Policy Gradient: Optimality

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Today

- Recap
- Today:
    - NPG convergence proof wrap up
    - What about function approximation?
      remember compatible function approximation
      log linear policy classes and neural policy classes

- PG methods have stronger guarantees (over approximate value function methods) when we have errors.

Recap

Things to remember

For all π, π', s_0:

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

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

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

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

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

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

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:

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

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

Today:

Natural Policy Gradient:
Global Convergence and Function Approximation

Global convergence for Softmax NPG

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

Global convergence for Softmax NPG

- 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 $T \geq \dfrac{2}{(1-\gamma)^2\epsilon}$.

Global convergence for Softmax NPG

- 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 $T \geq \dfrac{2}{(1-\gamma)^2 \epsilon}$.
- 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 Softmax NPG

- 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 $T \geq \dfrac{2}{(1-\gamma)^2 \epsilon}$.
- 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:
$$\begin{aligned}\log Z_t(s) &= \log \sum_a \pi^{(t)}(a \mid s) \exp(\eta A^{(t)}(s,a)/(1-\gamma)) \\ &\geq \sum_a \pi^{(t)}(a \mid s) \log \exp(\eta A^{(t)}(s,a)/(1-\gamma)) \\ &= \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a \mid s) A^{(t)}(s,a) = 0.\end{aligned}$$

(using Jensen's inequality on the concave function $\log x$.)

Lemma Proof: continued….

By the performance difference lemma,

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

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

$$= \frac{1}{\eta} E_{s\sim d_\mu^{(t+1)}} \mathsf{KL}(\pi_s^{(t+1)}\,||\,\pi_s^{(t)}) + \frac{1}{\eta} E_{s\sim d_\mu^{(t+1)}} \log Z_t(s)$$

$$\geq \frac{1}{\eta} E_{s\sim d_\mu^{(t+1)}} \log Z_t(s) \geq \frac{1-\gamma}{\eta} E_{s\sim\mu} \log Z_t(s),$$

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 $d_\rho^\star$; π_s as shorthand for the vector of $\pi(\cdot \mid s)$

NPG Conv. Proof, Part 1

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

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

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

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

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

NPG Conv. Proof, Part 2

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

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

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

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

NPG Conv. Proof, Part 2

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

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

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

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

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

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

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

NPG Conv. Proof, Part 3

$$
\begin{aligned}
V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) &\le \frac{E_{s\sim d^\star}\mathsf{KL}(\pi^\star_s \,||\, \pi^{(0)})}{\eta T} + \frac{1}{\eta T}\sum_{t=0}^{T-1} E_{s\sim d^\star}\log Z_t(s) \\
&\le \frac{E_{s\sim d^\star}\mathsf{KL}(\pi^\star_s \,||\, \pi^{(0)})}{\eta T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\Big(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\Big) \\
&= \frac{E_{s\sim d^\star}\mathsf{KL}(\pi^\star_s \,||\, \pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T} \\
&\le \frac{\log A}{\eta T} + \frac{1}{(1-\gamma)^2 T}\,.
\end{aligned}
$$

What about Function Approximation?

NPG and variants for log-linear policy classes

What about Function Approximation?

1. Softmax Policy for Tabular MDPs:

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

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

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

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

NPG & Log Linear Policy Classes

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

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

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

- Equivalently, for the same $w_\star$,
  $$\pi(a\,|\,s) \leftarrow \frac{\pi(a\,|\,s)\exp\left(\frac{\eta}{1-\gamma} w_\star \cdot \phi_{s,a}\right)}{Z_s}$$
  (Z_s is the normalizing constant.) Using $\overline{\phi}$ 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:

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

Q-NPG: use Q rather A

(a little nice to interpret for analysis)

- Still log linear class.

- The Q-NPG update:

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

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

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:

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

- 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:

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

Approximate Q-NPG + With a Starting Measure

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

- For a state-action distribution D, define:

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

- 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:

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

- Equivalently,

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

Generic Perturbation Analysis of NPG

NPG regret lemma

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

NPG regret lemma

- Set $\theta^{(0)} = 0$. Consider an arbitrary sequence of weights $w^{(0)}, \ldots, w^{(T)}$, 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 $w^{(0)}, \ldots, w^{(T)}$, 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 $\widetilde{\pi}$ and any state distribution ρ.
  Assume $\log \pi_\theta(a \mid s)$ (for all s, a) is a β-smooth function of θ.
  Define: $\mathrm{err}_t = E_{s \sim \widetilde{d}} E_{a \sim \widetilde{\pi}(\cdot \mid s)} \left[A^{(t)}(s, a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a \mid s) \right]$.
  We have that:
  $$\min_{t<T} \left\{ V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho) \right\} \leq \frac{1}{1-\gamma} \left(W \sqrt{\frac{2\beta \log A}{T}} + \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{err}_t \right).$$
  (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 $w^{(0)}, \ldots, w^{(T)}$, 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 $\widetilde{\pi}$ and any state distribution ρ.
  Assume $\log \pi_\theta(a \mid s)$ (for all s, a) is a β-smooth function of θ.
  Define: $\mathrm{err}_t = E_{s \sim \widetilde{d}} E_{a \sim \widetilde{\pi}(\cdot \mid s)} \left[A^{(t)}(s, a) - w^{(t)} \cdot \nabla_\theta \log \pi^{(t)}(a \mid s)\right]$.
  We have that:
  $$\min_{t<T} \left\{ V^{\widetilde{\pi}}(\rho) - V^{(t)}(\rho) \right\} \leq \frac{1}{1-\gamma} \left(W \sqrt{\frac{2\beta \log A}{T}} + \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{err}_t \right).$$
  (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:
  $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$, 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)

- The approximate version:
  $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$, where. $w^{(t)} \approx$ 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 $w_\star^{(t)} \in$ 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:
  $L(w^{(t)}; \theta^{(t)}, d^{(t)}) - L(w_\star^{(t)}; \theta^{(t)}, d^{(t)}) \le \epsilon_{\mathsf{stat}},$

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

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 $\pi^\star$ (not necessarily optimal).

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

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

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:

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

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

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

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

- Theorem: Fix any state distribution ρ; any comparator policy $\pi^\star$ (not necessarily optimal). With η set appropriately and under the above assumptions, we have that:

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

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

NPG & Neural Policy Classes

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

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

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

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

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

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

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