Policy Gradient: Approximation

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Today

- Recap
- Today:
  - NPG and function approximation
    - (for log linear policy classes and neural policy classes)
    - PG methods have stronger guarantees (over approximate value function methods) when we have errors.

  - Trust region methods and conservative policy iteration

Recap

Things to remember

For all π, π', s_0:

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

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

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

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

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

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

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

$$d^\pi_\mu(s) = E_{s_0 \sim \mu}[d^\pi_{s_0}(s)]$$

The Natural Policy Gradient

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

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

Compatible Function Approximation

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

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

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

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

Global convergence for Softmax NPG

- Lemma: (Softmax NPG as soft policy iteration) The NPG update is:
  $$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$
  and this leads to the update:
  $$\pi^{(t+1)}(a \mid s) = \pi^{(t)}(a \mid s) \frac{\exp\left(\eta A^{(t)}(s,a)/(1-\gamma)\right)}{Z_t(s)},$$

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

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

Today:

Function Approximation & Distribution Shift

What about Function Approximation?

1. Softmax Policy for Tabular MDPs:

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

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

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

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

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

3. Neural Policy:

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

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

NPG & Log Linear Policy Classes

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

NPG & Log Linear Policy Classes

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

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

NPG & Log Linear Policy Classes

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

NPG & Log Linear Policy Classes

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

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

- The NPG update:
  $\theta \leftarrow \theta + \dfrac{\eta}{1-\gamma} w_\star$, $\quad w_\star \in \text{argmin}_w E_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot \mid s)}\left[\left(A^{\pi_\theta}(s,a) - w \cdot \overline{\phi}^{\theta}_{s,a}\right)^2\right]$.

- Equivalently, for the same $w_\star$,

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

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

Generic Perturbation Analysis of NPG (for smooth policy classes)

Recall a function $f : R^d \to R$ is said to be β-smooth if for all $x, x' \in R^d$:
$\|\nabla f(x) - \nabla f(x')\|_2 \leq \beta\|x - x'\|_2$

NPG regret lemma

NPG regret lemma

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

NPG regret lemma

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

Lemma: (NPG Regret Lemma)

NPG regret lemma

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

Lemma: (NPG Regret Lemma)

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

NPG regret lemma

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

Lemma: (NPG Regret Lemma)

- Fix any comparison policy $\widetilde{\pi}$ and any state distribution ρ.
- Suppose $\pi^{(0)}$ is the uniform policy.

NPG regret lemma

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

Lemma: (NPG Regret Lemma)

- Fix any comparison policy $\widetilde{\pi}$ and any state distribution ρ.
- Suppose $\pi^{(0)}$ is the uniform policy.
- Assume for all s, a that $\log \pi_\theta(a \mid s)$ is a β-smooth function of θ.

NPG regret lemma

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

Lemma: (NPG Regret Lemma)

- Fix any comparison policy $\widetilde{\pi}$ and any state distribution ρ.
- Suppose $\pi^{(0)}$ is the uniform policy.
- Assume for all s, a that $\log \pi_\theta(a \mid s)$ is a β-smooth function of θ.

For an arbitrary sequence $w^{(0)}, \ldots, w^{(T)}$, s.t. $\|w^{(t)}\|_2 \leq W$,
where $\text{err}_t := E_{s \sim d_\rho^{\widetilde{\pi}}} E_{a \sim \widetilde{\pi}(\cdot \mid s)} \left[A^{(t)}(s,a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a \mid s) \right]$, we have:

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

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

Proof, part 1

Proof, part 1

- A function $f : R^d \to R$ is said to be β-smooth if for all $x, x' \in R^d$:
  $\|\nabla f(x) - \nabla f(x')\|_2 \leq \beta \|x - x'\|_2$

  and, due to Taylor's theorem, this implies:
  $| f(x') - f(x) - \nabla f(x) \cdot (x' - x) | \leq \frac{\beta}{2} \|x' - x\|_2^2 .$

Proof, part 1

- A function $f: R^d \to R$ is said to be β-smooth if for all $x, x' \in R^d$:
  $\|\nabla f(x) - \nabla f(x')\|_2 \leq \beta \|x - x'\|_2$

  and, due to Taylor's theorem, this implies:
  $|f(x') - f(x) - \nabla f(x) \cdot (x' - x)| \leq \frac{\beta}{2} \|x' - x\|_2^2.$

- By smoothness,

$$
\begin{aligned}
& \log \pi^{(t+1)}(a|s) \\
\geq\ & \log \pi^{(t)}(a|s) + \nabla_\theta \log \pi^{(t)}(a|s) \cdot \left(\theta^{(t+1)} - \theta^{(t)}\right) - \frac{\beta}{2} \|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \\
=\ & \log \pi^{(t)}(a|s) + \eta \nabla_\theta \log \pi^{(t)}(a|s) \cdot w^{(t)} - \eta^2 \frac{\beta}{2} \|w^{(t)}\|_2^2
\end{aligned}
$$

Proof, part 2

Proof, part 2

- Shorthand: $\widetilde{d}$ for $d_\rho^{\widetilde{\pi}}$ (note ρ and $\widetilde{\pi}$ are fixed); π_s for the distribution $\pi(\cdot \mid s)$.

Proof, part 2

- Shorthand: $\widetilde{d}$ for $d_\rho^{\widetilde{\pi}}$ (note ρ and $\widetilde{\pi}$ are fixed); π_s for the distribution $\pi(\cdot \mid s)$.
- By smoothness,

$$E_{s\sim\widetilde{d}}\left(KL(\widetilde{\pi}_s \,||\, \pi_s^{(t)}) - KL(\widetilde{\pi}_s \,||\, \pi_s^{(t+1)})\right) = E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\log \frac{\pi^{(t+1)}(a\mid s)}{\pi^{(t)}(a\mid s)}\right]$$

$$\geq \eta E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\nabla_\theta \log \pi^{(t)}(a\mid s)\cdot w^{(t)}\right] - \eta^2 \frac{\beta}{2}\|w^{(t)}\|_2^2$$

Proof, part 2

- Shorthand: $\widetilde{d}$ for $d_\rho^{\widetilde{\pi}}$ (note ρ and $\widetilde{\pi}$ are fixed); π_s for the distribution $\pi(\cdot \mid s)$.
- By smoothness,

$$E_{s\sim\widetilde{d}}\left(KL(\widetilde{\pi}_s \,||\, \pi_s^{(t)}) - KL(\widetilde{\pi}_s \,||\, \pi_s^{(t+1)})\right) = E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\log \frac{\pi^{(t+1)}(a\,|\,s)}{\pi^{(t)}(a\,|\,s)}\right]$$

$$\geq \eta E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\nabla_\theta \log \pi^{(t)}(a\,|\,s)\cdot w^{(t)}\right] - \eta^2\frac{\beta}{2}\|w^{(t)}\|_2^2$$

- By the performance difference lemma and def of err_t,

$$= \eta E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[A^{(t)}(s,a)\right] - \eta^2\frac{\beta}{2}\|w^{(t)}\|_2^2 + \eta E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\nabla_\theta \log \pi^{(t)}(a\,|\,s)\cdot w^{(t)} - A^{(t)}(s,a)\right]$$

$$= (1-\gamma)\eta\left(V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho)\right) - \eta^2\frac{\beta}{2}\|w^{(t)}\|_2^2 - \eta\ \mathrm{err}_t$$

Proof, part 2

- Shorthand: $\widetilde{d}$ for $d_\rho^{\widetilde{\pi}}$ (note ρ and $\widetilde{\pi}$ are fixed); π_s for the distribution $\pi(\,\cdot \mid s)$.
- By smoothness,

$$E_{s\sim\widetilde{d}}\left(KL(\widetilde{\pi}_s \,||\, \pi_s^{(t)}) - KL(\widetilde{\pi}_s \,||\, \pi_s^{(t+1)})\right) = E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\log \frac{\pi^{(t+1)}(a\,|\,s)}{\pi^{(t)}(a\,|\,s)}\right]$$

$$\geq \eta E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\nabla_\theta \log \pi^{(t)}(a\,|\,s)\cdot w^{(t)}\right] - \eta^2\frac{\beta}{2}\|w^{(t)}\|_2^2$$

- By the performance difference lemma and def of err_t,

$$= \eta E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[A^{(t)}(s,a)\right] - \eta^2\frac{\beta}{2}\|w^{(t)}\|_2^2 + \eta E_{s\sim\widetilde{d}}\, E_{a\sim\widetilde{\pi}(\cdot|s)}\left[\nabla_\theta \log \pi^{(t)}(a\,|\,s)\cdot w^{(t)} - A^{(t)}(s,a)\right]$$

$$= (1-\gamma)\eta\left(V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho)\right) - \eta^2\frac{\beta}{2}\|w^{(t)}\|_2^2 - \eta\ \mathrm{err}_t$$

- Rearranging,

$$V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho) \leq \frac{1}{1-\gamma}\left(\frac{1}{\eta}E_{s\sim\widetilde{d}}\left(KL(\widetilde{\pi}_s \,||\, \pi_s^{(t)}) - KL(\widetilde{\pi}_s \,||\, \pi_s^{(t+1)})\right) + \frac{\eta\beta}{2}W^2 + \mathrm{err}_t\right)$$

Proof, part 3

Proof, part 3

- Proceeding,

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

$$\leq \frac{1}{\eta T(1-\gamma)}\sum_{t=0}^{T-1} E_{s\sim\widetilde{d}}\,(KL(\widetilde{\pi}_s\,||\,\pi_s^{(t)}) - KL(\widetilde{\pi}_s\,||\,\pi_s^{(t+1)})) + \frac{1}{T(1-\gamma)}\sum_{t=0}^{T-1}\left(\frac{\eta\beta W^2}{2} + \mathrm{err}_t\right)$$

$$\leq \frac{E_{s\sim\widetilde{d}}\,KL(\widetilde{\pi}_s\,||\,\pi^{(0)})}{\eta T(1-\gamma)} + \frac{\eta\beta W^2}{2(1-\gamma)} + \frac{1}{T(1-\gamma)}\sum_{t=0}^{T-1}\mathrm{err}_t$$

$$\leq \frac{\log A}{\eta T(1-\gamma)} + \frac{\eta\beta W^2}{2(1-\gamma)} + \frac{1}{T(1-\gamma)}\sum_{t=0}^{T-1}\mathrm{err}_t$$

which completes the proof (after setting η).

What about Function Approximation?

NPG and variants for log-linear policy classes

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:

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

Q-NPG: use Q rather A

(a little nice to interpret for analysis)

- Still log linear class.

- The Q-NPG update:

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

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

Approximate Q-NPG + With a Starting Measure

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

- For a state-action distribution D, define:

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

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

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

Approximate Q-NPG + With a Starting Measure

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

- For a state-action distribution D, define:

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

- 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:

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

- Equivalently,

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

Approximate Q-NPG

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

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

Approximate Q-NPG

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

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

- 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 $w_\star^{(t)} \in \text{argmin}_{\|w\|_2 \le W} L(w; \theta^{(t)}, d^{(t)})$

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,

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

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:

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

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

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

- Theorem: Fix any state distribution ρ; any comparator policy $\widetilde{\pi}$ (not necessarily optimal). Suppose $\|\phi(s,a)\| \le B$. With η set appropriately and under the above assumptions,

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

Q-NPG Conv Rate with Approx+Est. Errors

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

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

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

Q-NPG Conv Rate with Approx+Est. Errors

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

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

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

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

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

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

- Theorem: Fix any state distribution ρ; any comparator policy $\widetilde{\pi}$ (not necessarily optimal). Suppose $\|\phi(s,a)\| \leq B$. With η set appropriately and under the above assumptions,

$$E\left[\min_{t<T}\left\{V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho)\right\}\right]$$

$$\leq \frac{BW}{1-\gamma}\sqrt{\frac{2\log A}{T}} + \sqrt{\frac{4A}{(1-\gamma)^3}\left(\kappa\cdot\epsilon_{\text{stat}} + \left\|\frac{\widetilde{d}}{\nu}\right\|_\infty\cdot\epsilon_{\text{approx}}\right)}$$

NPG & Neural Policy Classes

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

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

NPG & Neural Policy Classes

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

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

- We have:

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

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:

$\nabla_\theta \log \pi_\theta(a \,|\, s) = g_\theta(s,a)$, 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:

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

NPG & Neural Policy Classes

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

- We have:
  $\nabla_\theta \log \pi_\theta(a \mid s) = g_\theta(s,a)$, 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:
  $\theta \leftarrow \theta + \eta w_\star$, $w_\star \in \text{argmin}_w E_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)}\left[\left(A^{\pi_\theta}(s,a) - w \cdot g_\theta(s,a)\right)^2\right]$