# 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 $\pi$ 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 $\pi$ 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 $\pi$ 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 $s_0 \in S$

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

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

# 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:

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

# 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:

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

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

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

# 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 $\Pi$:

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

# Setting and Notation

Discounted infinite horizon MDP:

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

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

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 $\Pi$:

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

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

# Attempt One: Approximate Policy Iteration (API)

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

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

# Attempt One: Approximate Policy Iteration (API)

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

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

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

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

Given the current policy $\pi^t$, let's act greedily wrt $\pi$ 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 $\pi^t$, let's act greedily wrt $\pi$ 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 $A^{\pi^t}(s,a)$

# Attempt One: Approximate Policy Iteration (API)

Given the current policy $\pi^t$, let's act greedily wrt $\pi$ 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 $A^{\pi^t}(s,a)$

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

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

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

# Implementing Approximate Greedy Policy Selector via Regression

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

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

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

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

# 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:

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

# 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 $\hat{f} \approx A^{\pi^t}$):

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

# 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**

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

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

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

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

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

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

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

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

## Summary So Far:

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

# Summary So Far:

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

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

# Summary So Far:

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

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

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),

# Summary So Far:

**By reduction to Supervised Learning** (i.e., classification using $\Pi$ 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:

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

# Algorithm: Approximate Policy Iteration (API)

Iterate:

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

Question:

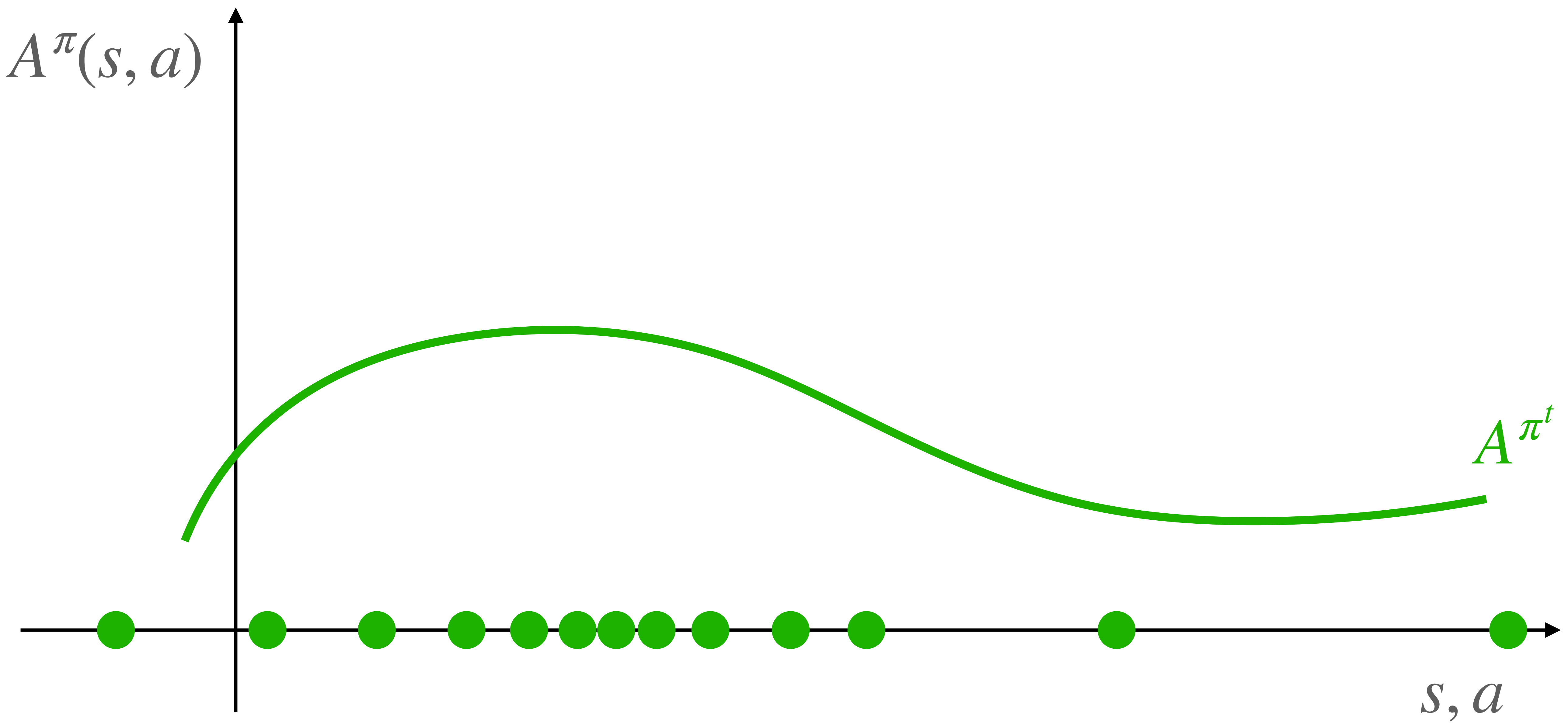
(1) Does API has monotonic improvement?

(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)

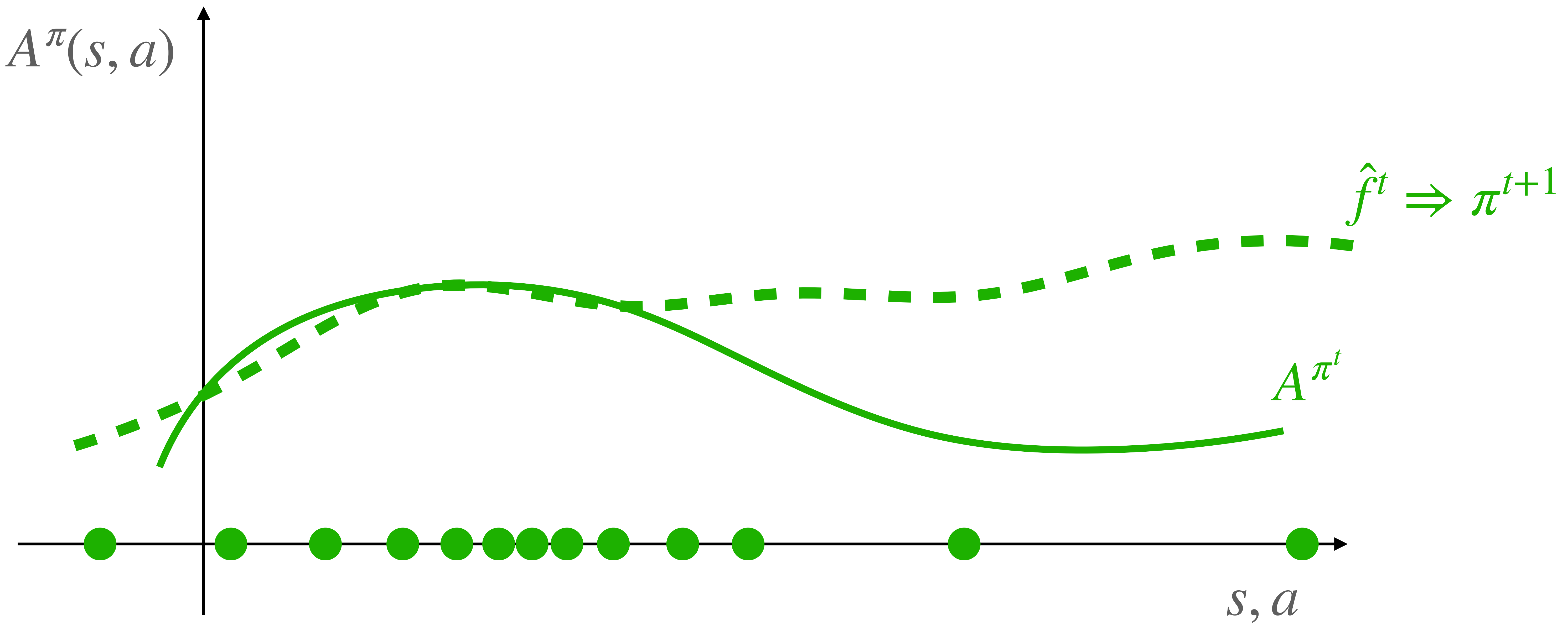
**Concrete example in Chapter 3**



# The Oscillation of API from Abrupt Distribution Change

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

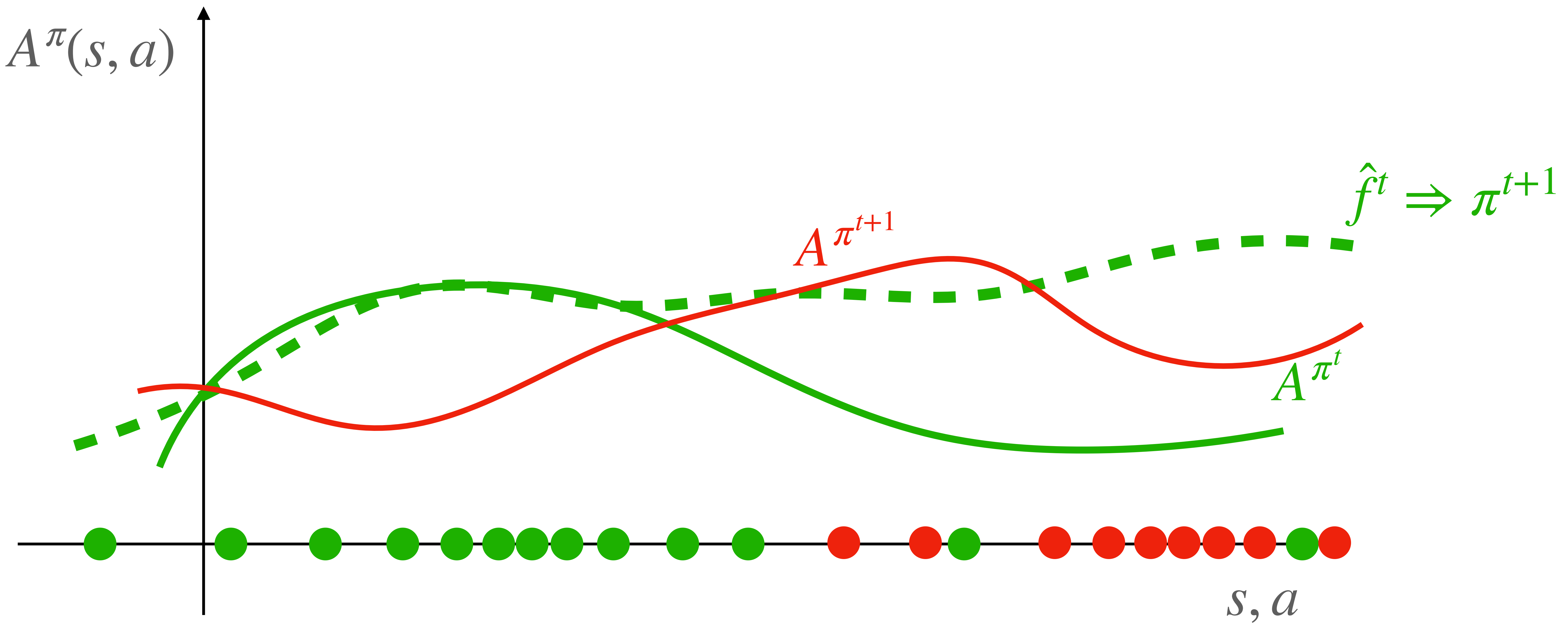
**Concrete example in Chapter 3**



# The Oscillation of API from Abrupt Distribution Change

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

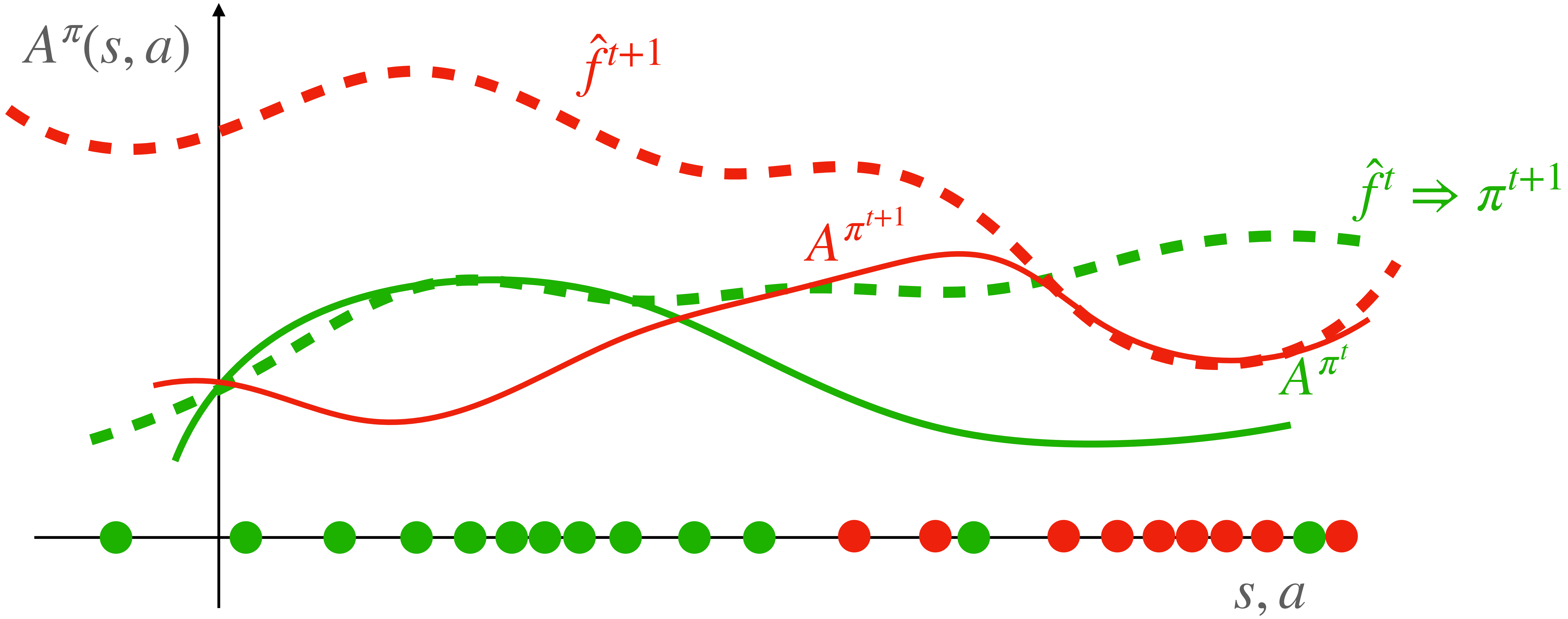
**Concrete example in Chapter 3**



# The Oscillation of API from Abrupt Distribution Change

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

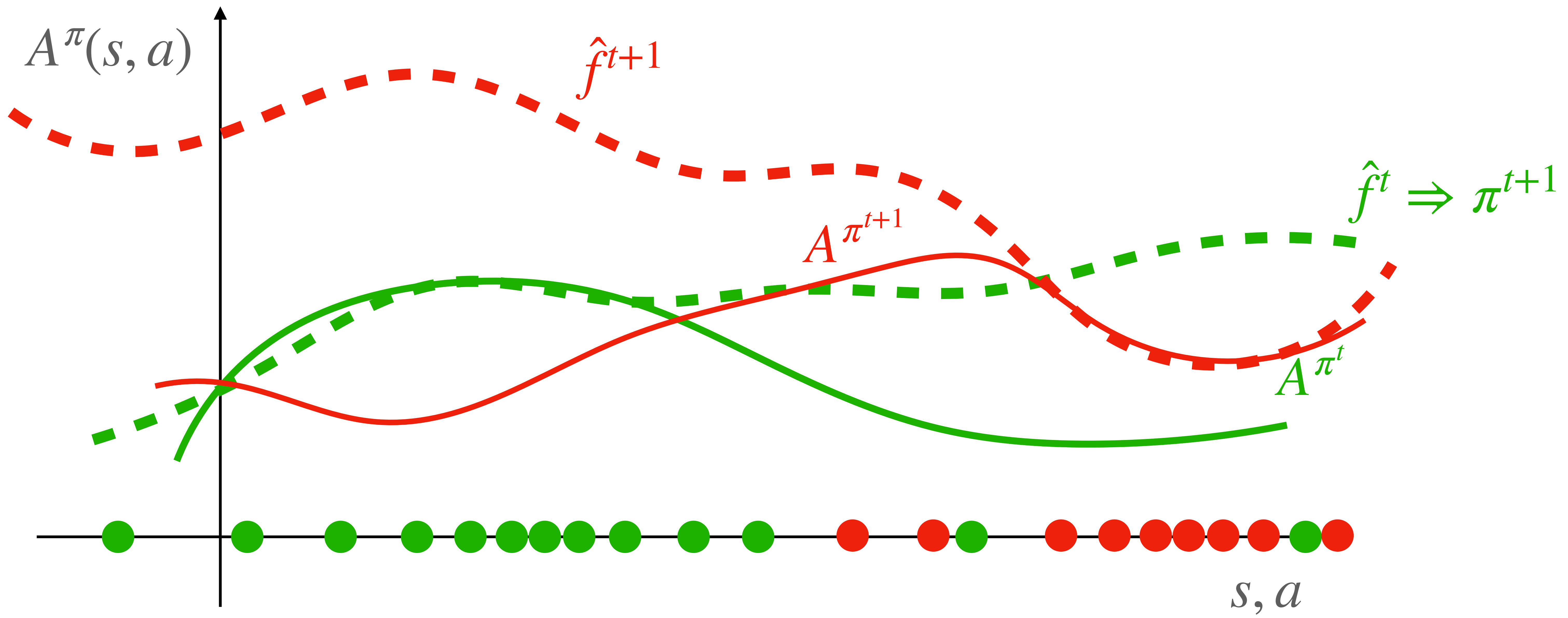
**Concrete example in Chapter 3**



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

# 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:** $C := \max_{\pi \in \Pi} \sup_s \dfrac{d_\mu^\pi(s)}{\mu(s)} < \infty$

## 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., $\mu$ covers **all** $d_\mu^\pi$, 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;

# Conservative Policy Iteration—An Incremental Policy Optimization Approach

(And the benefit of being incremental)

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

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

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

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

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

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

This we know how to optimize: the Greedy Policy Selector

# CPI Algorithm:

# CPI Algorithm:

1. Greedy Policy Selector:

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

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

   **Return** $\pi^t$

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

   **Return** $\pi^t$

3. Incremental Update:

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

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

   **Return** $\pi^t$

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:

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

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

# 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 $\pi$ and $\pi'$, 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.,** $\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^{t}}\|_1 \leq \dfrac{2\gamma\alpha}{1-\gamma}$

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** $\pi^t$

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**:

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

## 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** $\pi^t$

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** $\pi^t$

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 $V^{\pi^{t+1}} - V^{\pi^t}$?

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

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

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** $\pi^t$

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 $V^{\pi^{t+1}} - V^{\pi^t}$?

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

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

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

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** $\pi^t$

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 $V^{\pi^{t+1}} - V^{\pi^t}$?

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

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

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

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

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** $\pi^t$

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 $V^{\pi^{t+1}} - V^{\pi^t}$?

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

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

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

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

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

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** $\pi^t$

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 $V^{\pi^{t+1}} - V^{\pi^t}$?

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

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

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

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

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

(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** $\pi^t$

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 $V^{\pi^{t+1}} - V^{\pi^t}$?

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

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

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

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

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

(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** $\pi^t$

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

# Upon Termination we get a locally optimal solution (or globally optimal if $\mu$ 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 $\pi^t$**

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ 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** $\pi^t$

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

# Upon Termination we get a locally optimal solution (or globally optimal if $\mu$ 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 $\pi^t$**

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \dfrac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

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

# Upon Termination we get a locally optimal solution (or globally optimal if $\mu$ 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 $\pi^t$**

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \dfrac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall $\Pi$ 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 $\mu$ 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 $\pi^t$**

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall $\Pi$ 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 $\mu$ 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 $\pi^t$**

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \dfrac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

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

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

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

# Upon Termination we get a locally optimal solution (or globally optimal if $\mu$ 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** $\pi^t$

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

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

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

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

# Upon Termination we get a locally optimal solution (or globally optimal if $\mu$ 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 $\pi^t$**

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \dfrac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

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

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

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

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

# Upon Termination we get a locally optimal solution (or globally optimal if $\mu$ 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 $\pi^t$**

1. No more positive advantage by one-step deviation from $\pi^t$'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 $\mu$ covers $d_\mu^{\pi^\star}$, i.e., $C^\star := \sup_s \frac{d_\mu^{\pi^\star}(s)}{\mu(s)} < \infty$

Recall $\Pi$ 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 $\Pi$; but we cannot guarantee $\pi^\star \in \Pi$;

# 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 $\Pi$; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** $\Pi$

# 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 $\Pi$; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** $\Pi$

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

# 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 $\Pi$; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** $\Pi$

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

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 $\Pi$; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** $\Pi$

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

---

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

# 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 $\Pi$; but we cannot guarantee $\pi^\star \in \Pi$;

What we can hope is that we can find **the best classifier in the class** $\Pi$

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

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:

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 $\Pi$,
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 $\Pi$,
steady improvement via incremental update

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

1. Prior knowledge of how the optimal trajectories look like

2. Expert demonstrations (Imitation + RL)

# 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: $\mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla \ln \pi_\theta(a\,|\,s) A^{\pi_\theta}(s, a)$

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

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

# 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:

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

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

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

# 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:

$V^{\pi_\theta} + \lambda \sum_s \sum_a \ln \pi_\theta(a \mid s)$, 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:**

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