Conservative Policy Iteration

Recap

Recall Policy Iteration (PI) with known (P, r)

Recap

Recall Policy Iteration (PI) with known (P, r)

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

Recap

Recall Policy Iteration (PI) with known (P, r)

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

$$\pi_{new}(s) = \arg\max_a A^{\pi_{old}}(s, a)$$

Recap

Recall Policy Iteration (PI) with known (P, r)

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

$$\pi_{new}(s) = \arg\max_a A^{\pi_{old}}(s, a)$$

i.e., pick an action that has the largest advantage against π_{old} at every state s,

Recap

Recall Policy Iteration (PI) with known (P, r)

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

$$\pi_{new}(s) = \arg\max_a A^{\pi_{old}}(s, a)$$

i.e., pick an action that has the largest advantage against π_{old} at every state s,

Maximize advantage is great, as it gives monotonic improvement:
$Q^{\pi_{new}}(s, a) \geq Q^{\pi_{old}}(s, a), \forall s, a$

Recap

Recall Policy Iteration (PI):

$$\pi_{new}(s) = \arg\max_{a} A^{\pi_{old}}(s, a)$$

Recap

Recall Policy Iteration (PI):

$$\pi_{new}(s) = \arg\max_a A^{\pi_{old}}(s, a)$$

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

Recap

Recall Policy Iteration (PI):

$$\pi_{new}(s) = \arg\max_a A^{\pi_{old}}(s, a)$$

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

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

Recap

Recall Policy Iteration (PI):

$$\pi_{new}(s) = \arg\max_a A^{\pi_{old}}(s, a)$$

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

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

The advantage against π_{old} averaged over π_{new}'s own distribution

Today:
Conservative Policy Iteration

Q: How to enforce incremental policy update and ensure monotonic improvement

Outline

1. Greedy Policy Selection (via reduction to regression) and recap of API

2. Conservative Policy Iteration

3. Monotonic Improvement of CPI

Setting and Notation

Discounted infinite horizon MDP:

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

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

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

$$\Pi = \{\pi : S \mapsto A\}$$

Setting and Notation

Discounted infinite horizon MDP:

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

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

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

$$\Pi = \{\pi : S \mapsto A\}$$

(We can think about each policy as a classifier)

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

Setting and Notation

Discounted infinite horizon MDP:

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

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

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

$$\Pi = \{\pi : S \mapsto A\}$$

From now on, think about deterministic policy as a special stochastic policy

(We can think about each policy as a classifier)

Note that the optimal policy $\pi^{\star}$ may not be in Π

Recall Approximate Policy Iteration (API)

Recall <u>Approximate</u> Policy Iteration (API)

Given the current policy π^t, let's find a new policy that has large local adv over π^t under $d_\mu^{\pi^t}$

Recall <u>Approximate</u> Policy Iteration (API)

Given the current policy π^t, let's find a new policy that has large local adv over π^t under $d_\mu^{\pi^t}$

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

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

Recall Approximate Policy Iteration (API)

Given the current policy π^t, let's find a new policy that has large local adv over π^t 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

Recall Approximate Policy Iteration (API)

Given the current policy π^t, let's find a new policy that has large local adv over π^t 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

How to implement such greedy policy selector?
We talked about a regression process..

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

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

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

Implementing Approximate Greedy Policy Selector via Regression

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

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

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

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

Implementing Approximate Greedy Policy Selector via Regression

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

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

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

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

Regression oracle:

$$\widehat{A}^t = \arg\min_{f \in \mathscr{F}} \sum_i \left(f(s_i, a_i) - y_i\right)^2$$

Implementing Approximate Greedy Policy Selector via Regression

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

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

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

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

Regression oracle:

$$\widehat{A}^t = \arg\min_{f \in \mathscr{F}} \sum_i \left(f(s_i, a_i) - y_i\right)^2$$

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

Implementing Approximate Greedy Policy Selector via Regression

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

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

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

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

Regression oracle:

$$\widehat{A}^t = \arg\min_{f \in \mathscr{F}} \sum_i \left(f(s_i, a_i) - y_i\right)^2$$

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

$$\widehat{\pi}(s) = \arg\max_a \widehat{A}^t(s,a), \forall s$$

Successful Regression ensures approximate greedy operator

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

Regression oracle:

$$\widehat{A}^{\,t} = \arg\min_{f \in \mathcal{F}} \sum_i \left(f(s_i, a_i) - y_i\right)^2$$

Successful Regression ensures approximate greedy operator

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

Regression oracle:

$$\widehat{A}^t = \arg\min_{f \in \mathcal{F}} \sum_i \left(f(s_i, a_i) - y_i\right)^2$$

Assume this regression is successful, i.e.,

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}, a \sim U(A)} \left(\widehat{A}^t(s, a) - A^{\pi^t}(s, a) \right)^2 \leq \delta$$

Successful Regression ensures approximate greedy operator

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

Regression oracle:

$$\widehat{A}^t = \arg\min_{f \in \mathcal{F}} \sum_i \left(f(s_i, a_i) - y_i\right)^2$$

Assume this regression is successful, i.e.,

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}, a \sim U(A)} \left(\widehat{A}^t(s, a) - A^{\pi^t}(s, a) \right)^2 \leq \delta$$

Then, $\widehat{\pi}(s) = \arg\max_a \widehat{A}^t(s, a)$ is an approximate greedy policy:

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}} A^{\pi^t}(s, \widehat{\pi}(s)) \geq \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} A^{\pi^t}(s, \pi(s)) - O\left(\sqrt{\delta}\right)$$

Summary So Far:

By reduction to Supervised Learning (i.e., via Regression to approximate A^{π^t} under $d_\mu^{\pi^t}$), we can expect to find an approximate greedy optimizer $\widehat{\pi}$, s.t.,

Summary So Far:

By reduction to Supervised Learning (i.e., via Regression to approximate A^{π^t} under $d_\mu^{\pi^t}$), we can expect to find an approximate greedy optimizer $\widehat{\pi}$, s.t.,

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

Summary So Far:

By reduction to Supervised Learning (i.e., via Regression to approximate A^{π^t} under $d_\mu^{\pi^t}$), we can expect to find an approximate greedy optimizer $\widehat{\pi}$, s.t.,

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

Throughout this lecture,

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

Outline

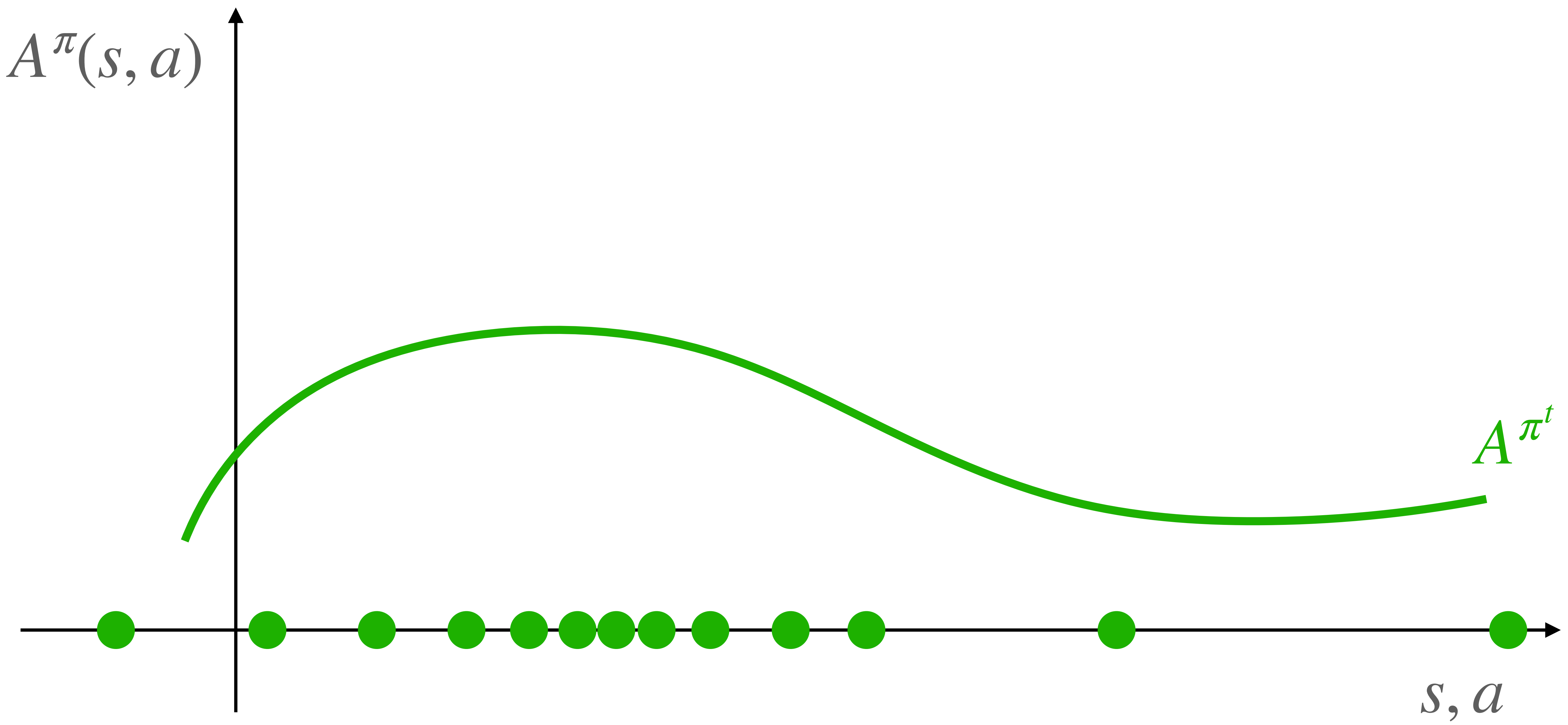
1. Greedy Policy Selection

2. Conservative Policy Iteration

3. Monotonic Improvement of CPI

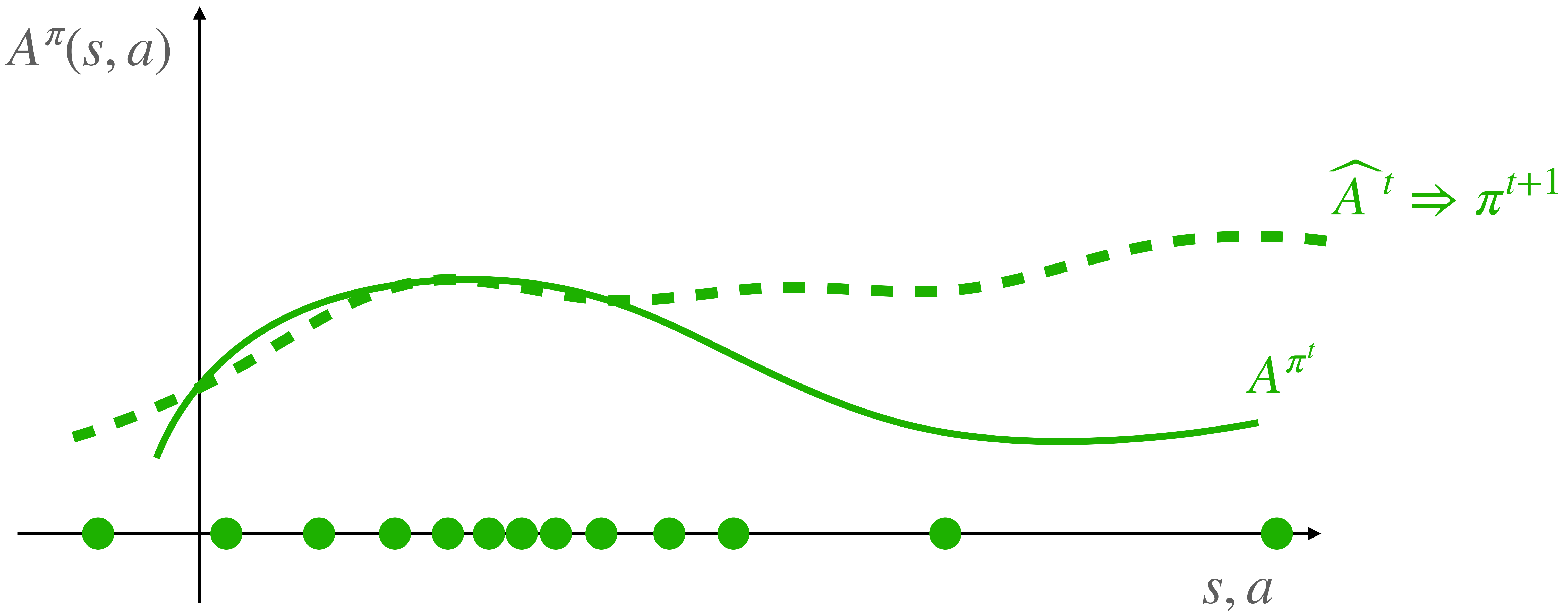
The Failure case of API: Abrupt distribution change

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



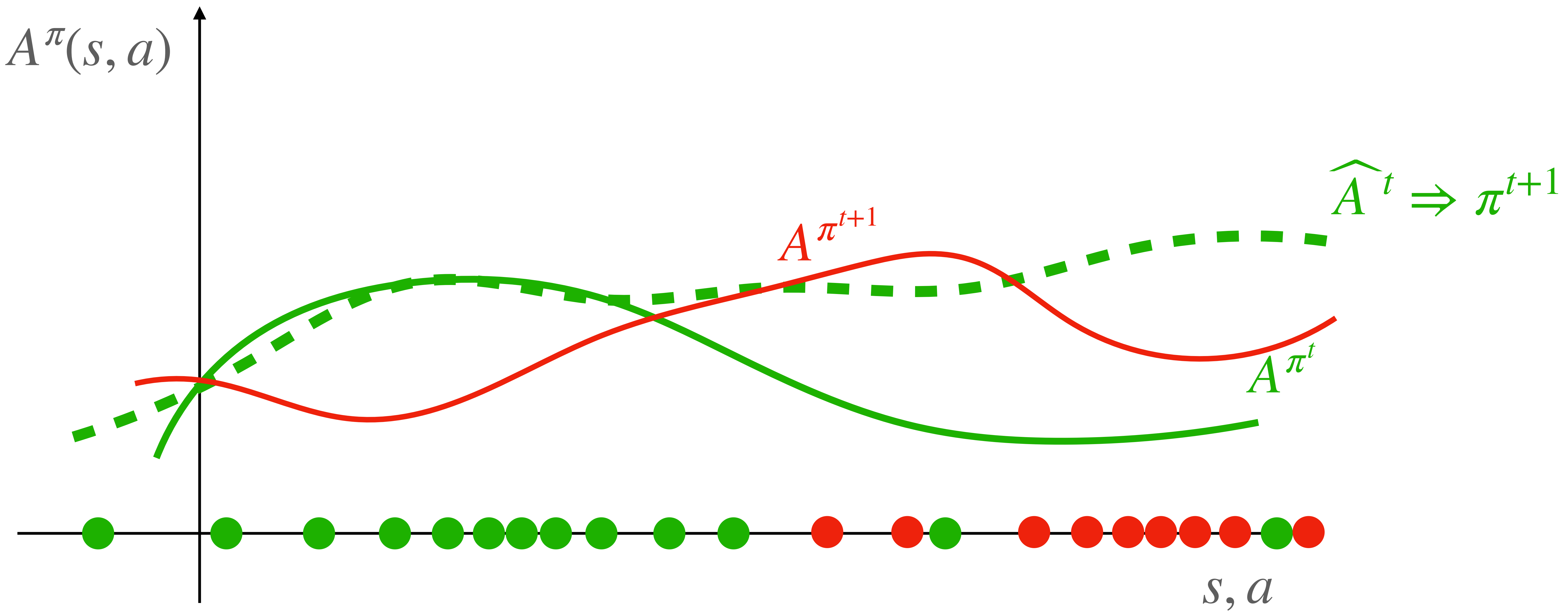
The Failure case of API: Abrupt distribution change

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



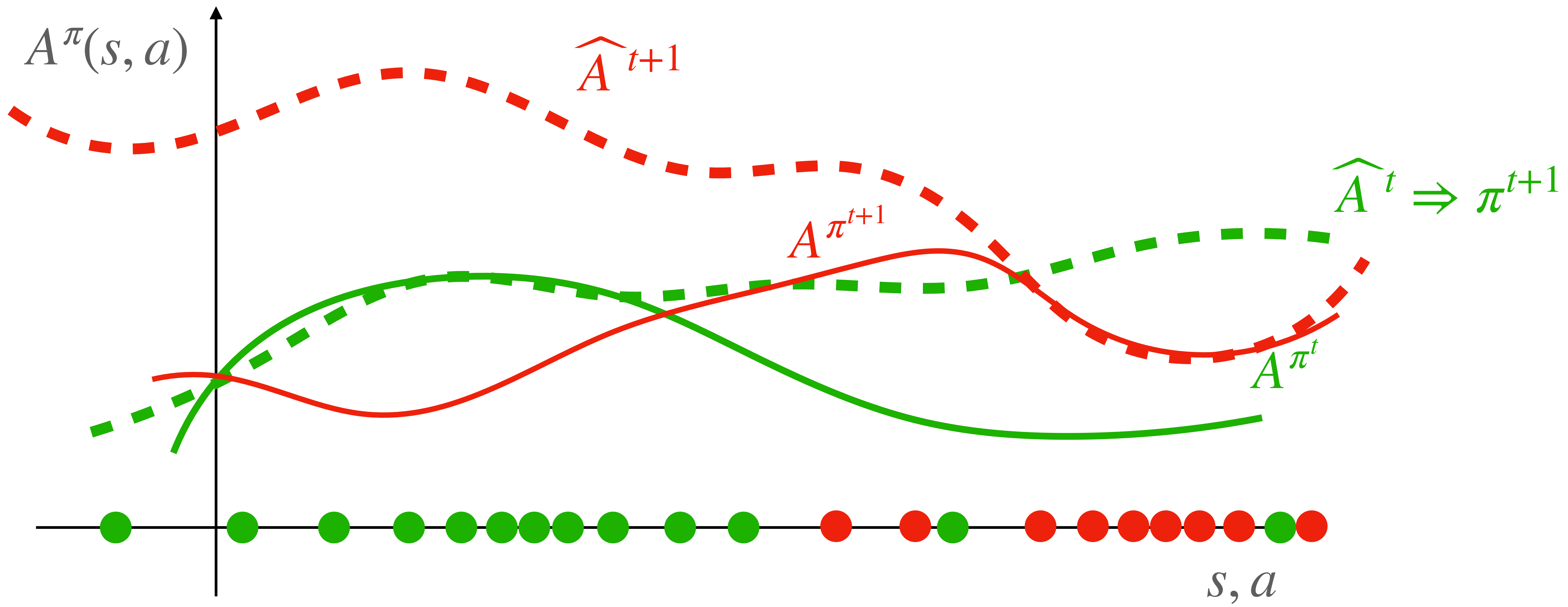
The Failure case of API: Abrupt distribution change

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



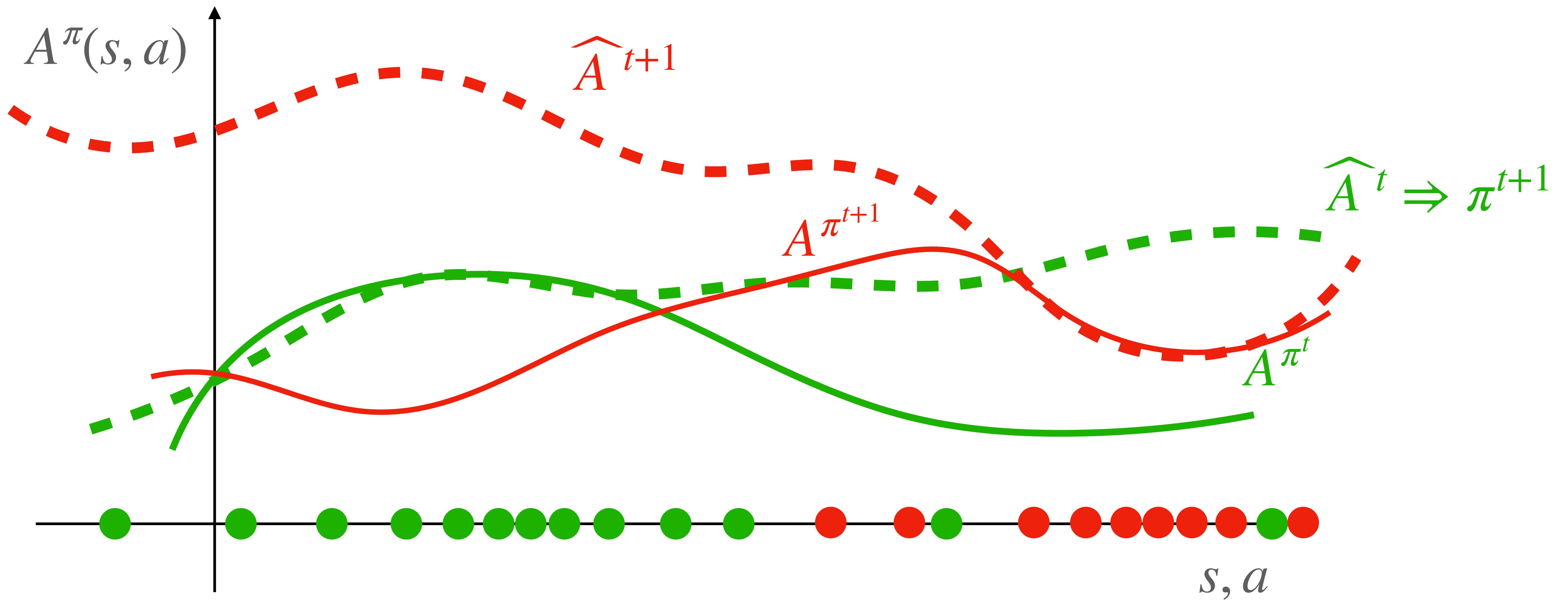
The Failure case of API: Abrupt distribution change

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



The Failure case of API: Abrupt distribution change

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



Oscillation between two updates: No monotonic improvement

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

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

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

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

Recall Performance Difference Lemma:

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

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

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

Let's design policy update rule such that $d_\mu^{\pi^{t+1}}$ and $d_\mu^{\pi^t}$ 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., $\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s, \pi^{t+1}(s))\right] \approx \mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\left[A^{\pi^t}(s, \pi^{t+1}(s))\right]$

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

Let's design policy update rule such that $d_\mu^{\pi^{t+1}}$ and $d_\mu^{\pi^t}$ 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., $\mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right] \approx \mathbb{E}_{s \sim d_\mu^{\pi^{t+1}}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right]$

This we know how to optimize: the Greedy Policy Selector

The CPI Algorithm

Initialize π^0

For t = 0 …

The CPI Algorithm

Initialize π^0

For t = 0 …

1. Greedy Policy Selector:

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

The CPI Algorithm

Initialize π^0

For t = 0 …

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

The CPI Algorithm

Initialize π^0

For t = 0 …

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:

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

The CPI Algorithm

Initialize π^0

For t = 0 …

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:

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

Q: Why this is incremental? In what sense?

Q: Can we get monotonic policy improvement?

Today: Policy Optimization

1. Greedy Policy Selection

2. Conservative Policy Iteration

3. Monotonic Improvement of CPI

Q1: Why this is incremental? In what sense?

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

Q1: Why this is incremental? In what sense?

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

Key observation 1:

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

Q1: Why this is incremental? In what sense?

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

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 (Lemma 12.1 in AJKS)

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

Q1: Why this is incremental? In what sense?

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

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 (Lemma 12.1 in AJKS)

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

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

Monotonic 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:

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

Monotonic 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 $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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:

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

Monotonic Improvement before Termination:

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

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

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

   Return π^t

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

Monotonic 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 $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

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:

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

Monotonic 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 $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

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

Return π^t

3. Incremental Update:

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

Monotonic Improvement before Termination:

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

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

Can we translate local advantage $\mathbb{A}$ to $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

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

Monotonic 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 $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

Why?

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

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

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:

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

Monotonic 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 $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

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

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

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

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:

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

Monotonic 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 $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

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

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

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

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

Monotonic 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 $V^{\pi^{t+1}} - V^{\pi^t}$? (Yes, by PDL)

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

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

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

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

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

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

Summary of CPI so far:

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \le \frac{2\gamma\alpha}{1-\gamma}$$

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

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

Local adv versus distribution change:

Summary of CPI so far:

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

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

Local adv versus distribution change:

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

Summary of CPI so far:

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:

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

1. Incremental update (Lemma 12.1 in AJKS)

$$\| d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t} \|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

Local adv versus distribution change:

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

(By setting step size α properly…)

Summary of CPI so far:

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly...)

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:

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

Local adv versus distribution change:

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:

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

Summary of CPI so far:

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \le \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \ge \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly…)

Local adv versus distribution change:

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}$$

Summary of CPI so far:

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly…)

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:

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

Local adv versus distribution change:

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}$$

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

Summary of CPI so far:

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly...)

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:

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

Local adv versus distribution change:

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}$$

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

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

Summary for today:

1. Algorithm: Conservative Policy Iteration:
Find the local greedy policy, and move towards it a little bit

2. Small change in policies results small change in state distributions

2. Unlike API, incremental policy update ensures monotonic improvement