

Policy Gradient: Optimality

Sham Kakade and Kianté Brantley

CS 2824: 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:
 - Landscape result & Exploration
 - Tabular case
 - Natural policy gradient

Recap++

Things to remember

$$J(\theta) = V^{\pi_\theta}(s_0)$$

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

$\mathbb{E} [\underbrace{\nabla_\theta \ln \pi_\theta(a | s)}_{\text{action}} \cdot f(s)] = 0$

Today: we use $d_{s_0}^\pi$ for a state distribution measure.
(for convenience, we overload notation)

for a distribution over states $s \sim \mu$, let:

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

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

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

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

Derivation of unbiased Stochastic Policy Gradient

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

Draw $h \propto \gamma^h$, **roll-in** π_{θ} to generate $s_h, a_h \sim \mathbb{P}_h^{\pi_{\theta}}$

Roll-out π_{θ} from (s_h, a_h) : terminate with prob $1 - \gamma$, $\widetilde{Q}^{\pi_{\theta}}(s_h, a_h) = \sum_{\tau=h}^{t \geq h} r_{\tau}$

Unbiased estimate: $\nabla_{\theta} \ln \pi_{\theta}(a_h | s_h) \widetilde{Q}^{\pi_{\theta}}(s_h, a_h)$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_{\theta}(a | s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (e.g., for linear MDPs):

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_{\theta}(a | s) = \frac{\exp(\theta^{\top} \phi(s, a))}{\sum_{a'} \exp(\theta^{\top} \phi(s, a'))}$$

3. Neural Policy:

Neural network $f_{\theta} : S \times A \mapsto \mathbb{R}$

$$\pi_{\theta}(a | s) = \frac{\exp(f_{\theta}(s, a))}{\sum_{a'} \exp(f_{\theta}(s, a'))}$$

PG as non-convex optimization

Convergence to Stationary Points of GD

$J(\pi_\theta)$ is non-convex

Convergence to Stationary Points of GD

$J(\pi_\theta)$ is non-convex

- Def of a β -smooth function F :

$$\|\nabla_\theta F(\theta) - \nabla_\theta F(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$$

which implies:

$$\left| F(\theta) - F(\theta_0) - \nabla_\theta F(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \|\theta - \theta_0\|_2^2$$

Convergence to Stationary Points of GD

$J(\pi_\theta)$ is non-convex

- Def of a β -smooth function F :

$$\|\nabla_\theta F(\theta) - \nabla_\theta F(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$$

which implies:

$$\left| F(\theta) - F(\theta_0) - \nabla_\theta F(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \|\theta - \theta_0\|_2^2$$

- **Proposition:** (stationary point convergence) Assume $F(\theta)$ is β -smooth. Suppose we run gradient ascent: $\theta_{t+1} = \theta_t + \eta \nabla_\theta F(\theta_t)$, with $\eta = 1/(2\beta)$. Then:

$$\min_{t \leq T} \|\nabla_\theta F(\theta_t)\|_2^2 \leq \frac{2\beta (\max_\theta F(\theta) - F(\theta_0))}{T}$$

Convergence to Stationary Point

$J(\pi_\theta)$ is non-convex (see example in the monograph)

Def of β -smooth:

$$\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$$

$$\left| J(\theta) - J(\theta_0) - \nabla_\theta J(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \|\theta - \theta_0\|_2^2, \forall \theta, \theta_0$$

[Theorem] If $J(\theta)$ is β -smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E} \left[\widetilde{\nabla}_\theta J(\theta_t) \right] = \nabla_\theta J(\theta_t)$, $\mathbb{E} \left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2 \right] \leq \sigma^2$,

then:

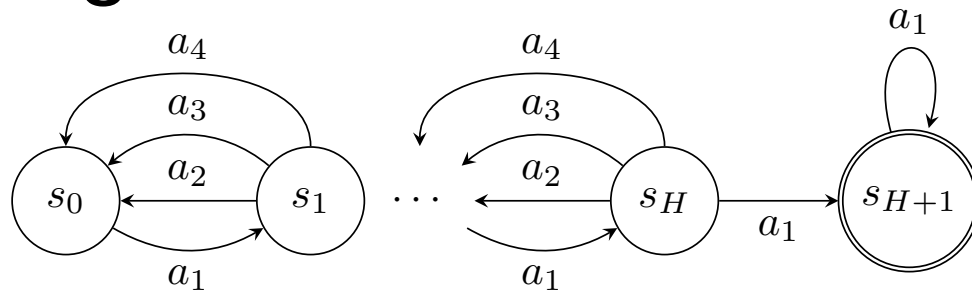
$$\mathbb{E} \left[\frac{1}{T} \sum_t \|\nabla_\theta J(\theta_t)\|_2^2 \right] \leq O \left(\sqrt{\beta \sigma^2 / T} \right)$$

Today:

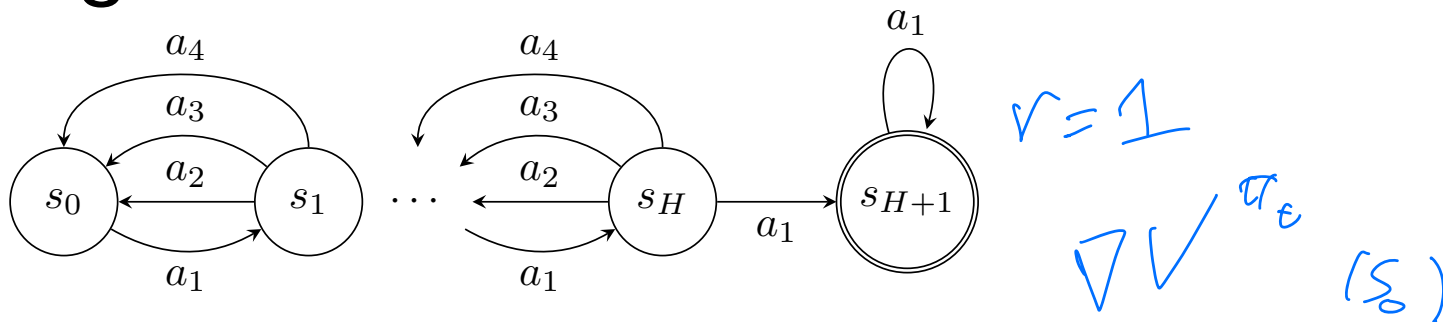
When do PG methods converge to a global optima?
(+ what about function approximation?)

A “landscape” result
(and “exploration”)

Vanishing Gradients and Saddle Points



Vanishing Gradients and Saddle Points

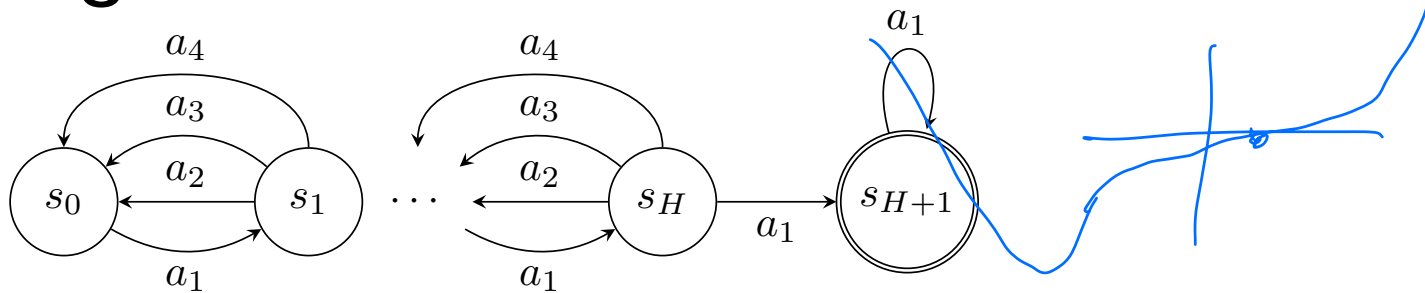


Set $\gamma = H/(H + 1)$. Policy param:

for $a = a_1, a_2, a_3$, $\pi_\theta(a | s) = \theta_{s,a}$, and $\pi_\theta(a_4 | s) = 1 - \theta_{s,a_1} - \theta_{s,a_2} - \theta_{s,a_3}$

(this a “direct” param, which is valid inside the simplex)

Vanishing Gradients and Saddle Points



Set $\gamma = H/(H + 1)$. Policy param:

for $a = a_1, a_2, a_3$, $\pi_\theta(a | s) = \theta_{s,a}$, and $\pi_\theta(a_4 | s) = 1 - \theta_{s,a_1} - \theta_{s,a_2} - \theta_{s,a_3}$

(this a “direct” param, which is valid inside the simplex)

Theorem: For $0 < \theta < 1$ (componentwise) and $\theta_{s,a_1} < 1/4$ (for all states s).

For all $k \leq O(H/\log(H))$, we have that

$$\|\nabla_\theta^k V^{\pi_\theta}(s_0)\| \leq (1/3)^{H/4}$$

(where $\|\nabla_\theta^k V^{\pi_\theta}(s_0)\|$ is the operator norm of the tensor $\nabla_\theta^k V^{\pi_\theta}(s_0)$).

“Vanilla” PG for the Softmax

Let's consider having a stationary state distribution with "coverage"

Let's consider having a starting state distribution with “coverage”

- Given our a starting distribution ρ over states, recall our objective is:

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

where $\{\pi_{\theta} \mid \theta \in \Theta \subset \mathbb{R}^d\}$ is some class of parametric policies.

Let's consider having a starting state distribution with “coverage”

- Given our a starting distribution ρ over states, recall our objective is:

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

where $\{\pi_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$ is some class of parametric policies.

- While we are interested in good performance under ρ , it is helpful to optimize under a different measure μ . Specifically, consider optimizing: $V^{\pi_\theta}(\mu)$, i.e.

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\mu),$$

even though our ultimate goal is performance under $V^{\pi_\theta}(\rho)$.

The Softmax Policy Class

The Softmax Policy Class

- $\pi_{\theta}(a | s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$,
(where the number of parameters is SA).

The Softmax Policy Class

- $$\pi_{\theta}(a | s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

(where the number of parameters is SA).

$$\begin{aligned} \log \pi_{\theta}(a | s) &= \theta_{s,a} - \log \sum_{a'} e^{\theta_{s,a'}} \end{aligned}$$

- We have that:

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

where $\mathbf{1}[\cdot]$ is the indicator function.

The Softmax Policy Class

- $$\pi_{\theta}(a | s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

(where the number of parameters is SA).

$$\sum_{a'} \pi(a'|s) A^{\pi}(s, a') = 0$$

- We have that:

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

where $\mathbf{1}[\cdot]$ is the indicator function.

- Lemma:** For the softmax policy class, we have:

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

$$E \left[\frac{\partial \log \pi}{\partial \theta} A^{\pi}(s, a) \right]$$

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

Proof

$$\begin{aligned}\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} &= E_{\tau \sim \text{Pr}_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \ln \pi_\theta(a | s) A^{\pi_\theta}(s, a) \right] \\ &= E_{\tau \sim \text{Pr}_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}[s_t = s] \left(\mathbf{1}[a_t = a] A^{\pi_\theta}(s, a) - \pi_\theta(a | s) A^{\pi_\theta}(s_t, a_t) \right) \right] \\ &= E_{\tau \sim \text{Pr}_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}[(s_t, a_t) = (s, a)] A^{\pi_\theta}(s, a) \right] + \pi_\theta(a | s) \sum_{t=0}^{\infty} \gamma^t E_{\tau \sim \text{Pr}_\mu^{\pi_\theta}} \left[\mathbf{1}[s_t = s] A^{\pi_\theta}(s_t, a_t) \right] \\ &= \frac{1}{1 - \gamma} E_{(s', a') \sim d^{\pi_\theta}} \left[\mathbf{1}[(s', a') = (s, a)] A^{\pi_\theta}(s, a) \right] + 0 \\ &= \frac{1}{1 - \gamma} d^{\pi_\theta}(s, a) A^{\pi_\theta}(s, a),\end{aligned}$$

Aside: The Performance Difference Lemma

For all π, π', s_0 :

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

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

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

Proof: The Performance Difference Lemma

$\pm V^{\pi'}(s_t)$

$$\begin{aligned} V^{\pi}(s) - V^{\pi'}(s) &= \mathbb{E}_{\tau \sim \text{Pr}^{\pi}(\cdot | s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] - V^{\pi'}(s) \\ &= \mathbb{E}_{\tau \sim \text{Pr}^{\pi}(\cdot | s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)) \right] \\ &= \mathbb{E}_{\tau \sim \text{Pr}^{\pi}(\cdot | s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + \gamma \mathbb{E}[V^{\pi'}(s_{t+1}) | s_t, a_t] - V^{\pi'}(s_t)) \right] \\ &= \mathbb{E}_{\tau \sim \text{Pr}^{\pi}(\cdot | s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t (Q^{\pi'}(s_t, a_t) - V^{\pi'}(s_t)) \right] \\ &= \mathbb{E}_{\tau \sim \text{Pr}^{\pi}(\cdot | s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}(s_t, a_t) \right]. \end{aligned}$$

Global Convergence

Global Convergence

- The update rule for gradient ascent is:

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

Global Convergence

- The update rule for gradient ascent is:

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

- Concerns:

- Non-convex

- Flat gradients if $\theta_t \rightarrow \infty$

(π_t becoming any deterministic policy implies θ_t approaches a stationary point)

Global Convergence

- The update rule for gradient ascent is:

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

- Concerns:

- Non-convex

- Flat gradients if $\theta_t \rightarrow \infty$

(π_t becoming any deterministic policy implies θ_t approaches a stationary point)

- **Theorem:** Assume the μ is strictly positive i.e. $\mu(s) > 0$ for all states s . For $\eta \leq (1 - \gamma)^3/8$, then we have that for all states s , $V^{(t)}(s) \rightarrow V^*(s)$, as $t \rightarrow \infty$.

Global Convergence

- The update rule for gradient ascent is:

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

- Concerns:

- Non-convex

- Flat gradients if $\theta_t \rightarrow \infty$

(π_t becoming any deterministic policy implies θ_t approaches a stationary point)

- **Theorem:** Assume the μ is strictly positive i.e. $\mu(s) > 0$ for all states s . For $\eta \leq (1 - \gamma)^3/8$, then we have that for all states s , $V^{(t)}(s) \rightarrow V^*(s)$, as $t \rightarrow \infty$.

- Comments:

- rate could be exponentially slow in S, H .
- need $\mu > 0$ is necessary.

PG+Log Barrier Regularization (for the softmax)

Log Barrier Regularization

Log Barrier Regularization

- Relative-entropy for distributions p, q is: $\text{KL}(p, q) := E_{x \sim p}[-\log q(x)/p(x)]$.

Log Barrier Regularization

- Relative-entropy for distributions p, q is: $\text{KL}(p, q) := E_{x \sim p}[-\log q(x)/p(x)]$.

- Consider the log barrier λ -regularized objective:

$$L_\lambda(\theta) := V^{\pi_\theta}(\mu) - \lambda E_{s \sim \text{Unif}_S}[\text{KL}(\text{Unif}_A, \pi_\theta(\cdot | s))]$$

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

Log Barrier Regularization

- Relative-entropy for distributions p, q is: $\text{KL}(p, q) := E_{x \sim p}[-\log q(x)/p(x)]$.

- Consider the log barrier λ -regularized objective:

$$\begin{aligned} L_\lambda(\theta) &:= V^{\pi_\theta}(\mu) - \lambda E_{s \sim \text{Unif}_S}[\text{KL}(\text{Unif}_A, \pi_\theta(\cdot | s))] \\ &= V^{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a | s) + \lambda \log A \end{aligned}$$

- Gradient Ascent:

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

Log Barrier Regularization

- Relative-entropy for distributions p, q is: $\text{KL}(p, q) := E_{x \sim p}[-\log q(x)/p(x)]$.
- Consider the log barrier λ -regularized objective:
$$L_\lambda(\theta) := V^{\pi_\theta}(\mu) - \lambda E_{s \sim \text{Unif}_S}[\text{KL}(\text{Unif}_A, \pi_\theta(\cdot | s))]$$
$$= V^{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a | s) + \lambda \log A$$
- Gradient Ascent:
$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta L_\lambda(\theta^{(t)})$$
- Do small gradients imply a globally optimal policy?

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:

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

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

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

Global Convergence with the Log Barrier

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).

Proof, part 1 (optional)

- 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 (optional)

- 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 (optional)

- 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 $\epsilon_{opt} \leq \lambda/(2SA)$,

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

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 $\{\pi_\theta(\cdot | s) | s \in S\}$ 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 $\{\pi_\theta(\cdot | s) | s \in S\}$ 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 $\{\pi_\theta(\cdot | s) | s \in S\}$ 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 .
- 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

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

Proof

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

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

- 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:

$$\theta_{\xi a}^{(t+1)} = \theta_{\xi a}^{(t)} + \frac{\eta}{1 - \gamma} A_{\xi a}^{(t)}$$

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

- Recall that:

$$F_{\mu}(\theta)^{\dagger} \nabla_{\theta} V^{\theta}(\mu) = \frac{1}{1 - \gamma} w^{\star}$$

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

- Recall that:

$$F_{\mu}(\theta)^{\dagger} \nabla_{\theta} V^{\theta}(\mu) = \frac{1}{1 - \gamma} w^{\star}$$

where

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

- So we want to show that: $[w^{\star}]_{s,a} = A^{\pi_{\theta}}(s, a)$

Proof

- Recall that:

$$F_{\mu}(\theta)^{\dagger} \nabla_{\theta} V^{\theta}(\mu) = \frac{1}{1 - \gamma} w^{\star}$$

where

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

- So we want to show that: $[w^{\star}]_{s,a} = A^{\pi_{\theta}}(s, a)$

- Also, recall that:

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

Proof

- Recall that:

$$F_{\mu}(\theta)^{\dagger} \nabla_{\theta} V^{\theta}(\mu) = \frac{1}{1 - \gamma} w^{\star}$$

where

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

- So we want to show that: $[w^{\star}]_{s,a} = A^{\pi_{\theta}}(s, a)$

- Also, recall that:

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

- What is a minimizer for this square loss problem?

Global convergence for NPG

- **Theorem:** Params: $\theta^{(0)} = 0$ and $\eta > 0$. For all ρ and $T > 0$, we have:
$$V^{(T)}(\rho) \geq V^*(\rho) - \frac{\log A}{\eta T} - \frac{1}{(1 - \gamma)^2 T}.$$

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,

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

NPG Conv. Proof, Part 2

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

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

- 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

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