Policy Gradient: Optimality

Wen Sun

CS 6789: Foundations of Reinforcement Learning

Recap: The Natural Policy Gradient

- Recall that the Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $E_{x\sim p_\theta}\left[\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top\right]$

Recap: The Natural Policy Gradient

- Recall that the Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $E_{x\sim p_\theta}\left[\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top\right]$
- Define $\mathscr{F}_\rho^\theta$ as the (average) Fisher matrix on the family of distributions $\{\pi_\theta(\,\cdot\,|\,s)\,|\,s \in S\}$ as: $\mathscr{F}_\rho^\theta := E_{s\sim d_\rho^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[(\nabla \log \pi_\theta(a\,|\,s))\nabla \log \pi_\theta(a\,|\,s)^\top\right]$.

Recap: The Natural Policy Gradient

- Recall that the Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $E_{x\sim p_\theta}\left[\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top\right]$
- Define $\mathscr{F}_\rho^\theta$ as the (average) Fisher matrix on the family of distributions $\{\pi_\theta(\cdot \mid s) \mid s \in S\}$ as: $\mathscr{F}_\rho^\theta := E_{s\sim d_\rho^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[(\nabla \log \pi_\theta(a \mid s))\nabla \log \pi_\theta(a \mid s)^\top\right]$.

- The NPG algorithm performs gradient updates in this induced geometry: $\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)})^\dagger \nabla_\theta V^{(t)}(\rho)$,

  where $M^\dagger$ denotes the Moore-Penrose pseudoinverse of M.

Recap: The Natural Policy Gradient

- Recall that the Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $E_{x\sim p_\theta}\left[\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top\right]$
- Define $\mathscr{F}_\rho^\theta$ as the (average) Fisher matrix on the family of distributions $\{\pi_\theta(\,\cdot\,|\,s)\,|\,s \in S\}$ as:
  $\mathscr{F}_\rho^\theta := E_{s\sim d_\rho^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[(\nabla \log \pi_\theta(a\,|\,s))\nabla \log \pi_\theta(a\,|\,s)^\top\right]$.

- The NPG algorithm performs gradient updates in this induced geometry:
  $\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)})^\dagger \nabla_\theta V^{(t)}(\rho),$
  where $M^\dagger$ denotes the Moore-Penrose pseudoinverse of M.

- Idea:
  - Travel faster and faster when approaching to the corners of the simplex (as opposed to the log-barrier which keeps us away)

Today:

Global Convergence

Another interpretation of NPG

- Let $w^\star$ denote the following minimizer of the least square problem:

$$w^\star \in \text{argmin}_w E_{s\sim d_\mu^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[\left(A^{\pi_\theta}(s,a) - w\cdot\nabla_\theta \log \pi_\theta(a\,|\,s)\right)^2\right]$$

Another interpretation of NPG

- Let $w^\star$ denote the following minimizer of the least square problem:

  $w^\star \in \text{argmin}_w E_{s\sim d_\mu^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[\left(A^{\pi_\theta}(s,a) - w\cdot\nabla_\theta \log \pi_\theta(a\,|\,s)\right)^2\right]$

- Lemma: We have that $F_\mu(\theta)^\dagger \nabla_\theta V^\theta(\mu) = \frac{1}{1-\gamma} w^\star,$

  The NPG direction is the weights $w^\star$

Proof

- The first order optimality conditions for $w^\star$ imply

$$E_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\left(A^{\pi_\theta}(s,a) - w^\star \cdot \nabla_\theta \log \pi_\theta(a\,|\,s)\right)\nabla_\theta \log \pi_\theta(a\,|\,s)\right] = 0$$

Proof

- The first order optimality conditions for $w^\star$ imply

$$E_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\left(A^{\pi_\theta}(s,a) - w^\star \cdot \nabla_\theta \log \pi_\theta(a\,|\,s)\right)\nabla_\theta \log \pi_\theta(a\,|\,s)\right] = 0$$

- Rearranging

$$E_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_\theta}(s,a)\,\nabla_\theta \log \pi_\theta(a\,|\,s)\right] = E_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a\,|\,s)\,\nabla_\theta \log \pi_\theta(a\,|\,s)^\top\right] w^\star$$

Proof

- The first order optimality conditions for $w^\star$ imply

$$E_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\left(A^{\pi_\theta}(s,a) - w^\star \cdot \nabla_\theta \log \pi_\theta(a\,|\,s)\right)\nabla_\theta \log \pi_\theta(a\,|\,s)\right] = 0$$

- Rearranging

$$E_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_\theta}(s,a)\,\nabla_\theta \log \pi_\theta(a\,|\,s)\right] = E_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a\,|\,s)\,\nabla_\theta \log \pi_\theta(a\,|\,s)^\top\right] w^\star$$

- By the definition of $\nabla_\theta V^\theta(\mu)$ and $F_\mu(\theta)$:

$$(1-\gamma)\,\nabla_\theta V^\theta(\mu) = F_\mu(\theta) w^\star$$

Softmax Case:
NPG and Global Convergence to Opt

NPG softmax case

(NPG as “soft” policy iteration)

- Lemma: (Softmax NPG as soft policy iteration) The NPG update is:

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

NPG softmax case

(NPG as “soft” policy iteration)

- Lemma: (Softmax NPG as soft policy iteration) The NPG update is:

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

- and this leads to the update:

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

where $Z_t(s) = \sum_a \pi^{(t)}(a \mid s) \exp\left(\eta A^{(t)}(s,a)/(1-\gamma)\right)$.

Proof

- Lemma: The NPG update is:
  $$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$

Proof

- Lemma: The NPG update is:
  $$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$

- Proof: Recall NPG update is $\frac{1}{1-\gamma} w^\star$ where
  $$w^\star \in \text{argmin}_w E_{s\sim d_\mu^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)} \left[\left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a \mid s) \right)^2 \right]$$

Proof

- Lemma: The NPG update is:
  $$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$
- Proof: Recall NPG update is $\frac{1}{1-\gamma} w^\star$ where
  $$w^\star \in \text{argmin}_w E_{s\sim d_\mu^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)} \left[\left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a \mid s) \right)^2 \right]$$
- What is a minimizer for the the softmax?

Global convergence for NPG

- Theorem: Params: $\theta^{(0)} = 0$ and $\eta > 0$. For all ρ and $T > 0$, we have:

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

Global convergence for NPG

- Theorem: Params: $\theta^{(0)} = 0$ and $\eta > 0$. For all ρ and $T > 0$, we have:

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

- Setting $\eta \geq (1-\gamma)^2 \log A$, NPG finds an ϵ-opt policy when $T \geq \frac{2}{(1-\gamma)^2 \epsilon}$.

Global convergence for NPG

- Theorem: Params: $\theta^{(0)} = 0$ and $\eta > 0$. For all ρ and $T > 0$, we have:

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

- Setting $\eta \geq (1-\gamma)^2 \log A$, NPG finds an ϵ-opt policy when $T \geq \frac{2}{(1-\gamma)^2 \epsilon}$.
- Iteration complexity has:
  - No dimension dependence (no dependence on S, A)
  - No dependence on start state measure ρ (and no “dist mismatch factor”)
  - No ‘flat gradient’ problem

Improvement Lower Bound

- Lemma: For the iterates $\pi^{(t)}$ generated by the NPG, we have for all distributions μ:

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{(1-\gamma)}{\eta} E_{s\sim\mu} \log Z_t(s) \geq 0.$$

Improvement Lower Bound

- Lemma: For the iterates $\pi^{(t)}$ generated by the NPG, we have for all distributions μ:

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) \geq \frac{(1-\gamma)}{\eta} E_{s\sim\mu} \log Z_t(s) \geq 0.$$

- Proof: First, let us show that $\log Z_t(s) \geq 0$. To see this, observe:

$$\begin{aligned}
\log Z_t(s) &= \log \sum_a \pi^{(t)}(a \,|\, s) \exp(\eta A^{(t)}(s,a)/(1-\gamma)) \\
&\geq \sum_a \pi^{(t)}(a \,|\, s) \log \exp(\eta A^{(t)}(s,a)/(1-\gamma)) \\
&= \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a \,|\, s) A^{(t)}(s,a) = 0.
\end{aligned}$$

(using Jensen's inequality on the concave function $\log x$.)

Lemma Proof: continued….

By the performance difference lemma,

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) = \frac{1}{1-\gamma} E_{s\sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a \mid s) A^{(t)}(s,a)$$

$$= \frac{1}{\eta} E_{s\sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a \mid s) \log \frac{\pi^{(t+1)}(a \mid s) Z_t(s)}{\pi^{(t)}(a \mid s)}$$

$$= \frac{1}{\eta} E_{s\sim d_\mu^{(t+1)}} \mathrm{KL}(\pi_s^{(t+1)} \,||\, \pi_s^{(t)}) + \frac{1}{\eta} E_{s\sim d_\mu^{(t+1)}} \log Z_t(s)$$

$$\geq \frac{1}{\eta} E_{s\sim d_\mu^{(t+1)}} \log Z_t(s) \geq \frac{1-\gamma}{\eta} E_{s\sim \mu} \log Z_t(s),$$

where the last step uses that $d_\mu^{(t+1)} \geq (1-\gamma)\mu$ and that $\log Z_t(s) \geq 0$.

NPG Conv. Proof, Part 1

- $d^\star$ as shorthand for $d_\rho^\star$; π_s as shorthand for the vector of $\pi(\,\cdot\,|\,s)$

NPG Conv. Proof, Part 1

- $d^\star$ as shorthand for $d_\rho^\star$; π_s as shorthand for the vector of $\pi(\cdot \mid s)$
- By the performance difference lemma,

$$V^{\pi^\star}(\rho) - V^{(t)}(\rho) = \frac{1}{1-\gamma} E_{s\sim d^\star} \sum_a \pi^\star(a \mid s) A^{(t)}(s,a)$$

$$= \frac{1}{\eta} E_{s\sim d^\star} \sum_a \pi^\star(a \mid s) \log \frac{\pi^{(t+1)}(a \mid s) Z_t(s)}{\pi^{(t)}(a \mid s)}$$

$$= \frac{1}{\eta} E_{s\sim d^\star} \left(\mathsf{KL}(\pi_s^\star \,||\, \pi_s^{(t)}) - \mathsf{KL}(\pi_s^\star \,||\, \pi_s^{(t+1)}) + \sum_a \pi^\star(a \mid s) \log Z_t(s) \right)$$

$$= \frac{1}{\eta} E_{s\sim d^\star} \left(\mathsf{KL}(\pi_s^\star \,||\, \pi_s^{(t)}) - \mathsf{KL}(\pi_s^\star \,||\, \pi_s^{(t+1)}) + \log Z_t(s) \right),$$

NPG Conv. Proof, Part 2

- By the improvement lemma $V^{(t+1)}(\rho) \geq V^{(t)}(\rho)$. Hence,

$$V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) \leq \frac{1}{T}\sum_{t=0}^{T-1}(V^{\pi^\star}(\rho) - V^{(t)}(\rho))$$

$$= \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star}(\mathsf{KL}(\pi_s^\star \,||\, \pi_s^{(t)}) - \mathsf{KL}(\pi_s^\star \,||\, \pi_s^{(t+1)})) + \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star} \log Z_t(s)$$

$$\leq \frac{E_{s\sim d^\star}\mathsf{KL}(\pi_s^\star \,||\, \pi^{(0)})}{\eta T} + \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star} \log Z_t(s)\,.$$

NPG Conv. Proof, Part 2

- By the improvement lemma $V^{(t+1)}(\rho) \geq V^{(t)}(\rho)$. Hence,

$$V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) \leq \frac{1}{T}\sum_{t=0}^{T-1}(V^{\pi^\star}(\rho) - V^{(t)}(\rho))$$

$$= \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star}(\mathrm{KL}(\pi_s^\star \,||\, \pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star \,||\, \pi_s^{(t+1)})) + \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star}\log Z_t(s)$$

$$\leq \frac{E_{s\sim d^\star}\mathrm{KL}(\pi_s^\star \,||\, \pi^{(0)})}{\eta T} + \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star}\log Z_t(s)\,.$$

- By the improvement lemma (applied with $d^\star$ as the distribution), we have:

$$\frac{1}{\eta} E_{s\sim d^\star}\log Z_t(s) \leq \frac{1}{1-\gamma}\left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\right)$$

which gives us a bound on $E_{s\sim d^\star}\log Z_t(s)$.

NPG Conv. Proof, Part 3

$$
\begin{aligned}
V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) &\leq \frac{E_{s\sim d^\star}\mathsf{KL}(\pi_s^\star \,||\, \pi^{(0)})}{\eta T} + \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star} \log Z_t(s) \\
&\leq \frac{E_{s\sim d^\star}\mathsf{KL}(\pi_s^\star \,||\, \pi^{(0)})}{\eta T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1} \left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star) \right) \\
&= \frac{E_{s\sim d^\star}\mathsf{KL}(\pi_s^\star \,||\, \pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T} \\
&\leq \frac{\log A}{\eta T} + \frac{1}{(1-\gamma)^2 T}\,.
\end{aligned}
$$