HW2 due ~~Fo~~ Friday

Approximate Policy Iteration
& Conservative Policy Iteration

Recap

Recall Policy Iteration (PI):

Recap

Recall Policy Iteration (PI):

Assume MDP is known, we compute $A^{\pi}(s, a)$ exactly for all s, a, PI updates policy as:

Recap

Recall Policy Iteration (PI):

Assume MDP is known, we compute $A^\pi(s,a)$ exactly for all s, a, PI updates policy as:

$$\pi'(s) = \arg\max_a A^\pi(s,a)$$

Recap

Recall Policy Iteration (PI):

Assume MDP is known, we compute $A^\pi(s, a)$ exactly for all s, a, PI updates policy as:

$$\pi'(s) = \arg\max_a A^\pi(s, a)$$

i.e., be greedy with respect to π at every state s,

Recap

Recall Policy Iteration (PI):

Assume MDP is known, we compute $A^{\pi}(s, a)$ exactly for all s, a, PI updates policy as:

$$\pi'(s) = \arg\max_{a} A^{\pi}(s, a)$$

i.e., be greedy with respect to π at every state s,

Monotonic improvement of PI: $Q^{\pi'}(s, a) \geq Q^{\pi}(s, a), \forall s, a$

Recap

Recall Policy Iteration (PI):

Assume MDP is known, we compute $A^\pi(s,a)$ exactly for all s, a, PI updates policy as:

$$\pi'(s) = \arg\max_a A^\pi(s,a)$$

i.e., be greedy with respect to π at every state s,

Monotonic improvement of PI: $Q^{\pi'}(s,a) \geq Q^\pi(s,a), \forall s, a$

However, for large scale, unknown MDP
there is no way we will be able to know $A^\pi(s,a)$ at all s, a,
so how can we do policy update?

Recap

Recall Policy Iteration (PI):

$$\pi'(s) = \arg\max_a A^{\pi}(s, a)$$

Monotonic improvement of PI: $Q^{\pi'}(s, a) \geq Q^{\pi}(s, a), \forall s, a$

Recap

Recall Policy Iteration (PI):

$$\pi'(s) = \arg\max_a A^\pi(s, a)$$

Monotonic improvement of PI: $Q^{\pi'}(s, a) \geq Q^\pi(s, a), \forall s, a$

Performance Difference Lemma (PDL): for all $s_0 \in S$

Recap

Recall Policy Iteration (PI):

$$\pi'(s) = \arg\max_a A^{\pi}(s, a)$$

Monotonic improvement of PI: $Q^{\pi'}(s, a) \geq Q^{\pi}(s, a), \forall s, a$

Performance Difference Lemma (PDL): for all $s_0 \in S$

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

Recap

Recall Policy Iteration (PI):

$$\pi'(s) = \arg\max_a A^\pi(s,a)$$

Monotonic improvement of PI: $Q^{\pi'}(s,a) \geq Q^\pi(s,a), \forall s, a$

Performance Difference Lemma (PDL): for all $s_0 \in S$

$\pi' = \arg\max_a A^\pi(s,a)$

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

$\max_a A^\pi(s,a) \geq 0$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi'}}\left[\underbrace{\max_{a\in A} A^\pi(s,a)}\right] \geq 0 \quad \checkmark$$

Recap

Recall Policy Iteration (PI):

$$\pi'(s) = \arg\max_a A^\pi(s,a)$$

Monotonic improvement of PI: $Q^{\pi'}(s,a) \geq Q^\pi(s,a), \forall s, a$

Performance Difference Lemma (PDL): for all $s_0 \in S$

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

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi'}}\left[\max_{a\in A} A^\pi(s,a)\right] \geq 0$$

However, for large scale, unknown MDP
there is no way we will be able to know $A^\pi(s,a)$ at all s, a,
so how can we do policy update?

Setting and Notation

Discounted infinite horizon MDP:

$$\mathcal{M} = \{S, A, \gamma, r, P, \mu\}$$

Δ

Setting and Notation

Discounted infinite horizon MDP:

$$\mathscr{M} = \{S, A, \gamma, r, P, \mu\}$$

State visitation: $d_\mu^\pi(s) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s; \mu)$

Setting and Notation

Discounted infinite horizon MDP:

$$\mathcal{M} = \{S, A, \gamma, r, P, \mu\}$$

State visitation: $d_\mu^\pi(s) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s; \mu)$

Unbiased estimate of $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$

$s \; a \quad \pi \quad Q^\pi(s,a)$

$\pi \quad V^\pi(s)$

Setting and Notation

Discounted infinite horizon MDP:

$$\mathscr{M} = \{S, A, \gamma, r, P, \mu\}$$

State visitation: $d_\mu^\pi(s) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s;\mu)$

Unbiased estimate of $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$

As we will consider large scale unknown MDP here, we start with a (restricted) function class Π:

$$\Pi = \{\pi : S \mapsto \Delta(A)\}$$

Δ

Setting and Notation

Discounted infinite horizon MDP:

$$\mathcal{M} = \{S, A, \gamma, r, P, \mu\}$$

State visitation: $d_\mu^\pi(s) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s;\mu)$

Unbiased estimate of $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$

As we will consider large scale unknown MDP here, we start with a (restricted) function class Π:

$$\Pi = \{\pi : S \mapsto \Delta(A)\}$$

Note that the optimal policy $\pi^\star$ may not be in Π $\qquad \pi^\star \notin \Pi$

Attempt One: Approximate Policy Iteration (API)

Attempt One: Approximate Policy Iteration (API)

π^t

Given the current policy π^t, let's act greedily wrt π under $d_\mu^{\pi^t}$

Δ

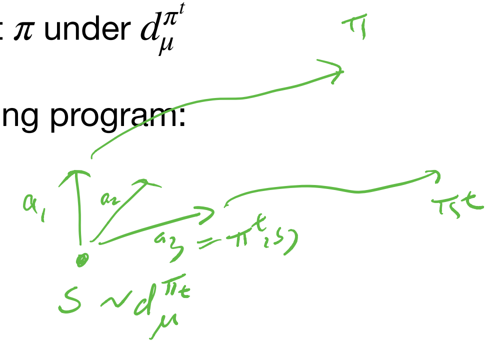
Attempt One: <u>Approximate</u> Policy Iteration (API)

Given the current policy π^t, let's act greedily wrt π under $d_\mu^{\pi^t}$

i.e., let's aim to (approximately) solve the following program:

$$\arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

A



Attempt One: <u>Approximate</u> Policy Iteration (API)

Given the current policy π^t, let's act greedily wrt π under $d_\mu^{\pi^t}$

i.e., let's aim to (approximately) solve the following program:

$$\arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$ **Greedy Policy Selector**

Attempt One: Approximate Policy Iteration (API)

Given the current policy π^t, let's act greedily wrt π under $d_\mu^{\pi^t}$

i.e., let's aim to (approximately) solve the following program:

$$\arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$ **Greedy Policy Selector**

But we can only sample from $d_\mu^{\pi^t}$, and we can only get an approximation of $A^{\pi^t}(s, a)$

A

Attempt One: Approximate Policy Iteration (API)

Given the current policy π^t, let's act greedily wrt π under $d_\mu^{\pi^t}$

i.e., let's aim to (approximately) solve the following program:

$\hat{E}$

$$\arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$ **Greedy Policy Selector**

a

But we can only sample from $d_\mu^{\pi^t}$, and we can only get an approximation of $A^{\pi^t}(s, a)$

We can hope for an Approximate Greedy Policy Selector a reduction to Regression

Implementing Approximate Greedy Policy Selector via Regression

We can do a **reduction to Regression** via Advantage function approximation

$$\mathscr{F} = \{f : S \times A \mapsto \mathbb{R}\} \quad (\approx A^{\pi^t})$$

Implementing Approximate Greedy Policy Selector via Regression

We can do a **reduction to Regression** via Advantage function approximation

$$\mathscr{F} = \{f : S \times A \mapsto \mathbb{R}\} \quad (\approx A^{\pi^t})$$

$$\Pi = \{\pi(s) = \arg\max_a f(s, a) : f \in \mathscr{F}\}$$

$f \approx A^{\pi^t}, \forall s, a$

$\arg\max_a f(s, a)$

Implementing Approximate Greedy Policy Selector via Regression

We can do a **reduction to Regression** via Advantage function approximation

$$\mathscr{F} = \{f : S \times A \mapsto \mathbb{R}\} \quad (\approx A^{\pi^t})$$

$$\Pi = \{\pi(s) = \arg\max_a f(s,a) : f \in \mathscr{F}\}$$

$$\{s_i, a_i, \widetilde{A}_i\}, s_i \sim d_\mu^{\pi^t}, a_i \sim U(A), \mathbb{E}\left[\widetilde{A}_i\right] = A^{\pi^t}(s_i, a_i)$$

unbiased from MC rollout

Implementing Approximate Greedy Policy Selector via Regression

We can do a **reduction to Regression** via Advantage function approximation

$$\mathcal{F} = \{f : S \times A \mapsto \mathbb{R}\} \quad (\approx A^{\pi^t})$$

$$\Pi = \{\pi(s) = \arg\max_a f(s, a) : f \in \mathcal{F}\}$$

$$\{s_i, a_i, \widetilde{A}_i\}, s_i \sim d_\mu^{\pi^t}, a \sim U(A), \mathbb{E}\left[\widetilde{A}_i\right] = A^{\pi^t}(s_i, a_i)$$

Regression oracle:

$A^{\pi^t}(s_i, a_i)$

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_i \left(f(s_i, a_i) - \widetilde{A}_i\right)^2$$

Implementing Approximate Greedy Policy Selector via Regression

We can do a **reduction to Regression** via Advantage function approximation

$$\mathscr{F} = \{f : S \times A \mapsto \mathbb{R}\} \quad (\approx A^{\pi^t})$$

$$\Pi = \{\pi(s) = \arg\max_a f(s,a) : f \in \mathscr{F}\}$$

$$\{s_i, a_i, \widetilde{A}_i\}, s_i \sim d_\mu^{\pi^t}, a \sim U(A), \mathbb{E}\left[\widetilde{A}_i\right] = A^{\pi^t}(s_i, a_i)$$

Regression oracle:

$$\hat{f} = \arg\min_{f \in \mathscr{F}} \sum_i \left(f(s_i, a_i) - \widetilde{A}_i\right)^2$$

Act greedily wrt the estimator $\hat{f}$ (as we hope $\hat{f} \approx A^{\pi^t}$):

Implementing Approximate Greedy Policy Selector via Regression

We can do a **reduction to Regression** via Advantage function approximation

$$\mathscr{F} = \{f : S \times A \mapsto \mathbb{R}\} \quad (\approx A^{\pi^t})$$

$$\Pi = \{\pi(s) = \arg\max_a f(s,a) : f \in \mathscr{F}\}$$

$$\{s_i, a_i, \widetilde{A}_i\}, s_i \sim d_\mu^{\pi^t}, a \sim U(A), \mathbb{E}\left[\widetilde{A}_i\right] = A^{\pi^t}(s_i, a_i)$$

Regression oracle:

$$\hat{f} = \arg\min_{f \in \mathscr{F}} \sum_i \left(f(s_i, a_i) - \widetilde{A}_i\right)^2$$

Act greedily wrt the estimator $\hat{f}$ (as we hope $\hat{f} \approx A^{\pi^t}$):

$$\widehat{\pi}(s) = \arg\max_a \hat{f}(s,a), \forall s$$

A

Implementing Approximate Greedy Policy Selector via Regression

We can do a **reduction to Regression** via Advantage function approximation

$$\mathscr{F} = \{f : S \times A \mapsto \mathbb{R}\} \quad (\approx A^{\pi^t})$$

$$\Pi = \{\pi(s) = \arg\max_a f(s,a) : f \in \mathscr{F}\}$$

$$\{s_i, a_i, \widetilde{A}_i\}, s_i \sim d_\mu^{\pi^t}, a \sim U(A), \mathbb{E}\left[\widetilde{A}_i\right] = A^{\pi^t}(s_i, a_i)$$

Do **finite sample analysis for Regression** first, and then transfer the guarantee to greedy policy selection

Regression oracle:

$$\hat{f} = \arg\min_{f \in \mathscr{F}} \sum_i \left(f(s_i, a_i) - \widetilde{A}_i\right)^2$$

Act greedily wrt the estimator $\hat{f}$ (as we hope $\hat{f} \approx A^{\pi^t}$):

$$\widehat{\pi}(s) = \arg\max_a \hat{f}(s,a), \forall s$$

Analyzing Approximation error via Regression

Greedy Policy Selector

$$\widetilde{\pi} := \arg\max_{\pi \in \Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

$$\{s_i, a_i, \widetilde{A}_i\}, s_i \sim d_\mu^{\pi^t}, a \sim U(A), \mathbb{E}\left[\widetilde{A}_i\right] = A^{\pi^t}(s_i, a_i)$$

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_i \left(f(s_i, a_i) - \widetilde{A}_i \right)^2$$

$$\hat{\pi}(s) = \arg\max_a \hat{f}(s, a), \forall s$$

$$\boxed{\mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \hat{\pi}(s))]} = \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[\hat{f}(s, \hat{\pi}(s)) + A^{\pi^t}(s, \hat{\pi}(s)) - \hat{f}(s, \hat{\pi}(s)) \right]$$

$$\geq \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[\hat{f}(s, \widetilde{\pi}(s)) + A^{\pi^t}(s, \hat{\pi}(s)) - \hat{f}(s, \hat{\pi}(s)) \right]$$

$$\geq \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \widetilde{\pi}(s)) + \hat{f}(s, \widetilde{\pi}(s)) - A^{\pi^t}(s, \widetilde{\pi}(s)) + A^{\pi^t}(s, \hat{\pi}(s)) - \hat{f}(s, \hat{\pi}(s)) \right]$$

$$\geq \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \widetilde{\pi}(s)) \right] + \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[\hat{f}(s, \widetilde{\pi}(s)) - A^{\pi^t}(s, \widetilde{\pi}(s)) + A^{\pi^t}(s, \hat{\pi}(s)) - \hat{f}(s, \hat{\pi}(s)) \right]$$

Summary So Far:

By reduction to Supervised Learning (i.e., classification using Π or Regression using $\mathscr{F}$), with high probability, we get:

Summary So Far:

By reduction to Supervised Learning (i.e., classification using Π or Regression using $\mathscr{F}$), with high probability, we get:

$$\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \widehat{\pi}(s))\right] \geq \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right] - \underbrace{\frac{A}{1-\gamma}\sqrt{\frac{\ln(|\mathscr{F}|/\delta)}{N}}}_{\text{statistical error}:\epsilon}$$

$\uparrow$

$\to 0$

N # of Training samples

Summary So Far:

By reduction to Supervised Learning (i.e., classification using Π or Regression using $\mathscr{F}$), with high probability, we get:

$$\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\widehat{\pi}(s))\right] \geq \max_{\pi\in\Pi}\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right] - \underbrace{\frac{A}{1-\gamma}\sqrt{\frac{\ln(|\mathscr{F}|/\delta)}{N}}}_{\text{statistical error}:\epsilon}$$

$\to \infty$

In the rest of the lecture, as we will focus on convergence rather than sample complexity, we ignore the statistical error (goes to zero as N increases),

Summary So Far:

By reduction to Supervised Learning (i.e., classification using Π or Regression using $\mathscr{F}$), with high probability, we get:

$$\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \widehat{\pi}(s))\right] \geq \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right] - \underbrace{\frac{A}{1-\gamma}\sqrt{\frac{\ln(|\mathscr{F}|/\delta)}{N}}}_{\text{statistical error}:\epsilon}$$

In the rest of the lecture, as we will focus on convergence rather than sample complexity, we ignore the statistical error (goes to zero as N increases),

i.e., we assume we can do the exact greedy policy selector: $\arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$

Algorithm: Approximate Policy Iteration (API)

Iterate: π_t

API: $$\underset{A}{\pi^{t+1}} \in \underbrace{\arg\max_{\pi \in \Pi} \mathbb{E}_{s,a \sim d_\mu^{\pi^t}} \left[A^{\pi_t}(s, \pi(s)) \right]}$$

← reduce Regression

$N \to \infty$

Algorithm: Approximate Policy Iteration (API)

Iterate:

$$\text{API:} \quad \pi^{t+1} \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s,a \sim d_\mu^{\pi^t}} \left[A^{\pi_t}(s, \pi(s)) \right]$$

Question:

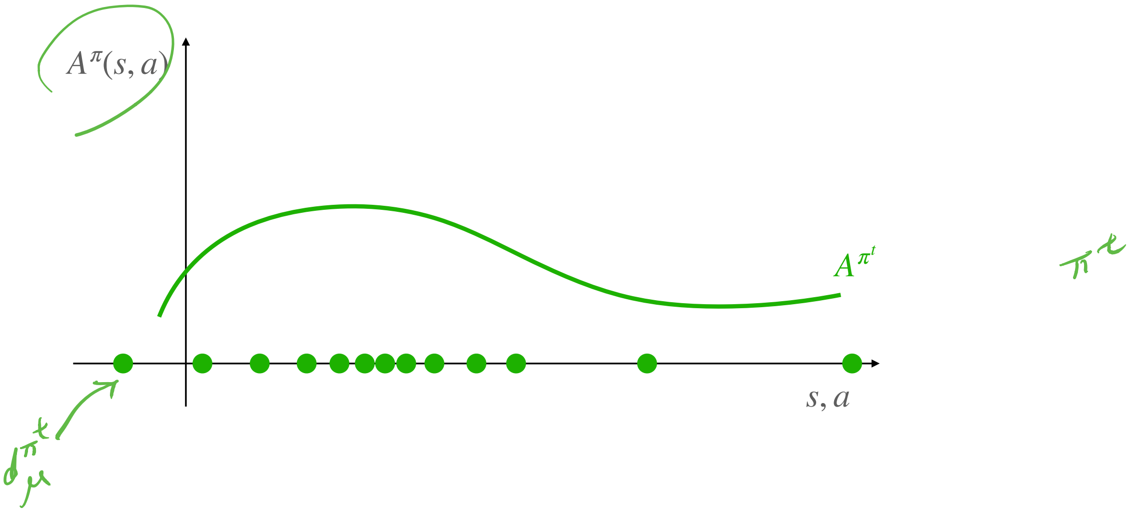
(1) Does API has monotonic improvement?

(2) Does it convergence?

The Oscillation of API from Abrupt Distribution Change

API cannot guarantee to succeed (let's think about advantage function approximation setting)

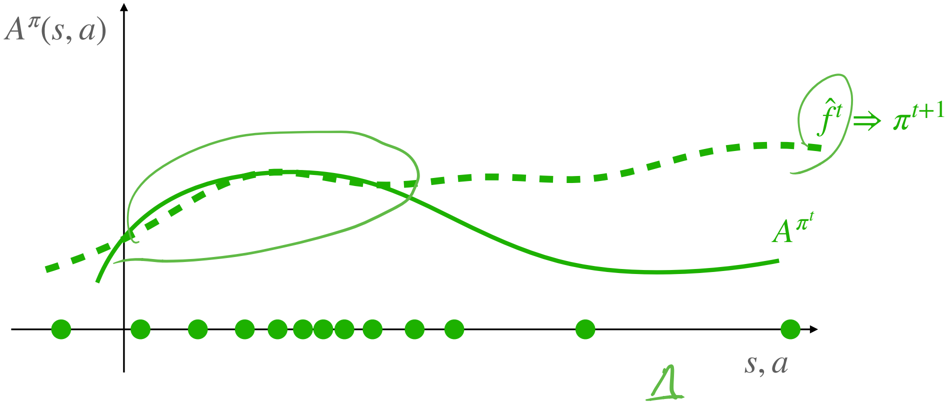
Concrete example in Chapter 3



The Oscillation of API from Abrupt Distribution Change

API cannot guarantee to succeed (let's think about advantage function approximation setting)

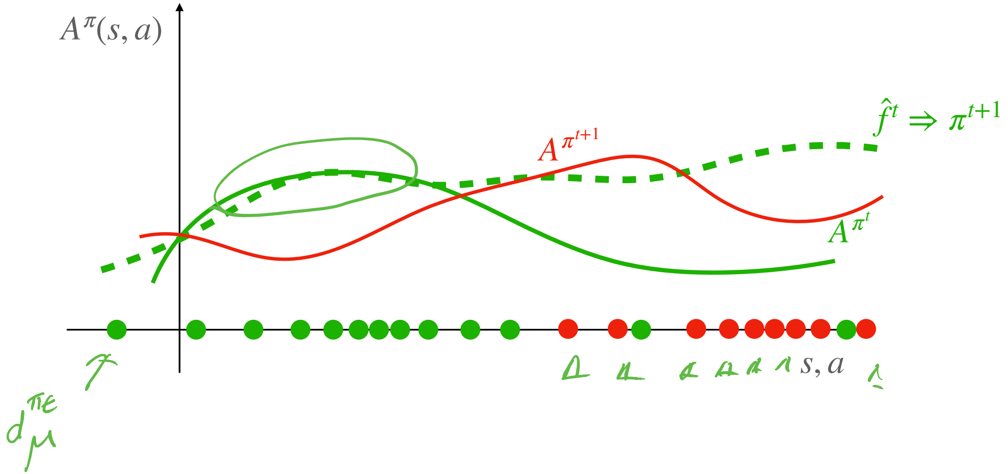
Concrete example in Chapter 3



The Oscillation of API from Abrupt Distribution Change

API cannot guarantee to succeed (let's think about advantage function approximation setting)

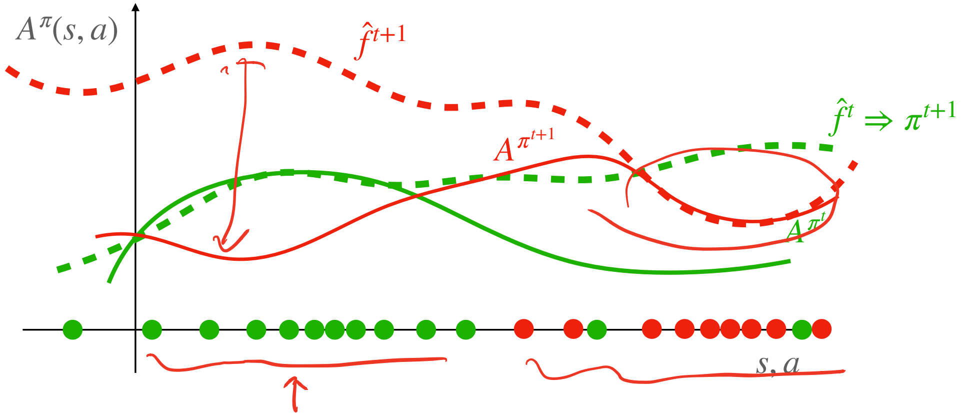
Concrete example in Chapter 3



The Oscillation of API from Abrupt Distribution Change

API cannot guarantee to succeed (let's think about advantage function approximation setting)

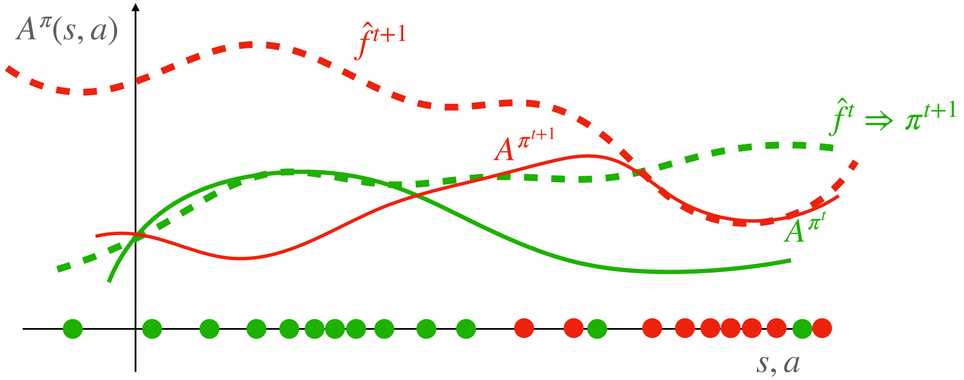
Concrete example in Chapter 3



The Oscillation of API from Abrupt Distribution Change

API cannot guarantee to succeed (let's think about advantage function approximation setting)

Concrete example in Chapter 3



Oscillation between two updates: No monotonic improvement

Key Issue: Abrupt Policy Change, i.e., $d_\mu^{\pi^{t+1}}$ and $d_\mu^{\pi^t}$ could be widely different

Key Issue: Abrupt Policy Change, i.e., $d_\mu^{\pi^{t+1}}$ and $d_\mu^{\pi^t}$ could be widely different

Our estimator $\hat{f}^t$ is only good under $d_\mu^{\pi^t}$, i.e.,

$\mathbb{E}_{s\sim d_\mu^{\pi^t}, a\sim U(A)}(\hat{f}^t(s,a) - A^{\pi^t}(s,a))^2$ small,

but $\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}, a\sim U(A)}(\hat{f}^t(s,a) - A^{\pi^t}(s,a))^2$ **might be arbitrarily big**

$d_\mu^{\pi^t} \neq d_\mu^{\pi^{t+1}}$

Key Issue: Abrupt Policy Change, i.e., $d_\mu^{\pi^{t+1}}$ and $d_\mu^{\pi^t}$ could be widely different

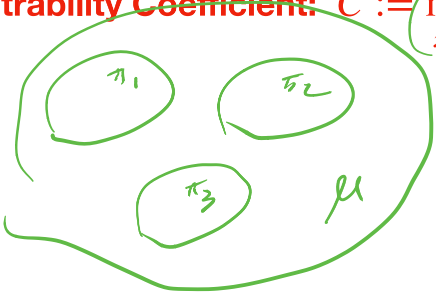
Our estimator $\hat{f}^t$ is only good under $d_\mu^{\pi^t}$, i.e.,

$\mathbb{E}_{s\sim d_\mu^{\pi^t}, a\sim U(A)}(\hat{f}^t(s,a) - A^{\pi^t}(s,a))^2$ small,

but $\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}, a\sim U(A)}(f^t(s,a) - A^{\pi^t}(s,a))^2$ **might be arbitrarily big**

To make API to make improvement, we need a much stronger coverage assumption:

A strong Concentrability Coefficient: $C := \max_{\pi\in\Pi} \sup_s \frac{d_\mu^\pi(s)}{\mu(s)} < \infty$



Key Issue: Abrupt Policy Change, i.e., $d_\mu^{\pi^{t+1}}$ and $d_\mu^{\pi^t}$ could be widely different

Our estimator $\hat f^t$ is only good under $d_\mu^{\pi^t}$, i.e.,

$\mathbb{E}_{s\sim d_\mu^{\pi^t}, a\sim U(A)}(\hat f^t(s,a) - A^{\pi^t}(s,a))^2$ small,

but $\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}, a\sim U(A)}(f^t(s,a) - A^{\pi^t}(s,a))^2$ **might be arbitrarily big**

To make API to make improvement, we need a much stronger coverage assumption:

A strong Concentrability Coefficient: $C := \max_{\pi\in\Pi} \sup_s \frac{d_\mu^\pi(s)}{\mu(s)} < \infty$

If $C < \infty$, i.e., μ covers **all** d_μ^π, then we can expect error

$\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}, a\sim U(A)}(\hat f^t(s,a) - A^{\pi^t}(s,a))^2$ is reasonably under control;

① $d_\mu^{\pi^t} \approx d_\mu^{\pi^{t+1}}$

$d^{\pi^{t+1}}$ μ

Conservative Policy Iteration—An Incremental Policy Optimization Approach

(And the benefit of being incremental)

NPG & Trust Region

$\Uparrow$

$KL\left(\Pr^{\pi^{t+1}} \,\|\, \Pr^{\pi^t}\right) \leq \delta$

Key Idea of CPI: Incremental Update—No Abrupt Distribution Change

Let's design policy update rule such that $d^{\pi^{t+1}}$ and d^{π^t} are not that different!

Δ Δ

Key Idea of CPI: Incremental Update—No Abrupt Distribution Change

Let’s design policy update rule such that $d^{\pi^{t+1}}$ and d^{π^t} are not that different!

Recall Performance Difference Lemma:

$$V^{\pi^{t+1}} - V^{\pi^t} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[A^{\pi^t}(s, \pi^{t+1}(s))\right]$$

Key Idea of CPI: Incremental Update—No Abrupt Distribution Change

Let's design policy update rule such that $d^{\pi^{t+1}}$ and d^{π^t} are not that different!

Recall Performance Difference Lemma:

$$V^{\pi^{t+1}} - V^{\pi^t} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[A^{\pi^t}(s,\pi^{t+1}(s))\right]$$

$$\underline{d^{\pi^t}} \approx \underline{d^{\pi^{t+1}}}$$

Key Idea of CPI: Incremental Update—No Abrupt Distribution Change

Let's design policy update rule such that $d^{\pi^{t+1}}$ and d^{π^t} are not that different!

Recall Performance Difference Lemma:

$$V^{\pi^{t+1}} - V^{\pi^t} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi^{t+1}}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right]$$

$$d^{\pi^t} \approx d^{\pi^{t+1}}$$

s.t., $\underbrace{\mathbb{E}_{s \sim d^{\pi^t}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right]}_{A} \approx \mathbb{E}_{s \sim d^{\pi^{t+1}}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right]$

Key Idea of CPI: Incremental Update—No Abrupt Distribution Change

Let's design policy update rule such that $d^{\pi^{t+1}}$ and d^{π^t} are not that different!

Recall Performance Difference Lemma:

$$V^{\pi^{t+1}} - V^{\pi^t} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi^{t+1}}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right] \quad \geq 0$$

$$d^{\pi^t} \approx d^{\pi^{t+1}}$$

$$\text{s.t., } \mathbb{E}_{s \sim d^{\pi^t}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right] \approx \mathbb{E}_{s \sim d^{\pi^{t+1}}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right]$$

≥ 0 ≥ 0

This we know how to optimize: the Greedy Policy Selector

Regression, $s \sim d^{\pi^t}$, $a \sim U(A)$, $\hat{A}^{\pi^t}(s,a)$

CPI Algorithm:

CPI Algorithm:

π^t

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

(Regression)

CPI Algorithm:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$ ← local optimal

Return π^t

CPI Algorithm:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \le \varepsilon$

   Return π^t

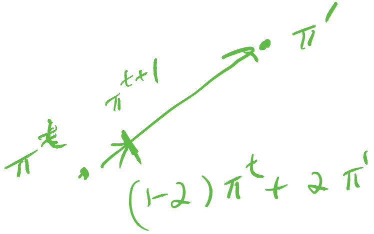
Greedy

3. Incremental Update:

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\underset{\Delta}{\pi^t(\cdot \mid s)} + \alpha \underset{\Delta}{\pi'(\cdot \mid s)}, \forall s$$

$$\pi^{t+1}(\cdot \mid s) = \begin{cases} \pi^t(\cdot \mid s) & \text{wp } 1-\alpha \\ \pi'(\cdot \mid s) & \text{wp } \alpha \end{cases}$$

$\alpha \to 0$



$\alpha \in (0,1)$

CPI Algorithm:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} [A^{\pi^t}(s, \pi(s))] \le \varepsilon$

   Return π^t

3. Incremental Update:

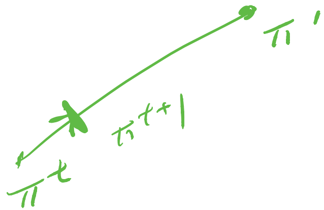
$$\pi^{t+1}(\cdot \mid s) = (1 - \alpha)\pi^t(\cdot \mid s) + \alpha \pi'(\cdot \mid s), \forall s$$

Q: Why this is incremental? In what sense?

Q: Can we get monotonic policy improvement?

The incremental Nature of CPI:

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$



The incremental Nature of CPI:

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

Key observation 1: $\forall s$

For any state s, we have $\|\pi^{t+1}(\cdot \mid s) - \pi^{t}(\cdot \mid s)\|_1 \leq 2\alpha$

$$\pi^{t+1}(\cdot \mid s) - \pi^{t}(\cdot \mid s)$$

$$= \alpha\left(\pi^{t}(\cdot \mid s) - \pi'(\cdot \mid s)\right)$$

small

The incremental Nature of CPI:

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

Key observation 1:

For any state s, we have $\|\pi^{t+1}(\cdot \mid s) - \pi^{t}(\cdot \mid s)\|_1 \leq 2\alpha$

Key observation 2:

For any two policies π and π', if $\|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1 \leq \delta$, $\forall s$ then $\|d_\mu^{\pi} - d_\mu^{\pi'}\|_1 \leq \dfrac{\gamma\delta}{1-\gamma}$

The incremental Nature of CPI:

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

Key observation 1:

For any state s, we have $\|\pi^{t+1}(\cdot \mid s) - \pi^{t}(\cdot \mid s)\|_1 \leq 2\alpha$

Key observation 2:

For any two policies π and π', if $\|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1 \leq \delta$, then $\|d_\mu^{\pi} - d_\mu^{\pi'}\|_1 \leq \dfrac{\gamma\delta}{1-\gamma}$

CPI ensures incremental update, i.e., $\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^{t}}\|_1 \leq \dfrac{2\gamma\alpha}{1-\gamma}$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] \geq \varepsilon$$

Recall CPI:

1. Greedy Policy Selector:
$\pi' \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:
$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^t(\cdot \mid s) + \alpha \pi'(\cdot \mid s), \forall s$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] \geq \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

Recall CPI:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^t(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] \geq \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

$$(1-\gamma)\left(V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}\right) = \mathbb{E}_{s \sim d_\mu^{\pi^{t+1}}}\left[\mathbb{E}_{a \sim \pi^{t+1}(\cdot|s)} A^{\pi^t}(s, a)\right]$$

Recall CPI:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$$

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^t(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi'(s))\right] \geq \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

$$(1-\gamma)\left(V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}\right) = \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)}A^{\pi^t}(s,a)\right]$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s,\pi'(s))\right] \quad (:= \alpha\mathbb{A})$$

$$\mathbb{E}_{a\sim\pi^t(\cdot|s)} A^{\pi^t}(s,a) = 0$$

Recall CPI:

1. Greedy Policy Selector:
$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:
$\pi^{t+1}(\cdot|s) = (1-\alpha)\pi^t(\cdot|s) + \alpha\pi'(\cdot|s), \forall s$

$$\pi^{t+1} = \begin{cases} \pi^t & 1-\alpha \\ \pi' & \alpha \end{cases}$$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] \geq \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

$$(1-\gamma)\left(V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}\right) = \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)} A^{\pi^t}(s,a)\right]$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] \quad (:= \alpha\mathbb{A})$$

$$= \underbrace{\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right]}_{\geq \varepsilon} + \underline{\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] - \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right]}$$

Recall CPI:

1. Greedy Policy Selector:
$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:
$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$

Recall CPI:

1. Greedy Policy Selector:
$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \le \varepsilon$

   Return π^t

3. Incremental Update:
$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi'(s))\right] \ge \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

$$(1-\gamma)\left(V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}\right) = \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)}A^{\pi^t}(s,a)\right]$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s,\pi'(s))\right] \quad (:= \alpha\mathbb{A})$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s,\pi'(s))\right] + \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s,\pi'(s))\right] - \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s,\pi'(s))\right]$$

$$\ge \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s,\pi'(s))\right] - \frac{\alpha}{1-\gamma}\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1$$

$$\left|\mathbb{E}_P(f(x)) - \mathbb{E}_Q(f(x))\right| \le \sup_x \|f(x)\| \, \|P-Q\|_{TV}$$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] \geq \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

$$(1-\gamma)\left(V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}\right) = \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)} A^{\pi^t}(s,a)\right]$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] \quad (:= \alpha\mathbb{A})$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] + \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] - \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right]$$

$$\geq \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] - \frac{\alpha}{1-\gamma}\underbrace{\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1}_{\leq \frac{2\alpha\gamma}{1-\gamma}}$$

$$\geq \underbrace{\alpha\mathbb{A}}_{A} - \frac{2\gamma\alpha^2}{(1-\gamma)^2}$$

Recall CPI:

1. Greedy Policy Selector:
$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:
$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^t(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] \geq \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

$$(1-\gamma)\left(V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}\right) = \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)} A^{\pi^t}(s,a)\right]$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] \quad (:= \alpha\mathbb{A})$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] + \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] - \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right]$$

$$\geq \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] - \frac{\alpha}{1-\gamma}\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1$$

$$\geq \alpha\mathbb{A} - \frac{2\gamma\alpha^2}{(1-\gamma)^2}$$

(Set $\alpha = \dfrac{(1-\gamma)^2\mathbb{A}}{4\gamma}$)

Recall CPI:

1. Greedy Policy Selector:
$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:
$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

Policy Improvement before Termination:

Before terminate, we have non-trivial avg **local advantage**:

$$\mathbb{A} := \mathbb{E}_{d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] \geq \varepsilon$$

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$?

$$(1-\gamma)\left(V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}\right) = \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)} A^{\pi^t}(s,a)\right]$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] \quad (:= \alpha\mathbb{A})$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] + \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] - \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right]$$

$$\geq \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\alpha A^{\pi^t}(s, \pi'(s))\right] - \frac{\alpha}{1-\gamma}\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1$$

$$\geq \alpha\mathbb{A} - \frac{2\gamma\alpha^2}{(1-\gamma)^2} \quad \geq \frac{\mathbb{A}^2(1-\gamma)}{8\gamma} \qquad \text{(Set } \alpha = \frac{(1-\gamma)^2\mathbb{A}}{4\gamma}\text{)}$$

Recall CPI:

1. Greedy Policy Selector:
$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:
$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

$$V^{\pi^{t+1}} - V^{\pi^t} \geq \frac{\epsilon^2(1-\gamma)}{8\gamma}$$

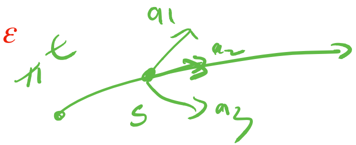
Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$$



Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \le \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \le \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Δ

Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \le \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \le \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall Π is restricted, denote $\epsilon_\Pi = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s,a)\right] - \underbrace{\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]}$

$\arg\max_a A^{\pi^t}(s,a) \notin \Pi$

$\Pi \uparrow \quad \epsilon_\Pi \searrow$

Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \dfrac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall Π is restricted, denote $\epsilon_\Pi = \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[\max_a A^{\pi^t}(s, a) \right] - \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi'(s)) \right] + \epsilon_\Pi \geq \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[\max_a A^{\pi^t}(s, a) \right] = \mathbb{E}_{s \sim d^\star} \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)} \right) \max_a A^{\pi^t}(s, a)$$

Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \dfrac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall Π is restricted, denote $\epsilon_\Pi = \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s, a)\right] - \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] + \epsilon_\Pi \geq \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s, a)\right] = \mathbb{E}_{s \sim d^\star}\left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right) \max_a A^{\pi^t}(s, a)$$

$$\geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right) \mathbb{E}_{s \sim d^\star} \max_a A^{\pi^t}(s, a)$$

Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} [A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} [A^{\pi^t}(s, \pi(s))] \leq \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall Π is restricted, denote $\epsilon_\Pi = \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[\max_a A^{\pi^t}(s, a) \right] - \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi'(s)) \right] + \epsilon_\Pi \geq \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[\max_a A^{\pi^t}(s, a) \right] = \mathbb{E}_{s \sim d^\star} \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)} \right) \max_a A^{\pi^t}(s, a)$$

$$\geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)} \right) \mathbb{E}_{s \sim d^\star} \max_a A^{\pi^t}(s, a) \quad \geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)} \right) \mathbb{E}_{s \sim d^\star} A^{\pi^t}(s, \pi^\star(s))$$

Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \le \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \le \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \dfrac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall Π is restricted, denote $\epsilon_\Pi = \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s,a)\right] - \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$

$$\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi'(s))\right] + \epsilon_\Pi \ge \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s,a)\right] = \mathbb{E}_{s\sim d^\star}\left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\max_a A^{\pi^t}(s,a)$$

$\Leftarrow$ PDL +

$$\ge \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\mathbb{E}_{s\sim d^\star}\max_a A^{\pi^t}(s,a) \ \ge \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\mathbb{E}_{s\sim d^\star}A^{\pi^t}(s,\pi^\star(s)) \ \ge \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\left(V^\star - V^{\pi^t}\right)(1-\gamma)$$

Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall Π is restricted, denote $\epsilon_\Pi = \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s, a)\right] - \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] + \epsilon_\Pi \geq \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s, a)\right] = \mathbb{E}_{s \sim d^\star}\left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\max_a A^{\pi^t}(s, a)$$

$$\geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\mathbb{E}_{s \sim d^\star}\max_a A^{\pi^t}(s, a) \ \geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\mathbb{E}_{s \sim d^\star}A^{\pi^t}(s, \pi^\star(s)) \ \geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right)\left(V^\star - V^{\pi^t}\right)(1-\gamma)$$

$$V^\star - V^{\pi^t} \leq \sup_s \left(\frac{d^\star(s)}{d_\mu^{\pi^t}(s)}\right)\frac{\varepsilon + \epsilon_\Pi}{1-\gamma}$$

Upon Termination we get a locally optimal solution (or globally optimal if μ is nice)

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

Return π^t

1. No more positive advantage by one-step deviation from π^t's own states

$$\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \varepsilon$$

2. Indeed, we can say more if μ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall Π is restricted, denote $\epsilon_\Pi = \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s, a)\right] - \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi(s))\right]$

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi'(s))\right] + \epsilon_\Pi \geq \mathbb{E}_{s \sim d_\mu^{\pi^t}}\left[\max_a A^{\pi^t}(s, a)\right] = \mathbb{E}_{s \sim d^\star}\left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right) \max_a A^{\pi^t}(s, a)$$

$$\geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right) \mathbb{E}_{s \sim d^\star} \max_a A^{\pi^t}(s, a) \geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right) \mathbb{E}_{s \sim d^\star} A^{\pi^t}(s, \pi^\star(s)) \geq \inf_s \left(\frac{d_\mu^{\pi^t}(s)}{d^\star(s)}\right) \left(V^\star - V^{\pi^t}\right)(1 - \gamma)$$

$$V^\star - V^{\pi^t} \leq \sup_s \left(\frac{d^\star(s)}{d_\mu^{\pi^t}(s)}\right) \frac{\varepsilon + \epsilon_\Pi}{1 - \gamma} \leq C^\star \frac{\varepsilon + \epsilon_\Pi}{(1 - \gamma)^2}$$

Connection to Agnostic Guarantees in Supervised Learning

Multi-class Classification (A many classes):

$$s \sim \nu, y \sim \pi^\star(s), y \in [A]$$

Connection to Agnostic Guarantees in Supervised Learning

Multi-class Classification (A many classes):

$$s \sim \nu, y \sim \pi^\star(s), y \in [A]$$

We start with a set of classifiers Π; but we cannot guarantee $\pi^\star \in \Pi$;

Δ

Connection to Agnostic Guarantees in Supervised Learning

Multi-class Classification (A many classes):

$$s \sim \nu, y \sim \pi^\star(s), y \in [A]$$

We start with a set of classifiers Π; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class Π**

Connection to Agnostic Guarantees in Supervised Learning

Multi-class Classification (A many classes):

$$s \sim \nu, y \sim \pi^\star(s), y \in [A]$$

We start with a set of classifiers Π; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** Π

$$\mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\hat{\pi}(s), y) \leq \underbrace{\min_{\pi \in \Pi} \mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\pi(s), y)}_{\epsilon_\Pi} + \underbrace{\sqrt{\frac{\ln(|\Pi|/\delta)}{N}}}_{\text{statistical error: } \epsilon}$$

N: # of samples

Connection to Agnostic Guarantees in Supervised Learning

Multi-class Classification (A many classes):

$$s \sim \nu, y \sim \pi^\star(s), y \in [A]$$

We start with a set of classifiers Π; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** Π

$$\mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\hat\pi(s), y) \leq \underbrace{\min_{\pi\in\Pi} \mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\pi(s), y)}_{\epsilon_\Pi} + \underbrace{\sqrt{\frac{\ln(|\Pi|/\delta)}{N}}}_{\text{statistical error: } \epsilon}$$

In RL:

Connection to Agnostic Guarantees in Supervised Learning

Multi-class Classification (A many classes):

$$s \sim \nu, y \sim \pi^\star(s), y \in [A]$$

We start with a set of classifiers Π; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** Π

$$\mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\hat{\pi}(s), y) \leq \underbrace{\min_{\pi \in \Pi} \mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\pi(s), y)}_{\epsilon_\Pi} + \underbrace{\sqrt{\frac{\ln(|\Pi|/\delta)}{N}}}_{\text{statistical error: } \epsilon}$$

In RL: $V^\star - V^{\hat{\pi}} \leq \sup_s \left(\frac{d^\star(s)}{\mu(s)} \right) \frac{\epsilon_\Pi + \epsilon}{(1-\gamma)^2}$

stat

Connection to Agnostic Guarantees in Supervised Learning

Multi-class Classification (A many classes):

$$s \sim \nu, y \sim \pi^\star(s), y \in [A]$$

We start with a set of classifiers Π; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** Π

$$\mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\hat\pi(s), y) \leq \underbrace{\min_{\pi\in\Pi} \mathbb{E}_{s\sim\nu, y=\pi^\star(s)} \ell(\pi(s), y)}_{\epsilon_\Pi} + \underbrace{\sqrt{\frac{\ln(|\Pi|/\delta)}{N}}}_{\text{statistical error: } \epsilon}$$

In RL: $V^\star - V^{\hat\pi} \leq \sup_s \left(\frac{d^\star(s)}{\mu(s)}\right) \frac{\epsilon_\Pi + \epsilon}{(1-\gamma)^2}$

1. Multi-step prediction (not i.i.d), 2. We don't get to see samples from $d^\star$

Compare the two Concentrability Coefficients from CPI and API:

$$\pi^{t+1} \Leftarrow \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left(A^{\pi_t}(s, \pi(s)) \right)$$

API: $\max_{\pi \in \Pi} \sup_s \frac{d^\pi(s)}{\mu(s)} < \infty$

Wide enough to cover all policies,

i.e.,making sure f $\widehat{A}$ is accurate at all places where any policy would go

Just need to cover the best in Π,

steady improvement via incremental update

CPI: $\sup_s \frac{d^\star(s)}{\mu(s)} < \infty$

Compare the two Concentrability Coefficients from CPI and API:

API: $\max_{\pi \in \Pi} \sup_{s} \frac{d^{\pi}(s)}{\mu(s)} < \infty$

Wide enough to cover all policies,
i.e.,making sure $\widehat{A}$ is accurate at all places where any policy would go

Just need to cover the best in Π,
steady improvement via incremental update

CPI: $\sup_{s} \frac{d^{\star}(s)}{\mu(s)} < \infty$

expert's distribution

1. Prior knowledge of how the optimal trajectories look like

2. Expert demonstrations (Imitation + RL)

$\sim \pi^{*}$

Summary of Policy Gradient Learning

1. PG formulation: $\mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla \ln \pi_\theta(a \mid s) A^{\pi_\theta}(s, a)$

Summary of Policy Gradient Learning

1. PG formulation: $\mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla \ln \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)$

2. For tabular MDP, gradient ascent on KL-regularized objective converges to global optimality:

$$V^{\pi_\theta} + \lambda \sum_s \sum_a \ln \pi_\theta(a \mid s), \text{ where } \pi_\theta(a \mid s) \propto \exp(\theta_{s,a})$$

Δ

$\ln \pi_\theta(a|s) \to -\infty \quad \pi_\theta(a|s) \to 0$

Summary of Policy Gradient Learning

1. PG formulation: $\mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla \ln \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)$

2. For tabular MDP, gradient ascent on KL-regularized objective converges to global optimality:

$$V^{\pi_\theta} + \lambda \sum_s \sum_a \ln \pi_\theta(a \mid s), \text{ where } \pi_\theta(a \mid s) \propto \exp(\theta_{s,a})$$

$KL(\Pr^{\pi^{t+1}} \| \Pr^{\pi^t}) \le \delta$

3. Natural Policy Gradient (trust region optimization) and its convergence (tabular, linear, & neural)

$$\widehat{w} \in \arg\min_w \mathbb{E}_{s,a\sim d_\nu^{\pi_\theta}}\left[\left(w^\top \nabla_\theta \ln \pi_\theta(a \mid s) - A^{\pi_\theta}(a \mid s)\right)^2\right], \quad \theta' = \theta + \eta \underset{A}{\widehat{w}}$$

$\sigma_{\min}\left(\mathbb{E}_{s,a\sim\mu_0} \phi(s,a)\phi(s,a)^\top\right) > c \in \mathbb{R}^+$

Summary of Policy Gradient Learning

1. PG formulation: $\mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla \ln \pi_\theta(a \mid s) A^{\pi_\theta}(s, a)$

2. For tabular MDP, gradient ascent on KL-regularized objective converges to global optimality:

$$V^{\pi_\theta} + \lambda \sum_s \sum_a \ln \pi_\theta(a \mid s), \text{ where } \pi_\theta(a \mid s) \propto \exp(\theta_{s,a})$$

3. Natural Policy Gradient (trust region optimization) and its convergence (tabular, linear, & neural)

$$\widehat{w} \in \arg\min_w \mathbb{E}_{s,a\sim d_\nu^{\pi_\theta}} \left[\left(w^\top \nabla_\theta \ln \pi_\theta(a \mid s) - A^{\pi_\theta}(a \mid s) \right)^2 \right], \quad \theta' = \theta + \eta \widehat{w}$$

4. The incremental nature of NPG/CPI/PPO and its advantage comparing to naive API

CPI (TRPO): $V^{\pi^{t+1}} > V^{\pi^t} > V^{\pi^{t-1}}$, thanks to $\| d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t} \|_1$ is small (i.e., incremental)

and API could oscillate and never converges

Next week on Control Theory:

$x_{t+1} = A x_t + B u_t$

Basics of Optimal Control on Linear Quadratic Regulators
(no learning, just planning/control)

$$\max_{\pi} \mathbb{E}_{s \sim d_{\mu}^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

s.t. $\forall s \in S$

$$\| \pi(\cdot|s) - \pi^t(\cdot|s) \|_{TV} \le \delta$$

① CPI : $\| \pi^{t+1}(\cdot|s) - \pi^t(\cdot|s) \|_{TV} \le \alpha$