

NGP: global convergence (and approximation)

Sham Kakade and Kianté Brantley

CS 2824: Foundations of Reinforcement Learning

Summary/Today

- Recap
- Today:
 - NPG analysis: (convergence and approximation)

Recap

Things to remember

For all π, π', s_0 :

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

$$A^{\pi'}(s, a) = Q^{\pi'}(s, a) - V^{\pi'}(s)$$

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

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

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

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

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

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 $\theta \in \mathbb{R}^d$

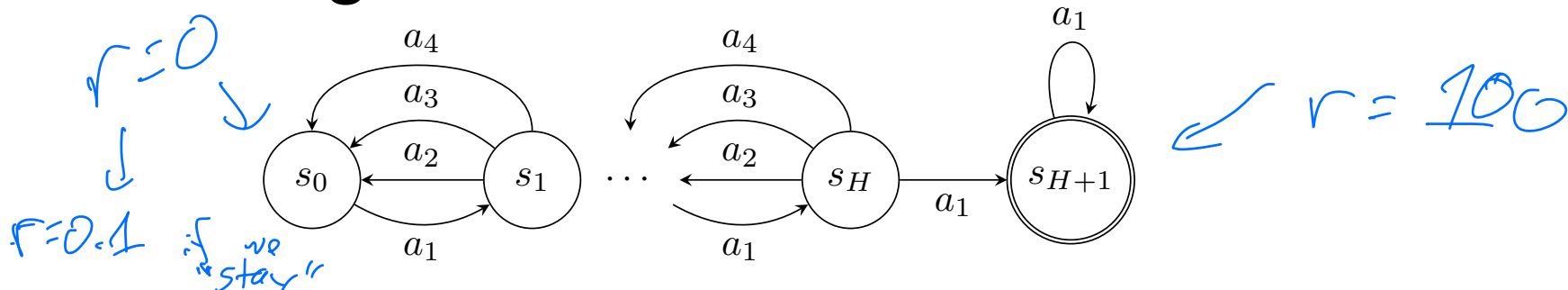
$$\pi_{\theta}(a | s) = \frac{\exp(\theta^{\top} \phi(s, a))}{\sum_{a'} \exp(\theta^{\top} \phi(s, a'))}$$

3. Neural Policy:

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

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

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

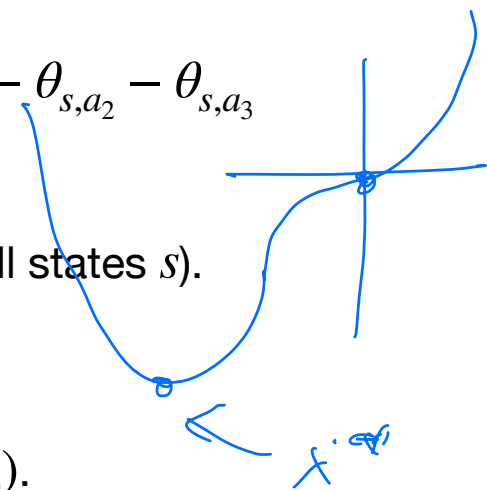
(this a “direct” param, which is valid inside the simplex)

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 $\nabla_\theta^k V^{\pi_\theta}(s_0)$).



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

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.

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?

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).

Today:

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

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)

$$J(\theta) = \mathbb{E}_{s \sim d^\pi} \left[A^\pi(s, a) \nabla \overrightarrow{\log \pi(a|s)} \right]$$

Compatible Function Approximation

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

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

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 $\widehat{A}^{\pi_\theta}(s, a)$,

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

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)

$$\pi_{\theta}(a|s) = \frac{e^{\theta_{sa}}}{\sum_a e^{\theta_{sa}}}$$

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

$$w_{s-a}^* = A^{(t)}(s-a)$$

Proof

- Recall that:

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

where

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

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

$$\sum_{a'} A^{\pi_{\theta}}(s, a') = D$$

- 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?

Improvement Lower Bound

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:

$$Z_t(s) = \sum_{\theta} \pi_{\theta}(g|s) e^{\frac{\eta}{1-\gamma} f_{\theta}(s)}$$

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:

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

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....

Lemma Proof: continued....

By the performance difference lemma,

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

$$A^t(s, a) = \frac{(1-\gamma)}{\gamma} \left(\log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)} + \log Z_t \right)$$

$\pi_{t+1}(a|s) = \frac{\pi_t(a|s) e^{\frac{\gamma}{1-\gamma} A^t(s, a)}}{Z_t}$

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,

$$\begin{aligned} V^{\pi^\star}(\rho) - V^{(t)}(\rho) &= \frac{1}{1 - \gamma} E_{s \sim d^\star} \sum_a \pi^\star(a | s) A^{(t)}(s, a) \\ &= \frac{1}{\eta} E_{s \sim d^\star} \sum_a \pi^\star(a | s) \log \frac{\pi^{(t+1)}(a | s) Z_t(s)}{\pi^{(t)}(a | s)} \end{aligned}$$

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,

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

$$\sum_a \pi^\star(a) \log \frac{V^{(t+1)}(s)}{\pi^\star(a)}$$

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

$$+ \sum_a \pi^\star(a) \log \frac{\pi^\star(a)}{\pi^{(t)}(a)}$$

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

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

NPG Conv. Proof, Part 2

NPG Conv. Proof, Part 2

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

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

NPG Conv. Proof, Part 2

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

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

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

•

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

NPG Conv. Proof, Part 2

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

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

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

•

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

- Again, 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

NPG Conv. Proof, Part 3

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

NPG Conv. Proof, Part 3

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

$O(\eta^2)$

$O\left(\frac{1}{\eta T} + \eta^2\right)$

easy.

What about Function Approximation?

NPG and variants for log-linear policy classes

What about Function Approximation?

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. Log 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
 $f_{\theta} : S \times A \mapsto \mathbb{R}$

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

NPG & Log Linear Policy Classes

- Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a | s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$

NPG & Log Linear Policy Classes

- Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a | s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$
- We have:
 $\nabla_\theta \log \pi_\theta(a | s) = \bar{\phi}_{s,a}^\theta$, where $\bar{\phi}_{s,a}^\theta = \phi_{s,a} - E_{a' \sim \pi_\theta(\cdot | s)}[\phi_{s,a'}]$.

NPG & Log Linear Policy Classes

- Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a | s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$
- We have:
 $\nabla_\theta \log \pi_\theta(a | s) = \bar{\phi}_{s,a}^\theta$, where $\bar{\phi}_{s,a}^\theta = \phi_{s,a} - E_{a' \sim \pi_\theta(\cdot | s)}[\phi_{s,a'}]$.
- The NPG update:
 $\theta \leftarrow \theta + \frac{\eta}{1 - \gamma} w_\star$, $w_\star \in \operatorname{argmin}_w E_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot | s)} [(A^{\pi_\theta}(s, a) - w \cdot \bar{\phi}_{s,a}^\theta)^2]$.

NPG & Log Linear Policy Classes

- Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a | s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$
- We have:
 $\nabla_\theta \log \pi_\theta(a | s) = \bar{\phi}_{s,a}^\theta$, where $\bar{\phi}_{s,a}^\theta = \phi_{s,a} - E_{a' \sim \pi_\theta(\cdot | s)}[\phi_{s,a'}]$.
- The NPG update:
 $\theta \leftarrow \theta + \frac{\eta}{1 - \gamma} w_\star$, $w_\star \in \operatorname{argmin}_w E_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot | s)} [(A^{\pi_\theta}(s, a) - w \cdot \bar{\phi}_{s,a}^\theta)^2]$.
- Equivalently, for the same w_\star ,
$$\pi(a | s) \leftarrow \frac{\pi(a | s) \exp\left(\frac{\eta}{1 - \gamma} w_\star \cdot \phi_{s,a}\right)}{Z_s}$$

(Z_s is the normalizing constant.) Using $\bar{\phi}$ or ϕ result in the same update for π .

Q-NPG: use Q rather A

(a little nice to interpret for analysis)

- Still log linear class.

Q-NPG: use Q rather A

(a little nice to interpret for analysis)

- Still log linear class.
- The Q-NPG update:

$$\theta \leftarrow \theta + \frac{\eta}{1 - \gamma} w_{\star}, \quad w_{\star} \in \operatorname{argmin}_w E_{s \sim d_{\rho}^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot|s)} \left[\left(Q^{\pi_{\theta}}(s, a) - w \cdot \phi_{s,a} \right)^2 \right].$$

Q-NPG: use Q rather A

(a little nice to interpret for analysis)

- Still log linear class.

- The Q-NPG update:

$$\theta \leftarrow \theta + \frac{\eta}{1 - \gamma} w_{\star}, \quad w_{\star} \in \operatorname{argmin}_w E_{s \sim d_{\rho}^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot | s)} \left[\left(Q^{\pi_{\theta}}(s, a) - w \cdot \phi_{s, a} \right)^2 \right].$$

- Equivalently, for the same w_{\star} ,

$$\pi(a | s) \leftarrow \frac{\pi(a | s) \exp\left(\frac{\eta}{1 - \gamma} w_{\star} \cdot \phi_{s, a}\right)}{Z_s}$$

(Z_s is the normalizing constant.)

Approximate Q-NPG + With a Starting Measure

(e.g. we use samples to estimate Q)

- For a state-action distribution D , define:

$$L(w; \theta, D) := E_{s,a \sim D} [(Q^{\pi_{\theta}}(s, a) - w \cdot \phi_{s,a})^2].$$

Approximate Q-NPG + With a Starting Measure

(e.g. we use samples to estimate Q)

- For a state-action distribution D , define:

$$L(w; \theta, D) := E_{s,a \sim D} [(Q^{\pi_\theta}(s, a) - w \cdot \phi_{s,a})^2].$$

- Let us consider using an on-policy state action measure starting with $s_0, a_0 \sim \nu$.
 - this will help with “exploration” and the flat gradient problem when there is approximation
 - shorthand:

$$d^{(t)}(s, a) := d_\nu^{\pi^{(t)}}(s, a)$$

Approximate Q-NPG + With a Starting Measure

(e.g. we use samples to estimate Q)

- For a state-action distribution D , define:

$$L(w; \theta, D) := E_{s,a \sim D} [(Q^{\pi_{\theta}}(s, a) - w \cdot \phi_{s,a})^2].$$

- Let us consider using an on-policy state action measure starting with $s_0, a_0 \sim \nu$.
 - this will help with “exploration” and the flat gradient problem when there is approximation
 - shorthand:

$$d^{(t)}(s, a) := d_{\nu}^{\pi^{(t)}}(s, a)$$

- The approximate version:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1 - \gamma} w^{(t)}, \text{ where } w^{(t)} \approx \operatorname{argmin}_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)}),$$

Approximate Q-NPG + With a Starting Measure

(e.g. we use samples to estimate Q)

- For a state-action distribution D , define:

$$L(w; \theta, D) := E_{s,a \sim D} [(Q^{\pi_\theta}(s, a) - w \cdot \phi_{s,a})^2].$$

- Let us consider using an on-policy state action measure starting with $s_0, a_0 \sim \nu$.
 - this will help with “exploration” and the flat gradient problem when there is approximation
 - shorthand:

$$d^{(t)}(s, a) := d_\nu^{\pi^{(t)}}(s, a)$$

- The approximate version:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} w^{(t)}, \text{ where } w^{(t)} \approx \operatorname{argmin}_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)}),$$

- Equivalently,

$$\pi^{(t+1)}(a | s) \leftarrow \frac{\pi^{(t)}(a | s) \exp\left(\frac{\eta}{1-\gamma} w^{(t)} \cdot \phi_{s,a}\right)}{Z_s}$$

Generic Perturbation Analysis of NPG

NPG regret lemma

- Set $\theta^{(0)} = 0$. Consider an arbitrary sequence of weights $w^{(0)}, \dots, w^{(T)}$, s.t. $\|w^{(t)}\|_2 \leq W$.

NPG regret lemma

- Set $\theta^{(0)} = 0$. Consider an arbitrary sequence of weights $w^{(0)}, \dots, w^{(T)}$, s.t. $\|w^{(t)}\|_2 \leq W$.
- Update rule: $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$

NPG regret lemma

- Set $\theta^{(0)} = 0$. Consider an arbitrary sequence of weights $w^{(0)}, \dots, w^{(T)}$, s.t. $\|w^{(t)}\|_2 \leq W$.
- Update rule: $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$

- **Lemma:** (NPG Regret Lemma)

Fix any comparison policy $\tilde{\pi}$ and any state distribution ρ .

Assume $\log \pi_\theta(a | s)$ (for all s, a) is a β -smooth function of θ .

Define: $\text{err}_t = E_{s \sim \tilde{d}} E_{a \sim \tilde{\pi}(\cdot | s)} [A^{(t)}(s, a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a | s)]$.

We have that:

$$\min_{t < T} \left\{ V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho) \right\} \leq \frac{1}{1 - \gamma} \left(W \sqrt{\frac{2\beta \log A}{T}} + \frac{1}{T} \sum_{t=0}^{T-1} \text{err}_t \right).$$

(where we set using $\eta = \sqrt{2 \log A / (\beta W^2 T)}$)

NPG regret lemma

- Set $\theta^{(0)} = 0$. Consider an arbitrary sequence of weights $w^{(0)}, \dots, w^{(T)}$, s.t. $\|w^{(t)}\|_2 \leq W$.
- Update rule: $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$

- **Lemma:** (NPG Regret Lemma)

Fix any comparison policy $\tilde{\pi}$ and any state distribution ρ .

Assume $\log \pi_\theta(a | s)$ (for all s, a) is a β -smooth function of θ .

Define: $\text{err}_t = E_{s \sim \tilde{d}} E_{a \sim \tilde{\pi}(\cdot | s)} [A^{(t)}(s, a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a | s)]$.

We have that:

$$\min_{t < T} \left\{ V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho) \right\} \leq \frac{1}{1 - \gamma} \left(W \sqrt{\frac{2\beta \log A}{T}} + \frac{1}{T} \sum_{t=0}^{T-1} \text{err}_t \right).$$

(where we set using $\eta = \sqrt{2 \log A / (\beta W^2 T)}$)

- Proof: Mirror descent style of analysis + Perf. Difference Lemma

Approximate Q-NPG

(e.g. we use samples to estimate Q)

- The approximate version:

$$\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}, \text{ where } w^{(t)} \approx \operatorname{argmin}_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)}),$$

Approximate Q-NPG

(e.g. we use samples to estimate Q)

- The approximate version:

$$\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}, \text{ where } w^{(t)} \approx \operatorname{argmin}_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)}),$$

- Error Decomposition:

$$L(w^{(t)}; \theta^{(t)}, d^{(t)}) = \underbrace{L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)})}_{\text{Excess risk}} + \underbrace{L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)})}_{\text{Approximation error}}$$

where $w_{\star}^{(t)} \in \operatorname{argmin}_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)})$

Q-NPG Conv Rate w/ Estimation Error (no approx error)

- Suppose **no approx error**: $L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) = 0$

Suppose the **excess risk**:

$$L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}},$$

Q-NPG Conv Rate w/ Estimation Error (no approx error)

- Suppose **no approx error**: $L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) = 0$

Suppose the **excess risk**:

$$L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}},$$

- **Conditioning**: suppose $\|\phi_{s,a}\|_2 \leq 1$ and, for the initial measure ν ,
 $\sigma_{\min}(E_{s,a \sim \nu}[\phi_{s,a} \phi_{s,a}^{\top}]) = \lambda_{\min}$, $\kappa = 1/\lambda$.

Q-NPG Conv Rate w/ Estimation Error (no approx error)

- Suppose **no approx error**: $L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) = 0$

Suppose the **excess risk**:

$$L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}},$$

- **Conditioning**: suppose $\|\phi_{s,a}\|_2 \leq 1$ and, for the initial measure ν ,
 $\sigma_{\min}(E_{s,a \sim \nu}[\phi_{s,a} \phi_{s,a}^{\top}]) = \lambda_{\min}$, $\kappa = 1/\lambda$.

- **Theorem**: Fix any state distribution ρ ; **any comparator policy π^{\star}** (not necessarily optimal).

With η set appropriately and under the above assumptions, we have that:

$$E \left[\min_{t < T} \left\{ V^{\pi^{\star}}(\rho) - V^{(t)}(\rho) \right\} \right] \leq \frac{W}{1-\gamma} \sqrt{\frac{2 \log A}{T}} + \sqrt{\frac{4A}{(1-\gamma)^3} (\kappa \cdot \epsilon_{\text{stat}})}$$

Q-NPG Conv Rate with Approx+Est. Errors

- Suppose the **excess risk** and **approx error** are bounded as:

$$L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}},$$

$$L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{approx}},$$

Q-NPG Conv Rate with Approx+Est. Errors

- Suppose the **excess risk** and **approx error** are bounded as:

$$L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}},$$

$$L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{approx}},$$

- Conditioning: suppose $\|\phi_{s,a}\|_2 \leq 1$ and, **for the initial measure ν** ,

$$\sigma_{\min}\left(E_{s,a \sim \nu}[\phi_{s,a} \phi_{s,a}^{\top}]\right) = \lambda_{\min}, \quad \kappa = 1/\lambda.$$

Q-NPG Conv Rate with Approx+Est. Errors

- Suppose the **excess risk** and **approx error** are bounded as:

$$L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}},$$

$$L(w_{\star}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{approx}},$$

- Conditioning: suppose $\|\phi_{s,a}\|_2 \leq 1$ and, **for the initial measure ν ,**

$$\sigma_{\min}\left(E_{s,a \sim \nu}[\phi_{s,a} \phi_{s,a}^{\top}]\right) = \lambda_{\min}, \quad \kappa = 1/\lambda.$$

- **Theorem:** Fix any state distribution ρ ; **any comparator policy π^{\star}** (not necessarily optimal).

With η set appropriately and under the above assumptions, we have that:

$$E \left[\min_{t < T} \left\{ V^{\pi^{\star}}(\rho) - V^{(t)}(\rho) \right\} \right] \\ \leq \frac{BW}{1-\gamma} \sqrt{\frac{2 \log A}{T}} + \sqrt{\frac{4A}{(1-\gamma)^3} \left(\kappa \cdot \epsilon_{\text{stat}} + \left\| \frac{d^{\star}}{\nu} \right\|_{\infty} \cdot \epsilon_{\text{approx}} \right)}$$

NPG & Neural Policy Classes

- Neural net $f_\theta : S \times A \mapsto \mathbb{R}$, Policy:

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

NPG & Neural Policy Classes

- Neural net $f_\theta : S \times A \mapsto \mathbb{R}$, Policy:

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

- We have:

$$\nabla_\theta \log \pi_\theta(a | s) = g_\theta(s, a), \quad \text{where} \quad g_\theta(s, a) = \nabla_\theta f_\theta(s, a) - E_{a' \sim \pi_\theta(\cdot | s)}[\nabla_\theta f_\theta(s, a')].$$

NPG & Neural Policy Classes

- Neural net $f_\theta : S \times A \mapsto \mathbb{R}$, Policy:

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

- We have:

$$\nabla_\theta \log \pi_\theta(a | s) = g_\theta(s, a), \quad \text{where } g_\theta(s, a) = \nabla_\theta f_\theta(s, a) - E_{a' \sim \pi_\theta(\cdot | s)}[\nabla_\theta f_\theta(s, a')].$$

- The NPG update rule is:

$$\theta \leftarrow \theta + \eta w_\star, \quad w_\star \in \operatorname{argmin}_w E_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot | s)} \left[\left(A^{\pi_\theta}(s, a) - w \cdot g_\theta(s, a) \right)^2 \right]$$

NPG & Neural Policy Classes

- Neural net $f_\theta : S \times A \mapsto \mathbb{R}$, Policy:

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

- We have:

$$\nabla_\theta \log \pi_\theta(a | s) = g_\theta(s, a), \quad \text{where} \quad g_\theta(s, a) = \nabla_\theta f_\theta(s, a) - E_{a' \sim \pi_\theta(\cdot | s)}[\nabla_\theta f_\theta(s, a')].$$

- The NPG update rule is:

$$\theta \leftarrow \theta + \eta w_\star, \quad w_\star \in \operatorname{argmin}_w E_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot | s)} \left[\left(A^{\pi_\theta}(s, a) - w \cdot g_\theta(s, a) \right)^2 \right]$$