Policy Gradient: Optimality

Wen Sun

CS 6789: Foundations of Reinforcement Learning
Recap
Policy Gradient Derivation

e.g., Reinforce, Natural Policy Gradient, TRPO, PPO:
(Williams 92, Kakade 02, Schulman et al 15, 17)

\[
\pi_\theta(a \mid s) = \pi(a \mid s; \theta) \quad J(\pi_\theta) = \mathbb{E}_{\pi_\theta} \left[ \sum_{h=0}^{\infty} \gamma^h r_h \right]
\]

\[
\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\pi_\theta) \big|_{\theta=\theta_t}
\]

Main question for today’s lecture:
how to compute the gradient?

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

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

Draw \( h \propto \gamma^h \), \textbf{roll-in} \( \pi_\theta \) to generate \( s_h, a_h \sim \mathbb{P}^\pi_\theta \)

\textbf{Roll-out} \( \pi_\theta \) from \( (s_h, a_h) \) : terminate with prob \( 1 - \gamma \), \( \hat{Q}^\pi_\theta(s_h, a_h) = \sum_{\tau = h}^{t \geq h} r_\tau \)

Unbiased estimate: \( \nabla_\theta \ln \pi_\theta(a_h | s_h) \hat{Q}^\pi_\theta(s_h, a_h) \)
Policy Gradient: Examples of Policy Parameterization (discrete actions)

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. 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 \( f_\theta : S \times A \mapsto \mathbb{R} \)

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

$J(\pi_\theta)$ is non-convex (see example in the monograph)

Def of $\beta$-smooth:

$$\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$$

$$\left| J(\theta) - J(\theta_0) - \nabla_\theta J(\theta_0)\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \|\theta - \theta_0\|_2^2, \forall \theta, \theta_0$$

[Theorem] If $J(\theta)$ is $\beta$-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\theta_t)$

where $\mathbb{E} \left[ \nabla_\theta J(\theta_t) \right] = \nabla_\theta J(\theta_t), \quad \mathbb{E} \left[ \|\nabla_\theta J(\theta_t)\|_2^2 \right] \leq \sigma^2$,

then:

$$\mathbb{E} \left[ \frac{1}{T} \sum_t \|\nabla_\theta J(\theta_t)\|_2^2 \right] \leq O \left( \sqrt{\beta \sigma^2 / T} \right)$$
Today (+future):
When do PG methods converge to a global optima?
(+ what about function approximation?)
Today:

• Let’s consider using exact gradients.
  • This allows us to ignore estimation issues
  • Let’s focus on “complete” parameterizations (e.g. the “tabular” case)
    \( \Pi \) contains all stochastic policies (e.g. softmax)

• I: Landscape of the problem
  • As a general non-convex optimization problem:
    do small gradients imply good performance?
  • what about “exploration”?
• II: Global convergence results
PG as non-convex optimization
Convergence to Stationary Points of GD

$J(\pi_\theta)$ is non-convex (see example in the AJKS)
Convergence to Stationary Points of GD

\( J(\pi_\theta) \) is non-convex (see example in the AJKS)

- Def of a \( \beta \)-smooth function \( F \):
  \[
  \| \nabla_\theta F(\theta) - \nabla_\theta F(\theta_0) \|_2 \leq \beta \| \theta - \theta_0 \|_2
  \]
  which implies:
  \[
  \left| F(\theta) - F(\theta_0) - \nabla_\theta F(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \| \theta - \theta_0 \|_2^2
  \]
Convergence to Stationary Points of GD

\[ J(\pi_\theta) \] is non-convex (see example in the AJKS)

- **Def of a \( \beta \)-smooth function \( F \):**
  \[ \| \nabla_\theta F(\theta) - \nabla_\theta F(\theta_0) \|_2 \leq \beta \| \theta - \theta_0 \|_2 \]

  which implies:
  \[
  \left| F(\theta) - F(\theta_0) - \nabla_\theta F(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \| \theta - \theta_0 \|_2^2
  \]

- **Proposition:** (stationary point convergence) Assume \( F(\theta) \) is \( \beta \)-smooth.
  Suppose we run gradient ascent: \( \theta_{t+1} = \theta_t + \eta \nabla_\theta F(\theta_t) \), with \( \eta = 1/(2\beta) \). Then:
  \[
  \min_{t \leq T} \| \nabla_\theta F(\theta_t) \|_2^2 \leq \frac{2\beta \left( \max_\theta F(\theta) - F(\theta_0) \right)}{T}
  \]
Convergence to Stationary Point

Proposition: (stationary point convergence) Assume $F(\theta)$ is $\beta$-smooth. Suppose we run gradient ascent: $\theta_{t+1} = \theta_t + \eta \nabla_{\theta} F(\theta_t)$, with $\eta = 1/(2\beta)$. Then:

$$
\min_{t \leq T} \| \nabla_{\theta} F(\theta_t) \|^2_2 \leq \frac{2\beta (F(\theta^*) - F(\theta_0))}{T}
$$

$$
\begin{align*}
|F(\theta_{t+1}) - F(\theta_t) - \nabla_{\theta} F(\theta_t)^\top (\theta_{t+1} - \theta_t)| &\leq \frac{\beta}{2} \| \theta_{t+1} - \theta_t \|^2 \\
\Rightarrow |F(\theta_{t+1}) - F(\theta_t) - \eta \nabla_{\theta} F(\theta_t)^\top \nabla_{\theta} F(\theta_t)| &\leq \frac{\beta}{2} \eta^2 \| \nabla_{\theta} F(\theta_t) \|^2 \\
\Rightarrow \eta \| \nabla_{\theta} F(\theta_t) \|^2 &\leq F(\theta_{t+1}) - F(\theta_t) + \frac{\beta}{2} \eta^2 \| \nabla_{\theta} F(\theta_t) \|^2 \\
\Rightarrow \frac{1}{2\beta} \| \nabla_{\theta} F(\theta_t) \|^2 &\leq F(\theta_{t+1}) - F(\theta_t) \quad \text{using } \eta \leq \frac{1}{\beta} \\
\Rightarrow \min_{t \leq T} \| \nabla_{\theta} F(\theta_t) \|^2 &\leq \frac{1}{T} \sum_t \| \nabla_{\theta} F(\theta_t) \|^2 \leq \sum_t (F(\theta_{t+1}) - F(\theta_t)) \leq \frac{2\beta (F(\theta^*) - F(\theta_0))}{T}
\end{align*}
$$
A “landscape” result
(and “exploration”)
Vanishing Gradients and Saddle Points

The following result not only shows that the gradient is exponentially small in \( |s-a| \) but also that the higher order derivatives are small.

We do not prove this lemma here (see Section 9.5). The lemma illustrates that lack of good exploration can indeed be a problem for MDPs.

Concretely, consider the chain MDP of length \( H+1 \) (obtained by permuting the actions) where all the probabilities for states, actions, and rewards are the same as those in Figure 0.1.

Note that this lemma is not about random exploration strategies; there is an exponentially small norm of the tensor.

The operator norm of a \( \mathcal{H}/d \)-order tensor \( \mathcal{H} \) is defined as

\[
\| \mathcal{H} \| = \max_{k=1,2,\ldots,\mathcal{H}} \sup_{i} \| \mathcal{H}^{(i)} \|_k
\]

Rewards are \( \mathcal{H}/d \) (componentwise) and \( \mathcal{H}/d+2 \) (for all \( i \)).

Furthermore, the discount factor is 1, i.e., \( \mathcal{H}/d=1 \). Rewards are defined as \( \mathcal{H}/d+1 \) everywhere other than

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1}
\]

Note we do not over-parameterize the policy. For this MDP, we consider a policy where

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

where all the probabilities for states, actions, and rewards are 1, and with the direct policy parameterization (with \( \mathcal{H}/d+1 \) parameters, as described in the text).

To understand the necessity of optimizing under a distribution

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

we have

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

Suppose

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

above). Suppose

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

The argument that some condition on the state distribution of

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

is such that

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

\( \mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho \) does not visit advantageous states often enough. Furthermore, this lemma also suggests that varied results

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

are highly suboptimal in terms of their value. Below, we give a more quantitative version of this intuition, which demonstrates that even if

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

states, a policy that does not visit rewarding states will have zero gradient, even though it is arbitrarily suboptimal.

The chain MDP of Figure 0.1, is a common example where

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

due to that the higher order derivatives are small.

While the chain MDP of Figure 0.1, is a common example where

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

we have

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

in the non-convex optimization literature, on escaping from saddle points, do not directly imply global convergence.

Vanishing gradients at suboptimal parameters

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]

Choosing all actions with reasonable probabilities (and hence the agent will visit all states if the MDP is connected),

\[
\mathcal{H}/s, a_{\mathcal{H}/d+1} = \rho
\]
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)
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 $\nabla_\theta^k V^{\pi_\theta}(s_0)$.)
“Vanilla” PG for the Softmax
Let’s consider having a staring state distribution with “coverage”
Let’s consider having a staring state distribution with “coverage”

• Given our a starting distribution $\rho$ over states, recall our objective is:
  \[
  \max_{\theta \in \Theta} V^{\pi_{\theta}}(\rho).
  \]
  where $\{ \pi_{\theta} \mid \theta \in \Theta \subset \mathbb{R}^d \}$ is some class of parametric policies.
Let’s consider having a staring state distribution with “coverage”

• Given our a starting distribution $\rho$ over states, recall our objective is:
  \[
  \max_{\theta \in \Theta} \ V^{\pi_\theta}(\rho),
  \]
  where $\{\pi_\theta | \theta \in \Theta \subset \mathbb{R}^d\}$ is some class of parametric policies.

• While we are interested in good performance under $\rho$, it is helpful to optimize under a different measure $\mu$. Specifically, consider optimizing: $V^{\pi_\theta}(\mu)$, i.e.
  \[
  \max_{\theta \in \Theta} V^{\pi_\theta}(\mu),
  \]
  even though our ultimate goal is performance under $V^{\pi_\theta}(\rho)$. 
notation (+ overloading)

Today: we will use $d_{s_0}^\pi$ for a state distribution measure.
(it should be clear from context how we use it).

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

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

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

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

Advantage function: $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$
The Softmax Policy Class
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).
The Softmax Policy Class

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

(where the number of parameters is SA).

We have that:

\[ \frac{\partial \log \pi_\theta(a \mid s)}{\partial \theta_{s',a'}} = 1[s = s'] \left( 1[a = a'] - \pi_\theta(a' \mid s) \right) \]

where \( 1[\cdot] \) is the indicator function.
The Softmax Policy Class

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

(where the number of parameters is SA).

- We have that:
  \[ \frac{\partial \log \pi_\theta(a \mid s)}{\partial \theta_{s',a'}} = \mathbf{1}[s = s'] \left( \mathbf{1}[a = a'] - \pi_\theta(a' \mid s) \right) \]
  where \( \mathbf{1}[\cdot] \) is the indicator function.

- **Lemma:** For the softmax policy class, we have:
  \[ \frac{\partial V_{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1 - \gamma} \partial_{\mu}^{\pi_\theta}(s) \pi_\theta(a \mid s) A^{\pi_\theta}(s, a) \]
Proof

\[
\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = E_{\tau \sim \Pr^{\pi_\theta}_\mu} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} \ln \pi_\theta(a | s) A^{\pi_\theta}(s, a) \right]
\]

\[
= E_{\tau \sim \Pr^{\pi_\theta}_\mu} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbf{1}[s_t = s] \left( \mathbf{1}[a_t = a] A^{\pi_\theta}(s, a) - \pi_\theta(a | s) A^{\pi_\theta}(s, a_t) \right) \right]
\]

\[
= E_{\tau \sim \Pr^{\pi_\theta}_\mu} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbf{1}[(s_t, a_t) = (s, a)] A^{\pi_\theta}(s, a) \right] + \pi_\theta(a | s) \sum_{t=0}^{\infty} \gamma^t E_{\tau \sim \Pr^{\pi_\theta}_\mu} \left[ \mathbf{1}[s_t = s] A^{\pi_\theta}(s, a_t) \right]
\]

\[
= \frac{1}{1 - \gamma} E_{(s', a') \sim d^{\pi_\theta}} \left[ \mathbf{1}[(s', a') = (s, a)] A^{\pi_\theta}(s, a) \right] + 0
\]

\[
= \frac{1}{1 - \gamma} d^{\pi_\theta}(s, a) A^{\pi_\theta}(s, a),
\]
Remember: The Performance Difference Lemma

For all $\pi, \pi', s_0$:

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

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

• The update rule for gradient ascent is:

\[ \theta^{(t+1)} = \theta^{(t)} + \eta \nabla_{\theta} V^{(t)}(\mu) \]
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 \( \theta_t \to \infty \)
    \( (\pi_t \text{ becoming any deterministic policy implies } \theta_t \text{ approaches a stationary point}) \)
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 \( \theta_t \to \infty \)
    (\( \pi_t \) becoming any deterministic policy implies \( \theta_t \) approaches a stationary point)

• **Theorem:** Assume the \( \mu \) 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 \( s \), \( V^{(t)}(s) \to V^*(s) \), as \( t \to \infty \).
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 \( \theta_t \to \infty \)
    \( (\pi_t \text{ becoming any deterministic policy implies } \theta_t \text{ approaches a stationary point}) \)

• **Theorem:** Assume the \( \mu \) 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 \( s, V^{(t)}(s) \to V^*(s) \), as \( t \to \infty \).

• Comments:
  • rate could be exponentially slow in \( S, H \).
  • need \( \mu > 0 \) is necessary.
PG+Log Barrier Regularization
(for the softmax)
Log Barrier Regularization
Log Barrier Regularization

- Relative-entropy for distributions $p, q$ is: $\text{KL}(p, q) := E_{x \sim p}[-\log q(x)/p(x)]$. 
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 $\lambda$-regularized objective:
  
  \[
  L_\lambda(\theta) := V_{\pi_\theta}(\mu) - \lambda E_{s \sim \text{Unif}_S} [\text{KL}(\text{Unif}_A, \pi_\theta(\cdot | s))] 
  \]

  \[
  = V_{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a | s) + \lambda \log A
  \]
Log Barrier Regularization

- Relative-entropy for distributions $p,q$ is: $KL(p, q) := E_{x \sim p}[-\log q(x)/p(x)]$.

- Consider the log barrier $\lambda$-regularized objective:
  
  $L_{\lambda}(\theta) := V_{\pi_{\theta}(\mu)} - \lambda E_{s \sim \text{Unif}_{S}}[KL(\text{Unif}_{A}, \pi_{\theta} \cdot | s))]$

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

- Gradient Ascent:
  
  $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_{\theta} L_{\lambda}(\theta^{(t)})$
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 $\lambda$-regularized objective:
  
  $$L_\lambda(\theta) := V_{\pi_\theta}(\mu) - \lambda E_{s \sim \text{Unif}_S} \left[ \text{KL}(\text{Unif}_A, \pi_{\theta}(\cdot | s)) \right]$$
  
  $$= V_{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a | s) + \lambda \log A$$

- Gradient Ascent:
  
  $$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_{\theta} L_\lambda(\theta^{(t)})$$

- Do small gradients imply a globally optimal policy?
Stationarity and Optimality
Stationarity and Optimality

• Log barrier regularized objective:

\[ L_\lambda(\theta) = V^{\pi_\theta}(\mu) + \frac{\lambda}{SA} \sum_{s,a} \log \pi_\theta(a | s) + \lambda \log A \]
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 \( \theta \) is such that:
  \[ \|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{opt} \text{ and } \epsilon_{opt} \leq \lambda/(2SA) \]
  then we have for all starting state distributions \( \rho \):
  \[ V^{\pi_\theta}(\rho) \geq V^*(\rho) - \frac{2\lambda}{1-\gamma} \left\| \frac{d\pi^*}{\rho} \right\|_\infty \]
  where the “distribution mismatch coefficient” is
  \[ \left\| \frac{d\pi^*}{\rho} \right\|_\infty = \max_s \left( \frac{d\pi^*(s)}{\mu(s)} \right) \]
  (componentwise division notation)
Global Convergence with the Log Barrier
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 $\lambda = \frac{\epsilon(1 - \gamma)}{2 \left\| \frac{d^*_\rho}{\mu} \right\|_\infty}$ and $\eta = 1/\beta_\lambda$. Starting from any initial $\theta^{(0)}$,

  then for all starting state distributions $\rho$, we have

  $$\min_{t<T} \{ V^*(\rho) - V^{(t)}(\rho) \} \leq \epsilon \quad \text{whenever} \quad T \geq c \frac{S^2A^2}{(1 - \gamma)^6 \epsilon^2} \left\| \frac{d^*_\rho}{\mu} \right\|_\infty^2$$

  (for constant $c$).
Remember: The Performance Difference Lemma

For all $\pi, \pi', s_0$:

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

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

- The proof consists of showing that: \( \max_a A^{\pi_0}(s, a) \leq \frac{2\lambda}{(\mu(s))} \) 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:

\[
V^*(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1 - \gamma} \sum_{s,a} d^\pi_\theta^*(s) \pi^*(a \mid s) A^\pi_\theta(s, a)
\]

\[
\leq \frac{1}{1 - \gamma} \sum_s d^\pi_\theta^*(s) \max_{a \in A} A^\pi_\theta(s, a)
\]

\[
\leq \frac{1}{1 - \gamma} \sum_s 2d^\pi_\theta^*(s) \lambda/\mu(s)S
\]

\[
\leq \frac{2\lambda}{1 - \gamma} \max_s \left( \frac{d^\pi_\theta^*(s)}{\mu(s)} \right).
\]

which would then complete the proof.
Proof, part 2
Proof, part 2

- need to show $A^{\pi_0}(s, a) \leq \frac{2\lambda}{\mu(s)S}$ for all $(s, a)$. consider $(s, a)$ where that $A^{\pi_0}(s, a) \geq 0$ (else claim is true).
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).

- Recall

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

- Recall

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

- Solving for $A^\pi_\theta(s, a)$ in the first step and using $\| \nabla_\theta L_\lambda(\theta) \|_2 \leq c_{opt} \leq \lambda/(2SA)$,

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

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

$$\leq \frac{1}{\mu(s)} \left( \frac{1}{\pi_\theta(a \mid s)} \frac{\lambda}{2SA} + \frac{\lambda}{S} \right) \quad \text{using that } d_{\mu}^\pi_\theta(s) \geq (1 - \gamma)\mu(s)$$
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).
- Recall $\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1 - \gamma}d_\mu^{\pi_\theta}(s)\pi_\theta(a \mid s)A^{\pi_\theta}(s, a) + \frac{\lambda}{S}\left(\frac{1}{A} - \pi_\theta(a \mid s)\right)$
- Solving for $A^{\pi_\theta}(s, a)$ in the first step and using $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon_{\text{opt}} \leq \lambda/(2SA)$,
  
  $A^{\pi_\theta}(s, a) = \frac{1 - \gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a \mid s)}\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{S}(1 - \frac{1}{\pi_\theta(a \mid s)A})\right)$

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

  $\leq \frac{1}{\mu(s)}\left(\frac{1}{\pi_\theta(a \mid s)}\frac{\lambda}{2SA} + \frac{\lambda}{S}\right)$

  using that $d_\mu^{\pi_\theta}(s) \geq (1 - \gamma)\mu(s)$

- Suppose we could show that $\pi_\theta(a \mid s) \geq 1/(2A)$, when $A^{\pi_\theta}(s, a) \geq 0$, then
  
  $\frac{1}{\mu(s)}\left(\frac{1}{\pi_\theta(a \mid s)}\frac{\lambda}{2SA} + \frac{\lambda}{S}\right) \leq \frac{1}{\mu(s)}\left(2A\frac{\lambda}{2SA} + \frac{\lambda}{S}\right) = \frac{2\lambda}{\mu(s)S}$

  and the proof is done!
Proof, part 3
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:

\[
\epsilon_{opt} \geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1 - \gamma}d_{\mu_\theta}(s)\pi_\theta(a \mid s)A^{\pi_\theta}(s, a) + \frac{\lambda}{S} \left( \frac{1}{A} - \pi_\theta(a \mid s) \right) \\
\geq 0 + \frac{\lambda}{S} \left( \frac{1}{A} - \pi_\theta(a \mid s) \right) \quad \text{using } A^{\pi_\theta}(s, a) \geq 0
\]
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:

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

\[
\geq 0 + \frac{\lambda}{S} \left( \frac{1}{A} - \pi_\theta(a \mid s) \right)
\]

using \(A^{\pi_\theta}(s, a) \geq 0\)

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

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