Policy Gradient: Optimality

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Summary/Today

- Do they PG methods globally converge to an optimal policy?
  $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{(t)}(\mu)$

- Recap
- Today:
  - Wrap up Log Barrier Proof
  - Natural policy gradient

Recap

Things to remember

For all π, π', s_0:

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

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

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

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

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

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

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

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

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

Softmax Gradients

- $\pi_\theta(a \mid s) = \dfrac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$,
- Lemma: For the softmax policy class, we have:

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

Stationarity and Optimality

Stationarity and Optimality

- Log barrier regularized objective:

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

Stationarity and Optimality

- Log barrier regularized objective:

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

- Theorem: (Log barrier regularization) Suppose θ is such that:

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

then we have for all starting state distributions ρ:

$$V^{\pi_\theta}(\rho) \ge V^\star(\rho) - \frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty$$

if $\mu \approx \frac{1}{S}$

where the "distribution mismatch coefficient" is

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

Global Convergence with the Log Barrier

- The smoothness of $L_\lambda(\theta)$ is $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{S}$

Global Convergence with the Log Barrier

- The smoothness of $L_\lambda(\theta)$ is $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{S}$

- Corollary: (Iteration complexity with log barrier regularization)
  Set $\lambda = \frac{\epsilon(1-\gamma)}{2\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty}$ and $\eta = 1/\beta_\lambda$. Starting from any initial $\theta^{(0)}$,
  then for all starting state distributions ρ, we have
  $$\min_{t<T}\left\{V^\star(\rho) - V^{(t)}(\rho)\right\} \le \epsilon \quad \text{whenever} \quad T \ge c\frac{S^2A^2}{(1-\gamma)^6\,\epsilon^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2$$
  (for constant c).

Wrapping up…

Proof, part 1

- The proof consists of showing that: $\max_a A^{\pi_\theta}(s,a) \leq 2\lambda/(\mu(s)S)$ for all states s.

Proof, part 1

- The proof consists of showing that: $\max_a A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all states s.

- To see that this is sufficient, observe that by the performance difference lemma:

$$V^\star(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^\star}(s)\pi^\star(a \mid s)A^{\pi_\theta}(s, a)$$

$$\leq \frac{1}{1-\gamma} \sum_{s} d_\rho^{\pi^\star}(s) \max_{a \in A} A^{\pi_\theta}(s, a)$$

$$\leq \frac{1}{1-\gamma} \sum_{s} 2 d_\rho^{\pi^\star}(s)\lambda/(\mu(s)S)$$

$$\leq \frac{2\lambda}{1-\gamma} \max_s \left(\frac{d_\rho^{\pi^\star}(s)}{\mu(s)} \right).$$

which would then complete the proof.

Proof, part 2

- need to show $A^{\pi_\theta}(s, a) \leq 2\lambda/(\mu(s)S)$ for all (s, a). consider (s, a) where that $A^{\pi_\theta}(s, a) \geq 0$ (else claim is true).

Proof, part 2

- need to show $A^{\pi_\theta}(s,a) \le 2\lambda/(\mu(s)S)$ for all (s,a). consider (s,a) where that $A^{\pi_\theta}(s,a) \ge 0$ (else claim is true).
- Recall $\dfrac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \dfrac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a \mid s)A^{\pi_\theta}(s,a) + \dfrac{\lambda}{S}\left(\dfrac{1}{A} - \pi_\theta(a \mid s)\right)$

Proof, part 2

- need to show $A^{\pi_\theta}(s,a) \le 2\lambda/(\mu(s)S)$ for all (s,a). consider (s,a) where that $A^{\pi_\theta}(s,a) \ge 0$ (else claim is true).
- Recall $\dfrac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \dfrac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a\,|\,s)A^{\pi_\theta}(s,a) + \dfrac{\lambda}{S}\left(\dfrac{1}{A} - \pi_\theta(a\,|\,s)\right)$
- Solving for $A^{\pi_\theta}(s,a)$ in the first step and using $\|\nabla_\theta L_\lambda(\theta)\|_2 \le \epsilon_{opt} \le \lambda/(2SA)$,

$$A^{\pi_\theta}(s,a) = \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a\,|\,s)}\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{S}\left(1 - \frac{1}{\pi_\theta(a\,|\,s)A}\right)\right)$$ — rearranging

$$\le \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a\,|\,s)}\frac{\lambda}{2SA} + \frac{\lambda}{S}\right)$$ — uses

$$\le \frac{1}{\mu(s)}\left(\frac{1}{\pi_\theta(a\,|\,s)}\frac{\lambda}{2SA} + \frac{\lambda}{S}\right)$$ using that $d_\mu^{\pi_\theta}(s) \ge (1-\gamma)\mu(s)$

Proof, part 2

- need to show $A^{\pi_\theta}(s,a) \le 2\lambda/(\mu(s)S)$ for all (s,a). consider (s,a) where that $A^{\pi_\theta}(s,a) \ge 0$ (else claim is true).
- Recall $\dfrac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \dfrac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a\,|\,s)A^{\pi_\theta}(s,a) + \dfrac{\lambda}{S}\left(\dfrac{1}{A} - \pi_\theta(a\,|\,s)\right)$
- Solving for $A^{\pi_\theta}(s,a)$ in the first step and using $\|\nabla_\theta L_\lambda(\theta)\|_2 \le \epsilon_{opt} \le \lambda/(2SA)$,

$$A^{\pi_\theta}(s,a) = \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a\,|\,s)}\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{S}\left(1 - \frac{1}{\pi_\theta(a\,|\,s)A}\right)\right)$$

$$\le \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a\,|\,s)}\frac{\lambda}{2SA} + \frac{\lambda}{S}\right)$$

$$\le \frac{1}{\mu(s)}\left(\frac{1}{\pi_\theta(a\,|\,s)}\frac{\lambda}{2SA} + \frac{\lambda}{S}\right)$$ using that $d_\mu^{\pi_\theta}(s) \ge (1-\gamma)\mu(s)$

- Suppose we could show that $\pi_\theta(a\,|\,s) \ge 1/(2A)$, when $A^{\pi_\theta}(s,a) \ge 0$, then

$$\frac{1}{\mu(s)}\left(\frac{1}{\pi_\theta(a\,|\,s)}\frac{\lambda}{2SA} + \frac{\lambda}{S}\right) \le \frac{1}{\mu(s)}\left(2A\frac{\lambda}{2SA} + \frac{\lambda}{S}\right) = \frac{2\lambda}{\mu(s)S}$$ and the proof is done!

Proof, part 3

- for (s, a) such that $A^{\pi_\theta}(s, a) \geq 0$, we want show $\pi_\theta(a \mid s) \geq 1/(2A)$.

Proof, part 3

- for (s, a) such that $A^{\pi_\theta}(s, a) \geq 0$, we want show $\pi_\theta(a \mid s) \geq 1/(2A)$.

- The gradient norm assumption $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt}$ implies that:

$$\epsilon_{opt} \geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a \mid s) A^{\pi_\theta}(s, a) + \frac{\lambda}{S}\left(\frac{1}{A} - \pi_\theta(a \mid s)\right)$$

$$\geq 0 + \frac{\lambda}{S}\left(\frac{1}{A} - \pi_\theta(a \mid s)\right) \qquad \text{using } A^{\pi_\theta}(s, a) \geq 0$$

by positivity of $A^{\theta}(s,a)$

Proof, part 3

- for (s, a) such that $A^{\pi_\theta}(s, a) \geq 0$, we want show $\pi_\theta(a \mid s) \geq 1/(2A)$.

- The gradient norm assumption $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt}$ implies that:

$$\epsilon_{opt} \geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a \mid s) A^{\pi_\theta}(s, a) + \frac{\lambda}{S}\left(\frac{1}{A} - \pi_\theta(a \mid s)\right)$$

$$\geq 0 + \frac{\lambda}{S}\left(\frac{1}{A} - \pi_\theta(a \mid s)\right) \qquad \text{using } A^{\pi_\theta}(s, a) \geq 0$$

- Rearranging and using our assumption $\epsilon_{opt} \leq \lambda/(2SA)$,

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

Recap

- Softmax policies with exact gradients:
  - Flat gradients could occur if we optimize $V^{\pi_\theta}(s_0)$
  - Coverage: considered optimizing $\max_{\theta \in \Theta} V^{\pi_\theta}(\mu)$
  - Convergence:
    (i) asymptotic convergence for GD
    (ii) poly rate with GD+log barrier regularization

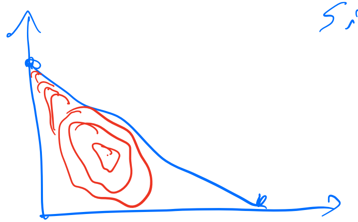
$\mu > 0$

Today:

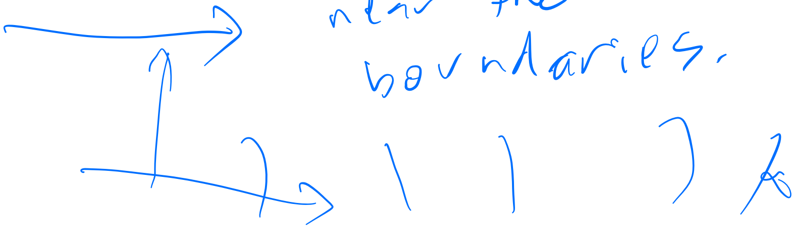
Natural Policy Gradient and Convergence

"π" lives in the simplex

$\|\theta\| \to \infty$ near boundary



pre-condition / variable metric to move more near the boundaries.



The Natural Policy Gradient

- Recall that the Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $E_{x \sim p_\theta}\left[\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top\right]$

The Natural Policy Gradient

$\{p_\theta \mid \theta \in \Theta\}$

- 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 $\{\pi_\theta(\cdot \mid s) \mid s \in S\}$ as:
$$\mathcal{F}_\rho^\theta := E_{s\sim d_\rho^{\pi_\theta}} \underbrace{E_{a\sim\pi_\theta(\cdot|s)}\left[(\nabla \log \pi_\theta(a\mid s)) \nabla \log \pi_\theta(a\mid s)^\top\right]}_{\text{Fisher conditioned on states } s,} .$$

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 $\{\pi_\theta(\cdot \mid s) \mid s \in S\}$ as:
$$\mathcal{F}_\rho^\theta := E_{s \sim d_\rho^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot \mid s)}\left[(\nabla \log \pi_\theta(a \mid s)) \nabla \log \pi_\theta(a \mid s)^\top\right] .$$

- The NPG algorithm performs gradient updates in this induced geometry:
$$\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)})^\dagger \nabla_\theta V^{(t)}(\rho),$$
← estimable in an on-policy manner

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

The Natural Policy Gradient

- Recall that the Fisher information matrix of a parameterized density $p_\theta(x)$ is defined as $E_{x \sim p_\theta}\left[\nabla \log p_\theta(x) \nabla \log p_\theta(x)^\top\right]$
- 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.
- 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)

NPG & Compatible Function Approximation

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

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

NPG & 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 $A^{\pi_\theta}(s,a)$.

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

due to 1st order opt. of $w^\star$

def of $\hat{A}$

NPG & Compatible Function Approximation

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

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

NPG & 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 $F_\mu(\theta)^\dagger \nabla_\theta V^\theta(\mu) = \dfrac{1}{1-\gamma} w^\star,$

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

$$\theta \leftarrow \theta + \eta \left(\frac{1}{1-\gamma} w^*(\theta) \right)$$

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

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

$$:= F_\mu(\theta)$$

- 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)

Vanilla PG. $\theta^{t+1}_{s,a} = \theta^{t}_{s,a} + \frac{\eta}{1-\gamma} \cdot d^{\theta}(s)\, \pi(a|s) A^{\theta}(s,a)$

- Lemma: (Softmax NPG as soft policy iteration) The NPG update is:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$

$\pi^{t}(a|s) \propto e^{\theta^{t}_{s,a}}$

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

Proof

- Lemma: The NPG update is:
  $\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$

Proof

- Lemma: The NPG update is:
  $\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$

- Proof: Recall NPG update is $\frac{1}{1-\gamma} w^\star$ where
  $w^\star \in \text{argmin}_w E_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)} \left[\left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a|s) \right)^2 \right]$

when $w^* = A^{(t)} \implies$ error is 0.

Proof

- Lemma: The NPG update is:
$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$

- Proof: Recall NPG update is $\frac{1}{1-\gamma} w^\star$ where
$$w^\star \in \text{argmin}_w E_{s\sim d_\mu^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)} \left[\left(A^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a \mid s) \right)^2 \right]$$

- What is a minimizer for the the softmax?

Global convergence for NPG

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

Global convergence for NPG

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

- 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 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 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:

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

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

Lemma Proof: continued….

By the performance difference lemma,

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

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

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

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

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 $d^\star_\rho$; 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,

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

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

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

NPG Conv. Proof, Part 2

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

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

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

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

- 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

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