Global Optimality Guarantees For Policy Gradient Methods

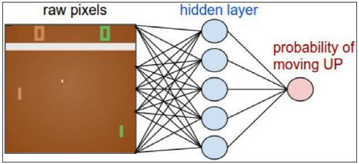
Daniel Russo

April 13, 2020

Columbia University

Table of Contents

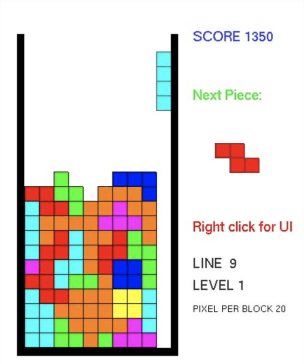
Policy Gradient Methods



REINFORCE:
Initialize policy parameters θ arbitrarily
for each episode $\{s_1, a_1, r_2, \cdots, s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$ **do**
  for $t = 1$ to $T - 1$ **do**
    $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) G_t$
  endfor
endfor
return θ

https://twitter.com/CovariantAI/status/1232396047948763136
https://youtu.be/kVmp0uGtShk

Policy Search In Tetris



- One approach is a Linear Approximation: $J(s) \approx \phi(s)^\top \theta$.
- $\phi(s)$ encodes features. E.g. column heights, inter-column height differences, max height etc.
- Each (s, a) associated with known successor state s'_a
- Could estimate θ, then play $\text{argmax}_a \, \phi(s'_a)^\top \theta$.
- Often better to directly search over the subclass of policies $\mu_\theta(s, a) \propto \exp(\phi(s'_a)^\top \theta)$

Policy Gradient Methods: Setup

Direct optimization over a policy class by local search.

- Cost-to-go function: $J_\mu(s) = \mathbb{E}_\mu \left[\sum_{t=0}^{\infty} \alpha^t g(s_t, \mu(s_t)) \mid s_0 = s\right]$
- Scalar loss: $\ell(\mu) = \mathbb{E}_{s\sim\nu}[J_\mu(s)]$.
  - Assume ν has full support. (*This is essential*!)
- Parameterized policies: $M_\Theta = \{\mu_\theta : \theta \in \Theta\} \subset M$.
  - $\Theta \subset \mathbb{R}^d$ is convex. $M = \{\text{Markov policies}\}$.
  - Overloaded notation: $\ell(\theta) \equiv \ell(\mu_\theta)$
- (Stochastic) Gradient Descent on $\ell(\cdot)$,

$$\theta_{t+1} = \theta_t - \gamma_t \left(\nabla \ell(\theta) + \text{noise}\right) \quad t = 1, 2, \ldots$$

Contrast to policy iteration

Policy gradient methods:

1. Make soft updates to policies
2. Aim to directly minimize a global loss function $\ell(\mu)$ rather than solve the changing surrogate problems $\min_u Q_{\mu_k}(s, u)$.
    - We'll see a connection, though, when we compute the gradients of $\ell(\cdot)$.
3. Approximation is through a policy class rather than a class of cost-to-go functions.
    - There is a connection to approximate PI when considering the set of policies that are (approximately) greedy with respect to a cost-to-go function. (See the tetris example above)
    - Actor-critic methods, which we'll get to later, are a hybrid.

Computing a stochastic gradient via simulation

Justified in the next two slides.

An unbiased estimate of $\nabla_\theta \ell(\mu_\theta)$ for a stochastic policy

Sample $T \sim \mathrm{Geometric}(1-\alpha)$.

Sample $s_0 \sim \nu$

Apply μ_θ for T periods from s_0.

Observe trajectory: $\tau = (s_0, u_0, c_0, \ldots, s_T, u_T, c_T, s_{T+1})$.

Observe total cost: $c(\tau) = c_0 + \cdots + c_T$.

Return $c(\tau) \sum_{t=0}^{T} \nabla_\theta \log \mu_\theta(u_t | s_t)$

Many variants and improvements:

- Infintesimal perturbation analysis
- Variance reduction by adding baselines
- Actor critic methods

Score function gradient estimator (Finite horizon case)

Consider a stochastic policy $\mu_\theta(u|s)$
For $\tau = (s_0, u_0, \cdots s_T, u_T)$, set $G(\tau) = \sum_{t=0}^{T} g(s_t, u_t)$
Let $P(\tau;\theta)$ denote its probability under μ_θ.

$$
\begin{aligned}
\nabla_\theta \, \mathbb{E}_\theta \left[\sum_{t=0}^{T} g(s_t, u_t) \right] &= \nabla_\theta \sum_\tau G(\tau) P(\tau;\theta) \\
&= \sum_\tau G(\tau) \nabla_\theta \nabla_\theta \, P(\tau;\theta) \\
&= \sum_\tau G(\tau) \left(\nabla_\theta \, \log P(\tau;\theta) \right) P(\tau;\theta) \\
&= \mathbb{E}_\theta \left[G(\tau) \left(\nabla_\theta \, \log P(\tau;\theta) \right) \right] \\
&= \mathbb{E}_\theta \left[G(\tau) \left(\nabla_\theta \, \log \prod_{t=0}^{T} \mu_\theta(u_t|s_t) P(s_{t+1}|u_t, s_t) \right) \right] \\
&= \mathbb{E}_\theta \left[G(\tau) \left(\sum_{t=0}^{T} \nabla_\theta \, \log \mu_\theta(u_t|s_t) \right) \right]
\end{aligned}
$$

Score function gradient estimator (infinite horizon case)

Take $T \sim \mathrm{Geom}(1-\alpha)$, so $\mathbb{P}(T \geq t) = \alpha^t$.

$$\ell(\theta) = \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} \alpha^t g(s_t, u_t) \right] = \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} \mathbb{P}(T \geq t) g(s_t, u_t) \right]$$
$$= \mathbb{E}_\theta \left[\sum_{t=0}^{T} g(s_t, u_t) \right]$$

One can show

$$\nabla \ell(\theta) = \mathbb{E}_\theta \left[\left(\sum_{t=0}^{T} g(s_t, u_t) \right) \left(\sum_{t=0}^{T} \nabla_\theta \log \mu_\theta(u_t | s_t) \right) \right]$$

You may encounter a slight improvement that recognizes some terms have mean zero and writes

$$\nabla \ell(\theta) = \mathbb{E}_\theta \left[\left(\sum_{t=0}^{T} \nabla_\theta \log \mu_\theta(u_t | s_t) \left(\sum_{i=t}^{T} g(s_t, u_t) \right) \right) \right]$$

Challenges with policy gradient methods

This class will cover issue (1). The rest next week.

1. Nonconvexity of the loss function $\ell(\mu)$.

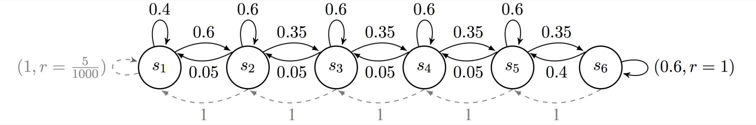
    Due to multi-period objective.

2. Unnatural policy parameterization $\theta \mapsto \mu_\theta$.

    Poor parameterization can lead to vanishing gradients in single period problems. Natural gradient methods are invariant to change of coordinates.

3. Insufficient exploration.

    We rely on the initial distribution $\nu(\cdot)$.



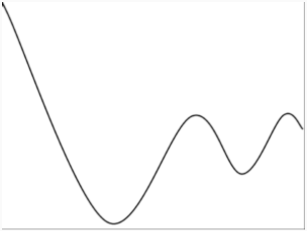
4. Large variance of stochastic gradients.

    Baselines and actor critic methods help.

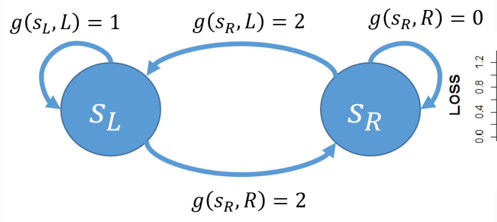
Table of Contents

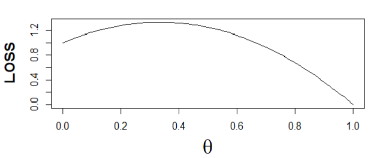
So we've reduced dynamic programming to black-box nonconvex optimization?

- PG is applicable to almost any decision problem.
- **But...** $\mu \mapsto \ell(\mu)$ is nonconvex.
  - Even for classical dynamic programming problems.
- Policy gradient is widely understood to converge toward first-order stationary points (i.e. w/ $\nabla_\theta \, \ell(\mu_\theta) = 0$).
  - Almost no guarantees on the quality of stationary points.
  - There are bad local minima in simple problems.



A Bad Example For Policy Gradient



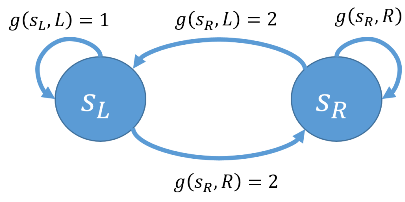


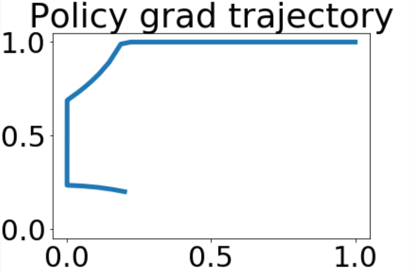
- $\mu_\theta(s_L) = \mu_\theta(S_R) = \theta =$ "Probability of Right."

Policy gradient gets stuck in a bad local minimum

Despite containing the optimal policy, the policy class is not rich enough to allow for a local improvement.

A Good Example For Policy Gradient





- $\mu_\theta(S_L) = \theta_L =$ "Probability of Right from left state."
- $\mu_\theta(S_R) = \theta_R =$ "Probability of Right from right state."

Policy gradient reaches the global optimum

The full policy class is rich enough to allow for local improvement... but this is unsatisfying.

Table of Contents

Warmup with the natural special case

- $\mathcal{S} = \{1, \cdots, n\}$ is a finite set.
- $U(s) = U = \Delta^{k-1}$ contains all probability distributions over k deterministic actions.
  - For $u \in U$, $g(s,u) = \sum_{i=1}^{k} g(s,i)$, $P_{ss'}(u) = \sum_{i=1}^{k} u_i P_{ss'}(i)$

Natural parameterization

- $\mu_\theta(s) = \theta_s \in \Delta^{k-1}$ associates each state with a probability distribution over actions.
- Drop the notational dependence on θ . . ., writing $\mu \in \mathbb{R}^{n \times k}$
- Stationary randomized policies: $M = \{\pi \in \mathbb{R}_{+}^{n \times k} : \sum_{i=1}^{k} \mu_{s,i} = 1 \ \forall s \in \{1, \cdots, n\}\}$.

First order methods

First order methods solve $\min_{\mu \in M} \ell(\mu)$ by iteratively minimizing (regularized) local linear approximations to $\ell(\cdot)$.

Example: Projected Gradient

$$\begin{aligned}\mu_{k+1} &= \Pi_{2,M}\left(\mu_k - \gamma_k \nabla \ell(\mu_k)\right) \\ &= \underset{\mu \in M}{\operatorname{argmin}} \;\; \ell(\mu_k) + \langle \nabla \ell(\mu_k)\,,\, \mu - \mu_k \rangle + \frac{1}{2\gamma_k}\left\|\mu - \mu_k\right\|_2^2\end{aligned}$$

Projection onto the probability simplex involves a simple soft-thresholding operation.

It is also natural to consider conditional gradient descent or exponentiated gradient descent (... a special case of mirror descent).

Convergence to first order stationary points (1)

Definition: $\ell(\cdot)$ is L–smooth if $\nabla^2 \ell(\mu) \preceq L^2 I$

Informal Proposition (Asymptotic convergence)

If $\ell(\cdot)$ is L–smooth, and stepsizes are set appropriately, then for any limit point μ_∞ of $\{\mu_k\}$.

$$\langle \nabla \ell(\mu_k)\,,\, \mu - \mu_\infty \rangle \geq 0 \qquad \forall \mu \in M$$

Proposition (Convergence rate)

Set $e_k = \min_{\mu \in M} \langle \nabla \ell(\mu_k)\,,\, \mu - \mu_k \rangle$. If $\gamma_k = 1/L$ then,

$$\min_{k \leq K} |e_k| \leq \sqrt{\frac{\ell(\mu_0) - \ell(\mu^*)}{2LR^2K}}$$

where $R = \sup_{\mu, \mu'} \|\mu - \mu'\|_2$.

Similar guarantees apply with stochastic gradient evaluations

Proof of rate to reach approximate stationary point (1)

Descent Lemma

If $\ell(\cdot)$ is L smooth, then $\ell(\bar{\mu}) \leq \ell(\mu) + \langle \nabla \ell(\mu)\,,\, \bar{\mu} - \mu \rangle + \frac{L}{2}\|\bar{\mu} - \mu\|_2^2$

For $\gamma_k \leq \frac{1}{L}$, projected gradient descent *minimizes a quadratic upper bound* on $\ell(\cdot)$. We find

$$\ell(\mu_{k+1}) \leq \min_{\mu \in M} \ \ell(\mu_k) + \langle \nabla \ell(\mu_k)\,,\, \mu - \mu_k \rangle + \frac{1}{2\gamma_k}\|\mu - \mu_k\|_2^2$$

Take $\mu^+ = \operatorname{argmin}_{\mu \in M} \langle \nabla \ell(\mu_k)\,,\, \mu - \mu_k \rangle$ and pick $\gamma_k = 1/L$.
Minimizing only over the line segment connecting μ_k, μ^+ gives

$$\begin{aligned}
\ell(\mu_{k+1}) &\leq \min_{t \in [0,1]} \ell(\mu_k) + t\langle \nabla \ell(\mu_k)\,,\, \mu^+ - \mu_k \rangle + \frac{Lt^2}{2}\left\|\mu^+ - \mu_k\right\|_2^2 \\
&\leq \min_{t \in [0,1]} \ell(\mu_k) + t\langle \nabla \ell(\mu_k)\,,\, \mu^+ - \mu_k \rangle + \frac{t^2}{L2} R^2 \\
&= \ell(\mu_k) + \frac{1}{2LR^2}\left(\min_{\mu \in M} \langle \nabla \ell(\mu_k)\,,\, \mu - \mu_k \rangle\right)^2
\end{aligned}$$

Proof of rate to reach approximate stationary point (2)

Set $e_k = \min_{\mu \in M} \langle \nabla \ell(\mu_k)\,,\, \mu - \mu_k \rangle$ to be the distance from stationarity. We showed

$$\min_{k \leq K} e_k^2 \leq \frac{1}{K} \sum_{k=1}^{K} e_k^2 \leq \frac{2LR^2}{K} \sum_{k=1}^{K} \left(\ell(\mu_{k+1}) - \ell(\mu_k) \right) \leq \frac{\ell(\mu_0) - \ell(\mu^*)}{2LR^2 K}$$

Or

$$\min_{k \leq K} |e_k| \leq \sqrt{\frac{\ell(\mu_0) - \ell(\mu^*)}{2LR^2 K}}$$

Despite nonconvexity, there are suboptimal stationary points for policy gradient with a full policy class

If $\nu(s) > 0$ for all s the RHS is zero if and only if $J_\mu = TJ_\mu$.

Lemma: If μ^+ is a policy iteration update to μ, then

$$\frac{d}{d\gamma}\ell(\mu + \gamma(\mu^+ - \mu))\Big|_{\gamma=0} \leq -\|J_\mu - TJ_\mu\|_{1,\nu}$$

Proof: Set $\mu_\gamma = \mu + \gamma(\mu^+ - \mu)$. One can show

$$T_{\mu_\gamma} J_\mu = (1-\gamma) T_\mu J_\mu + \gamma T_{\mu^+} J_\mu = J_\mu - \gamma \left(TJ_\mu - J_\mu \right) \preceq J_\mu$$

Monotonicity implies $J_{\mu_\gamma} \preceq \cdots \preceq T_{\mu_\gamma} J_\mu \preceq J_\mu$. Hence,

$$\frac{J_{\mu_\gamma} - J_\mu}{\gamma} \preceq \frac{T_{\mu_\gamma} J_\mu - J_\mu}{\gamma} = TJ_\mu - J_\mu$$

Left multiplying by ν, and using $TJ_\mu \preceq J_\mu$

$$\frac{\ell(\mu_\gamma) - \ell(\mu)}{\gamma} \leq \|TJ_\mu - J_\mu\|_{1,\nu}$$

Taking $\gamma \to 0$ gives the result.

Table of Contents

Weighted policy iteration

Policy iteration is

$$\mu_{k+1} \in \operatorname*{argmin}_{\mu \in M} T_\mu J_{\mu_k},$$

where the minimization is performed elementwize. As long as $w(s) > 0$, policy iteration can be written minimizing a scalarized objective:

$$\mu_{k+1} \in \operatorname*{argmin}_{\mu \in M} w \left(T_\mu J_{\mu_k} \right)$$

Define the weighted policy iteration cost function

$$\mathcal{B}(\bar{\mu}|w, J_\mu) = w T_{\bar{\mu}} J_\mu = \sum_s w(s)(T_{\bar{\mu}} J_\mu)(s) = \sum_s w(s) Q_\mu(s, \bar{\mu}(s))$$

We will connect policy gradient to the constrained policy iteration scheme

$$\mu_{\theta_{k+1}} = \operatorname*{argmin}_{\mu_\theta \in M_\Theta} \mathcal{B}(\mu_\theta | w, J_{\mu_{\theta_k}})$$

that restricts to the parameterized policy class when minimizing.

A policy gradient formula

*Overload notation: $\mathcal{B}(\bar{\theta}|\nu, J) \equiv \mathcal{B}(\mu_\theta|\nu, J)$.,
and $\nu_\mu := \sum_{k=0}^{\infty} \alpha^k \nu P_\mu^k$.

Under appropriate smoothness conditions

$$\nabla \ell(\theta) = \nabla_{\bar{\theta}} \mathcal{B}(\bar{\theta}|\nu_{\mu_\theta}, J_{\mu_\theta})\Big|_{\bar{\theta}=\theta} = \mathbb{E}_{s\sim\nu_{\mu_\theta}}\left[\nabla_{\bar{\theta}} Q_{\mu_\theta}(s, \mu_{\bar{\theta}}(s))\Big|_{\bar{\theta}=\theta}\right]$$

Policy gradient is the iteration

$$\theta_{i+1} = \theta_i - \nabla_{\bar{\theta}} \mathcal{B}(\bar{\theta}|\nu_{\mu_{\theta_i}}, J_{\mu_{\theta_i}})\Big|_{\bar{\theta}=\theta_i}$$

This is similar to weighted PI restricted to M_Θ, except:

1. PG takes incremental gradient steps rather than solving $\min_{\bar{\theta}} \mathcal{B}(\bar{\theta}|w, J_{\mu_{\theta_i}})$ to optimality.
2. PG uses state-relevance weights $\nu_{\mu_{\theta_i}}$ induced by applying μ_{θ_i}.

Deriving the policy gradient formula

A first-order approximation to J_{μ_θ}

Under appropriate smoothness conditions

$$J_{\mu_{\bar\theta}} = J_{\mu_\theta} + (I - \alpha P_{\mu_\theta})^{-1} \left(T_{\mu_{\bar\theta}} J_{\mu_\theta} - J_{\mu_\theta}\right) + O(\|\bar\theta - \theta\|_2^2),$$

Proof Set $T_\theta \equiv T_{\mu_\theta}$, $P_\theta \equiv P_{\mu_\theta}$, $g_\theta \equiv g_{\mu_\theta}$.

By the variational form of Bellman's equation,

$$\begin{aligned} J_{\bar\theta} =& J_\theta + \left(I - \alpha P_{\bar\theta}\right)^{-1} \left(T_{\bar\theta} J_\theta - J_\theta\right) \\ =& J_\theta + \left(I - \alpha P_\theta\right)^{-1} \left(T_{\bar\theta} J_\theta - J_\theta\right) + e_\theta \end{aligned}$$

where

$$\begin{aligned} e_\theta &= \left(\left[I - \alpha P_{\bar\theta}\right]^{-1} - \left[I - \alpha P_\theta\right]^{-1}\right) \left(T_{\bar\theta} J_\theta - J_\theta\right) \\ &= \left(\left[I - \alpha P_{\bar\theta}\right]^{-1} - \left[I - \alpha P_\theta\right]^{-1}\right) \left(g_{\bar\theta} - g_\theta + \alpha \left[P_{\bar\theta} - P_\theta\right] J_\theta\right) \end{aligned}$$

is $O(\|\bar\theta - \theta\|_2^2)$ assuming $\theta \mapsto P_\theta$ and $\theta \mapsto g_\theta$ are differentiable.

Deriving the policy gradient formula (2)

Under appropriate smoothness conditions

$$J_{\mu_{\bar{\theta}}} = J_{\mu_\theta} + (I - \alpha P_{\mu_\theta})^{-1} \left(T_{\mu_{\bar{\theta}}} J_{\mu_\theta} - J_{\mu_\theta} \right) + O(\|\bar{\theta} - \theta\|_2^2),$$

Left multiplying each side by ν gives the following:

Under appropriate smoothness conditions

$$\ell(\bar{\theta}) = \ell(\theta) + \nu_{\mu_\theta} \left(T_{\mu_{\bar{\theta}}} J_{\mu_\theta} - J_{\mu_\theta} \right) + O(\|\bar{\theta} - \theta\|_2^2),$$

Differentiating with respect to $\bar{\theta}$ gives,

$$\nabla \ell(\theta) = \nabla_{\bar{\theta}} \nu_{\mu_\theta} \left(T_{\mu_{\bar{\theta}}} J_{\mu_\theta} \right) \Big|_{\bar{\theta}=\theta} = \nabla_{\bar{\theta}} \mathcal{B}(\bar{\theta} | \nu_{\mu_\theta}, J_{\mu_\theta}) \Big|_{\bar{\theta}=\theta}$$

Table of Contents

Main insight of Bhandari and Russo (2020)

Two conditions ensure $\ell(\cdot)$ has no suboptimal stationary points

1. The policy class is closed under policy improvement
   - A PI update to $\mu \in M_\Theta$ can be solved within the policy class:

$$\min_{\mu_\theta \in M_\Theta} \mathcal{B}(\mu_\theta | \nu, J_\mu) = \min_{\bar{\mu} \in M} \mathcal{B}(\bar{\mu} | \nu, J_\mu)$$

   - A sense in which the policy class is sufficient for control.
   - Much weaker than requiring it represents (nearly) all policies
   - Necessarily stronger than requiring $\mu^* \in M_\Theta$.

2. First order methods can solve the PI problem
   - $\theta \mapsto \mathcal{B}(\mu_\theta \mid \nu, J_\mu)$ has no suboptimal stationary points.
   - E.g. this single period objective is often convex...

Example: linear quadratic control (deterministic for this talk)

States $s_t \in \mathbb{R}^n$ and actions $a_t \in \mathbb{R}^k$

$$\text{Minimize} \quad \sum_{t=0}^{\infty} \alpha^t \left(s_t^\top R s_t + a_t^\top U a_t \right)$$

$$\text{Subject to} \quad s_t = A s_{t-1} + B a_{t-1}$$

Linear policies: $\mu_\theta(s) = \theta s$.

Total cost $\ell(\mu_\theta) = \mathbb{E}_{s_0 \sim \nu}[J_{\mu_\theta}(s_0)]$ is messy and nonconvex:

$$\begin{aligned} J_{\mu_\theta}(s_0) &= \sum_{t=0}^{\infty} \alpha^t \left(s_t^\top R s_t + s_t^\top \theta^\top U \theta s_t \right) \\ &= s_0^\top \underbrace{\left(\sum_{t=0}^{\infty} \alpha^t \left[(A + B\theta)^t \right]^\top \left(R + \theta^\top U \theta \right) \left[(A + B\theta)^t \right] \right)}_{K_\theta} s_0 \end{aligned}$$

since $s_t = A s_{t-1} + B\theta s_{t-1} = \cdots = (A + B\theta)^t s_0$

Policy iteration for LQ control –Kleinman (1968)

The policy improvement objective is quadratic in a:

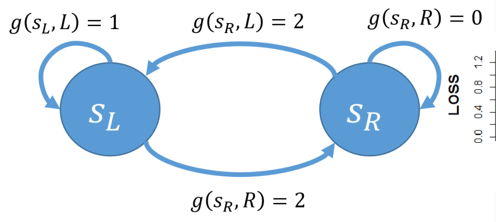
$$Q_{\mu_\theta}(s,a) = a^\top U a + \alpha(As+Ba)^\top K_\theta (As+Ba)$$

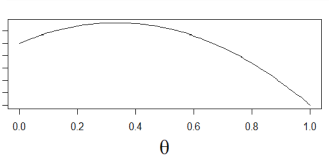
Condition 1: The policy class M_Θ is closed under PI updates?

- Yes. $\bar{\theta}s = \mathrm{argmin}_a\, Q(s,a)$ for all s
  (Updated parameter is $\bar{\theta} = \alpha\left(U + \alpha B^\top K_\theta B\right)^{-1} B^\top K_\theta A$)

Condition 2 The PI obj. has no suboptimal stationary points?

- Holds since $\bar{\theta} \mapsto \mathbb{E}_{s\sim\nu}\left[Q_{\mu_\theta}(s, \bar{\theta}s)\right]$ is quadratic





- $\mu_\theta(s_L) = \mu_\theta(S_R) = \theta =$ "Probability of Right."

The parameterized policy class is not closed

For $\theta = .1$,

$$L = \operatorname*{argmin}_{a \in \{L,R\}} Q_{\mu_\theta}(S_L, a)$$

whereas

$$R = \operatorname*{argmin}_{a \in \{L,R\}} Q_{\mu_\theta}(S_R, a)$$

Approximation with rich policy classes?

Punchline: if the policy iteration problem can be nearly solved within the policy class, then stationary points of the PG objective are nearly optimal.

Approximation with rich policy classes

- **Condition 1b**: Policy class is closed under approximate PI

$$\text{For any } \mu \in M_\Theta, \quad \min_{\mu_\theta \in M_\Theta} \mathcal{B}(\mu_\theta|\nu, J_\mu) \leq \min_{\bar{\mu} \in M} \mathcal{B}(\bar{\mu}|\nu, J_\mu) + \epsilon$$

- **Condition 2a** The PI objective has no suboptimal stationary points
  - $\theta \mapsto \mathbb{E}_{s\sim\nu}\left[Q_\mu(s, \mu_\theta(s))\right]$ has no suboptimal stationary points for any distribution ν over $\mathcal{S}$.

Theorem (Informal)

Under conditions 1b and 2a (and mild regularity conditions), if θ is a stationary point of $\ell(\cdot)$ then

$$\ell(\mu_\theta) \leq \min_{\mu \in M} \ell(\mu) + \frac{\kappa_\nu}{(1-\alpha)^2} \cdot \epsilon$$

where

$$k_\nu = \sup_{J_\mu \,:\, \mu \in M_\Theta} \frac{\|J_\mu - TJ_\mu\|_{1,\nu_{\mu^*}}}{\|J_\mu - TJ_\mu\|_{1,\nu}}$$