A Novel Framework for Policy Mirror Descent with General Parameterization and Linear Convergence

Carlo Alfano[1], Rui Yuan[2], Patrick Rebeschini[1]

1: Department of Statistics, University of Oxford; 2: Stellantis*

*The work was done prior to joining Stellantis, while the author was at Télécom Paris.

Introduction

Modern policy optimization algorithm, such as TRPO and PPO, owe their success to the use of parameterized policies such as

$$\pi(a|s) \propto \exp(f^\theta(s,a)),$$

where f^θ is a neural network. However, the use general parameterization schemes still lacks theoretical justification.

Contribution: A novel framework for policy optimization based on mirror descent that naturally accommodates general parameterizations and enjoys theoretical guarantees.

Preliminaries

Consider a discounted MDP $(\mathcal{S},\mathcal{A},P,r,\gamma,\mu)$. Given a policy π, define the value function

$$V^\pi(s) := \mathbb{E}_{a_t\sim\pi(\cdot|s_t), s_{t+1}\sim P(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\middle|\pi, s_0 = s\right]$$

and the Q-function

$$Q^\pi(s,a) := \mathbb{E}_{a_t\sim\pi(\cdot|s_t), s_{t+1}\sim P(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) \mid \pi, s_0 = s, a_0 = a\right].$$

Letting $V^\pi(\mu) := \mathbb{E}_{s\sim\mu}[V^\pi(s)]$, our objective is for the agent to find an optimal policy

$$\pi^\star \in \operatorname{argmax}_{\pi\in(\Delta(\mathcal{A}))^{\mathcal{S}}} V^\pi(\mu).$$

Define the discounted state visitation distribution by

$$d_\mu^\pi(s) := (1-\gamma)\mathbb{E}_{s_0\sim\mu}\left[\sum_{t=0}^{\infty}\gamma^t P(s_t = s \mid \pi, s_0)\right].$$

Mirror Descent. Let $\mathcal{Y} \subseteq \mathbb{R}^{|\mathcal{A}|}$ be a convex set. A *mirror map* $h : \mathcal{Y} \to \mathbb{R}$ is a strictly convex, continuously differentiable and essentially smooth function[a] such that $\nabla h(\mathcal{Y}) = \mathbb{R}^{|\mathcal{A}|}$. The convex conjugate of h, denoted by h^*, is given by

$$h^*(x^*) := \sup_{x\in\mathcal{Y}}\langle x^*, x\rangle - h(x), \quad x^* \in \mathbb{R}^{|\mathcal{A}|}.$$

The mirror map h induces a *Bregman divergence*, defined as

$$\mathcal{D}_h(x,y) := h(x) - h(y) - \langle\nabla h(y), x - y\rangle,$$

where $\mathcal{D}_h(x,y) \geqslant 0$ for all $x,y\in\mathcal{Y}$. Let $\mathcal{X}\subseteq\mathcal{Y}$ be a convex set and $V : \mathcal{X}\to\mathbb{R}$ be a differentiable function. To solve $\min_{x\in\mathcal{X}} V(x)$, MD consists in the updates: for all $t\geqslant 0$,

$$y^{t+1} = \nabla h(x^t) - \eta_t \nabla V(x)|_{x=x^t},$$

$$x^{t+1} = \text{Proj}_{\mathcal{X}}^h(\nabla h^*(y^{t+1})) = \operatorname{argmin}_{x\in\mathcal{X}} \mathcal{D}_h(x, \nabla h^*(y^{t+1})).$$

Notation. At each time t, let $\pi^t := \pi^{\theta_t}$, $f^t := f^{\theta_t}$, $V^t := V^{\pi^t}$, $Q^t := Q^{\pi^t}$, and $d_\mu^t := d_\mu^{\pi^t}$. Further, for any function $y : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and distribution v over $\mathcal{S}\times\mathcal{A}$, let $y_s := y(s,\cdot)\in\mathbb{R}^{|\mathcal{A}|}$ and $\|y\|_{L_2(v)}^2 = \mathbb{E}_v[(y(s,a))^2]$. Let $\mathcal{D}_0^\star = \mathbb{E}_{s\sim d_\mu^\star}[\mathcal{D}_h(\pi_s^\star, \pi_s^0)]$.

[a] h is essentially smooth if $\lim_{x\to\partial\mathcal{Y}}\|\nabla h(x)\|_2 = +\infty$, where $\partial\mathcal{Y}$ denotes the boundary of $\mathcal{Y}$.

Policy Mirror Descent

Given a parameterized function class $\mathcal{F}^\Theta = \{f^\theta : \mathcal{S}\times\mathcal{A}\to\mathbb{R}, \theta\in\Theta\}$, ideally, we would like to execute the exact MD-based algorithm: for all $t\geqslant 0$ and for all $s\in\mathcal{S}$,

$$f_s^{t+1} = \nabla h(\pi_s^t) + \eta_t(1-\gamma)\nabla_s V^t(\mu)/d_\mu^t(s) = \nabla h(\pi_s^t) + \eta_t Q_s^t, \qquad (1)$$

$$\pi_s^{t+1} = \text{Proj}_{\Delta(\mathcal{A})}^h(\nabla h^*(\eta_t f_s^{t+1})).$$

However, there may not be any $\theta^{t+1}\in\Theta$ such that (1) is satisfied for all $s\in\mathcal{S}$. To remedy this issue, we propose Approximate Mirror Policy Optimization (AMPO).

Approximate Mirror Policy Optimization

Algorithm 1: Approximate Mirror Policy Optimization

Input: Initial policy π^0, mirror map h, parameterization class $\mathcal{F}^\Theta$, iteration number T, step-size schedule $(\eta_t)_{t\geqslant 0}$, state-action distribution sequence $(v_t)_{t\geqslant 0}$.

1: For $t = 0,\dots,T-1$ do:

2: Obtain $\theta^{t+1}\in\Theta$ such that

$$\theta^{t+1} \in \operatorname{argmin}_{\theta\in\Theta}\left\|f^\theta - Q^t - \eta_t^{-1}\nabla h(\pi^t)\right\|_{L_2(v_t)}^2.$$

3: Update

$$\pi_s^{t+1} = \operatorname*{argmin}_{\pi'\in\Delta(\mathcal{A})} \mathcal{D}_h(\pi', \nabla h^*(\eta_t f_s^{\theta^{t+1}})), \ \forall s\in\mathcal{S}.$$

Output: $(\pi^1,\dots,\pi^T)$

Comparison with previous frameworks

Similarly to AMPO, previous approximations of PMD [1, 2] provide an expression to be optimized. For instance, [1] aim to maximize an expression equivalent to

$$\pi^{t+1} = \operatorname*{argmax}_{\pi^\theta\in\Pi(\Theta)} \mathbb{E}_{s\sim d_\mu^t}[\eta_t\langle Q_s^t, \pi_s^\theta\rangle - \mathcal{D}_h(\pi_s^\theta, \pi_s^t)], \qquad (2)$$

where $\Pi(\Theta)$ is a given parameterized policy class. The improvement of AMPO over this type of update is twofold.

▷ The parameterized policy class $\Pi(\Theta)$ is often non-convex with respect to θ in practice, which prevents the application of existing proof techniques that rely on the convexity of the tabular parameterization [3]. On the contrary, AMPO avoids this problem thanks to the Bregman projection and the update in Line 2 of Algorithm 1.

▷ AMPO involves a subroutine optimization procedure that is structurally different from the update in (2). Our approach employs a standard regression procedure, which has been extensively studied and benefits from established solving methods.

A practical class of mirror maps

For $a\in(-\infty,+\infty]$, $\omega\leqslant 0$, let an ω*-potential* be an increasing C^1-diffeomorphism $\phi : (-\infty,a)\to(\omega,+\infty)$ such that

$$\lim_{u\to-\infty}\phi(u) = \omega, \qquad \lim_{u\to a}\phi(u) = +\infty, \qquad \int_0^1 \phi^{-1}(u)du \leqslant \infty.$$

For any ω-potential ϕ, the associated mirror map h_ϕ is defined as

$$h_\phi(\pi_s) = \sum_{a\in\mathcal{A}}\int_1^{\pi(a|s)}\phi^{-1}(u)du.$$

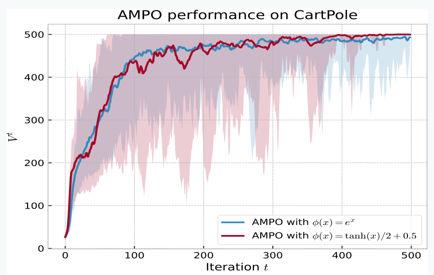
Thanks to [4, Proposition 2], the policy π^{t+1} in Line 3 induced by the ω-potential mirror map can be obtained with $\tilde{\mathcal{O}}(|\mathcal{A}|)$ computations and can be written as

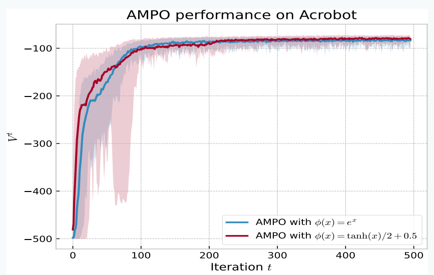
$$\pi^{t+1}(a|s) = \sigma(\phi(\eta_t f^{t+1}(s,a) + \lambda_s^{t+1})) \quad \forall s\in\mathcal{S}, a\in\mathcal{A},$$

where $\lambda_s\in\mathbb{R}$ is a normalization factor to ensure $\sum_{a\in\mathcal{A}}\pi^{t+1}(a|s) = 1$ for all $s\in\mathcal{S}$, and $\sigma(z) = \max(z,0)$ for $z\in\mathbb{R}$. The minimization problem in Line 2 is simplified to be

$$\theta^{t+1} \in \operatorname*{argmin}_{\theta\in\Theta}\left\|f^\theta - Q^t - \eta_t^{-1}\max(\eta_{t-1}f^t, \phi^{-1}(0) - \lambda_s^t)\right\|_{L_2(v_t)}^2.$$

When $\phi(x) = e^x$, we recover an approximation of NPG.





Convergence Rates

Assumption (A1) (*Approximation error*). There exists $\varepsilon_{\text{approx}}\geqslant 0$ such that, $\forall t\geqslant 0$,

$$\mathbb{E}\left[\left\|f^{t+1} - Q^t - \eta_t^{-1}\nabla h(\pi^t)\right\|_{L_2(v_t)}^2\right] \leqslant \varepsilon_{\text{approx}},$$

where $(v^t)_{t\geqslant 0}$ is a sequence of distributions over states and actions and the expectation is taken over the randomness of AMPO.

Assumption (A2) (*Concentrability coefficient*). There exists $C_v\geqslant 0$ such that, $\forall t\geqslant 0$,

$$\mathbb{E}_{(s,a)\sim v^t}\left[\left(\frac{d_\mu^\pi(s)\pi(a|s)}{v^t(s,a)}\right)^2\right] \leqslant C_v,$$

whenever (d_μ^π, π) is either $(d_\mu^\star, \pi^\star)$, (d_μ^{t+1}, π^{t+1}), $(d_\mu^\star, \pi^t)$, or (d_μ^{t+1}, π^t).

Assumption (A3) (*Distribution mismatch coefficient*). There exists $\nu_\mu\geqslant 0$ such that

$$\max_{s\in\mathcal{S}}\frac{d_\mu^\star(s)}{d_\mu^t(s)} \leqslant \frac{1}{1-\gamma}\max_{s\in\mathcal{S}}\frac{d_\mu^\star(s)}{\mu(s)} \leqslant \nu_\mu, \quad \text{for all times } t\geqslant 0.$$

Theorem 4.3. Let Assumptions (A1), (A2), and (A3) be true. If the step-size schedule is non-decreasing, i.e., $\eta_t\leqslant\eta_{t+1}$ for all $t\geqslant 0$, the iterates of Algorithm 1 satisfy: $\forall T\geqslant 0$,

$$V^\star(\mu) - \frac{1}{T}\sum_{t<T}\mathbb{E}\left[V^t(\mu)\right] \leqslant \frac{1}{T}\left(\frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + \frac{\nu_\mu}{1-\gamma}\right) + \frac{2(1+\nu_\mu)\sqrt{C_v\varepsilon_{\text{approx}}}}{1-\gamma}.$$

Furthermore, if the step-size schedule is geometrically increasing, i.e., satisfies

$$\eta_{t+1} \geqslant \frac{\nu_\mu}{\nu_\mu - 1}\eta_t \qquad \forall t\geqslant 0,$$

we have: for every $T\geqslant 0$,

$$V^\star(\mu) - \mathbb{E}\left[V^T(\mu)\right] \leqslant \frac{1}{1-\gamma}\left(1 - \frac{1}{\nu_\mu}\right)^T\left(1 + \frac{\mathcal{D}_0^\star}{\eta_0(\nu_\mu - 1)}\right) + \frac{2(1+\nu_\mu)\sqrt{C_v\varepsilon_{\text{approx}}}}{1-\gamma}.$$

▷ First result that establishes linear convergence for a PG-based method involving general policy parameterization and mirror maps.

▷ For the same setting, it is also the first result that establishes $O(1/T)$ convergence without regularization.

▷ First result that provides a convergence rate for a PMD-based algorithm that allows any mirror map and non-tabular policies.

Sample complexity for neural network parameterization

Let $\mathcal{F}^\Theta$ be a class of shallow neural networks. At each iteration t of AMPO, we set $v^t = d_\mu^t$ and solve the regression problem in Line 2 of Algorithm 1 through SGD. Then, thank to Theorem 4.3 and an existing analysis of neural networks [5, Theorem 1], we have the sample complexity of AMPO

Corollary 4.4. In the setting of Theorem 4.3, let the parameterization class $\mathcal{F}^\Theta$ consist of sufficiently wide shallow ReLU neural networks. Using an exponentially increasing step-size and solving the minimization problem in Line 2 with SGD, the number of samples required by AMPO to find an ε-optimal policy with high probability is $\tilde{\mathcal{O}}(C_v^2\nu_\mu^5/\varepsilon^4(1-\gamma)^6)$, where ε has to be larger than a non-vanishing error floor.

References

[1] S. Vaswani et al.. A general class of surrogate functions for stable and efficient reinforcement learning. In *AISTATS*, 2022.

[2] M. Tomar et al.. Mirror descent policy optimization. In *ICLR*, 2022.

[3] L. Xiao. On the convergence rates of policy gradient methods. *JMLR*, 2022.

[4] W. Krichene et al.. Efficient bregman projections onto the simplex. In *IEEE CDC*, 2015.

[5] Z. Allen-Zhu et al.. Learning and generalization in overparameterized neural networks, going beyond two layers. *NeurIPS*, 2019.