

Policy Gradient: Optimality

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Summary/Today

- Do they PG methods globally converge to an optimal policy?

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

- Recap
- Today:
 - Wrap up Log Barrier Proof
 - Natural policy gradient

Recap

Things to remember

For all π, π', s_0 :

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

$$\nabla_\theta J(\theta) := \frac{1}{1 - \gamma} \mathbb{E}_{s, a \sim d^{\pi_\theta}} [\nabla_\theta \ln \pi_\theta(a | s) Q^{\pi_\theta}(s, a)]$$

Today: we will use $d_{s_0}^\pi$ for a state distribution measure.

(it should be clear from context how we use it).

$$d_{s_0}^\pi(s) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s | s_0, \pi)$$

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s, a_h = a | s_0, \pi)$$

$$V^\pi(\mu) = E_{s \sim \mu}[V^\pi(s)]$$

$$d_\mu^\pi(s) = E_{s_0 \sim \mu}[d_{s_0}^\pi(s)]$$

Recap

- Softmax policies with exact gradients:
 - Flat gradients could occur if we optimize $V^{\pi_{\theta}}(s_0)$
 - Coverage: considered optimizing $\max_{\theta \in \Theta} V^{\pi_{\theta}}(\mu)$
 - Convergence:
 - (i) asymptotic convergence for GD
 - (ii) poly rate with GD+log barrier regularization

Softmax Gradients

- $\pi_{\theta}(a | s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$,
- **Lemma:** For the softmax policy class, we have:

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

Stationarity and Optimality

Stationarity and Optimality

- Log barrier regularized objective:

$$L_\lambda(\theta) = V^{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a | s) + \lambda \log A$$

Stationarity and Optimality

- Log barrier regularized objective:

$$L_\lambda(\theta) = V^{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a | s) + \lambda \log A$$

- **Theorem:** (Log barrier regularization) Suppose θ is such that:

$$\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt} \text{ and } \epsilon_{opt} \leq \lambda/(2SA)$$

then we have for all starting state distributions ρ :

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

where the “distribution mismatch coefficient” is

$$\left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty = \max_s \left(\frac{d_\rho^{\pi^*}(s)}{\mu(s)} \right) \text{ (componentwise division notation)}$$

Global Convergence with the Log Barrier

- The smoothness of $L_\lambda(\theta)$ is $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{S}$

Global Convergence with the Log Barrier

- The smoothness of $L_\lambda(\theta)$ is $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{S}$
- **Corollary:** (Iteration complexity with log barrier regularization)
Set $\lambda = \frac{\epsilon(1-\gamma)}{2 \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty}$ and $\eta = 1/\beta_\lambda$. Starting from any initial $\theta^{(0)}$,

then for all starting state distributions ρ , we have

$$\min_{t < T} \{ V^*(\rho) - V^{(t)}(\rho) \} \leq \epsilon \quad \text{whenever} \quad T \geq c \frac{S^2 A^2}{(1-\gamma)^6 \epsilon^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty^2$$

(for constant c).

Wrapping up...

Proof, part 1

- The proof consists of showing that: $\max_a A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all states s .

Proof, part 1

- The proof consists of showing that: $\max_a A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all states s .

- To see that this is sufficient, observe that by the performance difference lemma:

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

$$\leq \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*}(s) \max_{a \in A} A^{\pi_\theta}(s, a)$$

$$\leq \frac{1}{1-\gamma} \sum_s 2d_\rho^{\pi^*}(s) \lambda/(\mu(s)S)$$

$$\leq \frac{2\lambda}{1-\gamma} \max_s \left(\frac{d_\rho^{\pi^*}(s)}{\mu(s)} \right).$$

which would then complete the proof.

Proof, part 2

- need to show $A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all (s, a) . consider (s, a) where that $A^{\pi_\theta}(s, a) \geq 0$ (else claim is true).

Proof, part 2

- need to show $A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all (s, a) . consider (s, a) where that $A^{\pi_\theta}(s, a) \geq 0$ (else claim is true).
- Recall
$$\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s, a) + \frac{\lambda}{S} \left(\frac{1}{A} - \pi_\theta(a|s) \right)$$

Proof, part 2

- need to show $A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all (s, a) . consider (s, a) where that $A^{\pi_\theta}(s, a) \geq 0$ (else claim is true).

- Recall
$$\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s, a) + \frac{\lambda}{S} \left(\frac{1}{A} - \pi_\theta(a|s) \right)$$

- Solving for $A^{\pi_\theta}(s, a)$ in the first step and using $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt} \leq \lambda/(2SA)$,

$$A^{\pi_\theta}(s, a) = \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \left(\frac{1}{\pi_\theta(a|s)} \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{S} \left(1 - \frac{1}{\pi_\theta(a|s)A} \right) \right)$$

$$\leq \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \left(\frac{1}{\pi_\theta(a|s)} \frac{\lambda}{2SA} + \frac{\lambda}{S} \right)$$

$$\leq \frac{1}{\mu(s)} \left(\frac{1}{\pi_\theta(a|s)} \frac{\lambda}{2SA} + \frac{\lambda}{S} \right)$$

using that $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\mu(s)$

Proof, part 2

- need to show $A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all (s, a) . consider (s, a) where that $A^{\pi_\theta}(s, a) \geq 0$ (else claim is true).

- Recall
$$\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s, a) + \frac{\lambda}{S} \left(\frac{1}{A} - \pi_\theta(a|s) \right)$$

- Solving for $A^{\pi_\theta}(s, a)$ in the first step and using $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt} \leq \lambda/(2SA)$,

$$\begin{aligned} A^{\pi_\theta}(s, a) &= \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \left(\frac{1}{\pi_\theta(a|s)} \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{S} \left(1 - \frac{1}{\pi_\theta(a|s)A} \right) \right) \\ &\leq \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \left(\frac{1}{\pi_\theta(a|s)} \frac{\lambda}{2SA} + \frac{\lambda}{S} \right) \\ &\leq \frac{1}{\mu(s)} \left(\frac{1}{\pi_\theta(a|s)} \frac{\lambda}{2SA} + \frac{\lambda}{S} \right) \quad \text{using that } d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\mu(s) \end{aligned}$$

- Suppose we could show that $\pi_\theta(a|s) \geq 1/(2A)$, when $A^{\pi_\theta}(s, a) \geq 0$, then

$$\frac{1}{\mu(s)} \left(\frac{1}{\pi_\theta(a|s)} \frac{\lambda}{2SA} + \frac{\lambda}{S} \right) \leq \frac{1}{\mu(s)} \left(2A \frac{\lambda}{2SA} + \frac{\lambda}{S} \right) = \frac{2\lambda}{\mu(s)S} \quad \text{and the proof is done!}$$

Proof, part 3

- for (s, a) such that $A^{\pi_\theta}(s, a) \geq 0$, we want show $\pi_\theta(a | s) \geq 1/(2A)$.

Proof, part 3

- for (s, a) such that $A^{\pi_\theta}(s, a) \geq 0$, we want show $\pi_\theta(a | s) \geq 1/(2A)$.

- The gradient norm assumption $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt}$ implies that:

$$\begin{aligned} \epsilon_{opt} &\geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a | s) A^{\pi_\theta}(s, a) + \frac{\lambda}{S} \left(\frac{1}{A} - \pi_\theta(a | s) \right) \\ &\geq 0 + \frac{\lambda}{S} \left(\frac{1}{A} - \pi_\theta(a | s) \right) \quad \text{using } A^{\pi_\theta}(s, a) \geq 0 \end{aligned}$$

Proof, part 3

- for (s, a) such that $A^{\pi_\theta}(s, a) \geq 0$, we want show $\pi_\theta(a | s) \geq 1/(2A)$.

- The gradient norm assumption $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt}$ implies that:

$$\begin{aligned}\epsilon_{opt} &\geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a | s) A^{\pi_\theta}(s, a) + \frac{\lambda}{S} \left(\frac{1}{A} - \pi_\theta(a | s) \right) \\ &\geq 0 + \frac{\lambda}{S} \left(\frac{1}{A} - \pi_\theta(a | s) \right) \quad \text{using } A^{\pi_\theta}(s, a) \geq 0\end{aligned}$$

- Rearranging and using our assumption $\epsilon_{opt} \leq \lambda/(2SA)$,

$$\pi_\theta(a | s) \geq \frac{1}{A} - \frac{\epsilon_{opt} S}{\lambda} \geq \frac{1}{2A}.$$

Today:

Natural Policy Gradient and Convergence

The Natural Policy Gradient

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

The Natural Policy Gradient

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

The Natural Policy Gradient

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

The Natural Policy Gradient

- Recall that the Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $E_{x \sim p_\theta} [\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top]$
- Define \mathcal{F}_ρ^θ as the (average) Fisher matrix on the family of distributions $\{\pi_\theta(\cdot | s) | s \in \mathcal{S}\}$ as:
$$\mathcal{F}_\rho^\theta := E_{s \sim d_\rho^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot | s)} [(\nabla \log \pi_\theta(a | s)) \nabla \log \pi_\theta(a | s)^\top] .$$
- The NPG algorithm performs gradient updates in this induced geometry:
$$\theta^{(t+1)} = \theta^{(t)} + \eta \mathcal{F}_\rho(\theta^{(t)})^\dagger \nabla_\theta V^{(t)}(\rho),$$

where M^\dagger denotes the Moore-Penrose pseudoinverse of M .
- Idea:
 - ‘stretch’ the corners of the simplex out to travel faster
(as opposed to the log-barrier which keeps us away)

“Compatible Function Approximation” (and NPG)

Compatible Function Approximation

- Let w^\star denote the following minimizer of the “compatible function approximation” error:

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

Compatible Function Approximation

- Let w^\star denote the following minimizer of the “compatible function approximation” error:

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

- Lemma:** Let $\widehat{A}^{\pi_\theta}(s, a)$ be the best linear predictor of $A^{\pi_\theta}(s, a)$ using $\nabla_\theta \log \pi_\theta(a|s)$, i.e. $\widehat{A}^{\pi_\theta}(s, a) := w^\star \cdot \nabla_\theta \log \pi_\theta(a|s)$. We have:

$$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} E_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \widehat{A}^{\pi_\theta}(s, a) \right]$$

We can use $\widehat{A}^{\pi_\theta}(s, a)$ instead of $A^{\pi_\theta}(s, a)$.

Proof

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

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

Proof

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

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

- Rearranging and using the definition of $\widehat{A}^{\pi_\theta}(s, a)$,

$$\begin{aligned} \nabla_\theta V^{\pi_\theta}(\mu) &= \frac{1}{1 - \gamma} E_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a | s) \right] \\ &= \frac{1}{1 - \gamma} E_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)} \left[\left(w^\star \cdot \nabla_\theta \log \pi_\theta(a | s) \right) \nabla_\theta \log \pi_\theta(a | s) \right] \\ &= \frac{1}{1 - \gamma} E_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)} \left[\widehat{A}^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a | s) \right] \end{aligned}$$

NPG & Compatible Function Approximation

- Let w^\star denote the following minimizer of the “compatible function approximation” error:

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

NPG & Compatible Function Approximation

- Let w^\star denote the following minimizer of the “compatible function approximation” error:

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

- 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:

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

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

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

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

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

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

- 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:

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

Global convergence for NPG

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

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

- 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:

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

- 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

Global convergence for NPG

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

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

- 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
- What about approx/estimation errors? (next lecture)

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,

$$\begin{aligned} V^{(t+1)}(\mu) - V^{(t)}(\mu) &= \frac{1}{1-\gamma} E_{s \sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) A^{(t)}(s, a) \\ &= \frac{1}{\eta} E_{s \sim d_\mu^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) \log \frac{\pi^{(t+1)}(a|s) Z_t(s)}{\pi^{(t)}(a|s)} \\ &= \frac{1}{\eta} E_{s \sim d_\mu^{(t+1)}} \text{KL}(\pi_s^{(t+1)} || \pi_s^{(t)}) + \frac{1}{\eta} E_{s \sim d_\mu^{(t+1)}} \log Z_t(s) \\ &\geq \frac{1}{\eta} E_{s \sim d_\mu^{(t+1)}} \log Z_t(s) \geq \frac{1-\gamma}{\eta} E_{s \sim \mu} \log Z_t(s), \end{aligned}$$

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_ρ^\star ; π_s as shorthand for the vector of $\pi(\cdot | s)$

NPG Conv. Proof, Part 1

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

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

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

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

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

NPG Conv. Proof, Part 2

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

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

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

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

NPG Conv. Proof, Part 2

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

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

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

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

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

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

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

NPG Conv. Proof, Part 3

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