Policy Gradients: Optimality
Today

• Recap:
  • NPG convergence proof wrap up
  • remember compatible function approximation
Today

• Recap:
  • NPG convergence proof wrap up
  • remember compatible function approximation

• Today:
  • What about function approximation?
    log linear policy classes and neural policy classes
  • PG methods have stronger guarantees when we have errors.
Recap
For all $\pi, \pi ', s_0$:

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

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

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

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

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

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

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

$$d^\pi _{\mu}(s) = E_{s_0 \sim \mu} [d^\pi _{s_0}(s)]$$
The Natural Policy Gradient

• Define $\mathcal{F}_\rho^\theta$ as the (average) Fisher matrix on the family of distributions $\{\pi_\theta(\cdot | s) : s \in S\}$ as:

$$\mathcal{F}_\rho^\theta := E_{s \sim \rho \pi_\theta} E_{a \sim \pi_\theta(\cdot | s)} \left[ (\nabla \log \pi_\theta(a | s)) \nabla \log \pi_\theta(a | s)^T \right].$$

• The NPG algorithm performs gradient updates in this induced geometry:

$$\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)})^\dagger \nabla_\theta V^{(t)}(\rho),$$

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

• Idea:
  • ‘stretch’ the corners of the simplex out to travel faster
    (as opposed to the log-barrier which keeps us away)
NPG softmax case
(NPG as “soft” policy iteration)

- Lemma: (Softmax NPG as soft policy iteration) The NPG update is:
  \[ \theta^{(t+1)} = \theta^{(t)} + \eta \frac{A^{(t)}}{1 - \gamma} \]
  and so:
  \[ \pi^{(t+1)}(a | s) = \pi^{(t)}(a | s) \frac{\exp(\eta A^{(t)}(s, a) / (1 - \gamma))}{Z_t(s)} \]
  where \( Z_t(s) = \sum_a \pi^{(t)}(a | s) \exp(\eta A^{(t)}(s, a) / (1 - \gamma)) \).
Global convergence for NPG

- **Theorem:** Params: $\theta^{(0)} = 0$ and $\eta > 0$. For all $\rho$ 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 $\epsilon$-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 $\rho$ (and no “dist mismatch factor”)
  - No ‘flat gradient’ problem

- What about approx/estimation errors? (next lecture)
• Lemma: For the iterates $\pi^{(t)}$ generated by the NPG, we have for all distributions $\mu$:

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

- $d^*$ as shorthand for $d_p^*$; $\pi_s$ as shorthand for the vector of $\pi(\cdot \mid s)$
- By the performance difference lemma,

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

$$= \frac{1}{\eta} \mathbb{E}_{s \sim d^*} \sum_a \pi^*(a \mid s) \log \frac{\pi^{(t+1)}(a \mid s) Z_t(s)}{\pi^{(t)}(a \mid s)}$$

$$= \frac{1}{\eta} \mathbb{E}_{s \sim d^*} \left( KL(\pi^*_s \mid \pi^{(t)}_s) - KL(\pi^*_s \mid \pi^{(t+1)}_s) + \sum_a \pi^*(a \mid s) \log Z_t(s) \right)$$

$$= \frac{1}{\eta} \mathbb{E}_{s \sim d^*} \left( KL(\pi^*_s \mid \pi^{(t)}_s) - KL(\pi^*_s \mid \pi^{(t+1)}_s) + \log Z_t(s) \right)$$
NPG Conv. Proof, Part 2

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

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

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

$$\leq \frac{E_{s \sim d^*}KL(\pi_s^* || \pi_s^{(0)})}{\eta T} + \frac{1}{\eta T} \sum_{t=0}^{T-1} E_{s \sim d^*} \log Z_t(s).$$
By the improvement lemma $V^{(t+1)}(\rho) \geq V^{(t)}(\rho)$. Hence,

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

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

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

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 $E_{s \sim d^*} \log Z_t(s)$. 

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

\[
\leq \frac{E_{s \sim d^*} \text{KL}(\pi^*_s \mid \mid \pi^{(0)})}{\eta T} + \frac{1}{(1 - \gamma) T} \sum_{t=0}^{T-1} \left( V^{(t+1)}(d^*) - V^{(t)}(d^*) \right)
\]

\[
= \frac{E_{s \sim d^*} \text{KL}(\pi^*_s \mid \mid \pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^*) - V^{(0)}(d^*)}{(1 - \gamma) T}
\]

\[
\leq \frac{1}{\eta T} + \frac{1}{(1 - \gamma)^2 T}.
\]

\[
V \leq \frac{1}{1 - \gamma}.
\]
1. Softmax Policy for Tabular MDPs:

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

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

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

\[ \pi_\theta(a | s) = \frac{\exp(\theta^T \phi(s, a))}{\sum_{a'} \exp(\theta^T \phi(s, a'))} \]

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

3. Neural Policy:

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

Neural network \( f_\theta : S \times A \mapsto \mathbb{R} \)
NPG & Compatible Function Approximation

- Let \( w^* \) denote the following minimizer of the “compatible function approximation” error:
  \[
  w^* \in \arg\min_w E_{s \sim d^\pi_\theta} E_{a \sim \pi_\theta(s \mid s)} \left[ (A^\pi_\theta(s, a) - w \cdot \nabla_\theta \log \pi_\theta(a \mid s))^2 \right]
  \]

- **Lemma:** Let \( \tilde{A}^\pi_\theta(s, a) \) be the best linear predictor of \( A^\pi_\theta(s, a) \) using \( \nabla_\theta \log \pi_\theta(a \mid s) \), i.e. \( \tilde{A}^\pi_\theta(s, a) := w^* \cdot \nabla_\theta \log \pi_\theta(a \mid s) \). We have:
  \[
  \nabla_\theta V^\pi_\theta(\mu) = \frac{1}{1 - \gamma} E_{s \sim d^\pi_\theta} E_{a \sim \pi_\theta(s \mid s)} \left[ \nabla_\theta \log \pi_\theta(a \mid s) \tilde{A}^\pi_\theta(s, a) \right]
  \]
  We can use \( \tilde{A}^\pi_\theta(s, a) \) instead of \( A^\pi_\theta(s, a) \).

- **Lemma:** We have that
  \[
  F_\mu(\theta)^\dagger \nabla_\theta V^\theta(\mu) = \frac{1}{1 - \gamma} w^*,
  \]
  The NPG direction is the weights \( w^* \).
Today:
Natural Policy Gradient and Approximation
Examples:
NPG and variants for log-linear policy classes
Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a \mid s) = \frac{\exp(\theta^T \phi(s, a))}{\sum_{a'} \exp(\theta^T \phi(s, a'))}$
NPG & Log Linear Policy Classes

- Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a | s) = \frac{\exp(\theta^T \phi(s, a))}{\sum_{a'} \exp(\theta^T \phi(s, a'))}$

- We have: $\nabla_\theta \log \pi_\theta(a | s) = \overline{\phi}_{s,a}^\theta$, where $\overline{\phi}_{s,a}^\theta = \phi_{s,a} - E_{a' \sim \pi_\theta(\cdot | s)}[\phi_{s,a}]$. 
NPG & Log Linear Policy Classes

- Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a | s) = \frac{\exp(\theta^T \phi(s, a))}{\sum_{a'} \exp(\theta^T \phi(s, a'))}$

- We have:
  \[
  \nabla_\theta \log \pi_\theta(a | s) = \overline{\Phi}_{s,a}^\theta, \text{ where } \overline{\Phi}_{s,a}^\theta = \phi_{s,a} - E_{a' \sim \pi_\theta(\cdot | s)}[\phi_{s,a'}].
  \]

- The NPG update:
  \[
  \theta \leftarrow \theta + \eta w_*, \quad w_* \in \text{argmin}_w E_{s \sim d^\pi_\theta, a \sim \pi_\theta(\cdot | s)} \left[ (A_{\pi_\theta}(s, a) - w \cdot \overline{\Phi}_{s,a}^\theta)^2 \right].
  \]
NPG & Log Linear Policy Classes

- Feature vector $\phi(s, a) \in \mathbb{R}^d$, $\pi_\theta(a \mid s) = \frac{\exp(\theta^T \phi(s, a))}{\sum_{a'} \exp(\theta^T \phi(s, a'))}$

- We have:
  \[ \nabla_\theta \log \pi_\theta(a \mid s) = \bar{\phi}_{s,a}^\theta, \quad \text{where} \quad \bar{\phi}_{s,a}^\theta = \phi_{s,a} - E_{a' \sim \pi_\theta(\cdot \mid s)}[\phi_{s,a'}]. \]

- The NPG update:
  \[ \theta \leftarrow \theta + \eta w_\star, \quad w_\star \in \arg\min_w E_{s \sim d^{s\theta}, a \sim \pi_\theta(\cdot \mid s)}\left[(A^n_{\theta}(s, a) - w \cdot \bar{\phi}_{s,a}^\theta)^2\right]. \]

- Equivalently, for the same $w_\star$, $\pi(a \mid s) \leftarrow \frac{\pi(a \mid s) \exp(w_\star \cdot \phi_{s,a})}{Z_s}$

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

\[ \theta \leftarrow \theta + \eta w_\star, \quad w_\star \in \text{argmin}_w E_{s \sim d^s_\theta, a \sim \pi_\theta(\cdot|s)} \left[ (Q^{\pi_\theta}(s, a) - w \cdot \phi_{s,a})^2 \right]. \]
Q-NPG: use Q rather A
(a little nice to interpret for analysis)

• Still log linear class.

• The Q-NPG update:

\[
\theta \leftarrow \theta + \eta w_\star, \quad w_\star \in \text{argmin}_w E_{s \sim d^\pi_\theta, a \sim \pi_\theta(\cdot | s)} \left[ (Q_{\pi_\theta}(s, a) - w \cdot \phi_{s,a})^2 \right].
\]

• Equivalently, for the same \(w_\star\),

\[
\pi(a | s) \leftarrow \frac{\pi(a | s) \exp(w_\star \cdot \phi_{s,a})}{Z_s}
\]

\((Z_s\) is the normalizing constant.)
Approximate Q-NPG
(e.g. we use samples to estimate Q)

• For a state-action distribution \( \nu \), define:

\[
L(w; \theta, \nu) := E_{s,a} \left[ \left( Q_{\pi_{\theta}(s,a)} - w \cdot \phi_{s,a} \right)^2 \right].
\]
Approximate Q-NPG
(e.g. we use samples to estimate Q)

- For a state-action distribution \( \nu \), define:
  \[
  L(w; \theta, \nu) := E_{s,a \sim \nu}[(Q^{\pi_\theta}(s, a) - w \cdot \phi_{s,a})^2].
  \]

- With an on-policy state action measure starting with \( s_0, a_0 \sim \nu \). Shorthand:
  \[
  d^{(t)}(s, a) := d^{\pi_{(t)}}_\nu(s, a)
  \]
Approximate Q-NPG
(e.g. we use samples to estimate Q)

- For a state-action distribution $\nu$, define:
  \[ L(w; \theta, \nu) := E_{s,a \sim \nu} [(Q^{\pi}(s,a) - w \cdot \phi_{s,a})^2] . \]

- With an on-policy state action measure starting with $s_0, a_0 \sim \nu$. Shorthand:
  \[ d^{(t)}(s, a) := d^{\pi(t)}_\nu (s, a) \]

- The approximate version:
  \[ \theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)} , \text{ where } w^{(t)} \approx \text{argmin}_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)}) , \]

Approximate Q-NPG
(e.g. we use samples to estimate Q)

• For a state-action distribution $\nu$, define:

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

• With an on-policy state action measure starting with $s_0, a_0 \sim \nu$. Shorthand:

$$d_t(s, a) := d_{\pi_t}^{(t)}(s, a)$$

• The approximate version:

$$\theta_{t+1} = \theta_t + \eta w^{(t)}$$

where $w^{(t)} \approx \arg\min_{\|w\|_2 \leq W} L(w; \theta_t, d^{(t)})$.

• Equivalently,

$$\pi_{t+1}(a | s) \leftarrow \frac{\pi_t(a | s) \exp(w^{(t)} \cdot \phi_{s,a})}{Z_s}$$
Error Analysis of NPG
(and variants)
NPG regret lemma

- Consider the update rule: $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$

(starting with $\pi^{(0)}$ being the uniform policy).
NPG regret lemma

- Consider the update rule: \( \theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)} \) (starting with \( \pi^{(0)} \) being the uniform policy).

- **Lemma**: (NPG Regret Lemma)
  
  Fix any comparison policy \( \widetilde{\pi} \) and a state distribution \( \rho \).
  
  Assume \( \log \pi_{\theta}(a \mid s) \) (for all \( s, a \)) is a \( \beta \)-smooth function of \( \theta \).
  
  Consider an arbitrary sequence of weights \( w^{(0)}, \ldots, w^{(T)} \), s.t. \( \|w^{(t)}\|_2 \leq W \). Define:

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

  We have that:

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

  (using \( \eta = \sqrt{2 \log A/(\beta W^2 T)} \))
NPG regret lemma

• Consider the update rule: \( \theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)} \)
  (starting with \( \pi^{(0)} \) being the uniform policy).

• **Lemma:** (NPG Regret Lemma)
  Fix any comparison policy \( \tilde{\pi} \) and a state distribution \( \rho \).
  Assume \( \log \pi_\theta(a \mid s) \) (for all \( s, a \)) is a \( \beta \)-smooth function of \( \theta \).
  Consider an arbitrary sequence of weights \( w^{(0)}, \ldots, w^{(T)} \), s.t. \( \|w^{(t)}\|_2 \leq W \). Define:
  \[
  \text{err}_t = \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \tilde{\pi}(\cdot \mid s)} \left[ A^{(t)}(s, a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a \mid s) \right].
  \]
  We have that:
  \[
  \min_{t < T} \left\{ V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho) \right\} \leq \frac{1}{1 - \gamma} \left( W \sqrt{\frac{2\beta \log A}{T}} + \frac{1}{T} \sum_{t=0}^{T-1} \text{err}_t \right).
  \]
  (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:
\[
\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}, \quad \text{where. } w^{(t)} \approx \text{argmin}_{\|w\|_2 \leq \bar{w}} L(w; \theta^{(t)}, d^{(t)}),
\]
Approximate Q-NPG
(e.g. we use samples to estimate Q)

• The approximate version:
  \[ \theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}, \]
  where. \[ w^{(t)} \approx \arg\min_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)}), \]

• Error Decomposition:

\[
L(w^{(t)}; \theta^{(t)}, d^{(t)}) = L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w^{(t)}; \theta^{(t)}, d^{(t)}) + L(w^{(t)}; \theta^{(t)}, d^{(t)})
\]

\[
\text{Excess risk} \quad \text{Approximation error}
\]

where \( w^{(t)}_* \in \arg\min_{\|w\|_2 \leq W} L(w; \theta^{(t)}, d^{(t)}) \)
Q-NPG Conv Rate w/ Estimation Error (no approx error)

- Suppose no approx error: $L(w^{(i)}_*; \theta^{(i)}, d^{(i)}) = 0$
  
  Suppose the excess risk:
  
  $L(w^{(i)}; \theta^{(i)}, d^{(i)}) - L(w^{(i)}_*; \theta^{(i)}, d^{(i)}) \leq e_{\text{stat}}$. 
Q-NPG Conv Rate w/ Estimation Error
(no approx error)

• Suppose no approx error: \( L(w_*(t); \theta(t), d(t)) = 0 \)
  Suppose the excess risk:
  \[ L(w(t); \theta(t), d(t)) - L(w_*(t); \theta(t), d(t)) \leq \epsilon_{\text{stat}}. \]

• Conditioning: suppose \( \|\phi_{s,a}\|_2 \leq 1 \) and, for the initial measure \( \nu \),
  \[ \sigma_{\min}(E_{s,a \sim \nu}[\phi_{s,a}\phi_{s,a}^T]) = \lambda_{\min}, \quad \kappa = 1/\lambda. \]
Q-NPG Conv Rate w/ Estimation Error (no approx error)

- Suppose no approx error: $L(w_{**}^{(t)}; \theta^{(t)}, d^{(t)}) = 0$
  
  Suppose the excess risk:
  $$L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{**}^{(t)}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}}.$$  

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

- Theorem: Fix any state distribution $\rho$; any comparator policy $\pi^*$ (not necessarily optimal).
  With $\eta$ set appropriately and under the above assumptions, we have that:
  $$E \left[ \min_{t < T} \left\{ V_{\pi^*}(\rho) - V^{(t)}(\rho) \right\} \right] \leq W \sqrt{2 \log A/T} + \sqrt{4A \over (1 - \gamma)^3} \left( \kappa \cdot \epsilon_{\text{stat}} \right).$$
Q-NPG Conv Rate with Approx+Est. Errors

• Suppose the excess risk and approx error are bounded as:

\[
L(w_{(t)}; \theta^{(t)}, d^{(t)}) - L(w_{(t)}^{*}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}},
\]

\[
L(w_{(t)}^{*}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{approx}}.
\]
Q-NPG Conv Rate with Approx+Est. Errors

• Suppose the excess risk and approx error are bounded as:
  \[ L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w^{(t)}_{\ast}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}}, \]
  \[ L(w^{(t)}_{\ast}; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{approx}}, \]

• Conditioning: suppose \( \| \phi_{s,a} \|_2 \leq 1 \) and, for the initial measure \( \nu \),
  \[ \sigma_{\min}(E_{s,a \sim \nu}[\phi_{s,a} \phi_{s,a}^\top]) = \lambda_{\min}, \quad \kappa = 1/\lambda. \]
Q-NPG Conv Rate with Approx+Est. Errors

• Suppose the excess risk and approx error are bounded as:
  \[ L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w^{(t)}_\star; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{stat}}, \]
  \[ L(w^{(t)}_\star; \theta^{(t)}, d^{(t)}) \leq \epsilon_{\text{approx}}, \]

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

• Theorem: Fix any state distribution \( \rho \); any comparator policy \( \pi^* \) (not necessarily optimal).
  With \( \eta \) set appropriately and under the above assumptions, we have that:
  \[ E \left[ \min_{t<T} \left\{ V^{\pi^*}(\rho) - V^{(t)}(\rho) \right\} \right] \]
  \[ \leq \frac{BW}{1 - \gamma} \sqrt{\frac{2 \log A}{T}} + \sqrt{\frac{4A}{(1 - \gamma)^3} \left( \kappa \cdot \epsilon_{\text{stat}} + \|d^*/\nu\|_\infty \cdot \epsilon_{\text{approx}} \right)} \]
NPG & Neural Policy Classes

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

$$\pi_\theta(a \mid s) = \frac{\exp(f_\theta(s, a))}{\sum_{a'} \exp(f_\theta(s, a'))}$$
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:
  \[
  \nabla_\theta \log \pi_\theta(a | s) = g_\theta(s, a), \text{ where } g_\theta(s, a) = \nabla_\theta f_\theta(s, a) - E_{a' \sim \pi_\theta(\cdot | s)}[\nabla_\theta f_\theta(s, a')].
  \]
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:
  \[
  \theta \leftarrow \theta + \eta w_*, \quad w_* \in \arg\min_w E_{s \sim d^\pi_\theta, a \sim \pi_\theta(\cdot | s)} [(A^\pi_\theta(s, a) - w \cdot g_\theta(s, a))^2]
  \]
NPG & Neural Policy Classes

- Neural net $f_\theta : S \times A \mapsto \mathbb{R}$, Policy:
  \[ \pi_\theta(a \mid s) = \frac{\exp(f_\theta(s, a))}{\sum_{a'} \exp(f_\theta(s, a'))} \]

- We have:
  \[ \nabla_\theta \log \pi_\theta(a \mid s) = g_\theta(s, a), \text{ where } g_\theta(s, a) = \nabla_\theta f_\theta(s, a) - E_{a' \sim \pi_\theta(\cdot \mid s)}[\nabla_\theta f_\theta(s, a')] \]

- The NPG update rule is:
  \[ \theta \leftarrow \theta + \eta w_*, \quad w_* \in \text{argmin}_w E_{s \sim \pi_\theta(\cdot \mid s)} \left[ (A_{\pi_\theta}(s, a) - w \cdot g_\theta(s, a))^2 \right] \]
Ax = b

x = (ATA)^{-1} Atb
For linear maps

\[ \forall x, z \in V, \exists \alpha, \beta \in W \]

\[ \phi(x + z) = \phi(x) + \phi(z) \]

\[ \phi(c \cdot x) = c \cdot \phi(x) \]