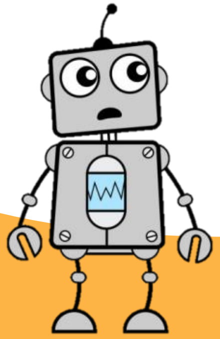
Policy Optimization in Adversarial MDPs:

Improved Exploration via Dilated Bonuses

Haipeng Luo, **Chen-Yu Wei**, Chung-Wei Lee

University of Southern California



Policy Optimization

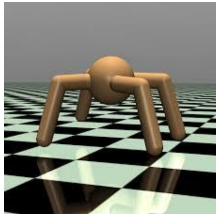
collect data using π_θ

$$\theta \leftarrow \theta - \eta \nabla_\theta \hat{V}(\theta)$$

repeat

estimated loss of π_θ

Wide empirical success ⟷ Theoretically less understood



in contrast to model-based (UCBVI)
or value-based (UCB-Q) approaches

Policy Optimization

Benefit: directly optimizes policies → less prone to modeling error (compared to model- or value-based methods)

In fact, standard policy optimization is based on the mirror descent framework, which can even handle adversarial losses.

Drawback: perform local policy search and lack exploration → slow/unable to find global optimum

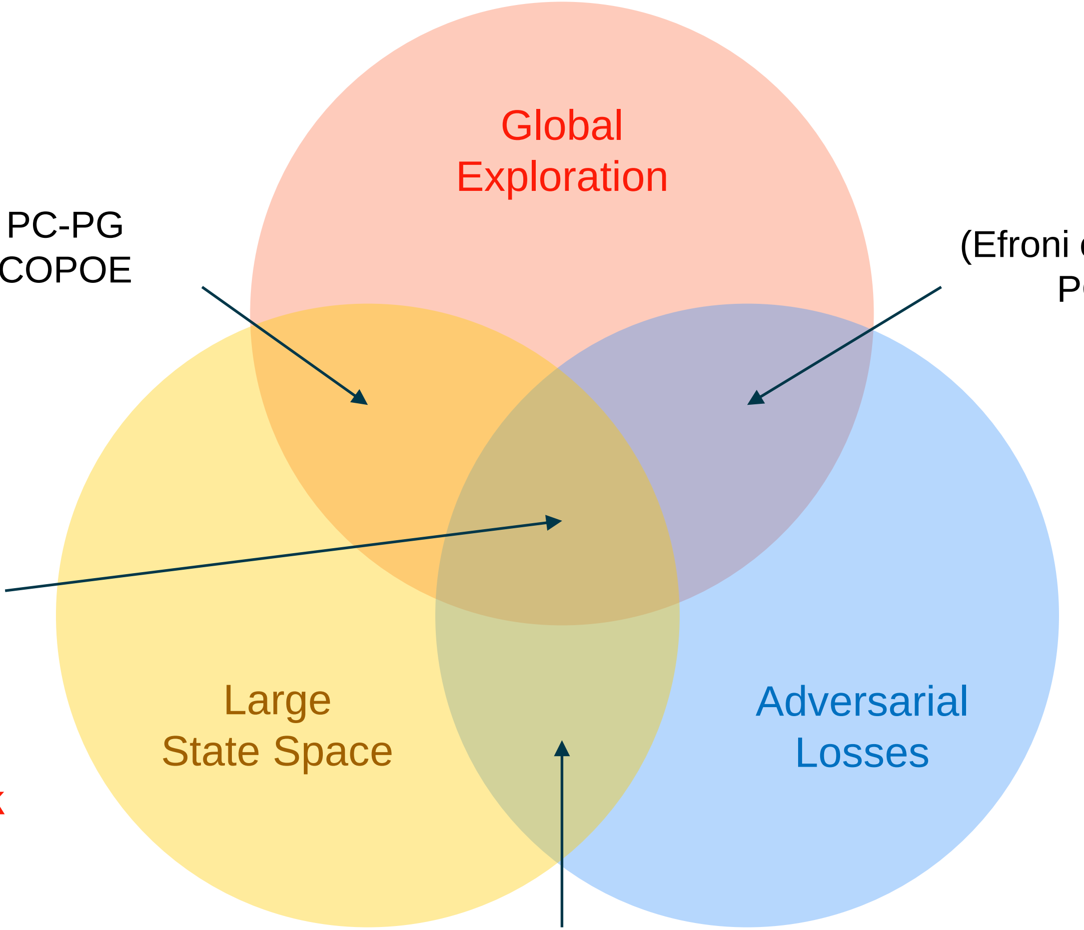
Q

Can Policy Optimization perform global exploration under adversarial losses?

Previous Work

(Agarwal et al. 2020) PC-PG
(Zanette et al. 2021) COPOE

(Efroni et al. 2020)
POMD



(Cai et al. 2020) OPPO
(He et al. 2021) POWER

Require full-information loss feedback

Our work: bandit feedback

(Neu and Olkhovskaya 2020)
MDP-LinExp3

Contributions

- A new general way of constructing exploration bonuses for policy optimization (suitable for adversarial loss + function approximation + bandit feedback)
- Applications to several settings:

Tabular MDP

regret $= \tilde{\mathcal{O}}(\sqrt{T})$

improving Efroni et al.'s $\tilde{\mathcal{O}}(T^{2/3})$ bound

Linear-Q MDP + simulator

regret $= \tilde{\mathcal{O}}(T^{2/3})$

matching Neu & Olkhovskaya's, but removing their requirement of an exploratory policy

Linear MDP + exploratory policy

regret $= \tilde{\mathcal{O}}(T^{6/7})$

first sublinear regret

Linear MDP

regret $= \tilde{\mathcal{O}}(T^{14/15})$

first sublinear regret

(only appearing in our arxiv version)

Setting and Algorithm

Finite-horizon MDP with horizon length H, state space $\mathcal{S}$, action space $\mathcal{A}$, and an unknown transition kernel $p(s'|s,a)$

For episode $t = 1, 2, \ldots, T$:

Adversary chooses a loss function $\ell_t(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to [0,1]$

Learner chooses a policy π_t

For step $h = 0, 1, \ldots, H-1$:

Learner observes s_h, and chooses $a_h \sim \pi_t(\cdot|s_h)$

Learner observes $\ell_t(s_h, a_h)$

Learner generates $\hat{Q}_t(\cdot,\cdot)$ (an estimator of $Q^{\pi_t}(\cdot,\cdot;\ell_t)$)

and perform mirror descent update $\pi_{t+1}(a|s) \propto \pi_t(a|s) \exp\left(-\eta \hat{Q}_t(s,a)\right)$

Q function under policy π_t and loss ℓ_t

Deriving Exploration Bonus for Policy Optimization

$$\text{regret} = \sum_{t=1}^{T} \left(V^{\pi^\star}(s_0; \ell_t) - V^{\pi_t}(s_0; \ell_t) \right)$$

$$= \sum_s \mu^{\pi^\star}(s) \underbrace{\sum_{t=1}^{T} \sum_a (\pi_t(a|s) - \pi^\star(a|s)) \, Q^{\pi_t}(s, a; \ell_t)}_{\text{A bandit problem on state s}}$$

Performance difference lemma

$$\leq \sum_s \mu^{\pi^\star}(s) \left(\frac{\log A}{\eta} + \eta \sum_{t=1}^{T} \sum_a \pi^\star(a|s) b_t(s, a) \right)$$

Mirror descent analysis $b_t(s, a) \approx \dfrac{c}{\mu^{\pi_t}(s, a)}$

$$= \tilde{O}\left(\frac{H}{\eta} \right) + \eta \sum_{t=1}^{T} V^{\pi^\star}(s_0; b_t)$$

$$\sum_s \mu^{\pi^\star}(s) \pi^\star(a|s) b_t(s, a) = V^{\pi^\star}(s_0; b_t)$$

$$\sum_{t=1}^{T}\left(V^{\pi_t}(s_0;\ell_t) - V^{\pi^\star}(s_0;\ell_t)\right) = \tilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \eta\underbrace{\sum_{t=1}^{T} V^{\pi^\star}(s_0;b_t)}$$

$$b_t(s,a) \approx \frac{c}{\mu^{\pi_t}(s,a)}$$

involves distribution mismatch coefficient $\kappa = \sup_{s,a,t} \frac{\mu^{\pi^\star}(s,a)}{\mu^{\pi_t}(s,a)}$ that is hard to handle

(so standard analysis of PO assumes that κ is bounded)

A simple trick to avoid this factor: using $\ell_t(s,a) - \eta b_t(s,a)$ as loss, instead of $\ell_t(s,a)$

$$\sum_{t=1}^{T}\left(V^{\pi_t}(s_0;\ell_t - \eta b_t) - V^{\pi^\star}(s_0;\ell_t - \eta b_t)\right) \lesssim \tilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \eta\sum_{t=1}^{T} V^{\pi^\star}(s_0;b_t)$$

assuming we can get the same bound for now

$$\Rightarrow \quad \sum_{t=1}^{T}\left(V^{\pi_t}(s_0;\ell_t) - V^{\pi^\star}(s_0;\ell_t)\right) \lesssim \tilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \eta\sum_{t=1}^{T} V^{\pi_t}(s_0;b_t)$$

rearranging

Change of measure: $V^{\pi^\star}(s_0;b_t) \longrightarrow V^{\pi_t}(s_0;b_t)$

(no longer involving distribution mismatch coefficient)

Standard bonus (e.g., UCBVI)

$$\bar{\ell}_t(s,a) - \frac{c}{\sqrt{n_t(s,a)}}$$

empirical mean of loss
in episode 1 to t

#visits to (s,a)
in episode 1 to t

Constructed from **Hoeffding's bound**

To compensate the **loss estimation error**

Find the optimal policy under the modified loss

Our bonus

$$\hat{\ell}_t(s,a) - \eta \frac{c}{\mu_t(s,a)}$$

biased loss estimator
in episode t

Prob { visiting (s,a) }
in episode t

Constructed from the **regret analysis of mirror descent**

To compensate the stability penalty
(≈ variance of the loss estimator)

Perform policy optimization over the modified loss

Dilated Bonus?

Recall that we made the following assumption in the previous derivation:

$$\sum_{t=1}^{T}\left(V^{\pi_t}(s_0;\ell_t) - V^{\pi^\star}(s_0;\ell_t)\right) \le \tilde{O}\left(\frac{H}{\eta}\right) + \eta\sum_{t=1}^{T} V^{\pi^\star}(s_0;b_t)$$

Really? Close, but not exactly.

$$\sum_{t=1}^{T}\left(V^{\pi_t}(s_0;\ell_t - \eta b_t) - V^{\pi^\star}(s_0;\ell_t - \eta b_t)\right) \lesssim \tilde{O}\left(\frac{H}{\eta}\right) + \eta\sum_{t=1}^{T} V^{\pi^\star}(s_0;b_t)$$

Our choice of bonus is slightly modified in order to resolve the above issue.
We call the modified bonus **dilated bonus.**

In fact, without modification, almost the same bounds (only slightly worse) can be achieved for tabular MDPs and linear MDPs.

Dealing with Linear Models

Linear-Q MDP: for any policy π, $Q^\pi(s, a; \ell_t)$ can be represented as $\phi(s,a)^\top w_t^\pi$ for some w_t^π (unknown to the learner)

Linear MDP: $\ell_t(s,a) = \phi(s,a)^\top \theta_t$ and $p(s'|s,a) = \phi(s,a)^\top \nu(s')$ for some θ_t and $\nu(\cdot)$ (both unknown to the learner)

Bonus in LSVI-UCB (Jin et al.)

$$\|\phi(s,a)\|_{\Lambda_t^{-1}}$$

$$\Lambda_t = \lambda I + \textstyle\sum_{\tau=1}^{t-1} \phi(s_\tau, a_\tau)\phi(s_\tau, a_\tau)^\top$$

Our bonus

$$\eta \|\phi(s,a)\|^2_{\Sigma_t^{-1}}$$

$$\Sigma_t = \lambda' I + \mathbb{E}\left[\phi(s,a)\phi(s,a)^\top \,\middle|\, (s,a) \sim \pi_t\right]$$

Summary

- A new general way of constructing exploration bonuses for policy optimization (suitable for adversarial loss + function approximation + bandit feedback)

- Applications to several settings:

Tabular MDP	Linear-Q MDP + simulator	Linear MDP + exploratory policy	Linear MDP
regret $= \tilde{\mathcal{O}}(\sqrt{T})$	regret $= \tilde{\mathcal{O}}(T^{2/3})$	regret $= \tilde{\mathcal{O}}(T^{6/7})$	regret $= \tilde{\mathcal{O}}(T^{14/15})$
improving Efroni et al.'s $\tilde{\mathcal{O}}(T^{2/3})$ bound	matching Neu & Olkhovskaya's, but removing their requirement of an exploratory policy	first sublinear regret	first sublinear regret (only appearing in our arxiv version)