Optimality and Approximation with Policy Gradient Methods

Alekh Agarwal

Microsoft Research AI

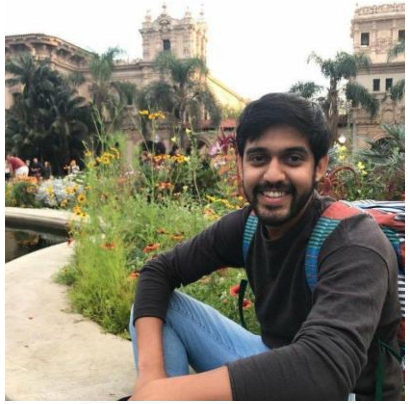
Collaborators

Sham Kakade

UW

Jason Lee

USC -> Princeton



Gaurav Mahajan

UCSD

Policy gradient methods in RL

- Widely used in **practice**
  - Directly optimize quantity of interest
  - Easily handle continuous and discrete states and actions
  - Apply to any differential policy parametrization

- Coarse-grained understanding in **theory**
  - Converge to a stationary point under sufficient smoothness

Can we sharpen our understanding of when and how well do policy gradient methods work?

Questions of interest

- When do policy gradient methods find a **globally optimal policy** with tabular parameterizations?

- What is the effect of **function approximation** on these guarantees?

- How does using **finitely many samples** effect convergence?

Main challenges

- The underlying maximization problem is typically **non-concave**
- Poor **exploration** leads to bad stationary points
- Role of **function approximation** tricky to quantify

Outline of the talk

- Policy gradient preliminaries
- Convergence in tabular settings
- Guarantees for restricted parameterizations

Outline of the talk

- Policy gradient preliminaries
- Convergence in tabular settings
- Guarantees for restricted parameterizations

MDP Preliminaries

- Discounted Markov Decision Process (S, A, r, P, γ)
- Policy $\pi : S \to \Delta(A)$

- State distribution of a policy π

$$d_{s_0}^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^{\pi}(s_t = s \,|s_0)$$

- Value functions of a policy π

$$V^{\pi}(s_0) = E_{s,a\sim d_{s_0}^{\pi}}\,[r(s,a)] \text{ and } Q^{\pi}(s,a) = E[r(s,a) + \gamma V^{\pi}(s') \mid s,a]$$

Policy parameterizations

- Policy class $\Pi = \{\pi_\theta : \theta \in \Theta\}$.
- Policy optimization: $\max_{\pi \in \Pi} \left[V^\pi(\rho) = E_{s\sim\rho}[V^\pi(s)] \right]$
- Example: Softmax parameterization

$$\Theta = R^{SA} \text{ and } \pi_\theta(a|s) \propto \exp(\theta_{s,a})$$

One parameter per state action, always contains optimal policy

- In general, Π need not contain the best unconstrained policy

Policy gradient algorithm

- Given a distribution μ over states
  - Can be different from ρ for better exploration

- First-order updates on value of policy

V^{π} with $\pi = \pi_{\theta_t}$

$$\theta_{t+1} = \theta_t + \eta \nabla \mathrm{V}^{(\mathrm{t})}(\mu)$$

- Policy gradient theorem [Williams '92, Sutton et al., '99]

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{s,a \sim d_{\mu}^{\pi_\theta}} \left[\nabla_\theta \log \pi_\theta(a|s) \; Q^{\pi_\theta}(s,a)\right]$$

- Can be estimated using trajectories from π_θ

Policy gradient example: Softmax parameterization

- Advantage function of π

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - E_{a\sim\pi(\cdot|s)}[Q^{\pi}(s,a)]$$

- Policy gradients (PG) for softmax:

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a)$$

Favor actions with a large advantage

Stationary if better actions are not explored

Outline of the talk

- Policy gradient preliminaries
- Convergence in tabular settings
- Guarantees for restricted parameterizations

Convergence of policy gradients for softmax

Theorem

Suppose the initial distribution μ satisfies $\mu(s) > 0$ for all $s \in S$. Using $\eta \leq \frac{(1-\gamma)^2}{5}$, we have for all states s:

$$V^{(t)}(s) \rightarrow V^{\star}(s) \text{ as } t \rightarrow \infty$$

- 👍 Converges as all states, actions have non-zero probability under softmax
- 👎 Can be slow as optimal policy is deterministic, θ grow to ∞

Entropy regularization

- Vanilla policy gradient slow to converge when probabilities are small

- Entropy regularization:

$$\max_{\theta \in R^{SA}} \left[L_\lambda(\theta) := V^{\pi_\theta}(\mu) - \frac{\lambda}{S} \sum_s \mathrm{KL}\big(\mathrm{Unif}, \pi_\theta(\cdot \mid s)\big) \right]$$

Entropy regularized PG

- Vanilla policy gradient slow to converge when probabilities are small
- Entropy regularization:

$$\max_{\theta \in R^{SA}} \left[L_\lambda(\theta) := V^{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a|s) \right]$$

- Different from more commonly used entropy of π
- Entropy regularized PG updates

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta L_\lambda(\theta_t)$$

Convergence of Entropy regularized PG

Distribution mismatch ratio:

$$M(\pi, \rho; \mu) = \max_{s \in S} \frac{d_\rho^\pi(s)}{\mu(s)}$$

Theorem

For appropriate choices of λ, η and for any state distribution ρ we have

$$\min_{t<T} V^\star(\rho) - V^{(t)}(\rho) = O\left(\frac{SA}{(1-\gamma)^3} \frac{M(\pi^\star, \rho; \mu)}{\sqrt{T}}\right)$$

Convergence of Entropy regularized PG

Theorem

For appropriate choices of λ, η and for any state distribution ρ we have

$$\min_{t<T} V^\star(\rho) - V^{(t)}(\rho) = O\left(\frac{SA}{(1-\gamma)^3} \frac{M(\pi^\star, \rho; \mu)}{\sqrt{T}}\right)$$

- $\mathrm{poly}\left(S, A, \frac{1}{1-\gamma}, \frac{1}{\epsilon}\right)$ convergence when distribution mismatch is small
- Counterexamples without dependence on $M(\pi^\star, \rho; \mu)$
- Exploration matters in PG even with exact gradients

Can we do better?

Algorithm	Iteration complexity
PG for softmax	Asymptotic
Entropy-regularized PG for softmax	$O\left(\frac{S^2A^2}{(1-\gamma)^6\epsilon^2}M(\pi^\star,\rho;\mu)^2\right)$

- Policy gradients (PG) for softmax:

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s,a)$$

- Distribution mismatch arises as PG depends on probability of visiting s under π

A natural solution

- Let us consider the Natural Policy Gradient algorithm [Kakade, 2001]
  - Uses Fisher information based preconditioner

- Simple form for softmax parameterization:

$$\theta_{t+1} = \theta_t + \frac{\eta}{1-\gamma} A^{(t)} \text{ and } \pi_{t+1}(a|s) \propto \pi_t(a|s) \exp(\eta A^{(t)})$$

- Updates do not depend on $d^{\pi_\theta}(s)$

- Like multiplicative weights, but in a non-concave maximization setting

Convergence of Natural Policy Gradients

Theorem

Using $\mu = \rho$ and $\theta_0 = 0$, setting $\eta = (1-\gamma)^2 \log A$, for all t we have

$$V^\star(\rho) - V^{(t)}(\rho) \le \frac{2}{(1-\gamma)^2\, t}$$

- Dimension free convergence, no dependence on S, A
- No dependence on distribution mismatch coefficient
- Similar results for approximate policy iteration in Even-Dar et al., [2009] and [Geist et al. [2019]

Proof ideas

- Performance difference lemma:

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} E_{s \sim d_{s_0}^{\pi}(s)} E_{a \sim \pi(\cdot|s)} \left[A^{\pi'}(s,a) \right]$$

- Linearize regret using above lemma instead of concavity
- Yields $\frac{1}{\sqrt{t}}$ rate almost immediately by multiplicative weights analysis
- Lower bound per-step improvement for fast rate

Recap so far

Algorithm	Iteration complexity
PG for softmax	Asymptotic
Entropy-regularized PG for softmax	$O\left(\frac{S^2A^2}{(1-\gamma)^6\epsilon^2}M(\pi^\star,\rho;\mu)^2\right)$
NPG for softmax	$O\left(\frac{1}{(1-\gamma)^2T}\right)$

We now study NPG with restricted policy parameterizations which need not contain the optimal policy

Outline of the talk

- Policy gradient preliminaries
- Convergence in tabular settings
- Guarantees for restricted parameterizations

Restricted parameterizations

- Policy class $\Pi = \{\pi_\theta : \theta \in \Theta\}$

- Want a policy $\pi \in \Pi$ to minimize

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\rho) - V^{\pi}(\rho)$$

- Example (linear softmax):

$$\pi_\theta(a|s) \propto \exp\left(\theta^T \phi_{s,a}\right) \quad \phi_{s,a} \in R^d \text{ for } d \ll SA$$

A closer look at Natural Policy Gradient

- NPG performs the update:

$$F(\theta) = E_{s,a\sim\pi_\theta}[g_\theta(s,a)g_\theta(s,a)^T] \text{ where } g_\theta(s,a) = \nabla_\theta \log \pi_\theta(a|s)$$

$$\theta_{t+1} = \theta_t + \eta F(\theta_t)^\dagger \nabla_\theta V^{(t)}$$

Compatible function approximation loss [Sutton et al., 99]

- Ordinary least squares solution under the loss:

$$L(w;\theta) = E_{s,a\sim\pi_\theta}[(A^{\pi_\theta}(s,a) - w \cdot g_\theta(s,a))^2]$$

- Example for linear softmax:

$$L(w;\theta) = E_{s,a\sim\pi_\theta}\left[\left(A^{\pi_\theta}(s,a) - w \cdot \phi_{s,a}\right)^2\right]$$

A natural update rule

- Pick any $w_t \in \mathrm{argmin}_w\, L(w; \theta_t)$

- Update $\theta_{t+1} = \theta_t + \eta w_t$

- Similar to Natural Actor Critic [Peters and Schaal, 2008]

Assumptions on policies

Policy Smoothness: $\nabla_\theta \log \pi_\theta(a|s)$ is β Lipschitz continuous for all s, a

Bounded updates: $||w_t|| \leq W$ for all iterations t

Bounded approximation error: $L(w_t; \theta_t) \leq \epsilon_{\text{apx}}$ for all iterations t

Minimum action probabilities: $\mu(a|s) \geq p_{\min}$ for all s, a

Convergence of NPG for smooth policies

- Let $\theta^\star = \operatorname*{argmax}_{\theta \in \Theta} V^{\pi_\theta}(\rho)$. Set $\eta = \sqrt{2 \log A \;/\beta W^2 T}$

Theorem

$$\min_{t<T} V^{\pi_{\theta^\star}}(\mu) - V^{(t)}(\mu)$$

$$\leq \frac{W\sqrt{2\beta \log A}}{1-\gamma} \frac{1}{\sqrt{T}} + \sqrt{\frac{M(\pi_{\theta^\star}, \rho; \mu)}{(1-\gamma)^3 p_{\min}} \epsilon_{\text{apx}}}$$

Convergence of NPG for smooth policies

Theorem

$$\mathrm{Regret}(\mu, T) \le \frac{W\sqrt{2\beta \log A}}{1-\gamma} \frac{1}{\sqrt{T}} + \sqrt{\frac{M(\pi_{\theta^\star}, \rho; \mu)}{(1-\gamma)^3 p_{\min}} \epsilon_{\mathrm{apx}}}$$

- Slower rate than the tabular case

Convergence of NPG for smooth policies

Theorem

$$\mathrm{Regret}(\mu, T) \leq \frac{W\sqrt{2\beta \log A}}{1-\gamma} \frac{1}{\sqrt{T}} + \sqrt{\frac{M(\pi_{\theta^\star}, \rho; \mu)}{(1-\gamma)^3 p_{\mathrm{min}}} \epsilon_{\mathrm{apx}}}$$

- Slower rate than the tabular case
- Distribution mismatch coefficient strikes back

Convergence of NPG for smooth policies

Theorem

$$\mathrm{Regret}(\mu, T) \leq \frac{W\sqrt{2\beta \log A}}{1-\gamma}\frac{1}{\sqrt{T}} + \sqrt{\frac{M(\pi_{\theta^\star}, \rho; \mu)}{(1-\gamma)^3 p_{\min}}\epsilon_{\mathrm{apx}}}$$

- Slower rate than the tabular case
- Distribution mismatch coefficient strikes back
- Effect of function approximation captured using min compatible function approximation loss

Extension to finite samples

- Approximately minimize $L(w; \theta_t)$ using samples
- Easy to obtain unbiased gradients
- Regret in loss minimization adds to ϵ_{apx}
- We show convergence guarantees using averaged SGD

Summary and other results

- Finite-time convergence analysis of policy gradient methods

- Distribution mismatch coefficient captures role of exploration
  - Assumption on algorithm, but not MDP dynamics

- Also analyze some projected policy gradient methods in the paper
  - E.g.: $\pi_\theta(a|s) = \theta_{s,a}$ as long as parameters lie in the simplex

- Characterize relevant notions of policy class expressivity

Looking ahead

- Empirical validation of theoretical prescriptions
  - KL vs. reverse KL, Actor-critic vs. Natural actor critic,…
- How do variance reduction techniques help?
- Sharper problem-dependent quantities instead of distribution mismatch coefficient
- Design of good exploratory distributions μ

Thank You!

http://arxiv.org/abs/1908.00261

Theorem

Let $\beta_\lambda = \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{S}$. Starting from any θ_0, using $\lambda = \frac{\epsilon(1-\gamma)}{2M(\pi^\star,\rho;\mu)}$ and $\eta = \frac{1}{\beta_\lambda}$, for any state distribution ρ we have

$$\min_{t<T} V^\star(\rho) - V^{(t)}(\rho) \leq \epsilon \quad \text{whenever} \quad T \geq \frac{320\, S^2 A^2}{(1-\gamma)^6 \epsilon^2} M(\pi^\star, \rho; \mu)^2$$