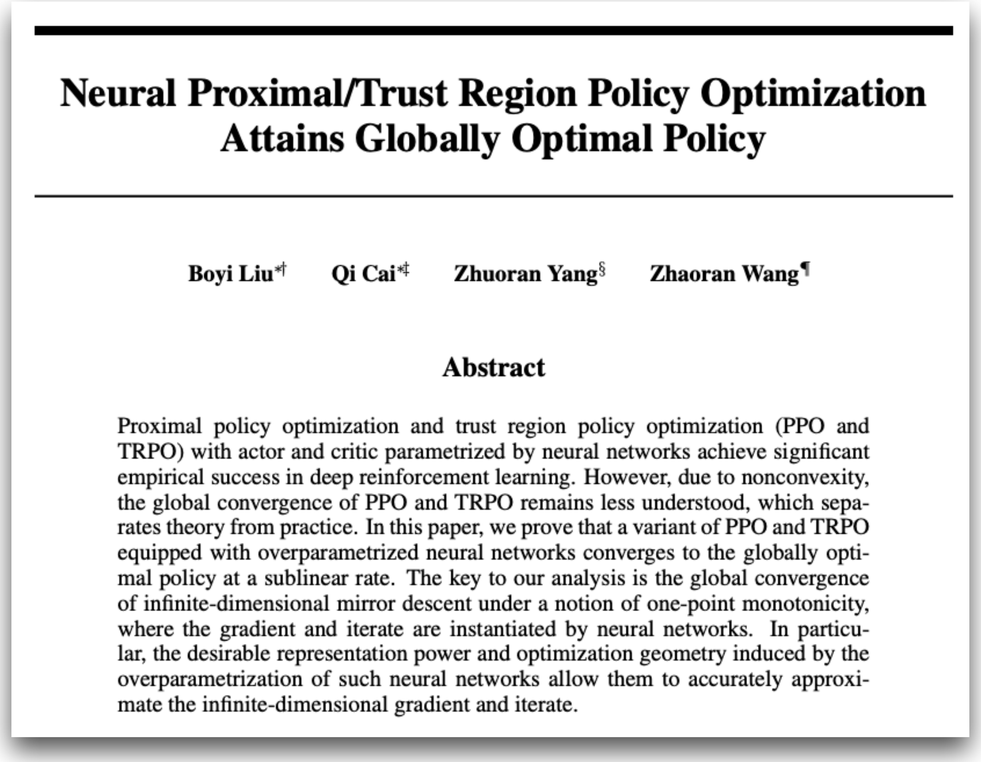
Function name	$\varphi(x)$	dom φ	$D_\varphi(x;y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty,+\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x-x$	$[0,+\infty)$	$x\log\frac{x}{y}-x+y$
Bit entropy	$x\log x+(1-x)\log(1-x)$	$[0,1]$	$x\log\frac{x}{y}+(1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0,+\infty)$	$\frac{x}{y}-\log\frac{x}{y}-1$
Hellinger	$-\sqrt{1-x^2}$	$[-1,1]$	$(1-xy)(1-y^2)^{-1/2}-(1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \quad (0<p<1)$	$[0,+\infty)$	$-x^p+p\,xy^{p-1}-(p-1)\,y^p$
ℓ_p norm	$\|x\|^p \quad (1<p<\infty)$	$(-\infty,+\infty)$	$\|x\|^p-p\,x\,\operatorname{sgn}y\,\|y\|^{p-1}+(p-1)\,\|y\|^p$
Exponential	$\exp x$	$(-\infty,+\infty)$	$\exp x-(x-y+1)\exp y$
Inverse	$1/x$	$(0,+\infty)$	$1/x+x/y^2-2/y$

A. Banerjee et al., “Clustering with Bregman Divergences,” JMLR 2005

I. Dillon and J. Tropp, “Matrix Nearness Problems with Bregman Divergences,” SIAM Journal on Matrix Analysis and Applications, 2008

Applications: PPO-KL/TRPO as Mirror Descent with KL Divergence

$$\max_{\theta} \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_k}}, a\sim\pi_{\theta_k}(\cdot|s)}\left[\frac{\pi_{\theta}(a\,|\,s)}{\pi_{\theta_k}(a\,|\,s)}A^{\theta_k}(s,a) - \beta_k D_{KL}(\pi_{\theta_k}(\,\cdot\,|\,s)\,||\,\pi_{\theta}(\,\cdot\,|\,s))\right]$$



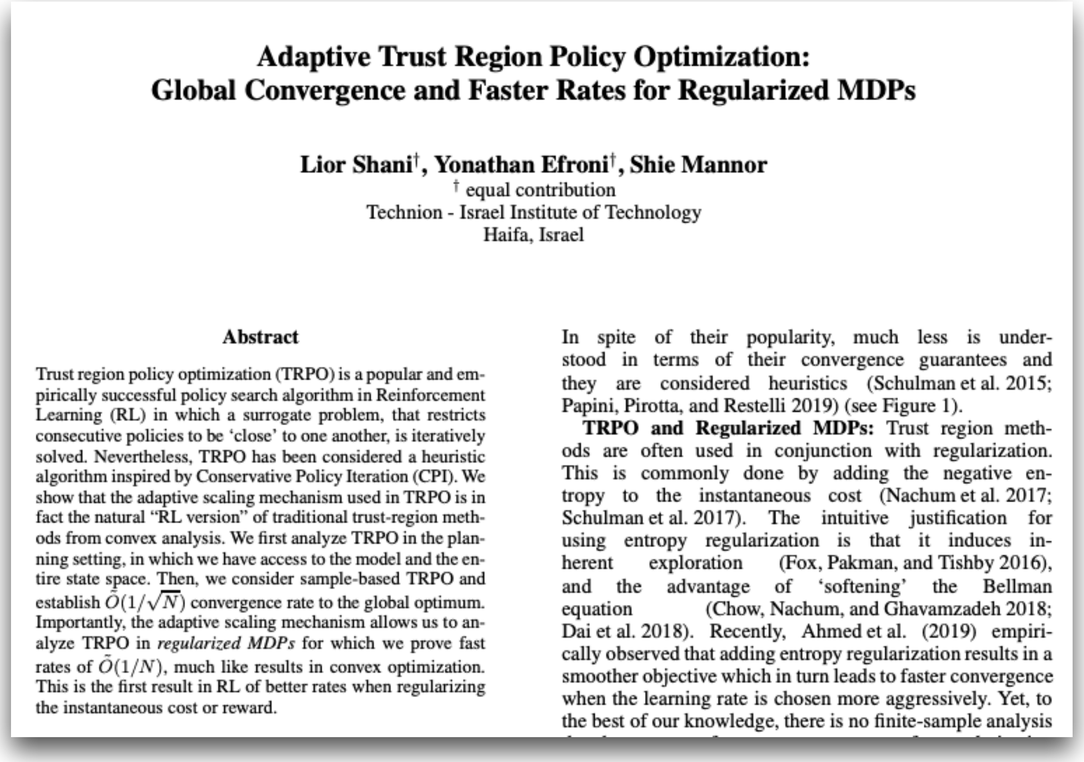
Neural Proximal/Trust Region Policy Optimization Attains Globally Optimal Policy

Boyi Liu*† Qi Cai*‡ Zhuoran Yang§ Zhaoran Wang¶

Abstract

Proximal policy optimization and trust region policy optimization (PPO and TRPO) with actor and critic parametrized by neural networks achieve significant empirical success in deep reinforcement learning. However, due to nonconvexity, the global convergence of PPO and TRPO remains less understood, which separates theory from practice. In this paper, we prove that a variant of PPO and TRPO equipped with overparametrized neural networks converges to the globally optimal policy at a sublinear rate. The key to our analysis is the global convergence of infinite-dimensional mirror descent under a notion of one-point monotonicity, where the gradient and iterate are instantiated by neural networks. In particular, the desirable representation power and optimization geometry induced by the overparametrization of such neural networks allow them to accurately approximate the infinite-dimensional gradient and iterate.

[NeurIPS 2019]



Adaptive Trust Region Policy Optimization: Global Convergence and Faster Rates for Regularized MDPs

Lior Shani†, Yonathan Efroni†, Shie Mannor
† equal contribution
Technion - Israel Institute of Technology
Haifa, Israel

Abstract

Trust region policy optimization (TRPO) is a popular and empirically successful policy search algorithm in Reinforcement Learning (RL) in which a surrogate problem, that restricts consecutive policies to be 'close' to one another, is iteratively solved. Nevertheless, TRPO has been considered a heuristic algorithm inspired by Conservative Policy Iteration (CPI). We show that the adaptive scaling mechanism used in TRPO is in fact the natural "RL version" of traditional trust-region methods from convex analysis. We first analyze TRPO in the planning setting, in which we have access to the model and the entire state space. Then, we consider sample-based TRPO and establish $\tilde{O}(1/\sqrt{N})$ convergence rate to the global optimum. Importantly, the adaptive scaling mechanism allows us to analyze TRPO in *regularized MDPs* for which we prove fast rates of $\tilde{O}(1/N)$, much like results in convex optimization. This is the first result in RL of better rates when regularizing the instantaneous cost or reward.

In spite of their popularity, much less is understood in terms of their convergence guarantees and they are considered heuristics (Schulman et al. 2015; Papini, Pirotta, and Restelli 2019) (see Figure 1).

TRPO and Regularized MDPs: Trust region methods are often used in conjunction with regularization. This is commonly done by adding the negative entropy to the instantaneous cost (Nachum et al. 2017; Schulman et al. 2017). The intuitive justification for using entropy regularization is that it induces inherent exploration (Fox, Pakman, and Tishby 2016), and the advantage of 'softening' the Bellman equation (Chow, Nachum, and Ghavamzadeh 2018; Dai et al. 2018). Recently, Ahmed et al. (2019) empirically observed that adding entropy regularization results in a smoother objective which in turn leads to faster convergence when the learning rate is chosen more aggressively. Yet, to the best of our knowledge, there is no finite-sample analysis

[AAAI 2020]

Basic Properties of Bregman Divergence

Let $\phi : X \to \mathbb{R}$ be a strictly convex and differentiable function

- 1. Non-negativity: $D_\phi(y\|x) \geq 0$
- 2. Distance between a point to itself is zero: $D_\phi(y\|x) = 0$ if and only if $x = y$
- 3. Convexity: $D_\phi(y\|x)$ is convex in y (but not necessarily in x)
- 4. Symmetry not guaranteed: $D_\phi(y\|x) \neq D_\phi(x\|y)$ in general

Basic Properties of Bregman Divergence (Cont.)

- 5. Linearity: For strictly convex ϕ_1 and ϕ_2 and $\lambda \geq 0$,

$$D_{\phi_1+\lambda\phi_2}(y\|x) = D_{\phi_1}(y\|x) + \lambda D_{\phi_2}(y\|x)$$

- 6. Gradient: $\nabla_y D_\phi(y\|x) = \nabla\phi(y) - \nabla\phi(x)$

Proof: HW2 problem

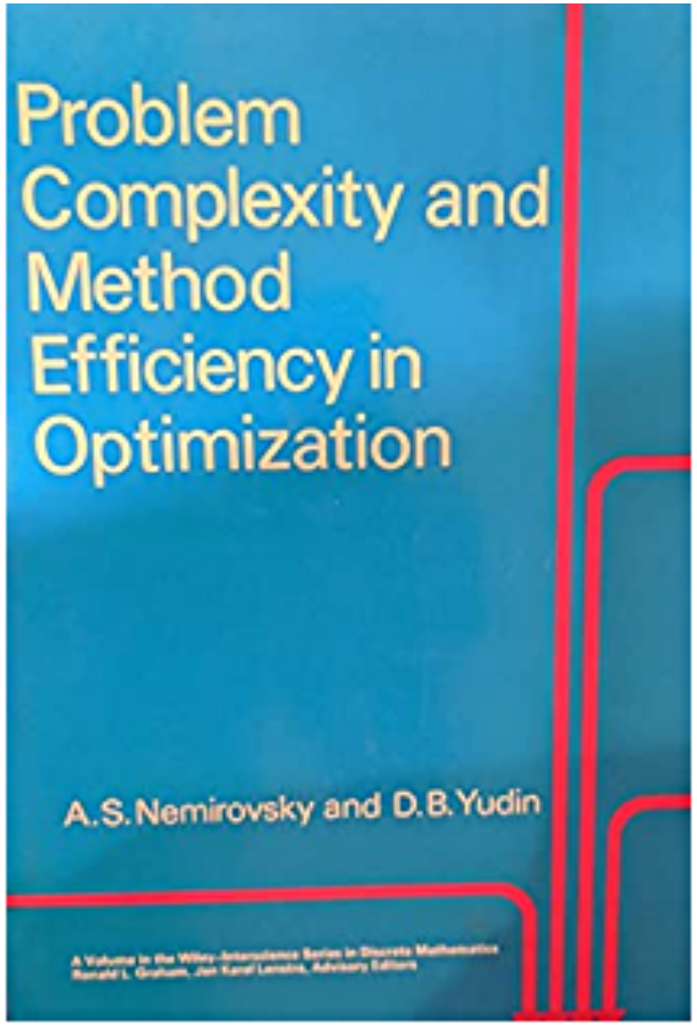
More properties of Mirror Descent will be discussed at the convergence analysis :)

Question: Why using this name “Mirror Descent”?

Mirror Descent involves the “mirror map”

A. Nemirovski and D. Yudin, ”Problem complexity and method efficiency in optimization,” Wiley, 1983.

Arkadi Nemirovski
(Professor @ Georgia Tech)



Warm-Up: MD Algorithm for Unconstrained Problems

- Mirror Descent (unconstrained cases):

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ \underbrace{f(x_t) + \nabla f(x_t)^\top (x - x_t)}_{\text{first-order approximation}} + \frac{1}{\eta_t} \underbrace{D_\phi(x \| x_t)}_{\text{Bregman divergence}} \right\}$$

- An Equivalent Update Rule of MD:

$$x_{t+1} = (\nabla \phi)^{-1} \Big(\nabla \phi(x_t) - \eta_t \nabla f(x_t) \Big)$$

- Does $(\nabla \phi)^{-1}(\cdot)$ always exist?
- Which update rule is easier to apply?

Derivation of Equivalent MD Update (Unconstrained Cases)

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \underbrace{\left\{ \nabla f(x_t)^\top (x - x_t) + \frac{1}{\eta_t} D_\phi(x \| x_t) \right\}}_{=: h(x)} \iff x_{t+1} = \nabla \phi^{-1}\left(\nabla \phi(x_t) - \eta_t \cdot \nabla f(x_t) \right)$$

---

1. By the necessary condition of optimality, we have

$$\nabla h(x_{t+1}) = 0, \quad \text{i.e.,} \quad \nabla f(x_t) + \frac{1}{\eta_t}\left(\nabla \phi(x_{t+1}) - \nabla \phi(x_t) \right) = 0$$

2. Equivalently, we have

$$\nabla \phi(x_{t+1}) = \nabla \phi(x_t) - \eta_t \cdot \nabla f(x_t)$$

Examples of MD Algorithms (Unconstrained cases)

$$x_{t+1} = \nabla\phi^{-1}\left(\nabla\phi(x_t) - \eta_t \cdot \nabla f(x_t)\right)$$

Example 1: $\phi(x) = \frac{1}{2}\|x\|^2$

- $\nabla\phi(x) = x$
- $\nabla\phi^{-1}(z) =$ _____

$$x_{t+1} = \nabla\phi(x_t) - \eta_t \cdot \nabla f(x_t)$$

$$= x_t - \eta_t \nabla f(x_t)$$

Example 2: $\phi(x) = \sum_{i=1}^{d} x_i \log x_i - x_i$

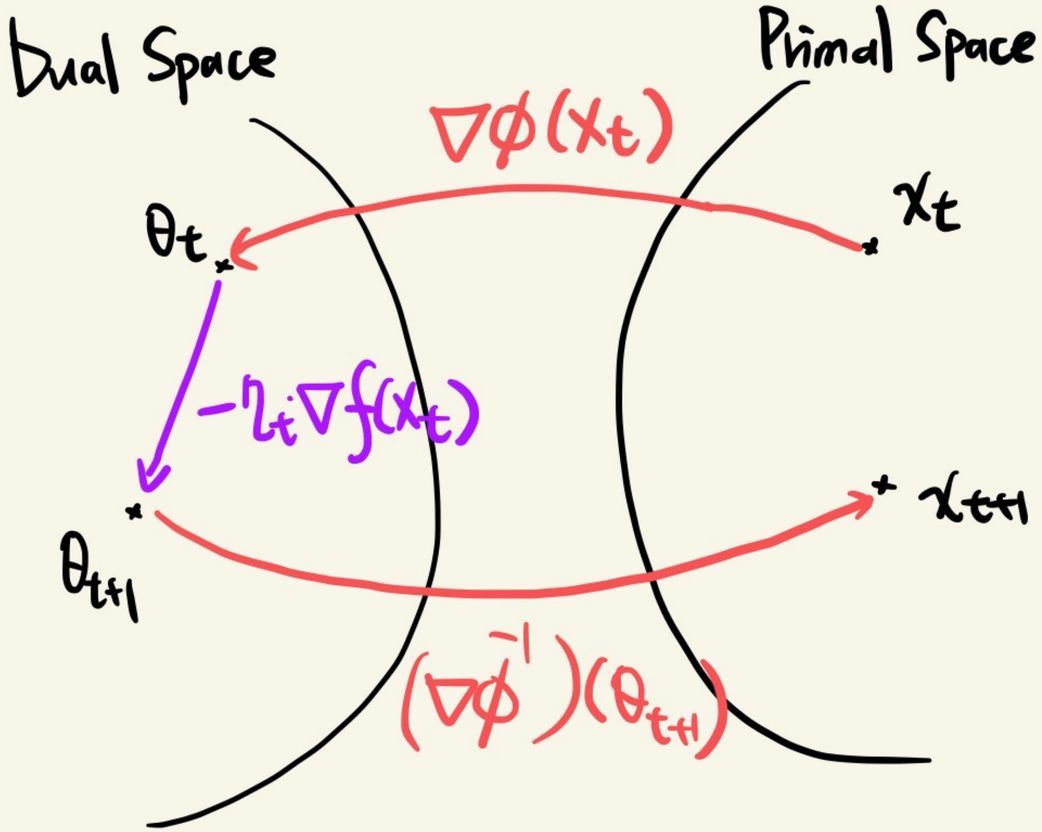
- $\nabla\phi(x) = (\log x_1, \cdots, \log x_d)$
- $\nabla\phi^{-1}(z) =$ _____

$$(x_{t+1})_i = \exp\left(\log(x_t)_i - \eta_t (\nabla f(x_t))_i\right)$$

$$= (x_t)_i \cdot \exp\left(-\eta_t (\nabla f(x_t))_i\right)$$

aka "exponentiated gradient"

Mirror Map Viewpoint (Nemirovski & Yudin, 1983)



Step 1: Map x_t to the "dual space"

$$\theta_t \leftarrow \nabla\phi(x_t)$$

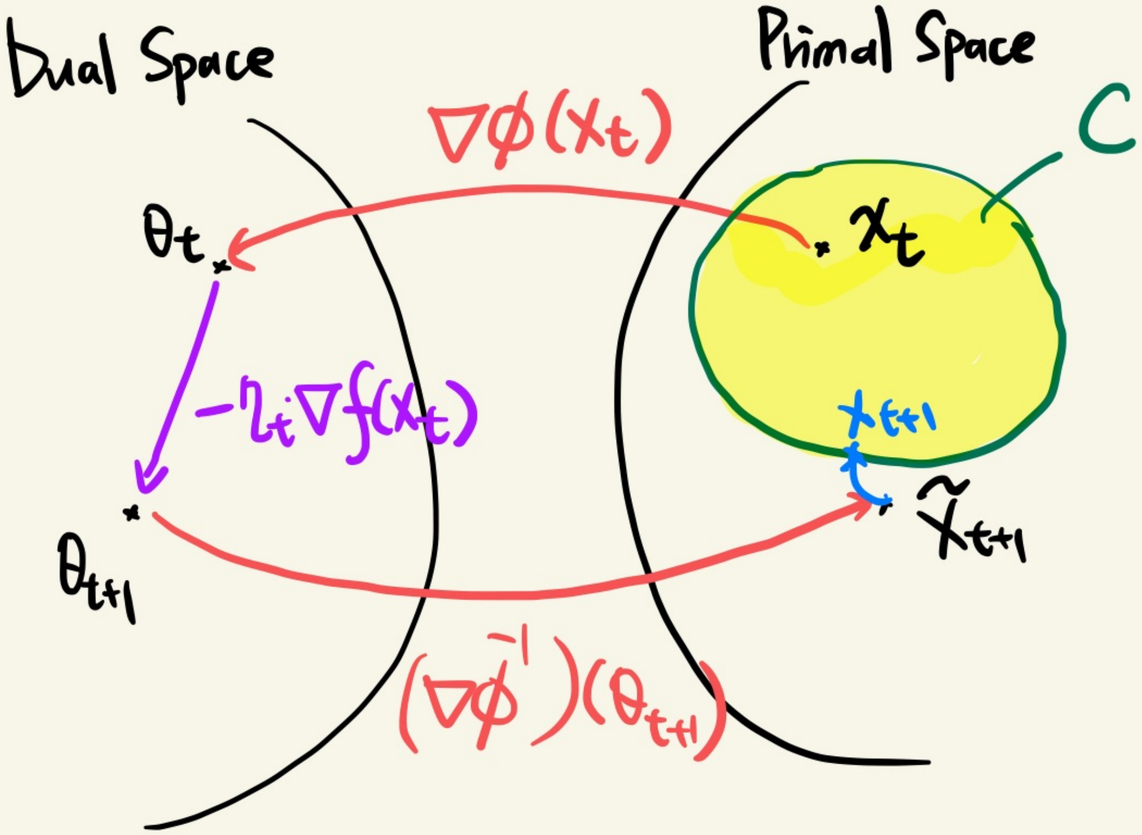
Step 2: Take a gradient step in the "dual space":

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \nabla f_t(x_t)$$

Step 3: Map θ_{t+1} back to the "primal space":

$$\tilde{x}_{t+1} \leftarrow \nabla\phi^{-1}(\theta_{t+1})$$

Mirror Map Viewpoint (Nemirovski & Yudin, 1983)



Step1: Map x_t to the "dual space"

$$\theta_t \leftarrow \nabla\phi(x_t)$$

Step2: Take a gradient step in the "dual space":

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \nabla f_t(x_t)$$

Step3: Map θ_{t+1} back to the "primal space":

$$\tilde{x}_{t+1} \leftarrow \nabla\phi^{-1}(\theta_{t+1})$$

Step4: Bregman Projection

$$x_{t+1} \leftarrow \min_{x \in C} D_\phi(x \| \tilde{x}_{t+1})$$

Equivalence Between "Mirror Map Viewpoint" and "Proximal Viewpoint"

$$x_{t+1} = \operatorname*{argmin}_{x \in C} D_\phi(x \| \tilde{x}_{t+1}) \quad \cdots\cdots \quad (\quad)$$

$$= \operatorname*{argmin}_{x \in C} \phi(x) - \phi(\tilde{x}_{t+1}) - \nabla\phi(\tilde{x}_{t+1})^T (x - \tilde{x}_{t+1}) \cdots \quad (\quad)$$

$$= \operatorname*{argmin}_{x \in C} \phi(x) - \nabla\phi(\tilde{x}_{t+1})^T x \quad \cdots \quad (\quad)$$

$$= \operatorname*{argmin}_{x \in C} \phi(x) - (\nabla\phi(x_t) - \eta_t \cdot \nabla f(x_t))^T x \quad \cdots \quad (\quad)$$

$$= \operatorname*{argmin}_{x \in C} \eta_t \cdot \nabla f(x_t)^T x + D_\phi(x \| x_t) \quad \cdots \quad (\quad)$$

A Popular MD Method: Entropic Mirror Descent

MD Algorithms for Constrained Problems

- Mirror Descent:

$$x_{t+1} = \arg\min_{x \in C} \left\{ \underbrace{f(x_t) + \nabla f(x_t)^\top (x - x_t)}_{\text{first-order approximation}} + \frac{1}{\eta_t} \underbrace{D_\phi(x \| x_t)}_{\text{Bregman divergence}} \right\}$$

- Is MD essentially a convex problem?

- The closed-form expression of MD depends on C or the constraints

A Popular MD Algorithm: Entropic Mirror Descent (EMD)

$$x_{t+1} = \arg\min_{x \in C} \left\{ f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{\eta_t} D_\phi(x \| x_t) \right\}$$

- When $D_\phi(y\|x) = \mathrm{KL}(y\|x)$ and C is the probability simplex, MD has a closed-form expression as follows: For each $i = 1, \cdots, d$,

$$[x_{t+1}]_i = \frac{[x_t]_i \cdot \exp(-\eta_t [\nabla f(x_t)]_j)}{\sum_{j=1}^{d} [x_t]_j \cdot \exp(-\eta_t [\nabla f(x_t)]_j)}$$

Entropic mirror descent / exponentiated gradient descent

Amir Beck and Marc Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” Operations Research Letters, 2003.

Application: EMD for Policy Optimization in RL

- Example: Proximal policy optimization

(ε is a hyperparameter, e.g. $\varepsilon = 0.2$)

$$\max_{\theta} \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_k}}, a\sim\pi_{\theta_k}(\cdot|s)}\left[\min\left\{\frac{\pi_\theta(a\,|\,s)}{\pi_{\theta_k}(a\,|\,s)}A^{\theta_k}(s,a), \text{clip}\left(\frac{\pi_\theta(a\,|\,s)}{\pi_{\theta_k}(a\,|\,s)}, 1-\varepsilon, 1+\varepsilon\right)A^{\theta_k}(s,a)\right\}\right]$$

- Suppose we use “tabular policies” and EMD:

$$[\theta_{t+1}]_i = \frac{[\theta_t]_i \cdot \exp(-\eta_t[\nabla f(\theta_t)]_j)}{\sum_{j=1}^{d}[\theta_t]_j \cdot \exp(-\eta_t[\nabla f(\theta_t)]_j)} =$$

- Why not using PGD?

Convergence Analysis of PPO-Clip With EMD

Neural PPO-Clip Attains Global Optimality: A Hinge Loss Perspective

Nai-Chieh Huang[1], Ping-Chun Hsieh[1], Kuo-Hao Ho[1], Hsuan-Yu Yao[2], Kai-Chun Hu[1], Liang-Chun Ouyang[1], I-Chen Wu[1,2]
[1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
{naich.cs09,pinghsieh}@nycu.edu.tw

Abstract

Policy optimization is a fundamental principle for designing reinforcement learning algorithms, and one example is the proximal policy optimization algorithm with a clipped surrogate objective (PPO-Clip), which has been popularly used in deep reinforcement learning due to its simplicity and effectiveness. Despite its superior empirical performance, PPO-Clip has not been justified via theoretical proof up to date. In this paper, we establish the first global convergence rate of PPO-Clip under neural function approximation. We identify the fundamental challenges of analyzing PPO-Clip and address them with the two core ideas: (i) We reinterpret PPO-Clip from the perspective of hinge loss, which connects policy improvement with solving a large-margin classification problem with hinge loss and offers a generalized version of the PPO-Clip objective. (ii) Based on the above viewpoint, we propose a two-step policy improvement scheme, which facilitates the convergence analysis by decoupling policy search from the complex neural policy parameterization with the help of entropic mirror descent and a regression-based policy update scheme. Moreover, our theoretical results provide the first characterization of the effect of the clipping mechanism on the convergence of PPO-Clip. Through experiments, we empirically validate the reinterpretation of PPO-Clip and the generalized objective with various classifiers on various RL benchmark tasks.

- EMD facilitates convergence analysis
- Extension to “neural policies” is made easy via EMD and a supervised learning argument

Nai-Chieh Huang, Ping-Chun Hsieh, Kuo-Hao Ho, Hsuan-Yu Yao, Kai-Chun Hu, Liang-Chun Ouyang, and I-Chen Wu “Hinge Policy Optimization: Reinterpreting PPO-Clip and Attaining Global Optimality,” arXiv preprint, March 2022.

Another Viewpoint: MD as *Follow-The-Regularized-Leader* in Online Linear Optimization

Chapter 28 of Csaba Sepesvari's textbook "Bandit Algorithms"

A Motivating Example of Online Linear Optimization

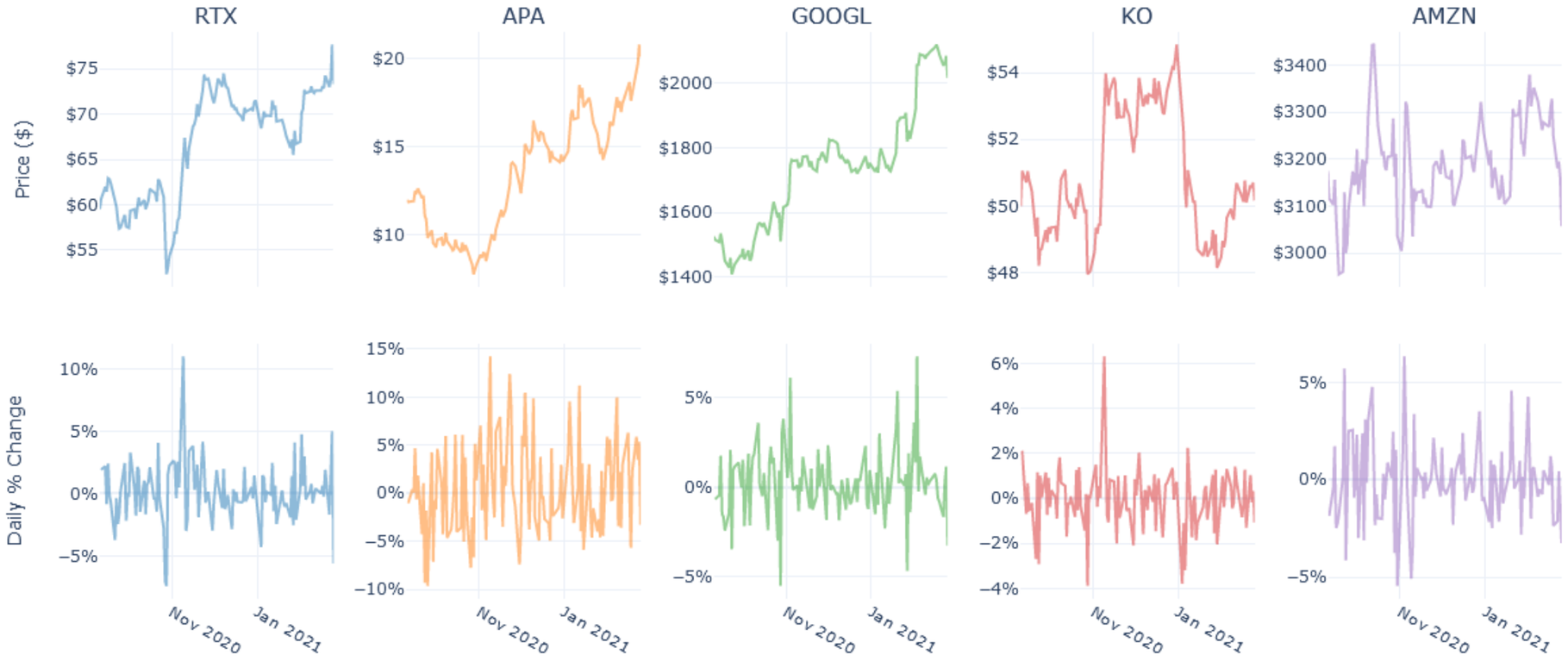
- Example: Expert Problem
  - A learner and n experts
  - At each time t, the learner chooses an action probability vector $a_t \in \Delta_n$ (indicating the distribution of expert-following behavior of the learner)
  - Let $y_t \in \mathbb{R}_+^n$ denote the cost vector of following the experts (known after choosing a_t)
  - Goal: Minimize "total cost" = Minimize "regret" (relative to a fixed comparator $a \in \Delta_n$)

$$R_T := \max_{a \in \Delta_n} \left\{ \sum_{t=1}^{T} (a_t - a)^\top y_t \right\}$$

Interesting Fact: $R_T = O(\sqrt{T \log n})$ under properly-designed algorithms!

More Online Linear/Convex Optimization Problem

Online Portfolio Selection



$$R_T = \max_{a \in \Delta_n} \sum_{t=1}^{T} \log(r_t^\top a) - \sum_{t=1}^{T} \log(r_t^\top a_t)$$

Price ratio

Wealth distribution over assets

MD and “Follow-The-Regularized-Leader” (FTRL)

Mirror Descent

Initially:

$$a_1 = \arg\min_{a \in A} \phi(a)$$

For $t \geq 1$:

$$a_{t+1} = \arg\min_{a \in A} \{a^\top y_t + \frac{1}{\eta} D_\phi(a \| a_t)\}$$

Follow-The-Regularized-Leader

Initially:

$$a_1 = \arg\min_{a \in A} \phi(a)$$

For $t \geq 1$:

$$a_{t+1} = \arg\min_{a \in A} \{\sum_{s=1}^{t} a^\top y_s + \frac{1}{\eta} \phi(a)\}$$

(Choosing the action that appears the best in hindsight under regularization)

Interesting Fact: MD and FTRL are equivalent! (under some mild conditions)

Equivalence Between MD and FTRL

MD $a_{t+1} = \operatorname{argmin}_{a} \{ \overbrace{a^T y_t + \frac{1}{\eta} \cdot D_\phi(a \| a_t)}^{F(a)} \}$

If a_{t+1} is in the interior of A, then

$\nabla F(a_{t+1}) = 0$ (and hence $-\eta \cdot y_t = \nabla \phi(a_{t+1}) - \nabla \phi(a_t)$)

Therefore, by taking "telescoping sum", we have

$$\nabla \phi(a_{t+1}) = -\eta \cdot \sum_{s=1}^{t} y_t$$

Equivalence Between MD and FTRL

FTRL $a_{t+1} = \operatorname{argmin}_{a} \left\{ \underbrace{\sum_{s=1}^{t} a^\top y_s + \frac{1}{\eta}\phi(a)}_{G(a)} \right\}$

If a_{t+1} is in the interior of A, then $\nabla G(a_{t+1}) = 0$

(Equivalently, $\nabla \phi(a_{t+1}) = -\eta \cdot \sum_{s=1}^{t} y_s$)

Why Not Using "Follow-The-Leader" (FTL)?

Follow-The-Leader: Choosing the action that appears the best in hindsight

$$a_{t+1} = \arg\min_{a \in A} \left\{ \sum_{s=1}^{t} a^\top y_s \right\}$$

- FTL (i.e., without a regularizer) suffers from linear regret!

- Counterexample:
  - Let action set $A = [-1,1]$ and initial action $a_1 = 0$
  - Cost vectors: $y_1 = 1/2$ and $y_s = (-1)^{s+1}, \forall s > 1$

Counterexample of "Follow-The-Leader"

$A = [-1, 1]$, $a_1 = 0$

$y_1 = \frac{1}{2} \Rightarrow$ $a_2 = \underset{a \in A}{\operatorname{argmin}}\ a^T y_1 =$ _____

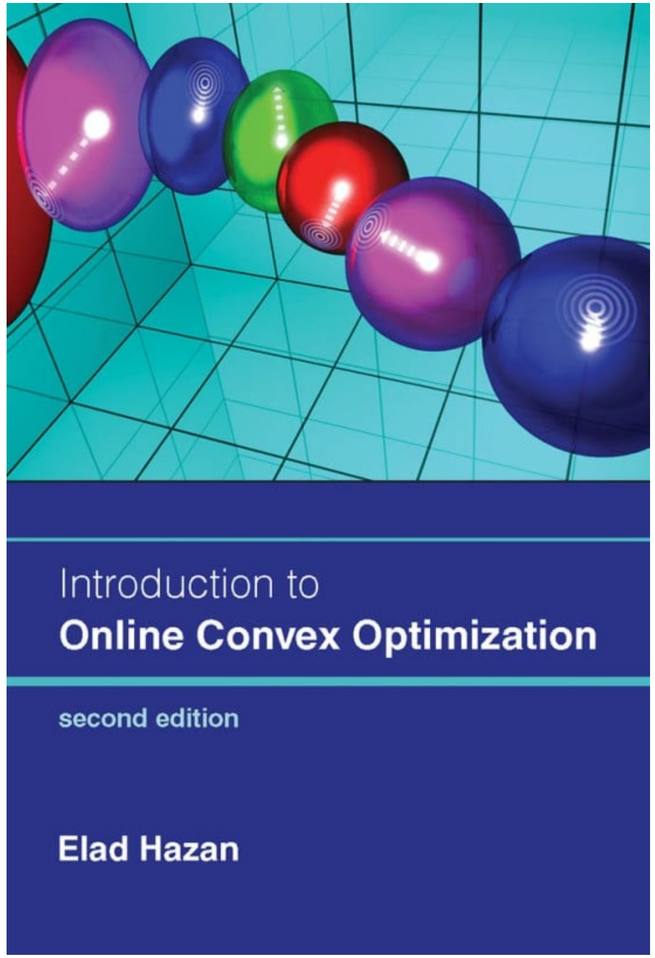
$y_2 = -1 \Rightarrow$ $a_3 = \underset{a \in A}{\operatorname{argmin}}\ a^T (y_1 + y_2) =$ _____

$y_3 = +1 \Rightarrow$ $a_4 = \underset{a \in A}{\operatorname{argmin}}\ a^T (y_1 + y_2 + y_3) =$ _____

Hence, $a_t =$ _____

"Best action in hindsight" is _____, Regret $R_T =$ _____

https://tor-lattimore.com/downloads/book/book.pdf



https://arxiv.org/abs/1909.05207

Convergence of Mirror Descent

Convergence of MD Algorithms

Theorem Suppose the following conditions hold:

① f is convex and L_f-Lipschitz continuous (in the sense that $\|\nabla f(x)\|_* \le L_f$)

② ϕ is ρ-strongly convex w.r.t. $\|\cdot\|$

Then, MD achieves

$$f_{best}^t - f^* \le \frac{\sup_{x \in C} D_\phi(x \| x_0)}{\sum_{\tau=0}^t \eta_\tau} + \frac{L_f^2}{2\rho} \cdot \frac{\sum_{\tau=0}^t \eta_\tau^2}{\sum_{\tau=0}^t \eta_\tau}$$

($\sup_{x \in C} D_\phi(x \| x_0) =: D_{max}$)

Moreover, by letting $\eta_t = \frac{\sqrt{2\rho \cdot D_{max}}}{L_f} \cdot \frac{1}{\sqrt{t}}$,

$$f_{best}^t - f^* = O\left(\frac{L_f \sqrt{D_{max}}}{\sqrt{\rho}} \cdot \frac{\log t}{t} \right)$$

Amir Beck and Marc Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” Operations Research Letters, 2003.

Example: Optimization over Probability Simplex

(For simplicity, let $X_0 = (\frac{1}{d}, \cdots, \frac{1}{d})$)

Suppose we consider: $\min f(x)$, subject to $x \geq 0$, $1^T x = 1$.

Let's compare MD with "Euclidean norm" vs "KL divergence"

---

① Euclidean norm:

- $\phi(x) = \frac{1}{2} \cdot \|x\|^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$

$$\sup_{x \in C} D_\phi(x \| x_0) = \sup_{x \in C} \frac{1}{2} \cdot \| x - (\tfrac{1}{d}, \cdots, \tfrac{1}{d}) \|^2 \leq \frac{1}{2}$$

Hence, $f_{best}^t - f^* = O\left(\frac{L_f \cdot \sqrt{D_{max}}}{\sqrt{\rho}} \cdot \frac{\log t}{\sqrt{t}} \right) = O\left(L_{f,2} \cdot \frac{\log t}{\sqrt{t}} \right)$

(Cont.)

② KL divergence:

- $\phi(x) = -\sum_{i=1}^{d} x^{(i)} \log x^{(i)}$ is 1-strongly convex w.r.t. $\|\cdot\|_1$

$$\sup_{x \in C} D_\phi(x \| x_0) = \sup_{x \in C} KL(x \| x_0) = \sup_{x \in C} \left(\sum_{i=1}^{d} x^{(i)} \log x^{(i)} - \sum_{i=1}^{d} x^{(i)} \cdot \log \frac{1}{n} \right)$$

$$= \log n + \sup_{x \in C} \sum_{i=1}^{d} x^{(i)} \log x^{(i)} \leq \log n$$

Hence, $f_{best}^t - f^* = O\left(\frac{L_f \cdot \sqrt{D_{max}}}{\sqrt{\rho}} \cdot \frac{\log t}{\sqrt{t}} \right) = O\left(L_{f,\infty} \cdot \sqrt{\log n} \cdot \frac{\log t}{\sqrt{t}} \right)$

Comparison:

	Euclidean	KL Divergence
Convergence Rate	$O\left(L_{f,2} \cdot \frac{\log t}{\sqrt{t}}\right)$	$O\left(L_{f,\infty} \cdot \sqrt{\log n} \cdot \frac{\log t}{\sqrt{t}}\right)$

Note that $\|\nabla f\|_\infty \le \|\nabla f\|_2 \le \sqrt{n} \cdot \|\nabla f\|_\infty$ (why?)

Hence, $\frac{1}{\sqrt{n}} \le \frac{L_{f,\infty}}{L_{f,2}} \le 1 \Rightarrow$ MD with KL divergence has a better convergence rate!

Proof of Convergence of Mirror Descent

Amir Beck and Marc Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” Operations Research Letters, 2003.

Hölder's Inequality

Lemma: Let $p, q \in [1, \infty]$ so that $\frac{1}{p} + \frac{1}{q} = 1$. Then, for any two vectors $x = (x_1, \cdots, x_n)$, $y = (y_1, \cdots, y_n)$, we have

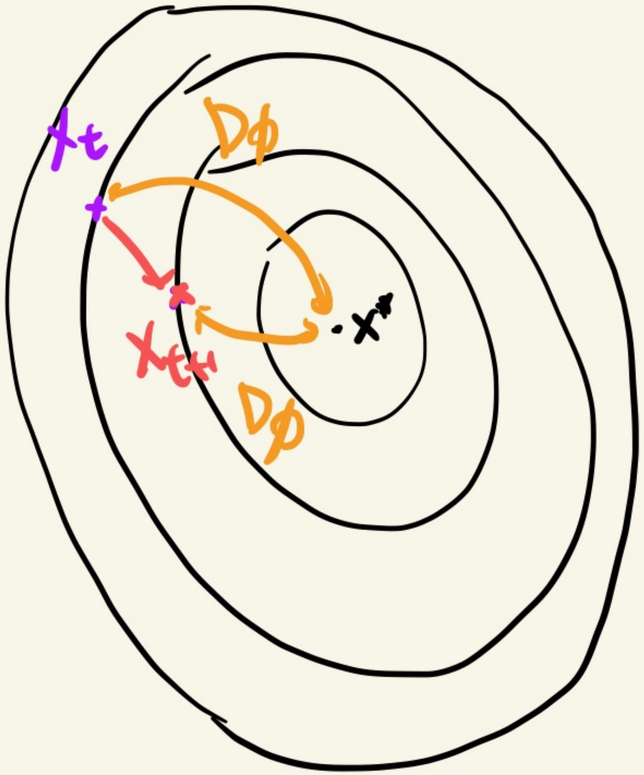
$$\left| \sum_{i=1}^{n} x_i \cdot y_i \right| \leq \|x\|_p \cdot \|y\|_q .$$

Remark: If $p=2, q=2$, Hölder's inequality is essentially Cauchy–Schwarz

A Fundamental Inequality of MD

Lemma $$\eta_t \cdot (f(x_t) - f^*) \le D_\phi(x^* \| x_t) - D_\phi(x^* \| x_{t+1}) + \frac{\eta_t^2 L_f^2}{2\cdot\rho}$$

Question: What's the intuition?



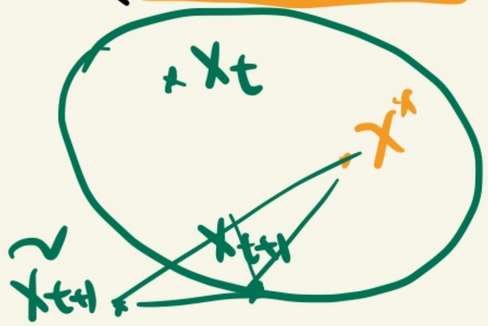
Proof of Lemma :

Step 1: $f(x_t) - f(x^*) \le \nabla f(x_t)^T (x_t - x^*)$ ... (

$= \frac{1}{\eta_t} \cdot \left(\nabla \phi(x_t) - \nabla \phi(\tilde{x}_{t+1}) \right)^T (x_t - x^*)$ ... (

$= \frac{1}{\eta_t} \cdot \left(D_\phi(x^* \| x_t) + D_\phi(x_t \| \tilde{x}_{t+1}) - D_\phi(x^* \| \tilde{x}_{t+1}) \right)$ ... (

$\le \frac{1}{\eta_t} \left(D_\phi(x^* \| x_t) + D_\phi(x_t \| \tilde{x}_{t+1}) - \left(D_\phi(x^* \| x_{t+1}) + D_\phi(x_{t+1} \| \tilde{x}_{t+1}) \right) \right)$ ... (



(Cont.).

Step2: It is sufficient to show the following claim:

$$D_\phi(x_t, \tilde{x}_{t+1}) - D_\phi(x_{t+1}, \tilde{x}_{t+1}) \leq \frac{(\eta_t \cdot L_f)^2}{2\cdot\rho} \quad \text{——(*)}$$

(which would naturally lead to our required lemma)

$$D_\phi(x_t, \tilde{x}_{t+1}) - D_\phi(x_{t+1}, \tilde{x}_{t+1})$$

$$= \phi(x_t) - \phi(x_{t+1}) - \nabla\phi(\tilde{x}_{t+1})^T(x_t - x_{t+1}) \cdots (\quad)$$

$$\leq \nabla\phi(x_t)^T(x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2 - \nabla\phi(\tilde{x}_{t+1})^T(x_t - x_{t+1})$$

$$\cdots (\quad)$$

(Cont.)

$$= \left(\nabla\phi(x_t) - \nabla\phi(\tilde{x}_{t+1})\right)^T (x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2$$

$$= \eta_t \cdot \nabla f(x_t)^T \cdot (x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2 \quad \cdots (\quad)$$

$$\leq \eta_t \cdot L_f \cdot \|x_t - x_{t+1}\| - \frac{\rho}{2}\|x_t - x_{t+1}\|^2 \quad \cdots (\quad)$$

$$\leq \frac{(\eta_t \cdot L_f)^2}{2\rho} \quad \cdots (\quad)$$

Proof of Main Theorem

Step1: By taking the telescoping sum of $\eta_\tau(f(x_\tau)-f^*)$, we have

$$\sum_{\tau=0}^{t}\eta_\tau\cdot(f(x_\tau)-f^*)\le D_\phi(x^*\|x_0)-D_\phi(x^*\|x_{t+1})+\frac{L_f^2\cdot\sum_{\tau=0}^{t}\eta_\tau^2}{2\cdot\rho}$$

$$\le \sup_{x\in C} D_\phi(x\|x_0)+\frac{L_f^2\cdot\sum_{\tau=0}^{t}\eta_\tau^2}{2\cdot\rho}$$

Step2: Moreover, we have

$$f_{best}^t-f^*\le\frac{\sum_{\tau=0}^{t}\eta_\tau(f(x_\tau)-f^*)}{\sum_{\tau=0}^{t}\eta_\tau}$$