Optimality

Presented by :
Atmani Hanan

University mohammed vi polytechnic

July 9

Presentation plan

General introduction and motivation

- In this section, we seek to understand the convergence properties of policy gradient methods
- In this part, we will use the softmax policy class with the parameters θ defined without constraints. Unconstrained optimization can be employed, noting that we are dealing with a non-convex (concave) optimization case. are also interested in the good performance depending on the choice of ρ. Therefore, we will slightly modify the problem 'P1' to:

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\mu) \qquad \text{(P2)}$$

  with μ being any distribution defined on the states
  we will see later how the choice of ρ influences convergence
- Another challenge is that the optimal (deterministic) policy is reached when $\theta \to \infty$.

Vanishing Gradients and Saddle Points

- To understand the necessity of optimizing under a distribution μ, we will consider the following MDP example.

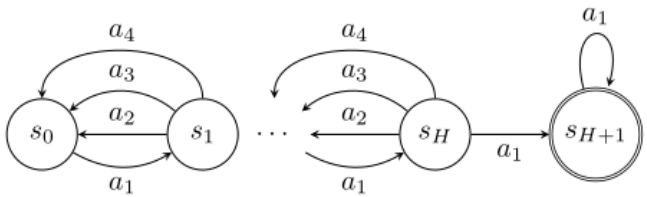

Figure: (Vanishing gradient example) A deterministic chain MDP of length $H + 2$. We consider a policy where $\pi(a \mid s_i) = \theta_{s_i,a}$ for $i = 1, 2, \ldots, H$. The rewards are zero everywhere else, except at $r(s_{H+1}, a_1) = 1$.

- in this example the agent is only rewarded upon viting some small set of states, then the politiy that does not visit any rewarding states will have zero gradient even if it is not optimal

Vanishing Gradients and Saddle Points

Proposition

Consider the example of the preceding MDP with $H+2$ states, $\sigma = \frac{H}{1+H}$ with direct policy parametrezation supppose $0 < \theta < 1$ and $\theta_{s,a_1} < \frac{1}{4}, \quad \forall s \in S$. Then $\forall k \leq \frac{H}{40 \log(2H)} - 1$, we have $\left\| \nabla_\theta^k V^{\pi_\theta}(s_0) \right\| \leq (1/3)^{H/4}$, where $\nabla_\theta^k V^{\pi_\theta}(s_0)$ is a tensor of the k_{th} order derivatives of $V^{\pi_\theta}(s_0)$ and $V^\star(s_0) - V^{\pi_\theta}(s_0) \geq (H+1)/8 - (H+1)^2/3^H$.

- This property explains that if we choose a strategy in which the agent does not sufficiently explore the environment, it will negatively impact the gradient algorithm. The gradient becomes very small not because we are close to optimal, but rather because the strategy visits advantageous states.

Policy Gradient Ascent

Proposition

For the softmax policy class:

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

- **Proof:** (Indecation)
  Observe that:

$$\frac{\partial \log \pi_\theta(a \mid s)}{\partial \theta_{s',a'}} = \mathbf{1}\left[s = s'\right]\left(\mathbf{1}\left[a = a'\right] - \pi_\theta\left(a' \mid s\right)\right)$$

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s',a'}} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_\theta}(s,a) \frac{\partial \log \pi_\theta(a \mid s)}{\partial \theta_{s',a'}} \right]$$

- **Remark:** An issue arises: for any policy π that becomes almost deterministic, its gradient tends to 0 even if it's not

Policy Gradient Ascent

The update rule for gradient ascent is:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_{\theta_t}}(\mu) \tag{1}$$

Theorem

Assume we follow the gradient ascent update rule as specified above (1) and $\forall s \in S,\ \mu(s) > 0$. *Suppose* $\eta \leq \frac{(1-\sigma)^3}{8}$ *then we have that for all states* s, $V^{\pi_{\theta_t}}(s) \to V^\star(s)$ *as* $t \to \infty$

Remark: If the condition $\forall s \in S,\ \mu(s) > 0$ is not satisfied, then we can find $s \in S$ such that $\mu(s) = 0$, and thus $\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s',a'}} \to 0$.

- The exponential nature of the Softmax expression quickly leads to deterministic policies, making the optimization problem very challenging.
- The convergence rate of this algorithm is exponential (very slow) in the number of states. To avoid this issue, we will adopt a regularization approach to ensure polynomial convergence in the number of states.

Log Barrier Regularization

Definition

Relative_entropy for distribution p and q is defined as:

$$\mathrm{KL}(p, q) := \mathbb{E}_{x\sim p}[-\log q(x)/p(x)]$$

Denote the uniform distribution over a set $\mathcal{X}$ by $\mathrm{Unif}_{\mathcal{X}}$, and define the following log barrier regularized objective as:

$$\begin{aligned} L_\lambda(\theta) &:= V^{\pi_\theta}(\mu) - \lambda \mathbb{E}_{s\sim \ \mathit{Unif}\ \mathcal{S}}\left[\mathrm{KL}\left(\mathrm{Unif}_{\mathcal{A}}, \pi_\theta(\cdot \mid s)\right)\right] \\ &= V^{\pi_\theta}(\mu) + \frac{\lambda}{|\mathcal{S}||\mathcal{A}|}\sum_{s,a}\log \pi_\theta(a \mid s) + \lambda \log|\mathcal{A}| \end{aligned}$$

where λ is a regularization parameter

Log Barrier Regularization

- The policy gradient ascent updates for $L_\lambda(\theta)$ are given by:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta L_\lambda \left(\theta^{(t)} \right). \tag{2}$$

Theorem

Suppose θ is such that:

$$\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt}$$

and $\epsilon_{opt} \leq \lambda/(2|\mathcal{S}||\mathcal{A}|)$. Then we have that for all starting state distributions ρ :

$$V^{\pi_\theta}(\rho) \geq V^\star(\rho) - \frac{2\lambda}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty$$

Log Barrier Regularization

- **Proof:** Let's show that $\max_a A^{\pi_\theta}(s,a) \leq \frac{2\lambda}{\mu(s)|\mathcal{S}|}$ for all states. Next, we use a lemma from the first chapter:

$$V^\star(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^\star}(s)\pi^\star(a \mid s) A^{\pi_\theta}(s,a)$$

- **Remark:** The importance of choosing the measure μ.

Iteration complexity with log barrier regularization

Corollary

Let $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. *Starting from any initial* $\theta^{(0)}$, *consider the updates 0.6) with* $\lambda = \frac{\epsilon(1-\gamma)}{2\left\| \frac{\|\pi^*_\beta}{\mu} \right\|_\infty}$ *and* $\eta = 1/\beta_\lambda$. *Then for all starting state distributions* ρ, *we have*

$$\min_{t<T} \{V^\star(\rho) - V^{\pi_{\theta_t}}(\rho)\} \leq \epsilon \quad \textit{whenever} \quad T \geq \frac{320|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\epsilon^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty^2 .$$

- This corollary emphasizes the importance of choosing λ carefully for the precision of ϵ, and μ to ensure global optimality.
- We can consider other regularization examples such as entropy vs log barrier regularization.

The Natural Policy Gradient NPG

- Observe that a policy constitutes a family of probability distributions $\{\pi_\theta(./s) \mid s \in S\}$

Definition

The Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $\mathbb{E}_{x\sim p_\theta}\left[\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top\right]$. Now we let us define $\mathcal{F}_\rho^\theta$ as an (average) Fisher information matrix on the family of distributions $\{\pi_\theta(\cdot \mid s) \mid s \in \mathcal{S}\}$ as follows:

$$\mathcal{F}_\rho^\theta := \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[(\nabla \log \pi_\theta(a \mid s))\, \nabla \log \pi_\theta(a \mid s)^\top\right].$$

- Note that the average is under the state-action visitation frequencies

The Natural Policy Gradient NPG

- The NPG algorithm performs gradient updates in the geometry induced by this matrix as follows:

$$\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho \left(\theta^{(t)}\right)^\dagger \nabla_\theta V^{\pi_{\theta_t}}(\rho) \tag{3}$$

  where $M^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix M.
- We restrict to using the initial state distribution ρ and we restrict attention to states reachable from ρ

Softmax NPG as soft policy iteration

Lemma

For the softmax parameterization, the NPG updates take the form:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{\pi_{\theta_t}} + \eta v$$

and

$$\pi^{(t+1)}(a \mid s) = \pi^{(t)}(a \mid s) \frac{\exp\left(\eta A^{\pi_{\theta_t}}(s,a)/(1-\gamma)\right)}{Z_t(s)},$$

where $Z_t(s) = \sum_{a \in \mathcal{A}} \pi^{(t)}(a \mid s) \exp\left(\eta A^{\pi_{\theta_t}}(s,a)/(1-\gamma)\right)$

- The update rule NGP does not depend on either ρ or $d_\rho^{\pi_{\theta_t}}$.
- The Fisher information nullifies the impact of the initial distribution (stabilizes the algorithm).

Global convergence for NPG

Theorem

Suppose we run the NPG updates using $\rho \in \Delta(\mathcal{S})$ *and with* $\theta^{(0)} = 0$. *Fix* $\eta > 0$. *For all* $T > 0$, *we have:*

$$V^{(T)}(\rho) \geq V^*(\rho) - \frac{\log|\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}.$$

- Le taux de convergance ne depand pas de ρ
- Now setting $\eta \geq (1-\gamma)^2 \log|\mathcal{A}|$, we see that NPG finds an ϵ-optimal policy in a number of iterations that is at most:

$$T \leq \frac{2}{(1-\gamma)^2 \epsilon},$$

- The number of iterations required for convergence does not depend on the number of states or actions.

Improvement lower bound for NPG

For the iterates $\pi^{(t)}$ generated by the NPG updates (3) , we have for all starting state distributions μ

$$V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) \geq \frac{(1-\gamma)}{\eta}\mathbb{E}_{s\sim\mu} \log Z_t(s) \geq 0.$$