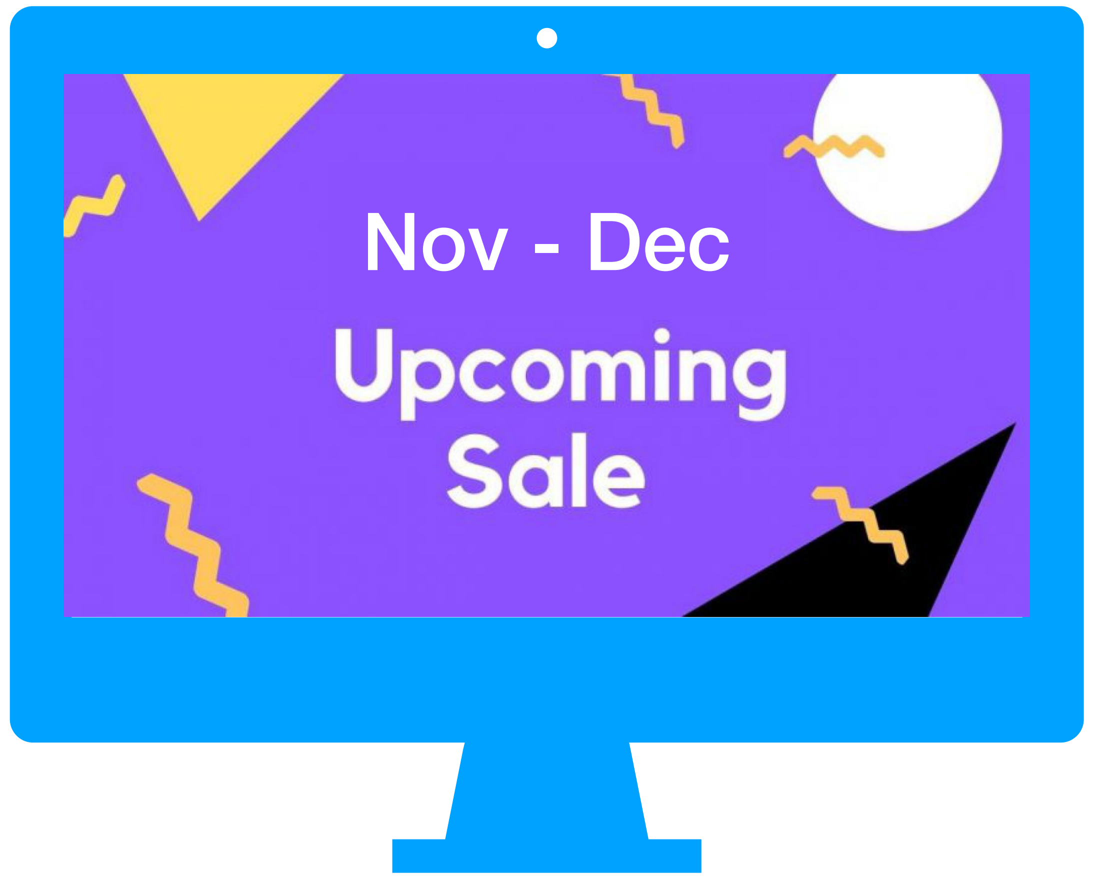
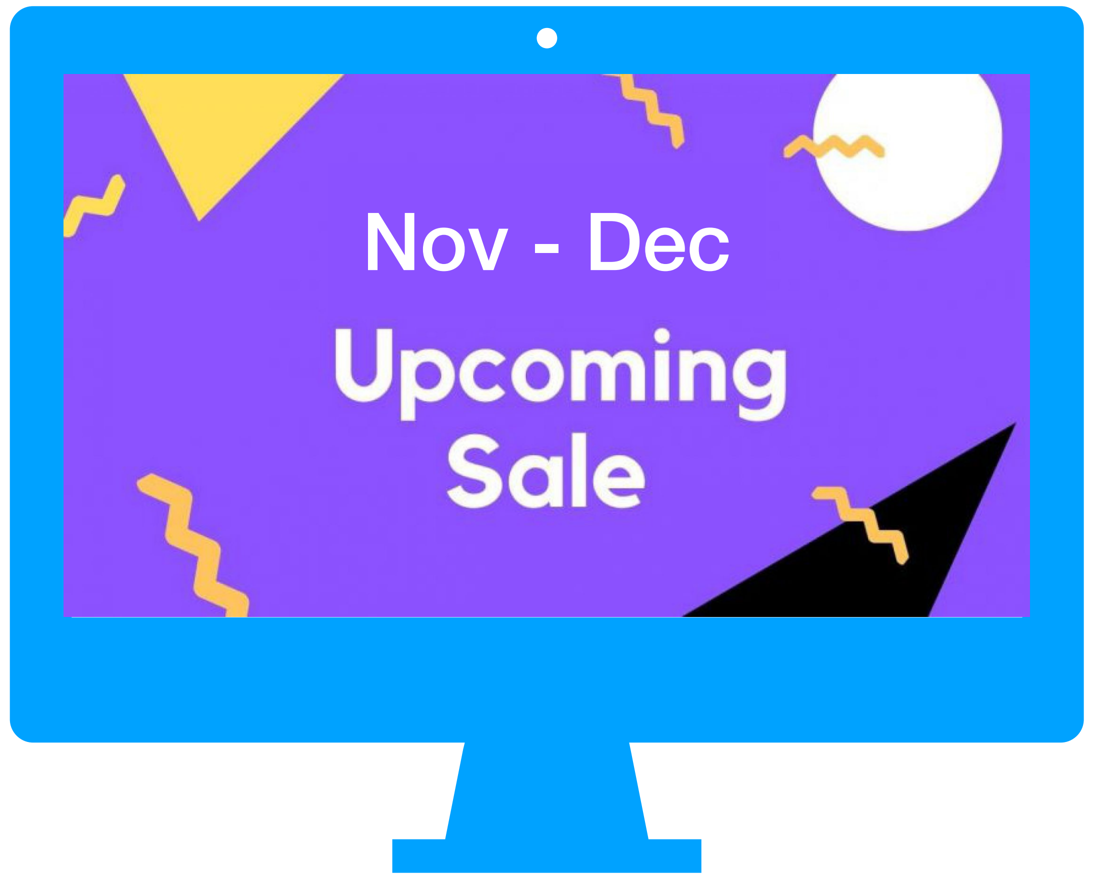
Estimation and Inference of Optimal Policies

Zhaoqi Li

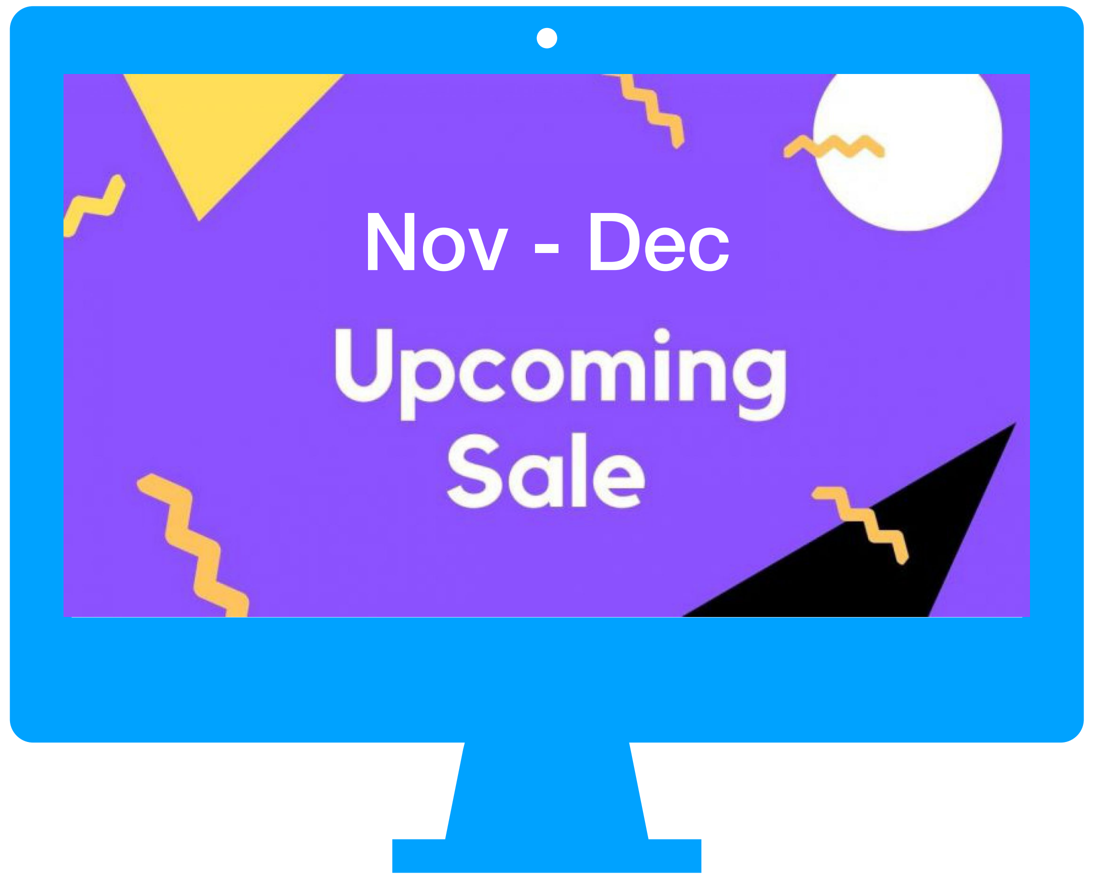
Motivation

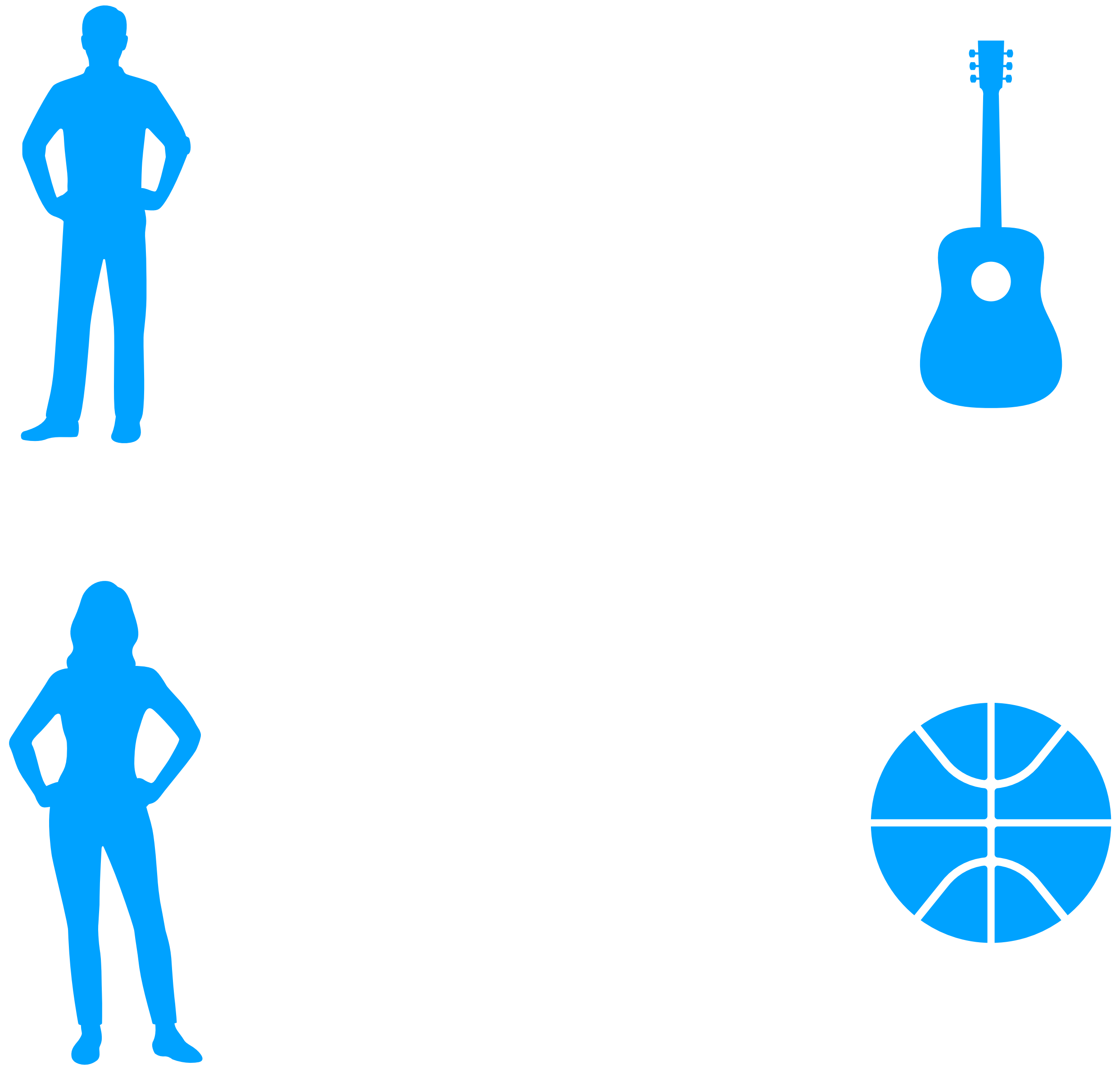


Motivation

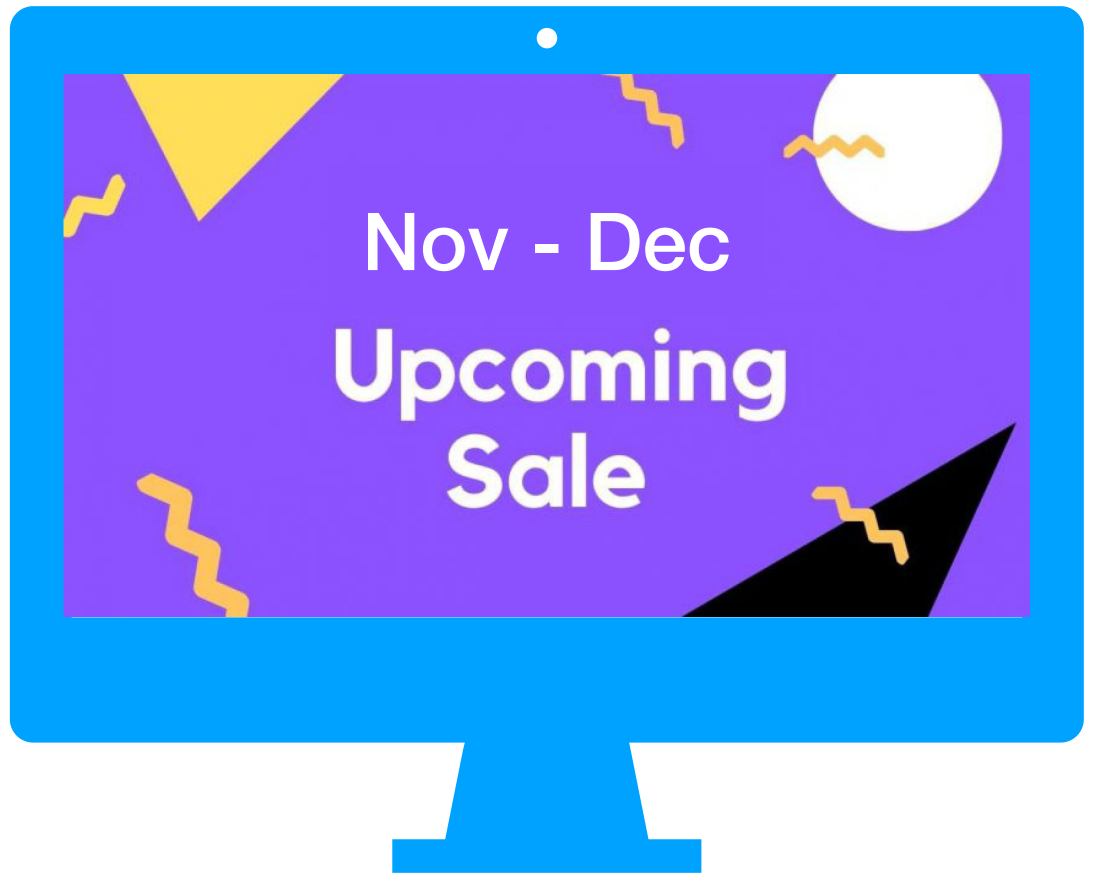


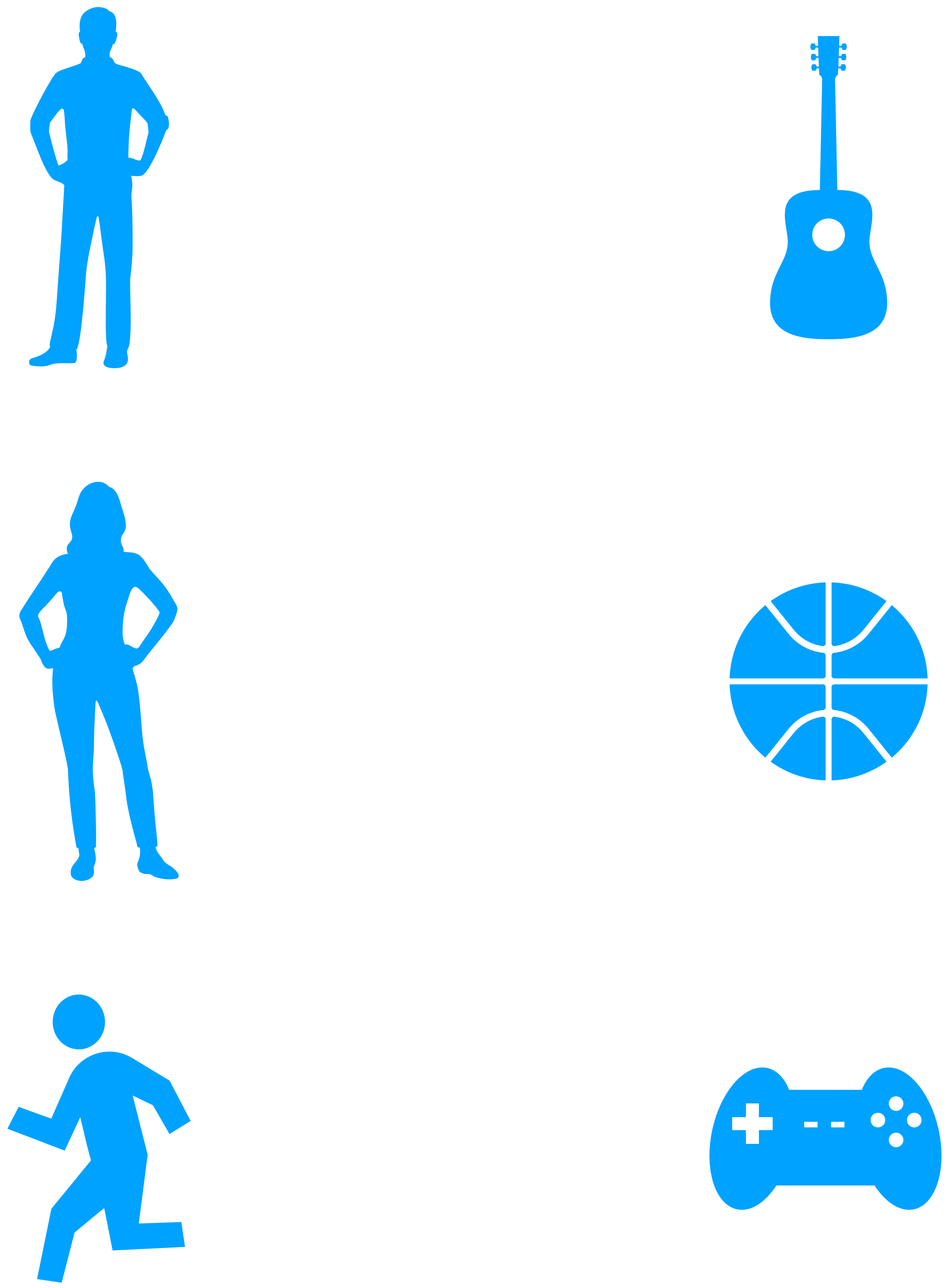
Motivation



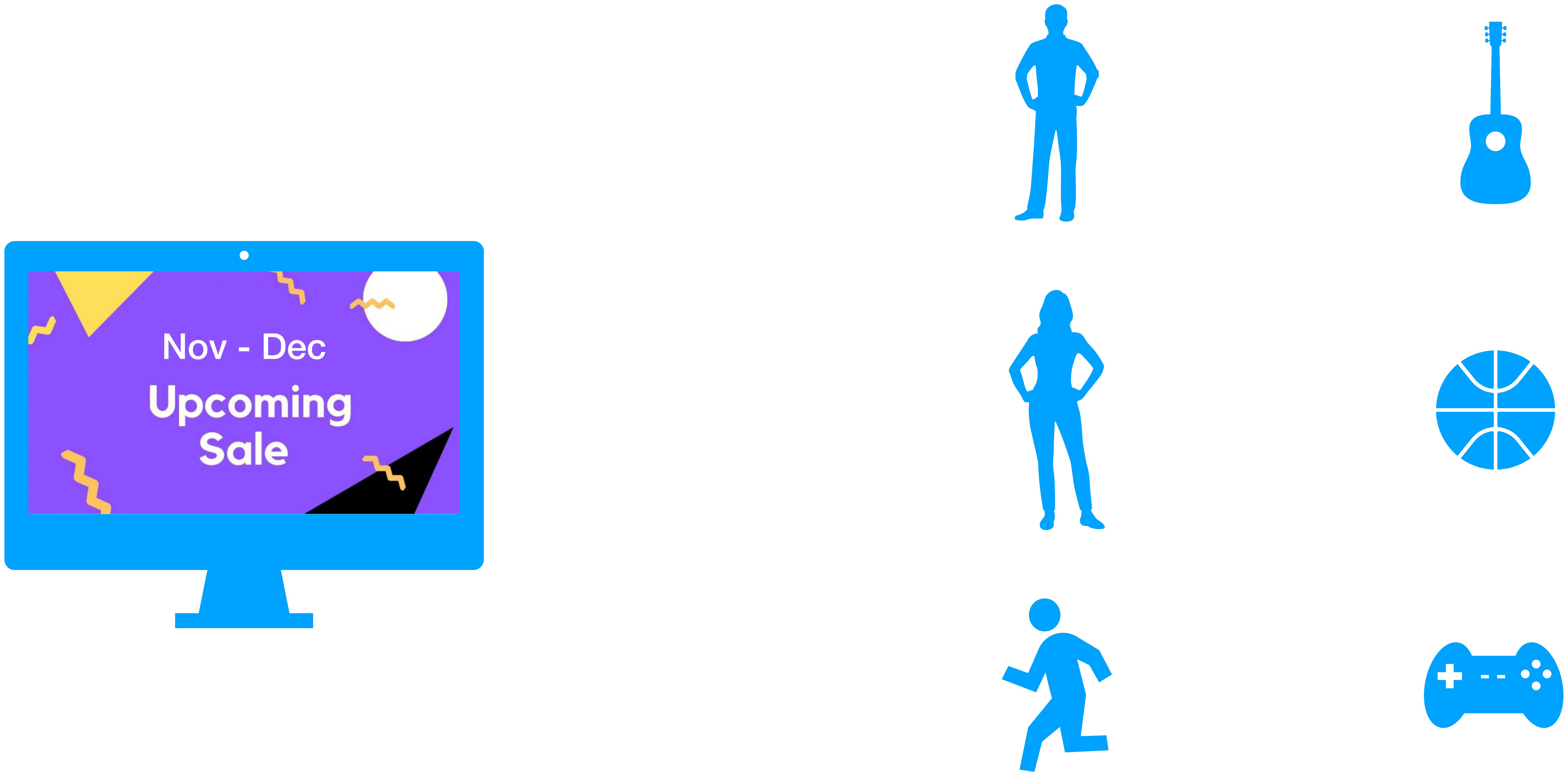


Motivation





Motivation

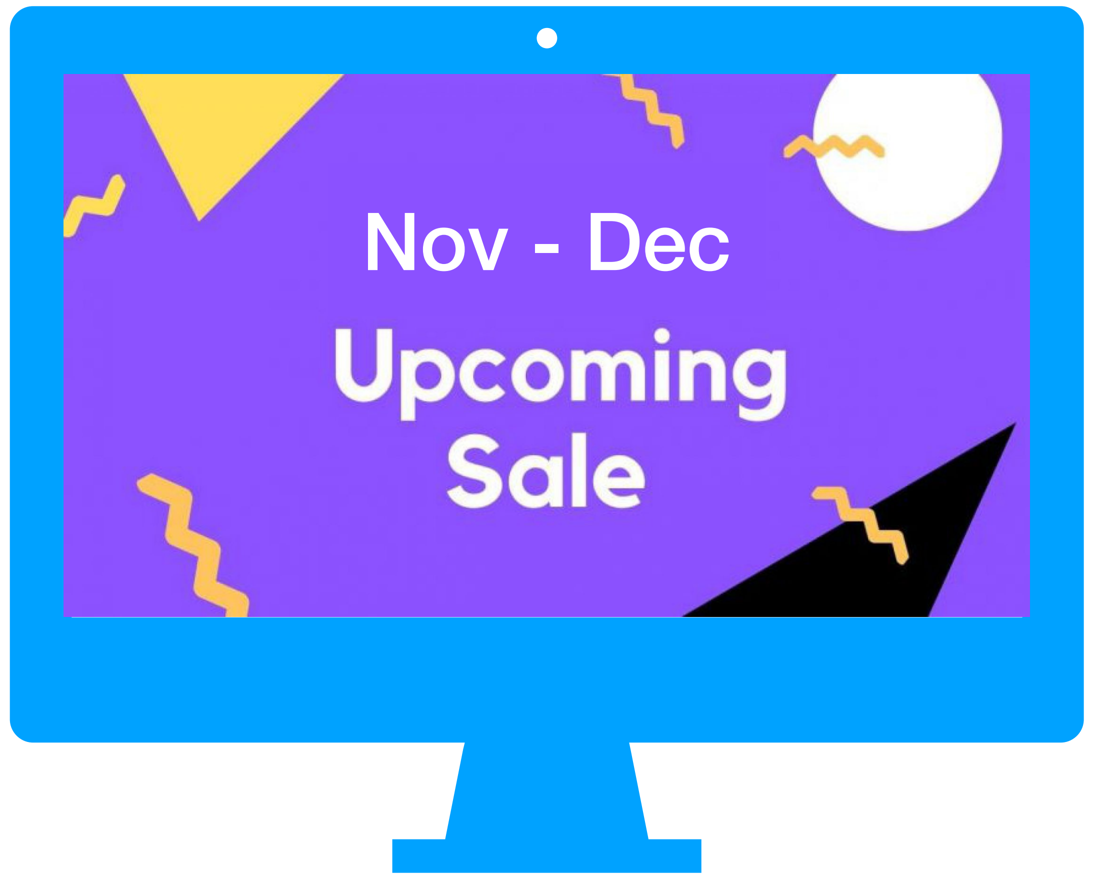


Question: What is the best way to give personalized recommendations to maximize revenue?

Motivation

policy

π

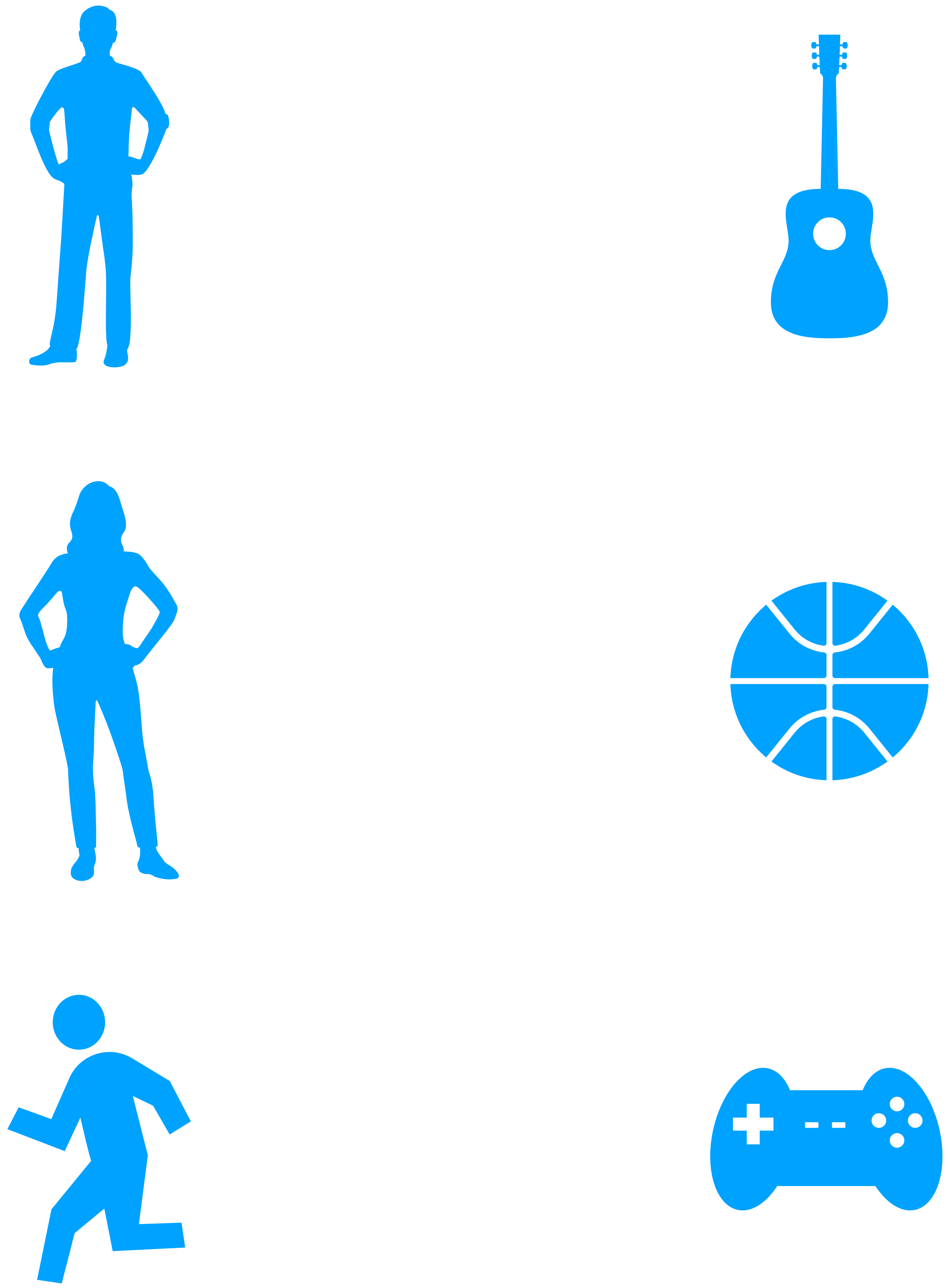


context

c

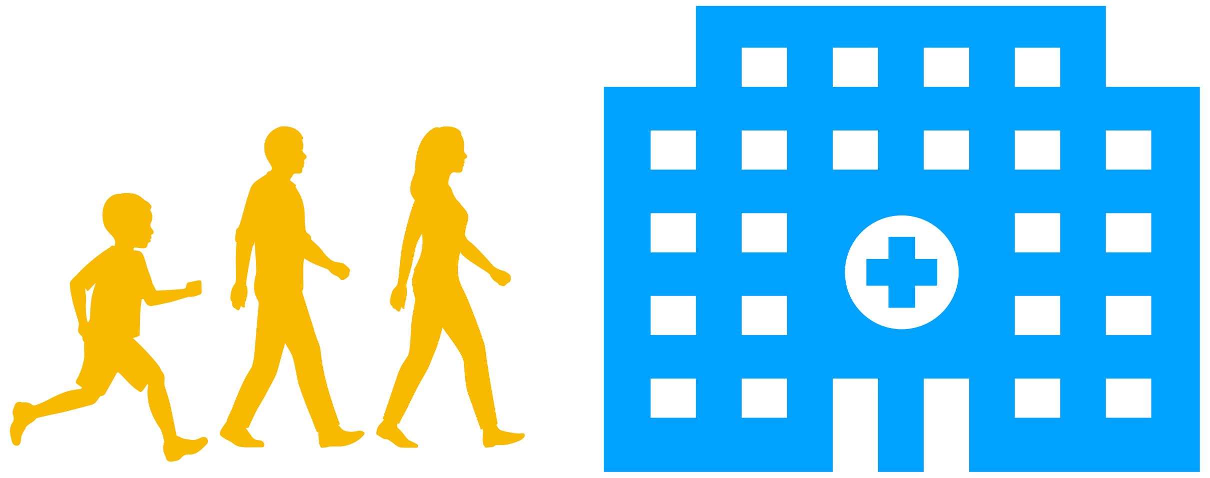
action

a

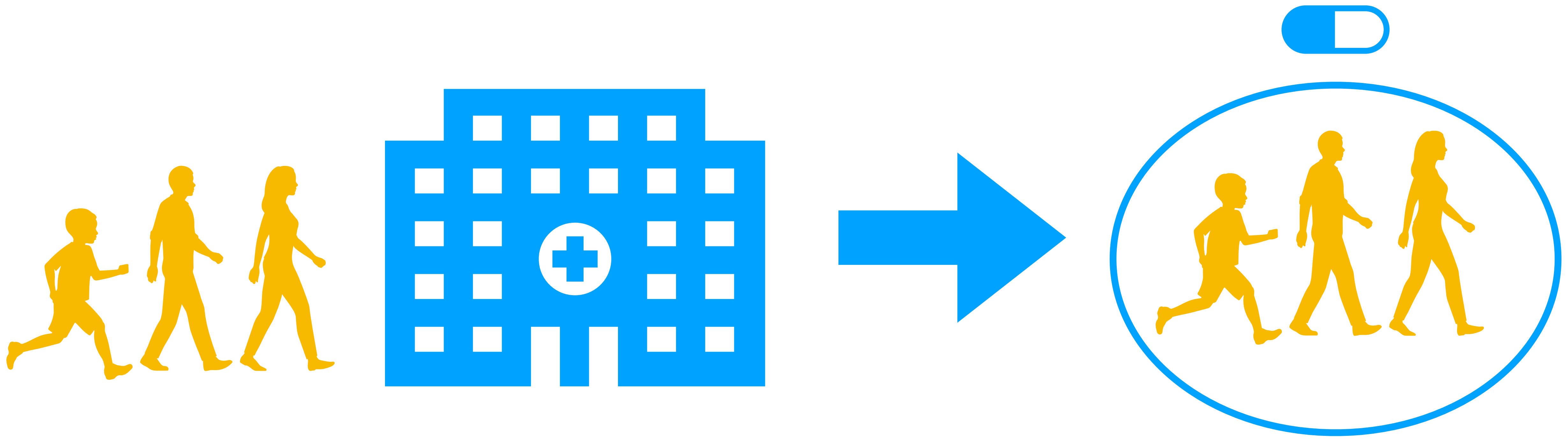


Question: What is the best way to give personalized recommendations to maximize revenue?

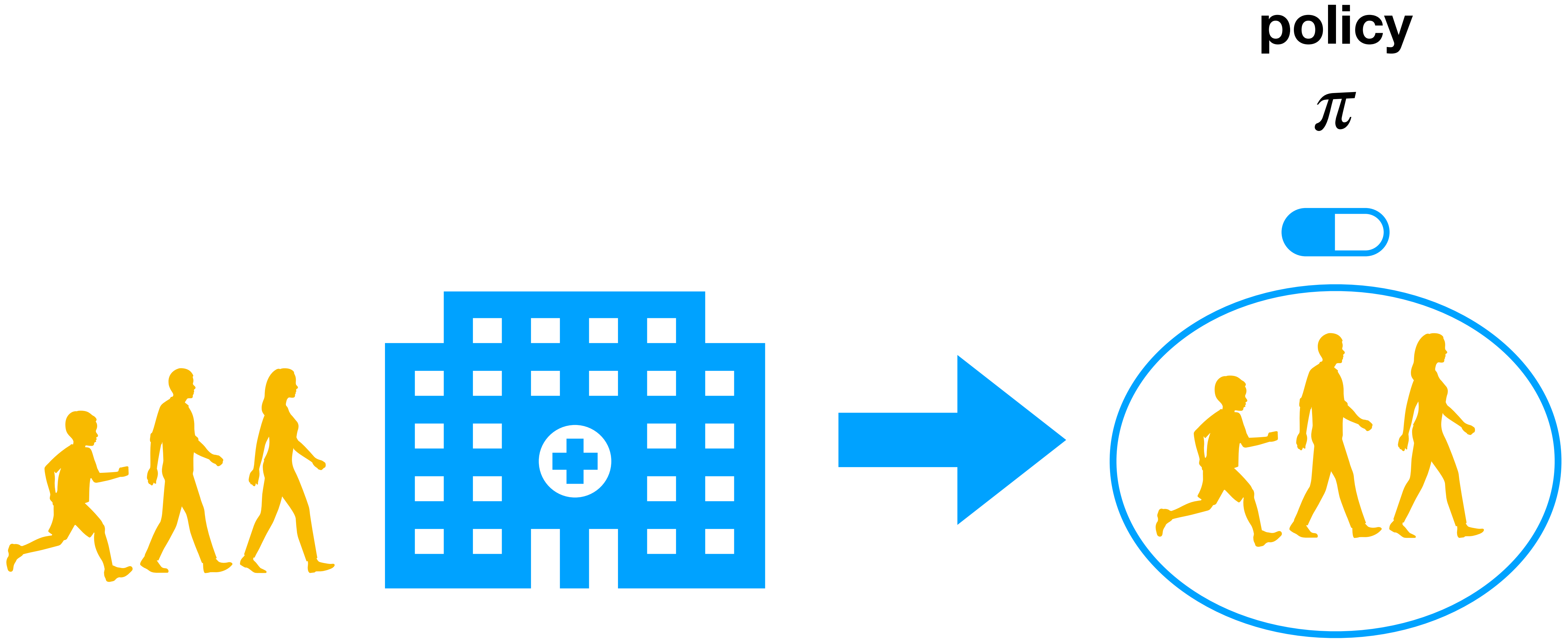
Motivation



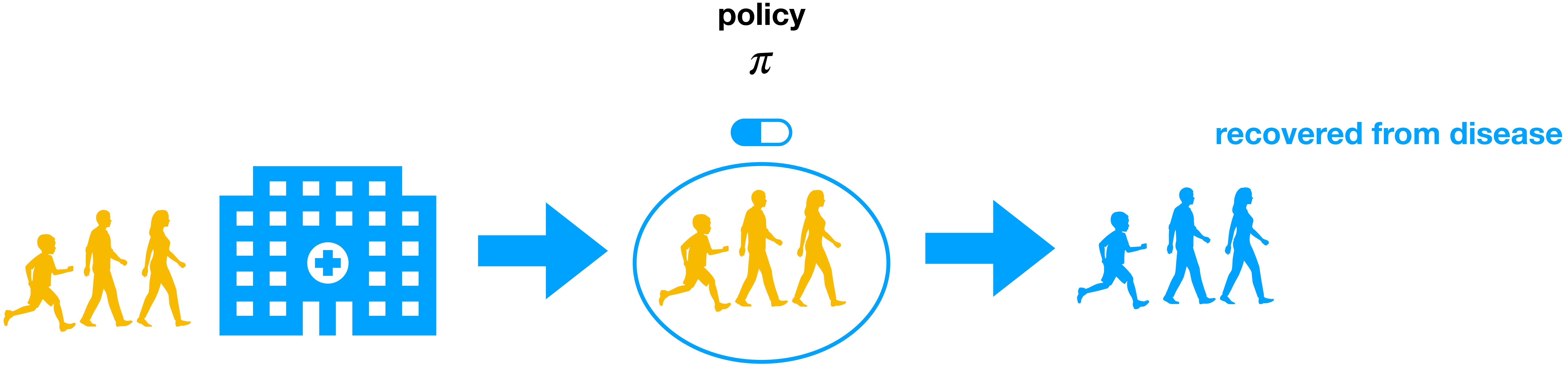
Motivation



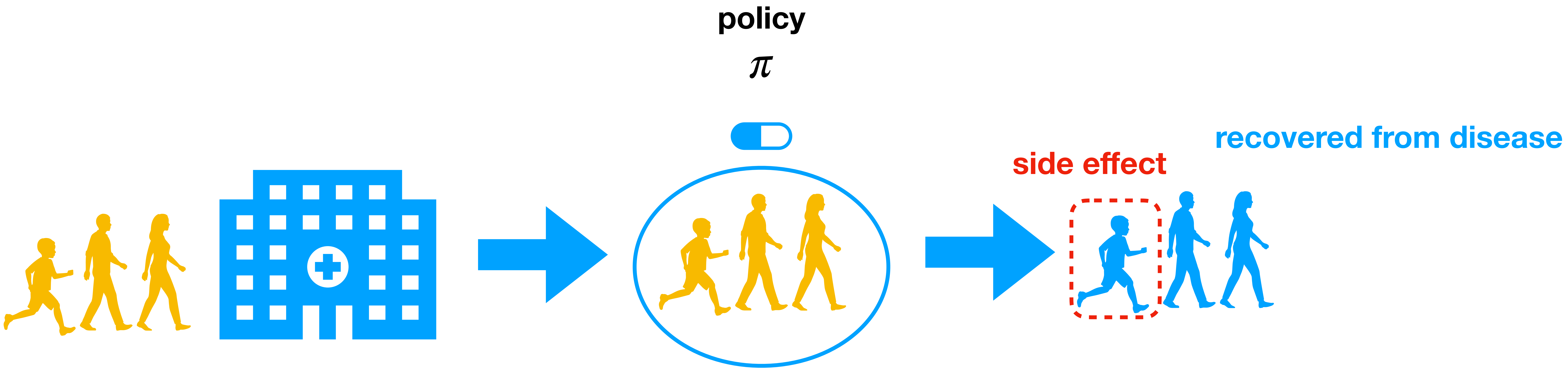
Motivation



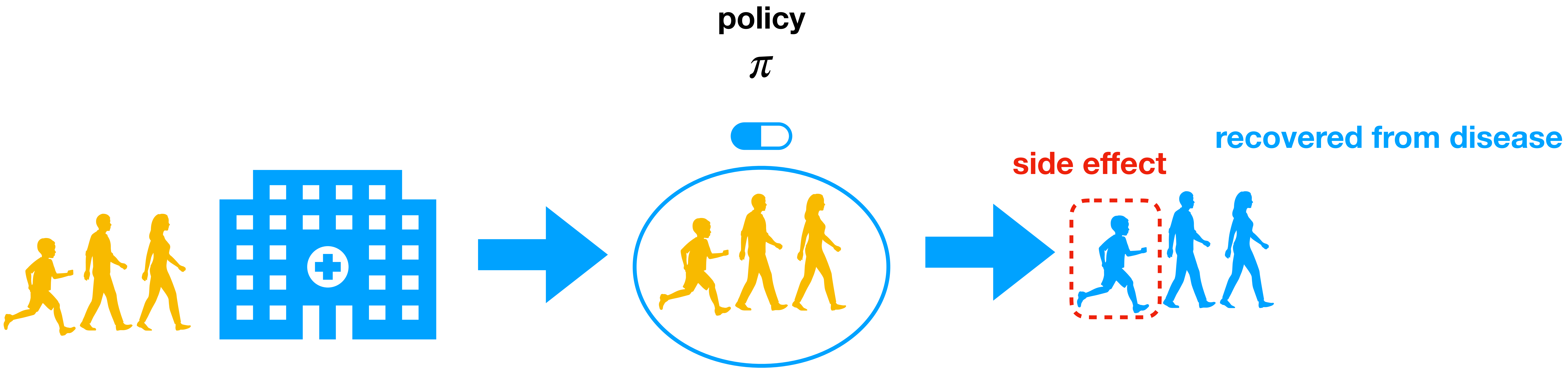
Motivation



Motivation



Motivation



policy

π

side effect

recovered from disease

Question: how do we characterize the amount of side effects when the treatment allocation is optimized for disease remission?

Outline

- Project 1: Instance-optimal PAC Contextual bandits

- Project 2: Estimation of the mean of subsidiary outcome

- Future Work

Instance-Optimal PAC Algorithms for Contextual Bandits

Zhaoqi Li*, Lillian Ratliff*, Houssam Nassif†, Kevin Jamieson*, Lalit Jain*

*University of Washington

†Amazon

Contextual Bandit Setting

- **At each time $t = 1,2,\cdots$:**
  - Context $c_t \in \mathrm{C}$ arrives, $c_t \sim \nu \in \Delta_{\mathrm{C}}$
  - Choose action $a_t \in \mathrm{A}$
  - Receive reward r_t, $\mathbb{E}[r_t \,|\, c_t, a_t] = r(c_t, a_t) \in \mathbb{R}$

Contextual Bandit Setting

- **At each time** $t = 1,2,\cdots$**:**
  - Context $c_t \in \mathsf{C}$ arrives, $c_t \sim \nu \in \Delta_{\mathsf{C}}$
  - Choose action $a_t \in \mathsf{A}$
  - Receive reward r_t, $\mathbb{E}[r_t \mid c_t, a_t] = r(c_t, a_t) \in \mathbb{R}$

- Policy class Π, each $\pi \in \Pi$, $\pi : \mathsf{C} \to \mathsf{A}$
- Value function: $V(\pi) := \mathbb{E}_{c \sim \nu}[r(c, \pi(c))]$
- Optimal policy: $\pi_* := \arg\max_{\pi \in \Pi} V(\pi)$

Contextual Bandit Setting

- **At each time $t = 1,2,\cdots$:**
  - Context $c_t \in$ C arrives, $c_t \sim \nu \in \Delta_{\mathrm{C}}$
  - Choose action $a_t \in$ A
  - Receive reward r_t, $\mathbb{E}[r_t \,|\, c_t, a_t] = r(c_t, a_t) \in \mathbb{R}$

- Policy class Π, each $\pi \in \Pi$, $\pi :$ C $\rightarrow$ A
- Value function: $V(\pi) := \mathbb{E}_{c \sim \nu}[r(c, \pi(c))]$
- Optimal policy: $\pi_* := \arg\max_{\pi \in \Pi} V(\pi)$

(ϵ, δ) – PAC Guarantee

Return $\hat{\pi}$ satisfying, $V(\hat{\pi}) \geq V(\pi_*) - \epsilon$ with probability greater than $1 - \delta$ in a minimum number of samples.

Regret Minimization vs. Policy Identification

Regret Minimization vs. Policy Identification

- Regret heavily studied:

$$R_T = \sum_{t=1}^{T} \left[r(c_t, \pi_*(c_t)) - r(c_t, a_t) \right]$$

Regret Minimization vs. Policy Identification

- Regret heavily studied:

$$R_T = \sum_{t=1}^{T} \left[r(c_t, \pi_*(c_t)) - r(c_t, a_t) \right]$$

- EXP4 achieves a minimax-optimal regret bound of $R_T = O(\sqrt{|\mathrm{A}|T\log(|\Pi|)})$, also achieved by ILOVETOCONBANDITS [Agarwal et al. 2014] and computationally efficient

Regret Minimization vs. Policy Identification

- Regret heavily studied:

$$R_T = \sum_{t=1}^{T} \left[r(c_t, \pi_*(c_t)) - r(c_t, a_t) \right]$$

- EXP4 achieves a minimax-optimal regret bound of $R_T = O(\sqrt{|\mathrm{A}| T \log(|\Pi|)})$, also achieved by ILOVETOCONBANDITS [Agarwal et al. 2014] and computationally efficient
- Modification gives (ϵ, δ)- PAC algorithm w/ sample complexity $O(|\mathrm{A}| \log(|\Pi|/\delta)/\epsilon^2)$, also see [Zanette et al. 2021]

Problems with Regret Minimization

Problems with Regret Minimization

- Minimax result! Does not adapt to hardness of instance

Problems with Regret Minimization

- Minimax result! Does not adapt to hardness of instance
  - $O(|A|\log(|\Pi|/\delta)/\epsilon^2)$: true for any policy class, does not capture the difficulty for learning π_*

Problems with Regret Minimization

- Minimax result! Does not adapt to hardness of instance
    - $O(|\mathsf{A}|\log(|\Pi|/\delta)/\epsilon^2)$: true for any policy class, does not capture the difficulty for learning π_*
- We are interested in **instance optimality**, i.e. optimal for each instance Π

Problems with Regret Minimization

- Minimax result! Does not adapt to hardness of instance
    - $O(|\mathsf{A}|\log(|\Pi|/\delta)/\epsilon^2)$: true for any policy class, does not capture the difficulty for learning π_*
- We are interested in **instance optimality**, i.e. optimal for each instance Π
- Can construct an example, where any optimal regret algorithm won't be instance optimal!

Problems with Regret Minimization

- Minimax result! Does not adapt to hardness of instance
  - $O(|A| \log(|\Pi|/\delta)/\epsilon^2)$: true for any policy class, does not capture the difficulty for learning π_*
- We are interested in **instance optimality**, i.e. optimal for each instance Π
- Can construct an example, where any optimal regret algorithm won't be instance optimal!

Theorem [Li et al. 2022] There exists an instance μ such that for any minimax regret algorithm that is $(0,\delta)$-PAC, the stopping time satisfies $\mathbb{E}_\mu[\tau] \geq |\Pi|^2 \log^2(1/(2.4\delta))/4$, which is the lower bound *squared.*

Challenges

Challenges

- What is the statistical limits of learning, i.e. the **instance-dependent** lower bound?

Challenges

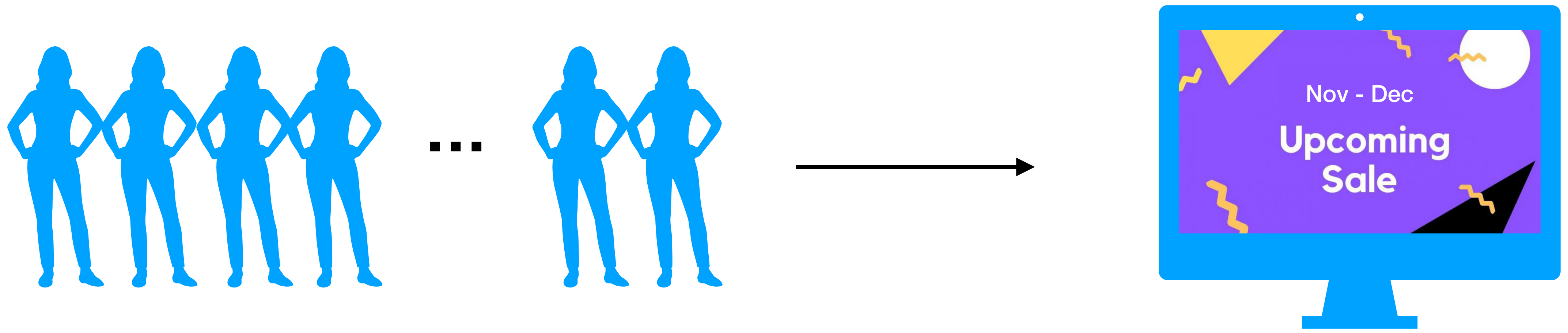
- What is the statistical limits of learning, i.e. the **instance-dependent** lower bound?
- Can we design sampling procedure to achieve this?

Challenges

- What is the statistical limits of learning, i.e. the **instance-dependent** lower bound?
- Can we design sampling procedure to achieve this?
- Computational efficiency - context space C could be **infinite** and Π could be large!

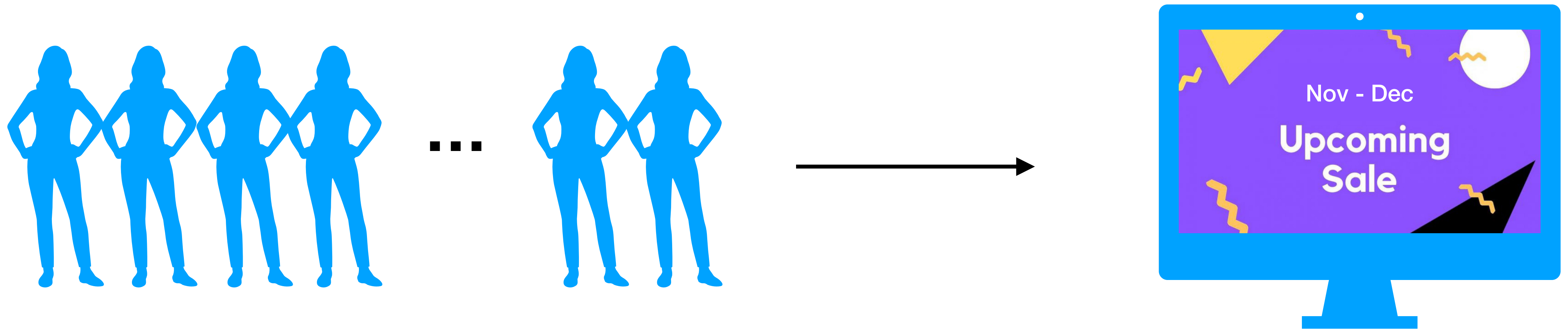
Challenges

- What is the statistical limits of learning, i.e. the **instance-dependent** lower bound?
- Can we design sampling procedure to achieve this?
- Computational efficiency - context space C could be **infinite** and Π could be large!



Challenges

- What is the statistical limits of learning, i.e. the **instance-dependent** lower bound?
- Can we design sampling procedure to achieve this?
- Computational efficiency - context space $\mathcal{C}$ could be **infinite** and Π could be large!



Question: what is possible?

Our Contribution

- Show the first **instance-dependent** lower bound for PAC contextual bandit
- Present a simple algorithm that achieves this lower bound
- Design a **computational efficient** algorithm that also achieves this lower bound

Towards Lower Bound: Estimators

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: ϕ : C × A $\to \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta* \rangle$ for $\theta* \in \Theta \subset \mathbb{R}^d$

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: ϕ : C × A $\rightarrow \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta* \rangle$ for $\theta* \in \Theta \subset \mathbb{R}^d$
- Given dataset D $= \{(c_t, a_t, r_t)\}_{t=1}^{n}$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t)r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta*] = \sum_c \nu_c \sum_a p_{c,a}\phi(c, a)\phi(c, a)^\top \theta*$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta^*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c, a)\phi(c, a)^\top}}_{A(p)} \theta^*$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta^*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c, a)\phi(c, a)^\top}}_{A(p)} \theta^*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} A(p)^{-1} \sum_{t=1}^{n} \phi(c_t, a_t) r_t$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
    - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c,a) = \langle \phi(c,a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t)r_t] = \mathbb{E}_{c,a}[\phi(c,a)\phi(c,a)^\top \theta^*] = \underbrace{\sum_c \nu_c \sum_a p_{c,a}\phi(c,a)\phi(c,a)^\top}_{A(p)} \theta^*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} A(p)^{-1} \sum_{t=1}^{n} \phi(c_t, a_t) r_t$$

IPW estimator!

A Lower Bound

A Lower Bound

- For each $\pi, \pi' \in \Pi$, define the gap $\Delta(\pi, \pi') := V(\pi') - V(\pi)$, let $\Delta(\pi) := \Delta(\pi, \pi_*)$

A Lower Bound

- For each $\pi, \pi' \in \Pi$, define the gap $\Delta(\pi, \pi') := V(\pi') - V(\pi)$, let $\Delta(\pi) := \Delta(\pi, \pi_*)$
- Let $\phi_\pi := \mathbb{E}_{c \sim \nu}[\phi(c, \pi(c))]$, an estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_\pi)^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_\pi) = \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{n}$$

A Lower Bound

- For each $\pi, \pi' \in \Pi$, define the gap $\Delta(\pi, \pi') := V(\pi') - V(\pi)$, let $\Delta(\pi) := \Delta(\pi, \pi_*)$
- Let $\phi_\pi := \mathbb{E}_{c\sim\nu}[\phi(c, \pi(c))]$, an estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_\pi)^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_\pi) = \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{n}$$

Theorem [Li et al. 2022] Let τ be the stopping time of the algorithm. Any $(0,\delta)$-PAC algorithm satisfies $\mathbb{E}[\tau] \geq \rho_{\Pi,0} \log(1/2.4\delta)$ where

$$\rho_{\Pi,\epsilon} := \min_{p_c \in \triangle_A, \forall c \in C} \max_{\pi \in \Pi \setminus \pi_*} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{(\Delta(\pi) \vee \epsilon)^2}.$$

variance

gap

A Lower Bound

- For each $\pi, \pi' \in \Pi$, define the gap $\Delta(\pi, \pi') := V(\pi') - V(\pi)$, let $\Delta(\pi) := \Delta(\pi, \pi_*)$
- Let $\phi_\pi := \mathbb{E}_{c\sim\nu}[\phi(c, \pi(c))]$, an estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_\pi)^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_\pi) = \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{n}$$

Theorem [Li et al. 2022] Let τ be the stopping time of the algorithm. Any $(0,\delta)$-PAC algorithm satisfies $\mathbb{E}[\tau] \geq \rho_{\Pi,0} \log(1/2.4\delta)$ where

$$\rho_{\Pi,\epsilon} := \min_{p_c \in \triangle_A, \forall c \in C} \max_{\pi \in \Pi \backslash \pi_*} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{(\Delta(\pi) \vee \epsilon)^2}.$$

variance

gap

instance-dependent!

An Instance-Optimal Algorithm

An Instance-Optimal Algorithm

An Instance-Optimal Algorithm

Input: Π

An Instance-Optimal Algorithm

```
**Input:** $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
```

An Instance-Optimal Algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
```

An Instance-Optimal Algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
```

An Instance-Optimal Algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
        $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$
```

An Instance-Optimal Algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
        $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \le 2^{-l}$
```

An Instance-Optimal Algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
        $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \boxed{\sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}}} \right) \leq 2^{-l}$
```

CI width for estimated gap

An Instance-Optimal Algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
                                                        CI width for estimated gap
        $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \boxed{\sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}}} \right) \leq 2^{-l}$
    2. For $t \in [n_l]$, for each context $c_t$, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate
       $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
```

An Instance-Optimal Algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
                                                    CI width for estimated gap
       $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \boxed{\sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}}} \right) \leq 2^{-l}$
    2. For $t \in [n_l]$, for each context $c_t$, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate
       $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
```

An Instance-Optimal Algorithm

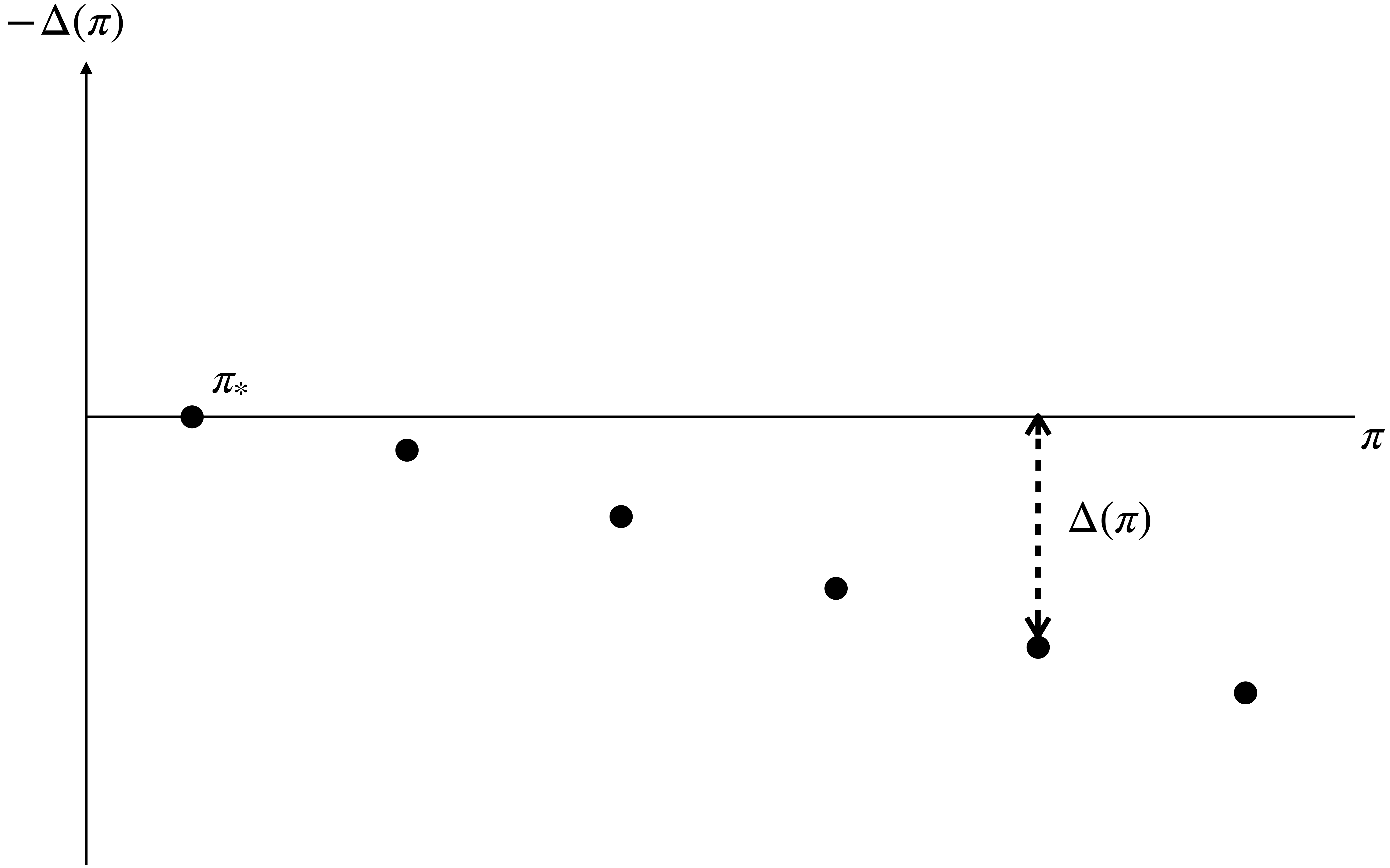
```
**Input:** $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
        $$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \overset{\text{CI width for estimated gap}}{\boxed{\sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}}}} \right) \le 2^{-l}$$
    2. For $t \in [n_l]$, for each context $c_t$, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate
       $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
        $$\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$$
```

An Instance-Optimal Algorithm

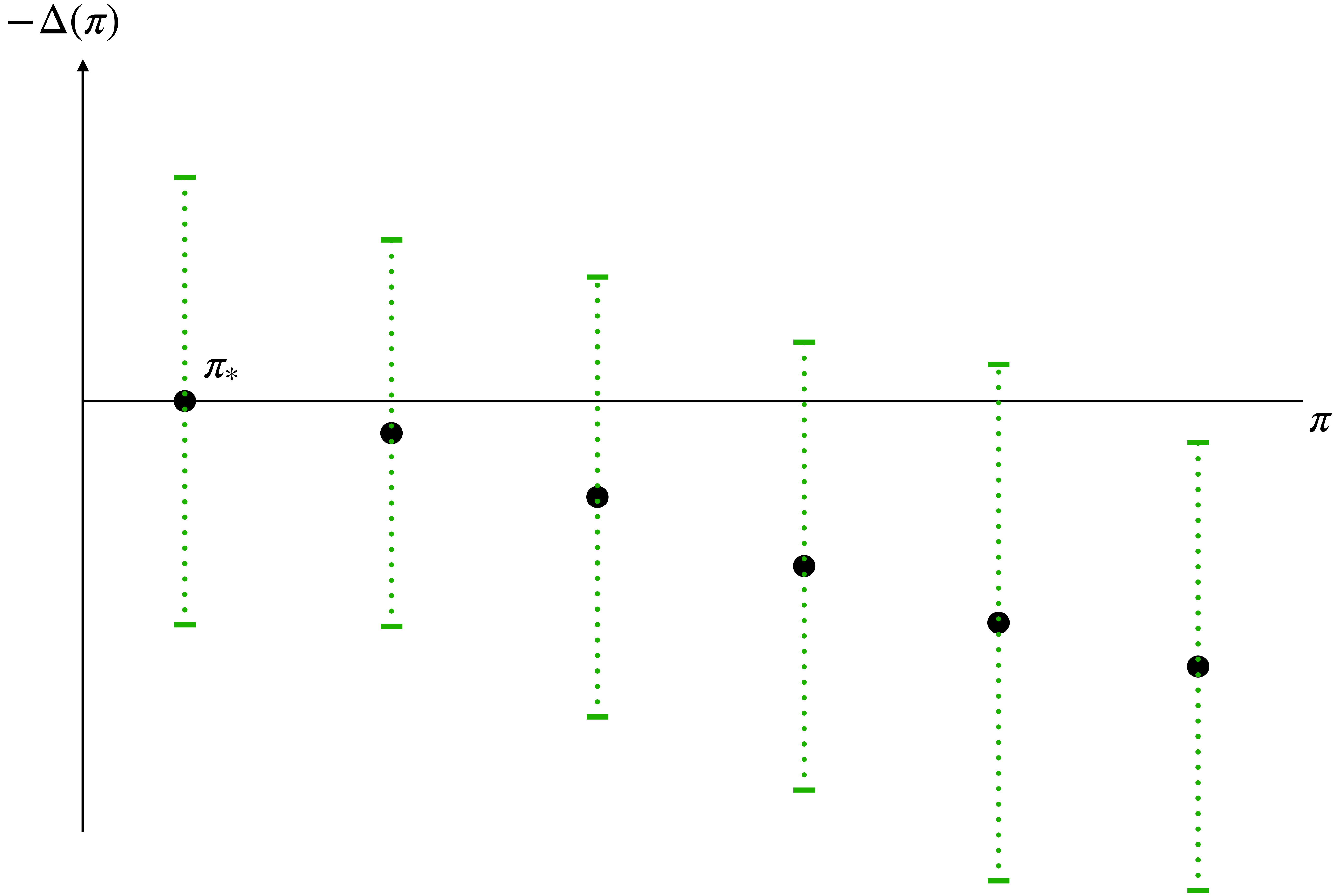
```
**Input:** $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and $n_l$ such that
                                                        CI width for estimated gap
       $$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \boxed{\sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}}} \right) \leq 2^{-l}$$
    2. For $t \in [n_l]$, for each context $c_t$, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate
       $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
       $$\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$$
```

Theorem [Li et al. 2022] The above algorithm returns an (ϵ, δ)-PAC policy with at most $O(\rho_{\Pi,\epsilon} \log(|\Pi|/\delta) \log_2(1/\epsilon))$ samples.

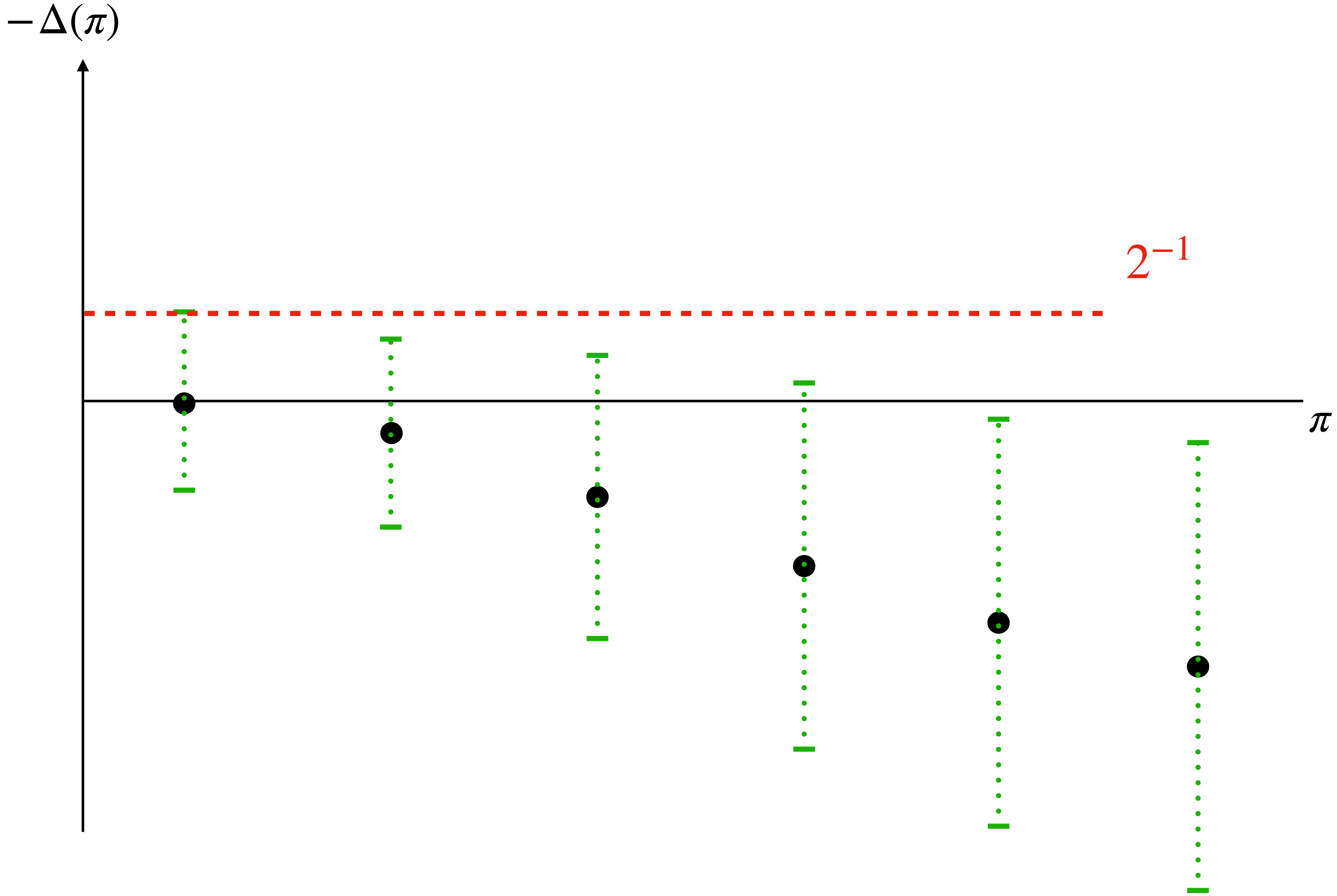
An Instance-Optimal Algorithm



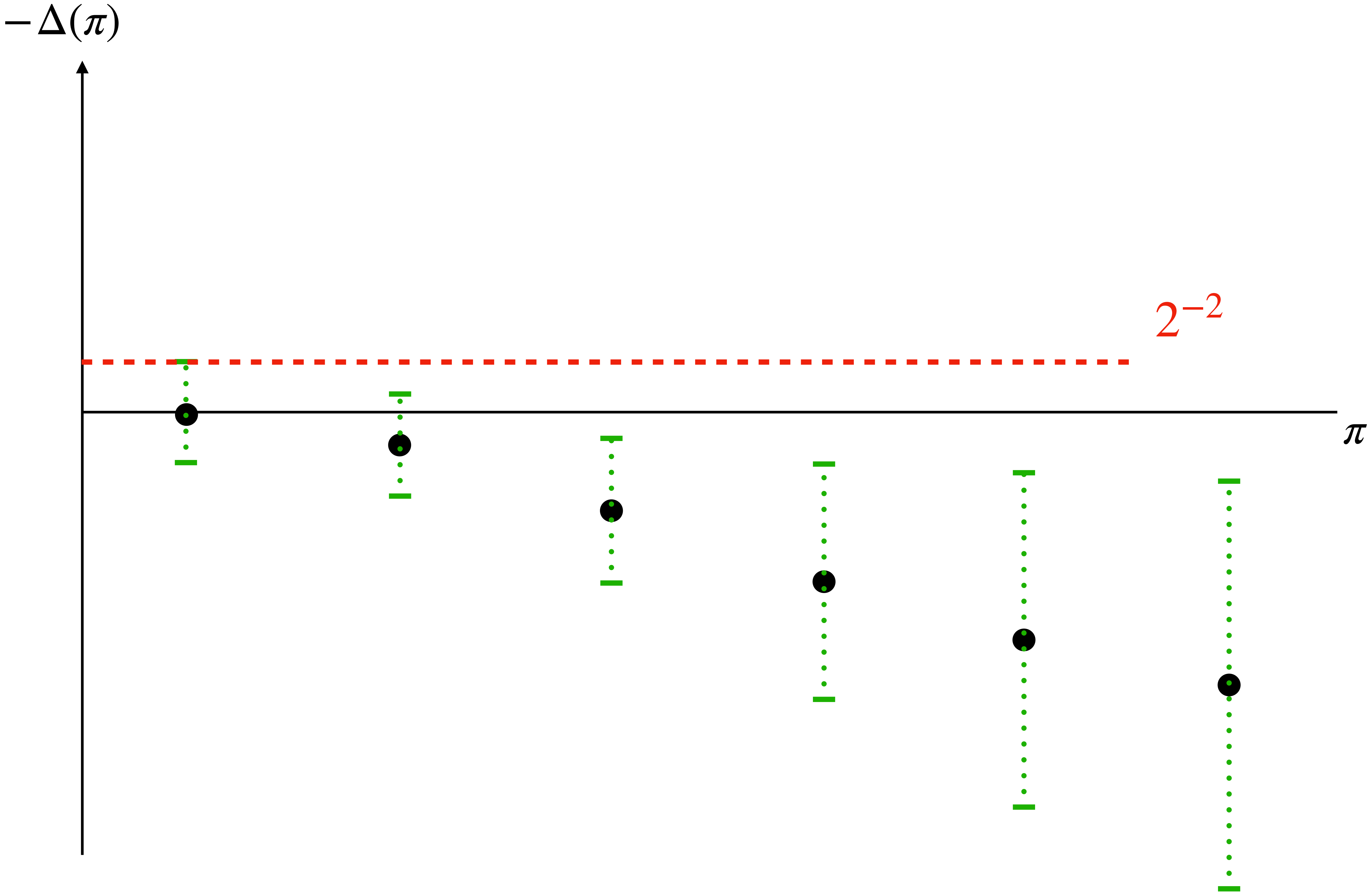
An Instance-Optimal Algorithm



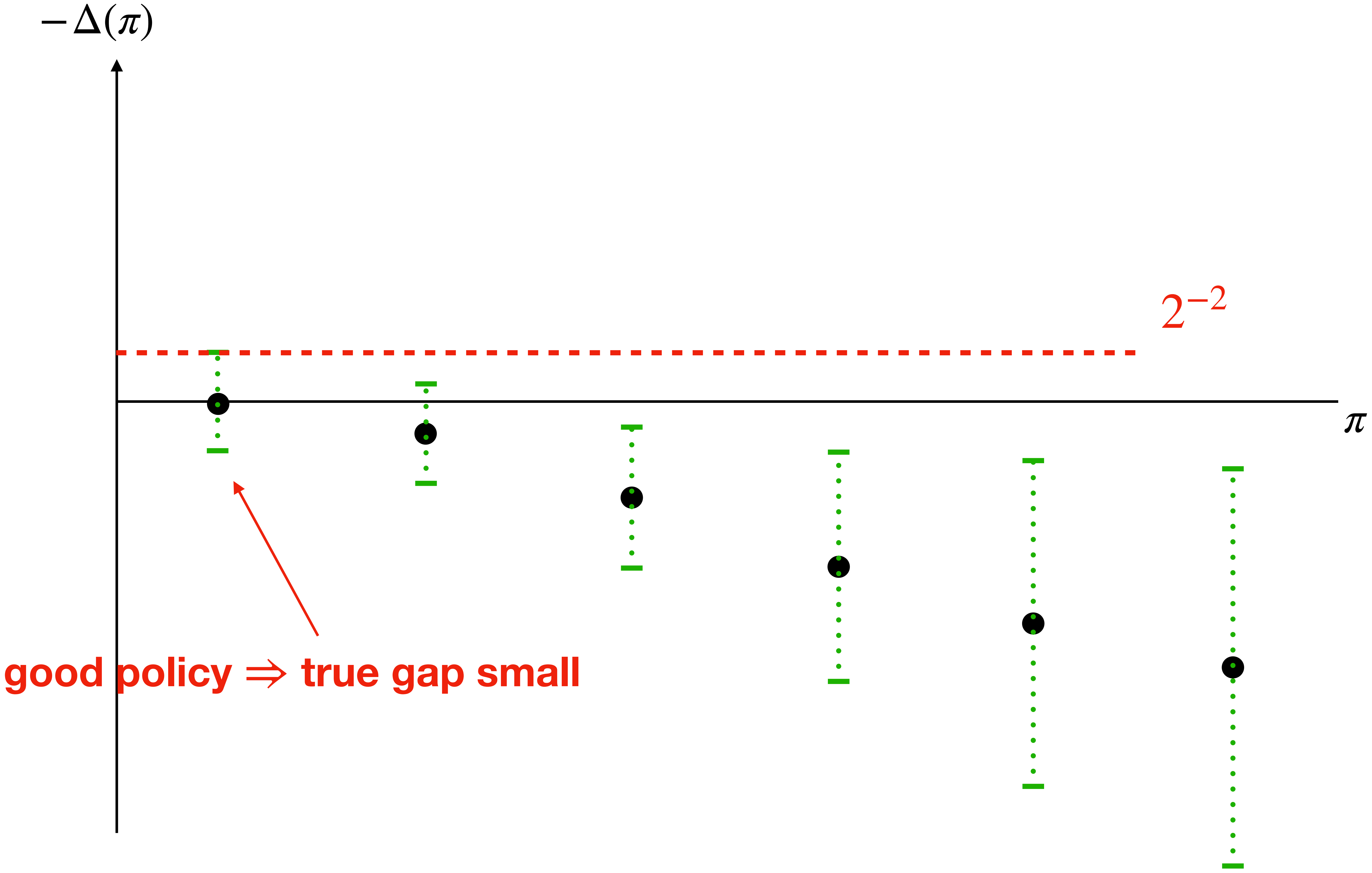
An Instance-Optimal Algorithm



An Instance-Optimal Algorithm

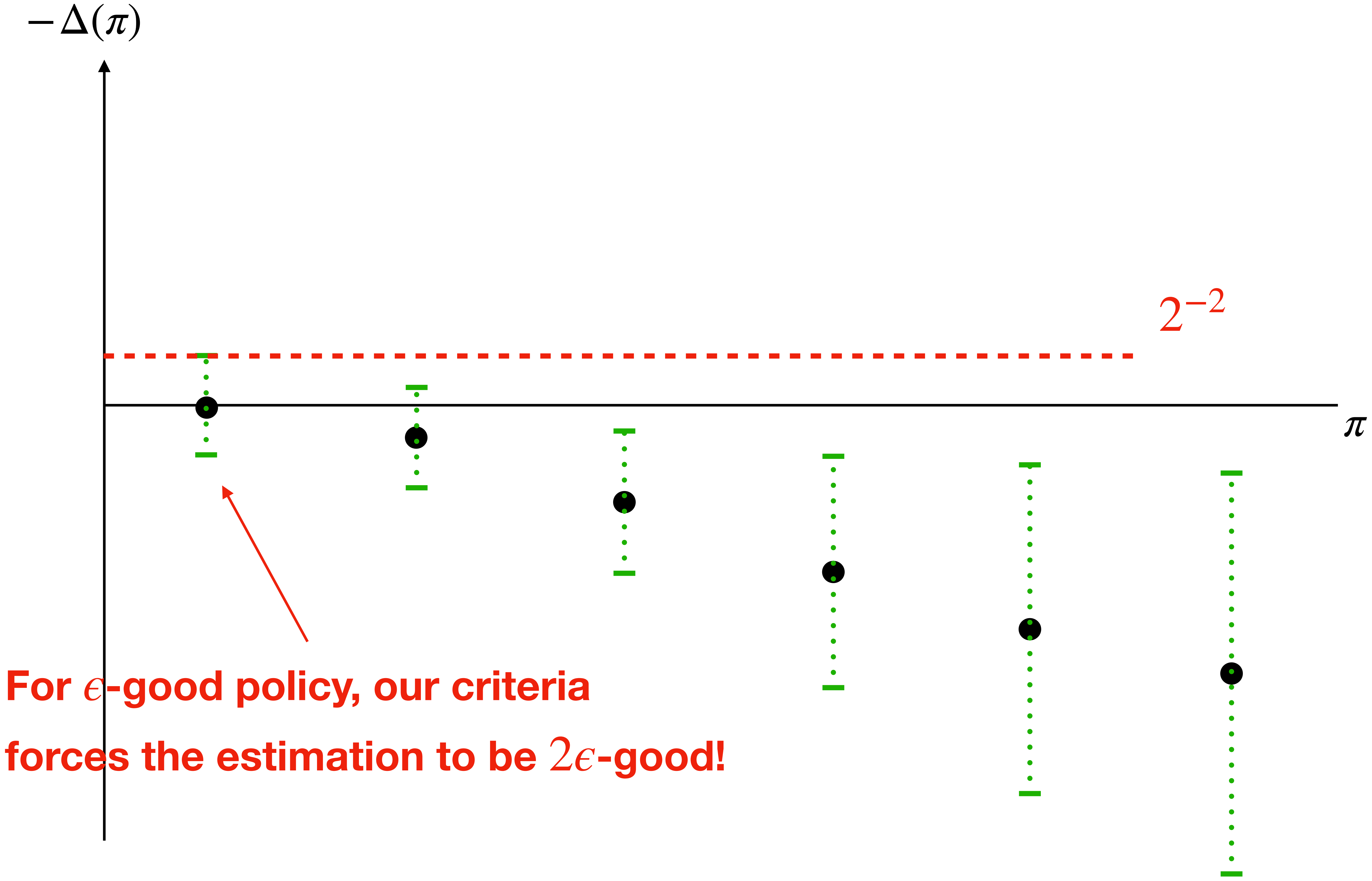


An Instance-Optimal Algorithm



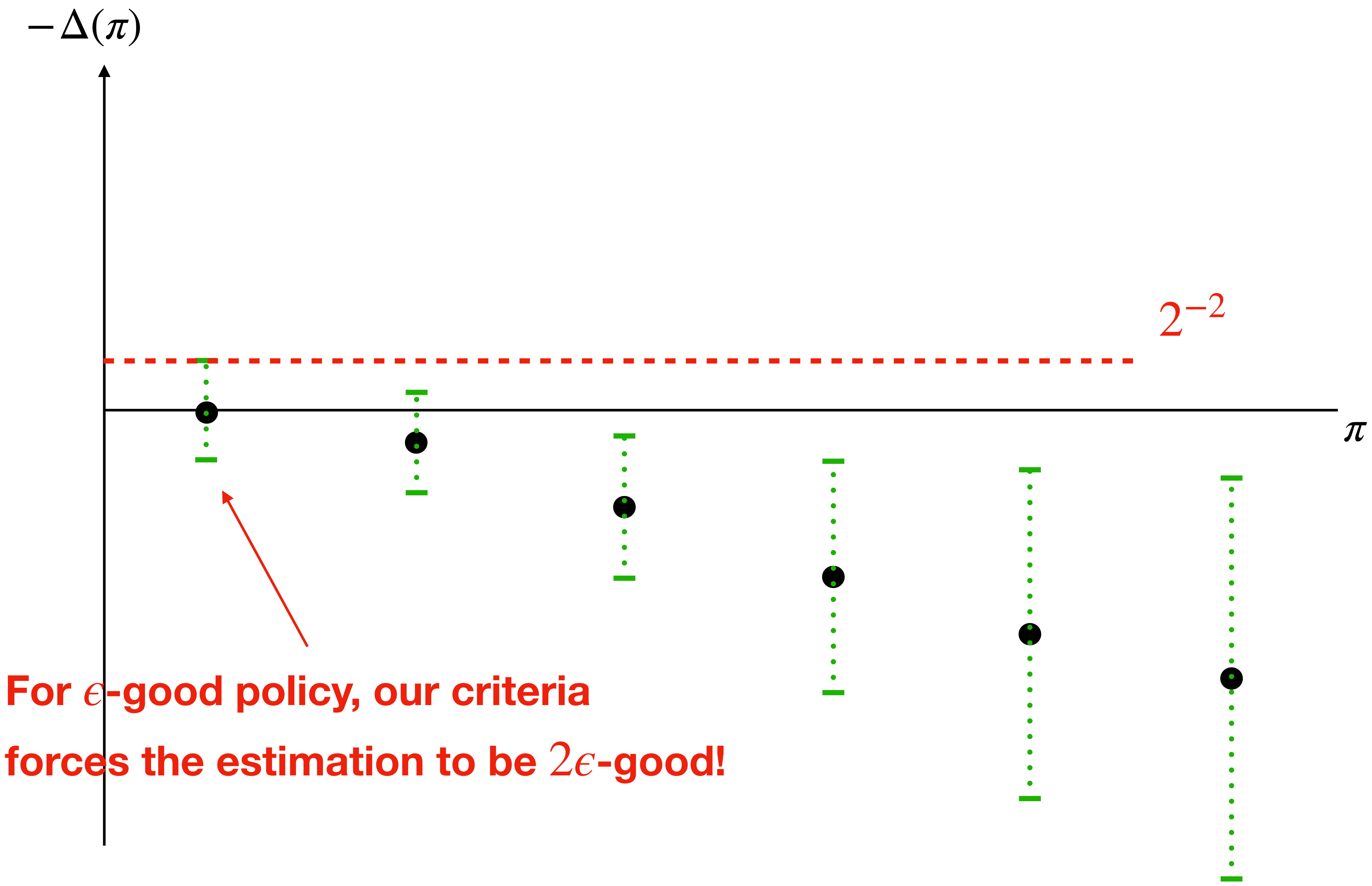
good policy ⇒ true gap small

An Instance-Optimal Algorithm



For ϵ-good policy, our criteria forces the estimation to be 2ϵ-good!

An Instance-Optimal Algorithm



For ϵ-good policy, our criteria forces the estimation to be 2ϵ-good!

Returning the empirical best policy at the end ⇒ at least 2ϵ-good

Towards an efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)}$ and $n_l$ such that
        $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$
    2. For $t \in [n_l]$, for each context $c_t$, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate
       $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
        $\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$
```

Towards an efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)}$ and $n_l$ such that
        $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \le 2^{-l}$
    2. For $t \in [n_l]$, for each context $c_t$, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate
       $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
        $\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$
```

Towards an efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Choose $p_c^{(l)}$ and $n_l$ such that
        $\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left( -\hat{\Delta}_l(\pi, \hat{\pi}_{l-1}) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$
    2. For $t \in [n_l]$, for each context $c_t$, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate
       $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
        $\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$
```

not efficient since cannot hold on to p_c for all $c \in \mathsf{C}$, also Π large!

Dual Problem

- Consider the dual formulation:

Primal

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} - \Delta(\pi, \pi_*) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}$$

Dual Problem

- Consider the dual formulation:

Primal

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} - \Delta(\pi, \pi_*) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}$$

$$= \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \min_{\gamma_\pi \geq 0} - \Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{\gamma_\pi n}$$

$$2\sqrt{ab} = \min_{\gamma > 0} \left[\gamma a + \frac{b}{\gamma}\right]$$

Dual Problem

- Consider the dual formulation:

Primal

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} - \Delta(\pi, \pi_*) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}$$

$$= \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \min_{\gamma_\pi \geq 0} - \Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{\gamma_\pi n}$$

$$2\sqrt{ab} = \min_{\gamma > 0} \left[\gamma a + \frac{b}{\gamma}\right]$$

Dual

$$\max_{\lambda \in \triangle_\Pi} \min_{\gamma_\pi \geq 0} \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \sum_{\pi \in \Pi} \lambda_\pi \left(-\Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{2\gamma_\pi n} \right).$$

Dual Problem

- Consider the dual formulation:

Primal

$$\boxed{\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} -\Delta(\pi, \pi_*) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}}$$

convex in $p_c, \forall c \in \mathsf{C}$ and KKT conditions hold $\Rightarrow$ strong duality holds!

$$= \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \min_{\gamma_\pi \geq 0} -\Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{\gamma_\pi n}$$

$$2\sqrt{ab} = \min_{\gamma > 0} \left[\gamma a + \frac{b}{\gamma}\right]$$

Dual

$$\max_{\lambda \in \triangle_\Pi} \min_{\gamma_\pi \geq 0} \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \sum_{\pi \in \Pi} \lambda_\pi \left(-\Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{2\gamma_\pi n} \right).$$

Dual Problem

- Consider the dual formulation:

Primal

$$\boxed{\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} -\Delta(\pi, \pi_*) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}}$$

convex in $p_c, \forall c \in \mathsf{C}$ and KKT conditions hold ⇒ strong duality holds!

$$2\sqrt{ab} = \min_{\gamma > 0} \left[\gamma a + \frac{b}{\gamma}\right]$$

$$= \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \min_{\gamma_\pi \geq 0} -\Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{\gamma_\pi n}$$

Dual

$$\max_{\lambda \in \triangle_\Pi} \min_{\gamma_\pi \geq 0} \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} g(\lambda, \gamma, p_c)\,.$$

Dual Problem

- Consider the dual formulation:

Primal

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} - \Delta(\pi, \pi_*) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}$$

convex in $p_c, \forall c \in \mathsf{C}$ and KKT conditions hold $\Rightarrow$ strong duality holds!

$$= \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \min_{\gamma_\pi \geq 0} - \Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{\gamma_\pi n}$$

$$2\sqrt{ab} = \min_{\gamma > 0} \left[\gamma a + \frac{b}{\gamma} \right]$$

problem of dimension $|\mathsf{C}| \times |\mathsf{A}|$

Dual

$$\max_{\lambda \in \triangle_\Pi} \min_{\gamma_\pi \geq 0} \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} g(\lambda, \gamma, p_c) \, .$$

Dual Problem

- Consider the dual formulation:

Primal

$$\boxed{\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} -\Delta(\pi, \pi_*) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}}$$

convex in $p_c, \forall c \in \mathsf{C}$ and KKT conditions hold $\Rightarrow$ strong duality holds!

$$2\sqrt{ab} = \min_{\gamma>0} \left[\gamma a + \frac{b}{\gamma}\right]$$

$$= \min_{\boxed{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}}} \max_{\pi \in \Pi} \min_{\gamma_\pi \geq 0} -\Delta(\pi, \pi_*) + \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{\gamma_\pi n}$$

problem of dimension $|\mathsf{C}| \times |\mathsf{A}|$

Dual

$$\max_{\boxed{\lambda \in \triangle_\Pi}} \min_{\gamma_\pi \geq 0} \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} g(\lambda, \gamma, p_c)\,.$$

problem of dimension $|\Pi|$

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \quad g(\lambda, \gamma, p_c) = h(\lambda, \gamma)$$

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \sum_{\pi \in \Pi} \lambda_\pi \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} = \mathbb{E}_{c \sim \nu} \left[\left(\sum_{a \in \mathsf{A}} \sqrt{\sum_{\pi \in \Pi} \lambda_\pi \gamma_\pi (\mathbf{1}\{\pi(c) = a\} + \mathbf{1}\{\pi_*(c) = a\} - 2\mathbf{1}\{\pi(c) = \pi_*(c)\})} \right)^2 \right]$$

$$=: \mathbb{E}_{c \sim \nu} \left[\left(\sum_{a \in \mathsf{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right]$$

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \sum_{\pi \in \Pi} \lambda_\pi \gamma_\pi \|\phi_\pi - \phi_{\pi_*}\|^2_{A(p)^{-1}} = \mathbb{E}_{c \sim \nu}\left[\left(\sum_{a \in \mathsf{A}} \sqrt{\sum_{\pi \in \Pi} \lambda_\pi \gamma_\pi (\mathbf{1}\{\pi(c) = a\} + \mathbf{1}\{\pi_*(c) = a\} - 2\mathbf{1}\{\pi(c) = \pi_*(c)\})}\right)^2\right]$$

$$=: \mathbb{E}_{c \sim \nu}\left[\left(\sum_{a \in \mathsf{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}}\right)^2\right]$$

Implicitly maintain p_c for all $c \in \mathsf{C}$ simultaneously!

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \quad g(\lambda, \gamma, p_c) = h(\lambda, \gamma)$$

Implicitly maintain p_c for all $c \in \mathsf{C}$ simultaneously!

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \; g(\lambda, \gamma, p_c) = h(\lambda, \gamma)$$

Implicitly maintain p_c for all $c \in \mathsf{C}$ simultaneously!

- Dual becomes

$$\max_{\lambda \in \Delta_\Pi} \min_{\gamma} \; h(\lambda, \gamma)$$

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \quad g(\lambda, \gamma, p_c) = h(\lambda, \gamma)$$

Implicitly maintain p_c for all $c \in \mathsf{C}$ simultaneously!

- Dual becomes

$$\max_{\lambda \in \Delta_\Pi} \min_{\gamma} \sum_{\pi \in \Pi} \lambda_\pi \left(-\Delta(\pi, \pi_*) + \frac{\log(1/\delta)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu} \left[\left(\sum_{a \in \mathsf{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right]$$

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c\in\triangle_{\mathsf{A}},\forall c\in\mathsf{C}} \quad g(\lambda,\gamma,p_c) = h(\lambda,\gamma)$$

Implicitly maintain p_c for all $c \in \mathsf{C}$ simultaneously!

- Dual becomes

$$\max_{\lambda\in\Delta_\Pi}\min_{\gamma}\sum_{\pi\in\Pi}\lambda_\pi\left(-\Delta(\pi,\pi_*)+\frac{\log(1/\delta)}{\gamma_\pi n}\right)+\mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathsf{A}}\sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}}\right)^2\right]$$

concave in λ and locally strongly convex in γ!

Compute Action Distribution

- If we solve for p_c for all c, we have an analytical solution:

$$\min_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} g(\lambda, \gamma, p_c) = h(\lambda, \gamma)$$

Implicitly maintain p_c for all $c \in \mathcal{C}$ simultaneously!

- Dual becomes

$$\max_{\lambda \in \Delta_\Pi} \min_{\gamma} \sum_{\pi \in \Pi} \lambda_\pi \left(-\Delta(\pi, \pi_*) + \frac{\log(1/\delta)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu} \left[\left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right]$$

concave in λ and locally strongly convex in γ!

distribution over Π, but $|\Pi|$ could still be large!

Frank-Wolfe

minimize $f(x)$

subject to $x \in \mathcal{X}$

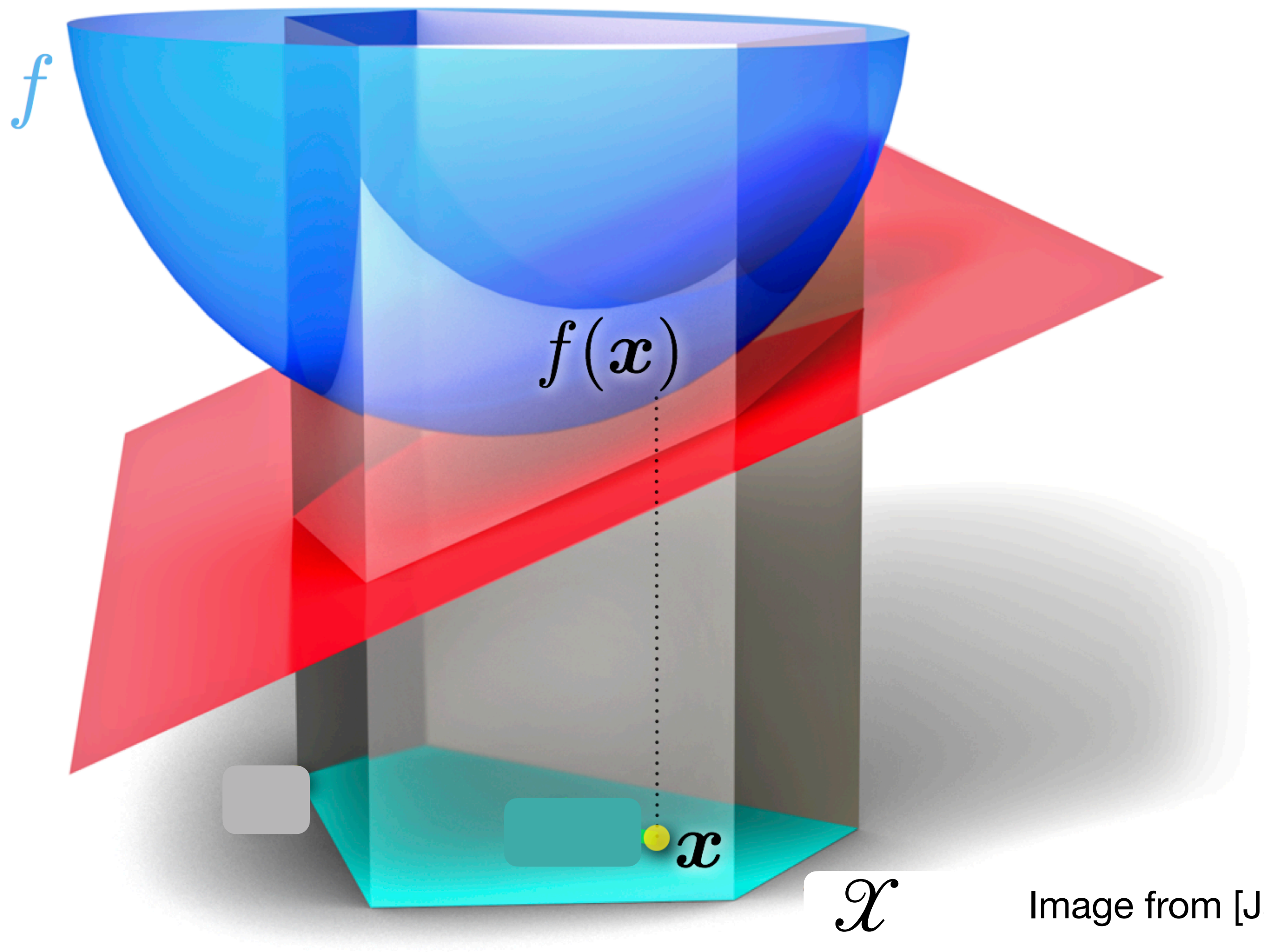


Image from [Jaggi]

Frank-Wolfe

minimize $f(x)$

subject to $x \in \mathcal{X}$

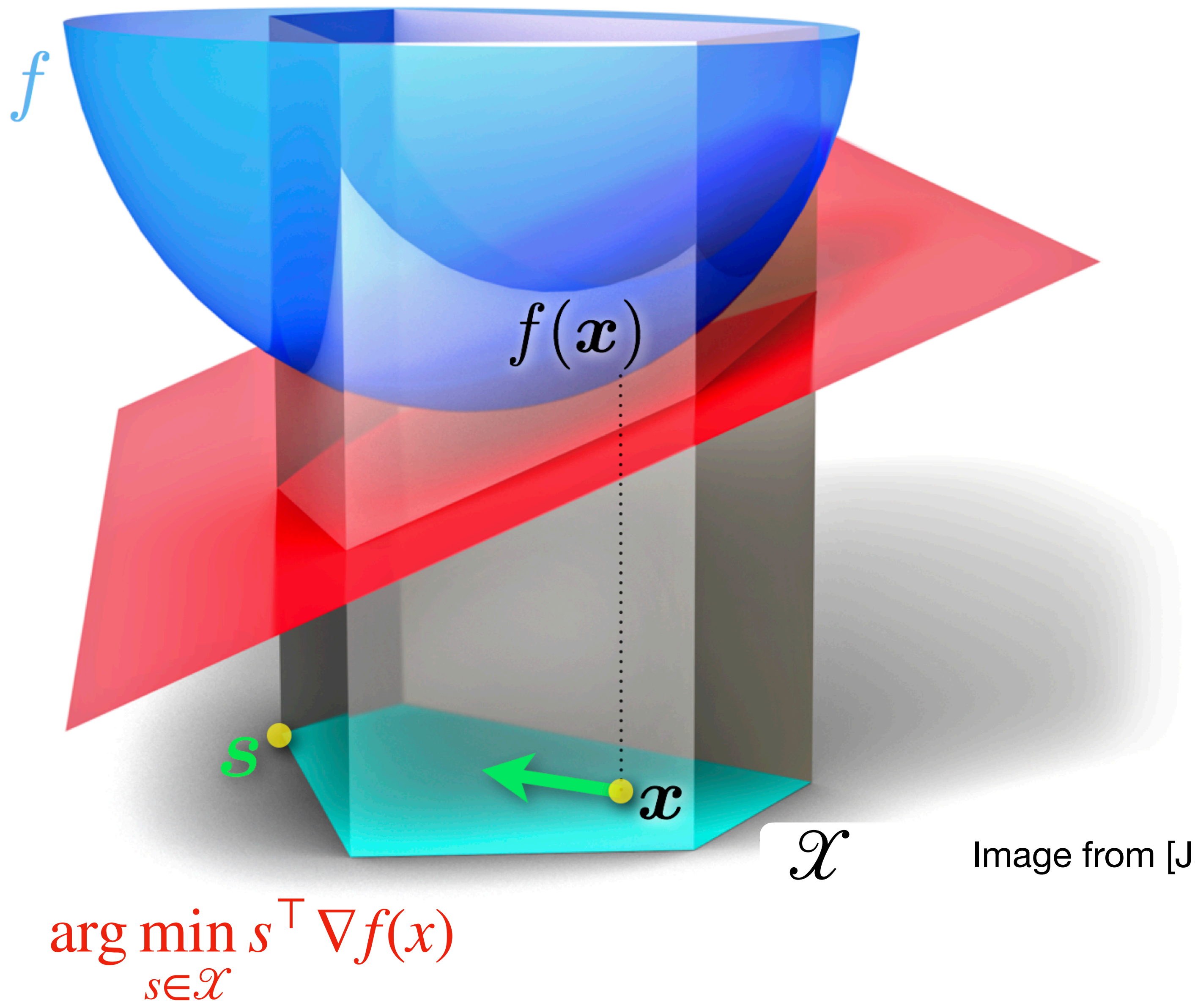


Image from [Jaggi]

Frank-Wolfe

minimize $f(x)$

subject to $x \in \mathcal{X}$

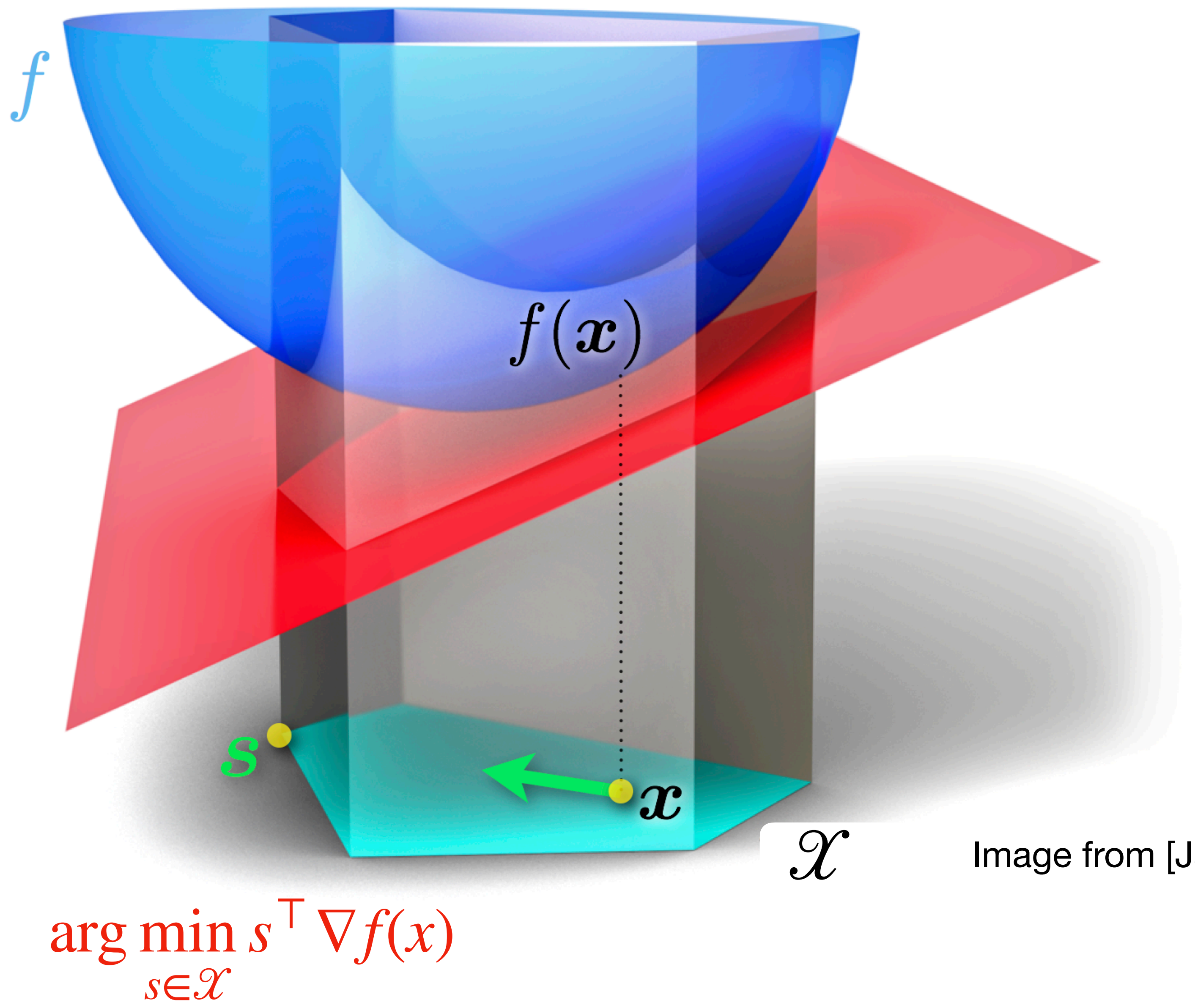


Image from [Jaggi]

- Update one coordinate at a time
- Gives us a sparse yet good enough solution λ
- Plug in solution λ in the closed-form gives us $p_c \in \triangle_\mathsf{A}$

Frank-Wolfe

minimize $f(x)$

subject to $x \in \mathcal{X}$

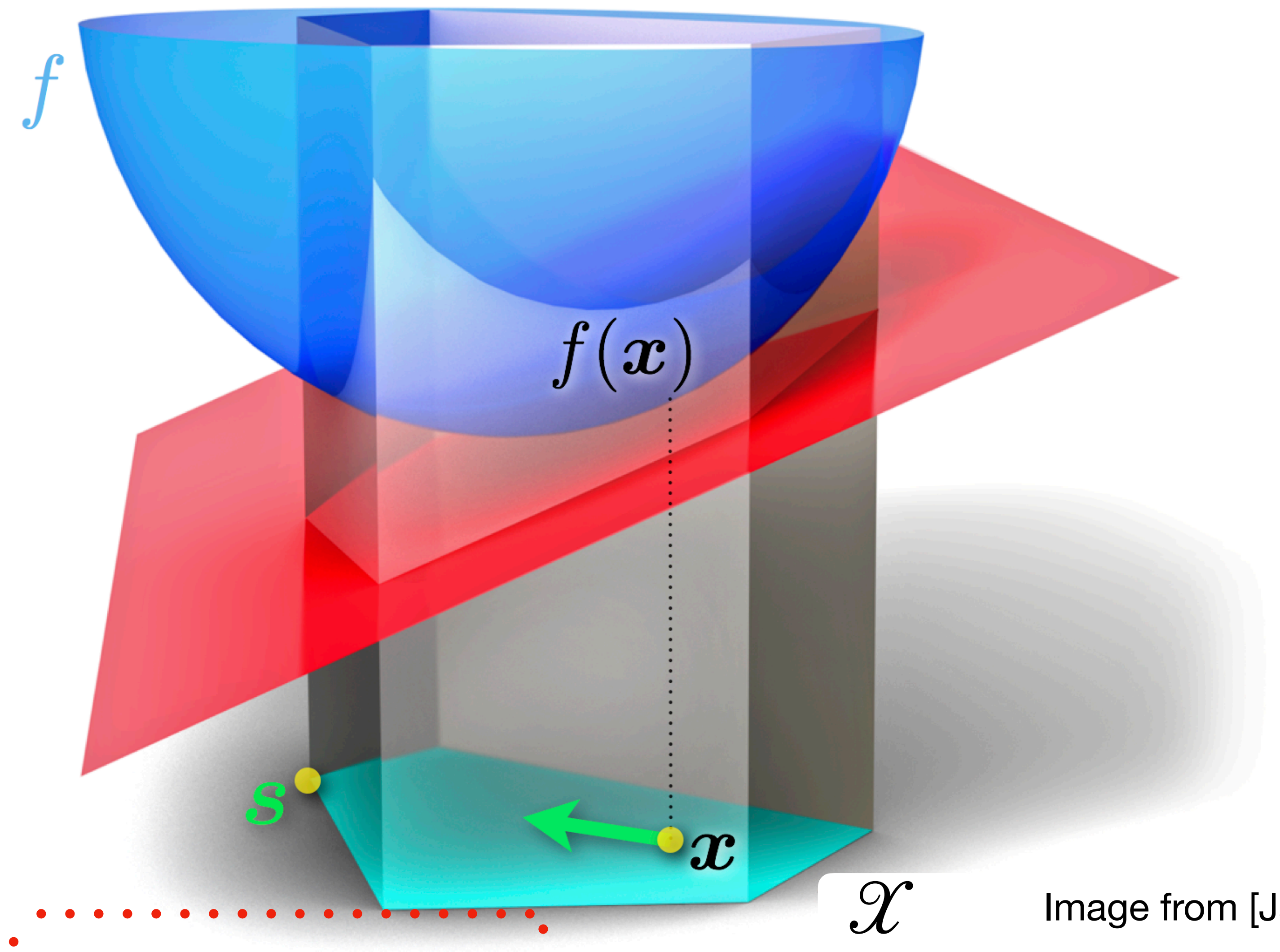


Image from [Jaggi]

$\arg\min_{s \in \mathcal{X}} s^\top \nabla f(x)$

argmax oracle

- Update one coordinate at a time
- Gives us a sparse yet good enough solution λ
- Plug in solution λ in the closed-form gives us $p_c \in \triangle_{\mathsf{A}}$

Towards an efficient algorithm

Towards an efficient algorithm

- ***argmax*** oracle: given contexts and cost vectors $(c_1, v_1), \cdots, (c_n, v_n) \in \mathsf{C} \times \mathbb{R}^{|\mathsf{A}|}$, returns $\arg\max_{\pi \in \Pi} \sum_{t=1}^{n} v_t(\pi(c_t))$

Towards an efficient algorithm

- ***argmax*** oracle: given contexts and cost vectors $(c_1, v_1), \cdots, (c_n, v_n) \in \mathsf{C} \times \mathbb{R}^{|\mathsf{A}|}$, returns $\arg\max_{\pi \in \Pi} \sum_{t=1}^{n} v_t(\pi(c_t))$
- Can be computed using cost-sensitive classification

Towards an efficient algorithm

- ***argmax*** oracle: given contexts and cost vectors $(c_1, v_1), \cdots, (c_n, v_n) \in \mathsf{C} \times \mathbb{R}^{|\mathsf{A}|}$, returns $\arg\max_{\pi \in \Pi} \sum_{t=1}^{n} v_t(\pi(c_t))$
- Can be computed using cost-sensitive classification
- Can estimate the context distribution using offline data $\mathscr{D}$

Towards an efficient algorithm

- ***argmax*** oracle: given contexts and cost vectors $(c_1, v_1), \cdots, (c_n, v_n) \in \mathsf{C} \times \mathbb{R}^{|\mathsf{A}|}$, returns $\arg\max_{\pi \in \Pi} \sum_{t=1}^{n} v_t(\pi(c_t))$

- Can be computed using cost-sensitive classification

- Can estimate the context distribution using offline data $\mathscr{D}$

- Final design we're solving:

$$\max_{\lambda \in \Delta_\Pi} \min_{\gamma} \sum_{\pi \in \Pi} \lambda_\pi \left(-\hat{\Delta}(\pi, \pi_*) + \frac{\log(1/\delta)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu_{\mathscr{D}}} \left[\left(\sum_{a \in \mathscr{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right]$$

An efficient algorithm

An efficient algorithm

Input: Π

An efficient algorithm

Input: Π

Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
```

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
        1. Solve $\lambda_l, \gamma_l$ and choose $n_l$ such that
```

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Solve $\lambda_l, \gamma_l$ and choose $n_l$ such that
        $\sum_{\pi \in \Pi} \lambda_\pi \left( -\hat{\Delta}_{l-1}(\pi, \hat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu_{\mathscr{D}}} \left[ \left( \sum_{a \in \mathsf{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right] \leq 2^{-l}$
```

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Solve $\lambda_l, \gamma_l$ and choose $n_l$ such that
        $\sum_{\pi \in \Pi} \lambda_\pi \left( -\hat{\Delta}_{l-1}(\pi, \hat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu_{\mathcal{D}}} \left[ \left( \sum_{a \in \mathsf{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right] \leq 2^{-l}$
```

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Solve $\lambda_l, \gamma_l$ and choose $n_l$ such that
        $$\sum_{\pi \in \Pi} \lambda_\pi \left( -\hat{\Delta}_{l-1}(\pi, \hat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu_{\mathcal{D}}} \left[ \left( \sum_{a \in \mathsf{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right] \leq 2^{-l}$$
    2. For $s \in [n_l]$, for each context $c_s$, sampling $a_s \sim p_{c_s}^{(l)}$ where $p_{c_s,a_s}^{(l)} \propto \sqrt{(\lambda_l \odot \gamma_l)^\top t_{a_s}^{(c_s)}}$
    and compute IPW estimate $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
```

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Solve $\lambda_l, \gamma_l$ and choose $n_l$ such that
       $$\sum_{\pi\in\Pi} \lambda_\pi \left(-\hat{\Delta}_{l-1}(\pi, \hat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n}\right) + \mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in A}\sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}}\right)^2\right] \le 2^{-l}$$
    2. For $s \in [n_l]$, for each context $c_s$, sampling $a_s \sim p_{c_s}^{(l)}$ where $p_{c_s,a_s}^{(l)} \propto \sqrt{(\lambda_l \odot \gamma_l)^\top t_{a_s}^{(c_s)}}$
    and compute IPW estimate $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
```

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1,2,\cdots$
    1. Solve $\lambda_l, \gamma_l$ and choose $n_l$ such that
        $$\sum_{\pi\in\Pi} \lambda_\pi \left( -\hat{\Delta}_{l-1}(\pi, \hat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[ \left( \sum_{a\in A} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right] \leq 2^{-l}$$
    2. For $s \in [n_l]$, for each context $c_s$, sampling $a_s \sim p_{c_s}^{(l)}$ where $p_{c_s,a_s}^{(l)} \propto \sqrt{(\lambda_l \odot \gamma_l)^\top t_{a_s}^{(c_s)}}$
    and compute IPW estimate $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
        $$\hat{\pi}_l = \arg\min_{\pi\in\Pi} \hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$$
```

An efficient algorithm

```
Input: $\Pi$
Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$
for $l = 1, 2, \cdots$
    1. Solve $\lambda_l, \gamma_l$ and choose $n_l$ such that
        $\sum_{\pi \in \Pi} \lambda_\pi \left( -\hat{\Delta}_{l-1}(\pi, \hat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu_{\mathscr{D}}}\left[ \left( \sum_{a \in A} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right] \leq 2^{-l}$
    2. For $s \in [n_l]$, for each context $c_s$, sampling $a_s \sim p_{c_s}^{(l)}$ where $p_{c_s, a_s}^{(l)} \propto \sqrt{(\lambda_l \odot \gamma_l)^\top t_{a_s}^{(c_s)}}$
    and compute IPW estimate $\hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
    3. Update
        $\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}_l(\pi, \hat{\pi}_{l-1})$
```

Theorem [Li et al. 2022] The above algorithm returns an (ϵ, δ)-PAC policy with at most $O(\rho_{\Pi,\epsilon} \log(|\Pi|/\delta) \log_2(1/\epsilon))$ samples and $\mathrm{poly}(|A|, \epsilon^{-1}, \log(1/\delta), \log(|\Pi|))$ calls to argmax oracle.

Conclusion

- Propose a new instance-dependent lower bound for PAC contextual bandits
- Design a computationally efficient algorithm and show that it is instance-optimal

Outline

- Project 1: Instance-optimal PAC Contextual bandits
- Project 2: Estimation of the mean of subsidiary outcome
- Future Work

Estimation of the mean of subsidiary outcome under an optimal policy for primary outcome

Zhaoqi Li, Alex Luedtke

Motivation

- In biomedical trials, it is of interest to identify the best treatment to induce disease remission, i.e. identifying the optimal policy
- However, side effects of certain medicine are also concerns
- Important to investigate subsidiary outcomes

Problem Notations

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \to \{0,1\}$

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \to \{0,1\}$

- Suppose $Y = (Y*, Y^{\dagger})$ is a primary-subsidiary outcome pair,

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \rightarrow \{0,1\}$

remission rate

- Suppose $Y = (Y*, Y^{\dagger})$ is a primary-subsidiary outcome pair,

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \to \{0,1\}$

remission rate side effects

- Suppose $Y = (Y^*, Y^\dagger)$ is a primary-subsidiary outcome pair,

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \to \{0,1\}$

remission rate side effects

- Suppose $Y = (Y^*, Y^\dagger)$ is a primary-subsidiary outcome pair,
  - let Φ_π be some primary performance metric for π, e.g. $\mathbb{E}[\mathbb{E}[Y^* \,|\, A = \pi(X), X]]$

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \to \{0,1\}$

remission rate side effects

- Suppose $Y = (Y^*, Y^\dagger)$ is a primary-subsidiary outcome pair,
  - let Φ_π be some primary performance metric for π, e.g. $\mathbb{E}[\mathbb{E}[Y^* \mid A = \pi(X), X]]$
  - let Ψ_π be some subsidiary performance metric for π, e.g. $\mathbb{E}[\mathbb{E}[Y^\dagger \mid A = \pi(X), X]]$

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \to \{0,1\}$

remission rate side effects

- Suppose $Y = (Y^*, Y^\dagger)$ is a primary-subsidiary outcome pair,
  - let Φ_π be some primary performance metric for π, e.g. $\mathbb{E}[\mathbb{E}[Y^* \,|\, A = \pi(X), X]]$
  - let Ψ_π be some subsidiary performance metric for π, e.g. $\mathbb{E}[\mathbb{E}[Y^\dagger \,|\, A = \pi(X), X]]$
  - let Π^* be the set of optimal policies with respect to Φ_π

Problem Notations

- More formally, let $X \in \mathscr{X}$ be some covariates, $A \in \{0,1\}$ be a binary action, $Y \in \mathscr{Y}$ be an observed outcome, and policy $\pi : \mathscr{X} \rightarrow \{0,1\}$

remission rate side effects

- Suppose $Y = (Y^*, Y^\dagger)$ is a primary-subsidiary outcome pair,
  - let Φ_π be some primary performance metric for π, e.g. $\mathbb{E}[\mathbb{E}[Y^* \,|\, A = \pi(X), X]]$
  - let Ψ_π be some subsidiary performance metric for π, e.g. $\mathbb{E}[\mathbb{E}[Y^\dagger \,|\, A = \pi(X), X]]$
  - let Π^* be the set of optimal policies with respect to Φ_π

Goal: conduct inference on $\{\Psi_\pi : \pi \in \Pi^*\}$!

Related Work

Related Work

- Estimate the mean outcome under an optimal policy:

Related Work

- Estimate the mean outcome under an optimal policy:
  - corresponding to conducting inference on $\{\Phi_{\pi} : \pi \in \Pi^*\}$

Related Work

- Estimate the mean outcome under an optimal policy:
  - corresponding to conducting inference on $\{\Phi_\pi : \pi \in \Pi^*\}$
  - the standard one-step estimator is efficient [Luedtke et al. 2016]

Related Work

- Estimate the mean outcome under an optimal policy:
  - corresponding to conducting inference on $\{\Phi_{\pi} : \pi \in \Pi^*\}$
  - the standard one-step estimator is efficient [Luedtke et al. 2016]
- Estimate $(Y^*, Y^\dagger)$ simultaneously:

Related Work

- Estimate the mean outcome under an optimal policy:
  - corresponding to conducting inference on $\{\Phi_\pi : \pi \in \Pi^*\}$
  - the standard one-step estimator is efficient [Luedtke et al. 2016]
- Estimate $(Y^*, Y^\dagger)$ simultaneously:
  - Multi-objective optimization

Related Work

- Estimate the mean outcome under an optimal policy:
  - corresponding to conducting inference on $\{\Phi_\pi : \pi \in \Pi^*\}$
  - the standard one-step estimator is efficient [Luedtke et al. 2016]
- Estimate $(Y^*, Y^\dagger)$ simultaneously:
  - Multi-objective optimization
  - Efficient algorithms exist to find the solution [Gunantara et al. 2018]

Objective

Objective

- Can we use a similar one-step estimator to estimate $\{\Psi_\pi : \pi \in \Pi^*\}$ and show that it is efficient, i.e. with provably minimum variance?

Objective

- Can we use a similar one-step estimator to estimate $\{\Psi_\pi : \pi \in \Pi^*\}$ and show that it is efficient, i.e. with provably minimum variance?

Yes, under certain (strong) margin conditions.

Objective

- Can we use a similar one-step estimator to estimate $\{\Psi_\pi : \pi \in \Pi^*\}$ and show that it is efficient, i.e. with provably minimum variance?

Yes, under certain (strong) margin conditions.

- Can we perform inference without conditions assumed previously?

Objective

- Can we use a similar one-step estimator to estimate $\{\Psi_\pi : \pi \in \Pi^*\}$ and show that it is efficient, i.e. with provably minimum variance?

Yes, under certain (strong) margin conditions.

- Can we perform inference without conditions assumed previously?

Yes, using a uniform band approach.

Objective

- Can we use a similar one-step estimator to estimate $\{\Psi_\pi : \pi \in \Pi^*\}$ and show that it is efficient, i.e. with provably minimum variance?

Yes, under certain (strong) margin conditions.

- Can we perform inference without conditions assumed previously?

Yes, using a uniform band approach.

- Can we improve on the previous method to provide a tighter confidence interval?

Objective

- Can we use a similar one-step estimator to estimate $\{\Psi_\pi : \pi \in \Pi^*\}$ and show that it is efficient, i.e. with provably minimum variance?

Yes, under certain (strong) margin conditions.

- Can we perform inference without conditions assumed previously?

Yes, using a uniform band approach.

- Can we improve on the previous method to provide a tighter confidence interval?

Yes, using a joint approach.

Towards an efficient estimator

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

$$\hat{\psi}_{\hat{\pi}} - \Psi_{\pi^*} = \left[\hat{\psi}_{\hat{\pi}} - \Psi_{\hat{\pi}}\right] \qquad + \qquad \left[\Psi_{\hat{\pi}} - \Psi_{\pi^*}\right]$$

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

$$\hat{\psi}_{\hat{\pi}} - \Psi_{\pi^*} = \underbrace{\left[\hat{\psi}_{\hat{\pi}} - \Psi_{\hat{\pi}}\right]}_{\textbf{small if } \hat{\psi} \textbf{ is good}} \qquad + \qquad \left[\Psi_{\hat{\pi}} - \Psi_{\pi^*}\right]$$

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

$$\hat{\psi}_{\hat{\pi}} - \Psi_{\pi^*} = \underbrace{\boxed{\left[\hat{\psi}_{\hat{\pi}} - \Psi_{\hat{\pi}}\right]}}_{\text{small if } \hat{\psi} \text{ is good}} + \underbrace{\boxed{\left[\Psi_{\hat{\pi}} - \Psi_{\pi^*}\right]}}_{\text{small if } \hat{\pi} \text{ is good and } \Psi \text{ is flat around } \pi^*}$$

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

$$\hat{\psi}_{\hat{\pi}} - \Psi_{\pi^*} = \underset{\substack{\downarrow \\ \textbf{small if } \hat{\psi} \textbf{ is good}}}{\boxed{\left[\hat{\psi}_{\hat{\pi}} - \Psi_{\hat{\pi}}\right]}} \qquad + \qquad \underset{\substack{\downarrow \\ \textbf{small if } \hat{\pi} \textbf{ is good and } \Psi \textbf{ is flat around } \pi^*}}{\boxed{\left[\Psi_{\hat{\pi}} - \Psi_{\pi^*}\right]}}$$

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

$$\hat{\psi}_{\hat{\pi}} - \Psi_{\pi^*} = \underbrace{\boxed{\left[\hat{\psi}_{\hat{\pi}} - \Psi_{\hat{\pi}}\right]}}_{\text{small if } \hat{\psi} \text{ is good}} + \underbrace{\boxed{\left[\Psi_{\hat{\pi}} - \Psi_{\pi^*}\right]}}_{\text{small if } \hat{\pi} \text{ is good and } \Psi \text{ is flat around } \pi^*}$$

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

$$\hat{\psi}_{\hat{\pi}} - \Psi_{\pi^*} = \underset{\substack{\downarrow \\ \textbf{small if } \hat{\psi} \textbf{ is good}}}{\boxed{\left[\hat{\psi}_{\hat{\pi}} - \Psi_{\hat{\pi}}\right]}} \quad + \quad \underset{\substack{\downarrow \\ \textbf{small if } \hat{\pi} \textbf{ is good and } \Psi \textbf{ is flat around } \pi^*}}{\boxed{\left[\Psi_{\hat{\pi}} - \Psi_{\pi^*}\right]}}$$

- When estimating Φ_{π^*}, since π^* optimizes Φ, we only need guarantees for the behavior of Φ on regions where estimation for π^* is hard

Towards an efficient estimator

- Let $\pi^* \in \Pi^*$ be a Φ-optimal policy

- Let $\hat{\pi}$ be some estimate of π^* and $\hat{\psi}_{\hat{\pi}}$ be some estimator of Ψ_{π^*}, then

$$\hat{\psi}_{\hat{\pi}} - \Psi_{\pi^*} = \underbrace{\boxed{\left[\hat{\psi}_{\hat{\pi}} - \Psi_{\hat{\pi}}\right]}}_{\text{small if } \hat{\psi} \text{ is good}} + \underbrace{\boxed{\left[\Psi_{\hat{\pi}} - \Psi_{\pi^*}\right]}}_{\text{small if } \hat{\pi} \text{ is good and } \Psi \text{ is flat around } \pi^*}$$

- When estimating Φ_{π^*}, since π^* optimizes Φ, we only need guarantees for the behavior of Φ on regions where estimation for π^* is hard

- Since π^* is not necessarily an optimizer for Ψ, we need much stronger conditions to guarantee the behavior of Ψ on the entire space

Estimation of the optimal policy

Estimation of the optimal policy

- Define the CATE function $q_b(x) := \mathbb{E}[Y^* \mid A = 1, X = x] - \mathbb{E}[Y^* \mid A = 0, X = x]$
- $q_b(x) = 0 \Rightarrow A$ has no impact on $Y^* \Rightarrow$ estimation of π^* hard!

Estimation of the optimal policy

- Define the CATE function $q_b(x) := \mathbb{E}[Y^* \mid A = 1, X = x] - \mathbb{E}[Y^* \mid A = 0, X = x]$
- $q_b(x) = 0 \Rightarrow A$ has no impact on $Y^* \Rightarrow$ estimation of π^* hard!

Condition 1 (Margin Condition of Y^*). For some $\beta > 0$,

$$\Pr\left(0 \leq \left|q_b(X)\right| \leq t\right) \lesssim t^{\beta} \qquad \forall t > 0$$

Estimation of the optimal policy

- Define the CATE function $q_b(x) := \mathbb{E}[Y^* \,|\, A=1, X=x] - \mathbb{E}[Y^* \,|\, A=0, X=x]$
- $q_b(x) = 0 \Rightarrow A$ has no impact on $Y^* \Rightarrow$ estimation of π^* hard!

Condition 1 (Margin Condition of Y^*). For some $\beta > 0$,

$$\Pr\left(0 \leq \left|q_b(X)\right| \leq t\right) \lesssim t^{\beta} \qquad \forall t > 0$$

- Condition 1 ensures that the mass of $q_b(X)$ concentrated around zero is small

Guarantees for flatness of Ψ

Guarantees for flatness of Ψ

- Similarly, let $s_b(x) := \mathbb{E}[Y^\dagger \mid A = 1, X = x] - \mathbb{E}[Y^\dagger \mid A = 0, X = x]$
- Condition 2 quantifies the amount of "flatness" that Ψ needs by relating the shape of Ψ with Φ

Guarantees for flatness of Ψ

- Similarly, let $s_b(x) := \mathbb{E}[Y^\dagger \,|\, A = 1, X = x] - \mathbb{E}[Y^\dagger \,|\, A = 0, X = x]$

- Condition 2 quantifies the amount of "flatness" that Ψ needs by relating the shape of Ψ with Φ

Condition 2 (Margin Condition between $Y^\dagger$ and Y^*). For some $\alpha > 2$,

$$\Pr_0\left(\left|s_b(X)\right| \geq t \left|q_b(X)\right|\right) \leq t^{-\alpha}, \forall t > 1.$$

Guarantees for flatness of Ψ

- Similarly, let $s_b(x) := \mathbb{E}[Y^\dagger \mid A = 1, X = x] - \mathbb{E}[Y^\dagger \mid A = 0, X = x]$

- Condition 2 quantifies the amount of “flatness” that Ψ needs by relating the shape of Ψ with Φ

Condition 2 (Margin Condition between $Y^\dagger$ and Y^*). For some $\alpha > 2$,

$$\Pr_0\left(\left|s_b(X)\right| \geq t\left|q_b(X)\right|\right) \leq t^{-\alpha}, \forall t > 1.$$

- It ensures that when estimation problem is hard (i.e. $q_b(x)$ small for some x), $|s_b(x)|$ is not too large, i.e. the impact of a policy on this x is controlled

Efficient estimator

- Under these conditions (plus some regularity conditions), we can show that the similar one-step estimator for Ψ_{π^*} is efficient given dataset $\mathsf{D} := \{x_i, a_i, y_i\}_{i=1}^n$
- Let $s(a, x) = \mathbb{E}[Y^\dagger \,|\, A = a, X = x]$ be the expected subsidiary outcome, $p(a \,|\, x) = \Pr(A = a \,|\, X = x)$ be the conditional probability, and π_n^* be the best policy under D

Efficient estimator

- Under these conditions (plus some regularity conditions), we can show that the similar one-step estimator for Ψ_{π^*} is efficient given dataset D $:= \{x_i, a_i, y_i\}_{i=1}^n$
- Let $s(a,x) = \mathbb{E}[Y^\dagger \mid A = a, X = x]$ be the expected subsidiary outcome, $p(a \mid x) = \Pr(A = a \mid X = x)$ be the conditional probability, and π_n^* be the best policy under D

Theorem (Efficient estimator of Ψ_{π^*}). Under conditions including Condition 1 and 2, the one-step estimator

$$\hat{\psi}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}\{a_i = \pi_n^*(x_i)\}}{p(a_i \mid x_i)} (y_i^\dagger - s(a_i, x_i)) + s(\pi_n^*(x_i), x_i)$$

is an efficient estimator of Ψ_{π^*}.

Inference without margin condition

Inference without margin condition

- Suppose we have good estimates $\hat{\phi}_\pi$ of Φ_π, $\hat{\psi}_\pi$ of Ψ_π

Inference without margin condition

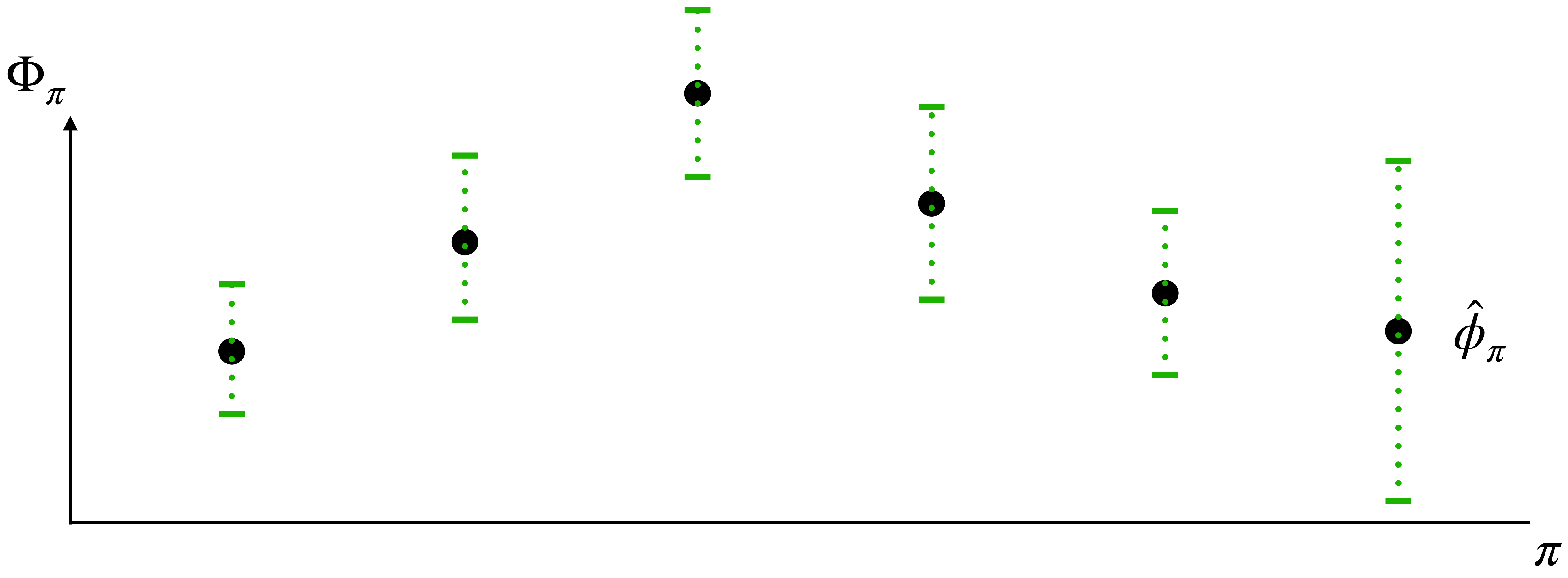
- Suppose we have good estimates $\hat{\phi}_\pi$ of Φ_π, $\hat{\psi}_\pi$ of Ψ_π
- Two-stage uniform confidence band

Inference without margin condition

- Suppose we have good estimates $\hat{\phi}_\pi$ of Φ_π, $\hat{\psi}_\pi$ of Ψ_π
- Two-stage uniform confidence band
    - First stage: eliminate policies that are unlikely to be optimal

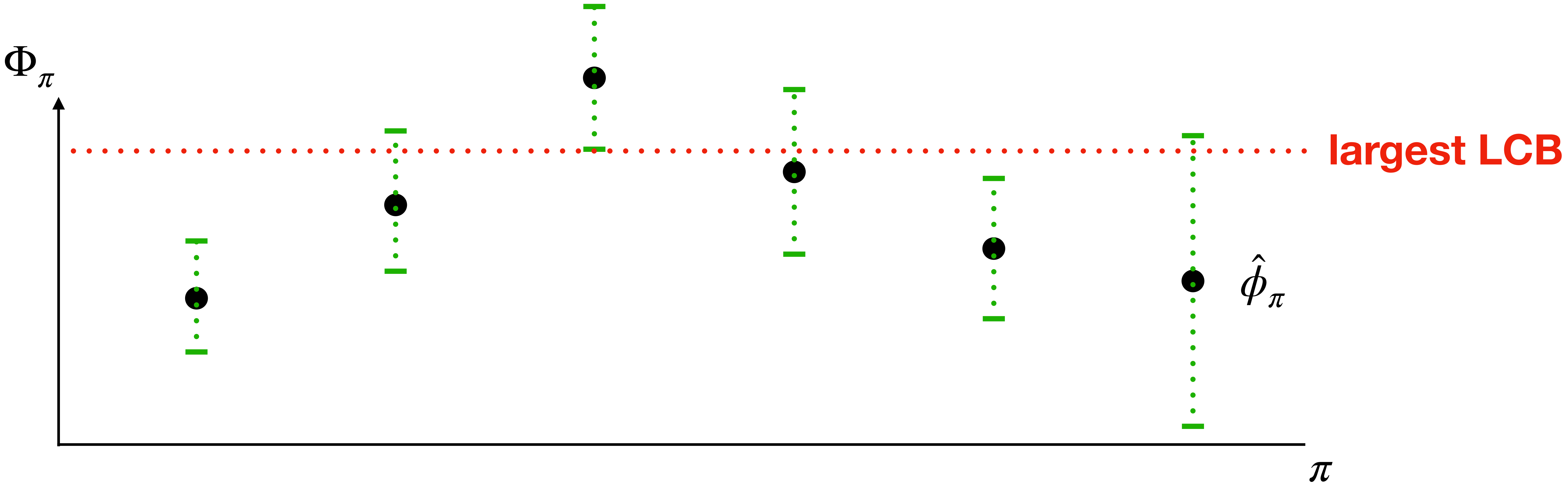
Inference without margin condition

- Suppose we have good estimates $\hat{\phi}_\pi$ of Φ_π, $\hat{\psi}_\pi$ of Ψ_π
- Two-stage uniform confidence band
  - First stage: eliminate policies that are unlikely to be optimal



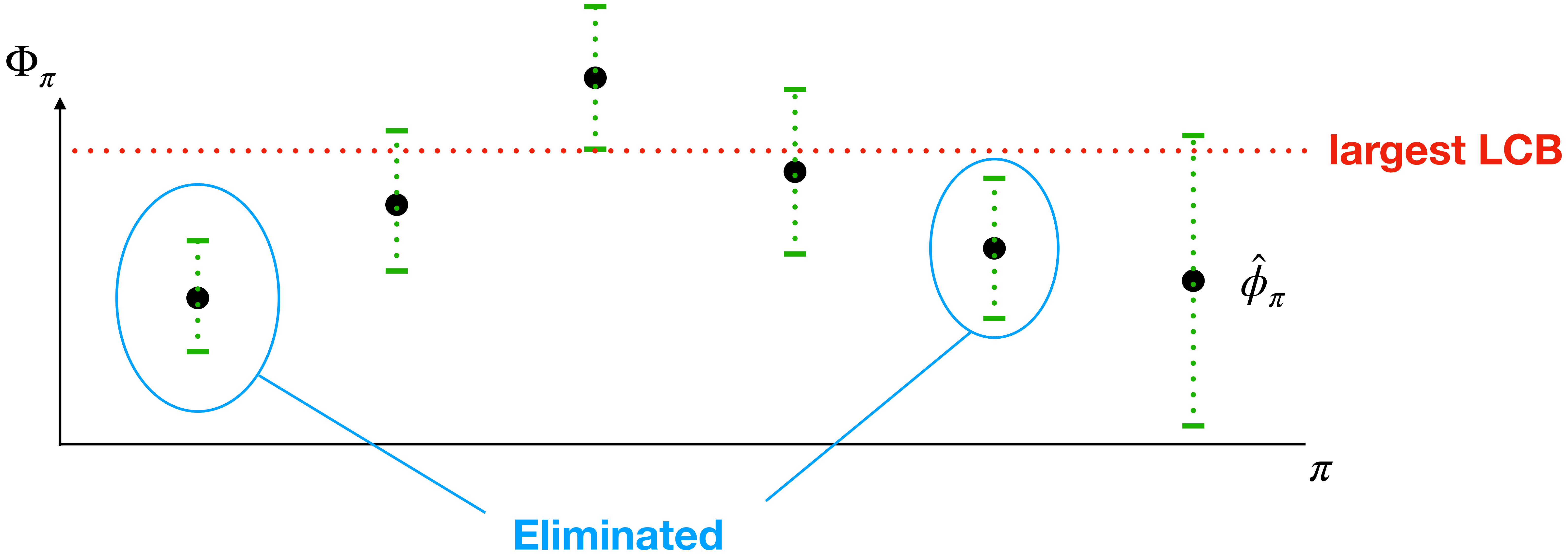
Inference without margin condition

- Suppose we have good estimates $\hat{\phi}_\pi$ of Φ_π, $\hat{\psi}_\pi$ of Ψ_π
- Two-stage uniform confidence band
  - First stage: eliminate policies that are unlikely to be optimal



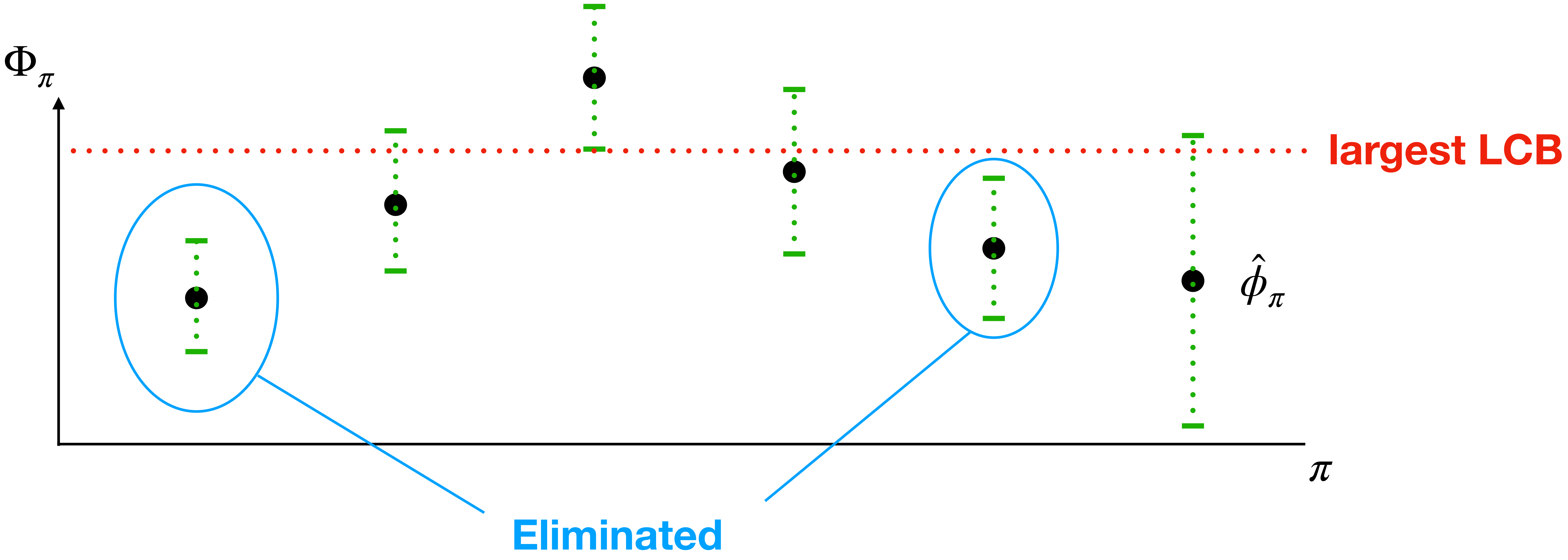
Inference without margin condition

- Suppose we have good estimates $\hat{\phi}_\pi$ of Φ_π, $\hat{\psi}_\pi$ of Ψ_π
- Two-stage uniform confidence band
  - First stage: eliminate policies that are unlikely to be optimal



Inference without margin condition

- Suppose we have good estimates $\hat{\phi}_\pi$ of Φ_π, $\hat{\psi}_\pi$ of Ψ_π
- Two-stage uniform confidence band
  - First stage: eliminate policies that are unlikely to be optimal



  - Second stage: construct a uniform confidence interval for the remaining policies

Uniform Confidence Band

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \hat{\phi}_\pi + \frac{\sigma_\pi t_{1-\beta/2}}{n^{1/2}} \geq \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \right\}.$

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

standard deviation w.r.t. $\hat{\phi}$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \hat{\phi}_\pi + \frac{\boxed{\sigma_\pi} t_{1-\beta/2}}{n^{1/2}} \geq \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \right\}.$

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

standard deviation w.r.t. $\hat{\phi}$ $1-\beta/2$ **quantile**

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \hat{\phi}_{\pi} + \frac{\boxed{\sigma_{\pi}}\boxed{t_{1-\beta/2}}}{n^{1/2}} \geq \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \right\}.$

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

standard deviation w.r.t. $\hat{\phi}$ $1 - \beta/2$ **quantile**

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \hat{\phi}_\pi + \frac{\boxed{\sigma_\pi} \boxed{t_{1-\beta/2}}}{n^{1/2}} \geq \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \right\}.$

- Second stage: construct a uniform confidence interval for the remaining policies

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

standard deviation w.r.t. $\hat{\phi}$ $1-\beta/2$ **quantile**

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \hat{\phi}_\pi + \frac{\sigma_\pi t_{1-\beta/2}}{n^{1/2}} \geq \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \right\}.$

- Second stage: construct a uniform confidence interval for the remaining policies

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\beta}} (\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi u_{1-(\alpha-\beta)/2}}{n^{1/2}}), \sup_{\pi \in \hat{\Pi}_{1-\beta}} (\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi u_{1-(\alpha-\beta)/2}}{n^{1/2}}) \right]$$

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

standard deviation w.r.t. $\hat{\phi}$ $1 - \beta/2$ quantile

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \hat{\phi}_{\pi} + \frac{\sigma_{\pi} t_{1-\beta/2}}{n^{1/2}} \geq \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \right\}.$

- Second stage: construct a uniform confidence interval for the remaining policies

standard deviation w.r.t. $\hat{\psi}$

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\beta}} (\hat{\psi}_{\pi} - \frac{\tilde{\sigma}_{\pi} u_{1-(\alpha-\beta)/2}}{n^{1/2}}),\ \sup_{\pi \in \hat{\Pi}_{1-\beta}} (\hat{\psi}_{\pi} + \frac{\tilde{\sigma}_{\pi} u_{1-(\alpha-\beta)/2}}{n^{1/2}}) \right]$$

Uniform Confidence Band

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

standard deviation w.r.t. $\hat{\phi}$ $1-\beta/2$ **quantile**

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \hat{\phi}_\pi + \frac{\sigma_\pi t_{1-\beta/2}}{n^{1/2}} \geq \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \right\}.$

- Second stage: construct a uniform confidence interval for the remaining policies

standard deviation w.r.t. $\hat{\psi}$ $1-(\alpha-\beta)/2$ **quantile**

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi u_{1-(\alpha-\beta)/2}}{n^{1/2}}\right), \sup_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi u_{1-(\alpha-\beta)/2}}{n^{1/2}}\right) \right]$$

A Joint Approach

A Joint Approach

- Replace the quantiles $(t_{1-\beta/2}, u_{1-(\alpha-\beta)/2})$ by $(t_{1-\alpha/2}, u_{1-\alpha/2})$ by considering the joint distribution of (Φ, Ψ)

A Joint Approach

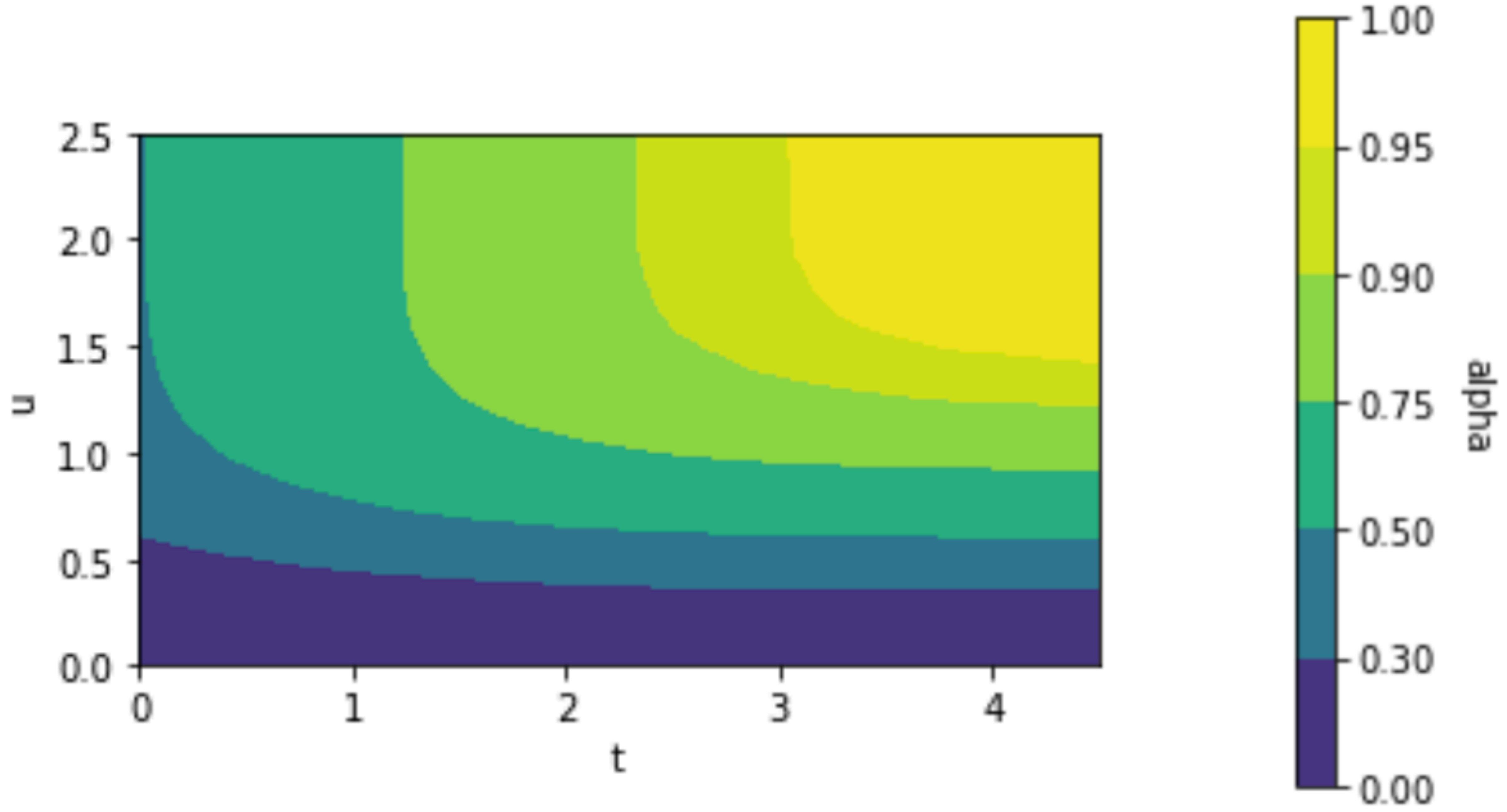
- Replace the quantiles $(t_{1-\beta/2}, u_{1-(\alpha-\beta)/2})$ by $(t_{1-\alpha/2}, u_{1-\alpha/2})$ by considering the joint distribution of (Φ, Ψ)

Theorem (confidence interval for Ψ_π). The following confidence interval contains $\{\Psi_\pi : \pi \in \Pi^*\}$ with probability at least $1-\alpha$ asymptotically:

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\alpha}} \left[\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi(P) u_{1-\alpha/2}}{n^{1/2}} \right], \sup_{\pi \in \hat{\Pi}_{1-\alpha}} \left[\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi(P) u_{1-\alpha/2}}{n^{1/2}} \right] \right].$$

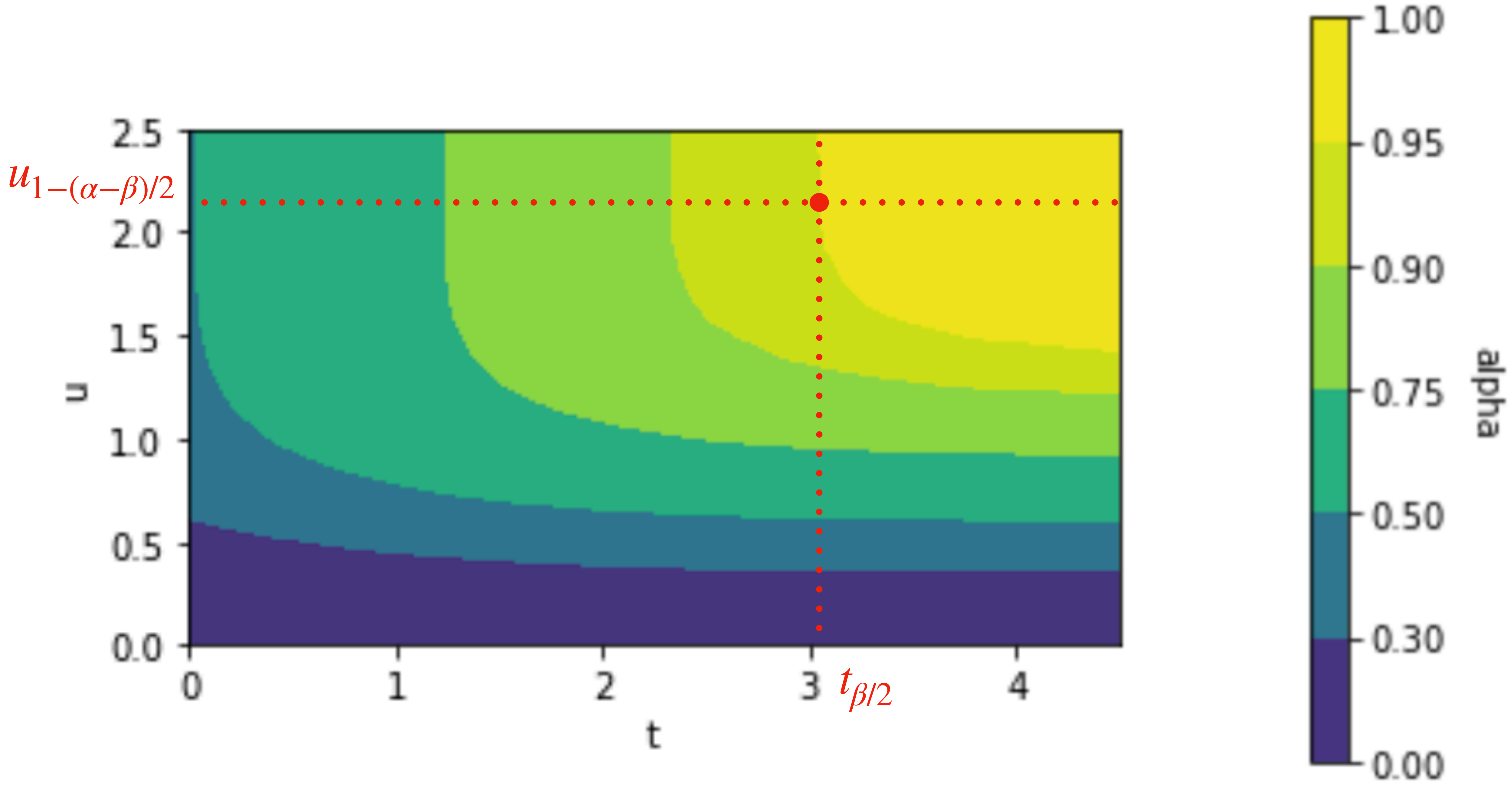
Why Joint Approach is Better

- We first demonstrate why the joint approach gives tighter confidence interval than the two-stage approach



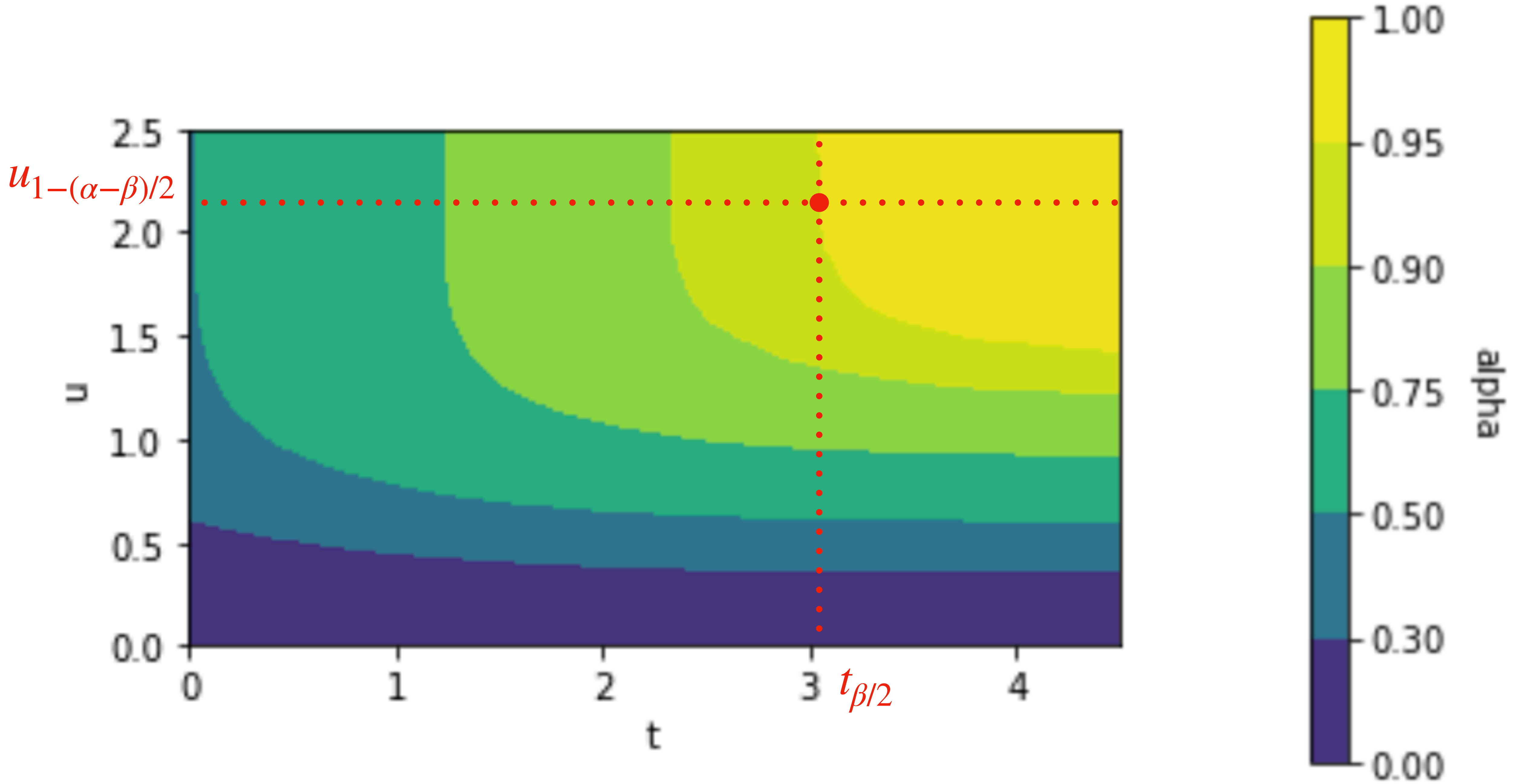
Why Joint Approach is Better

- We first demonstrate why the joint approach gives tighter confidence interval than the two-stage approach



Why Joint Approach is Better

- We first demonstrate why the joint approach gives tighter confidence interval than the two-stage approach

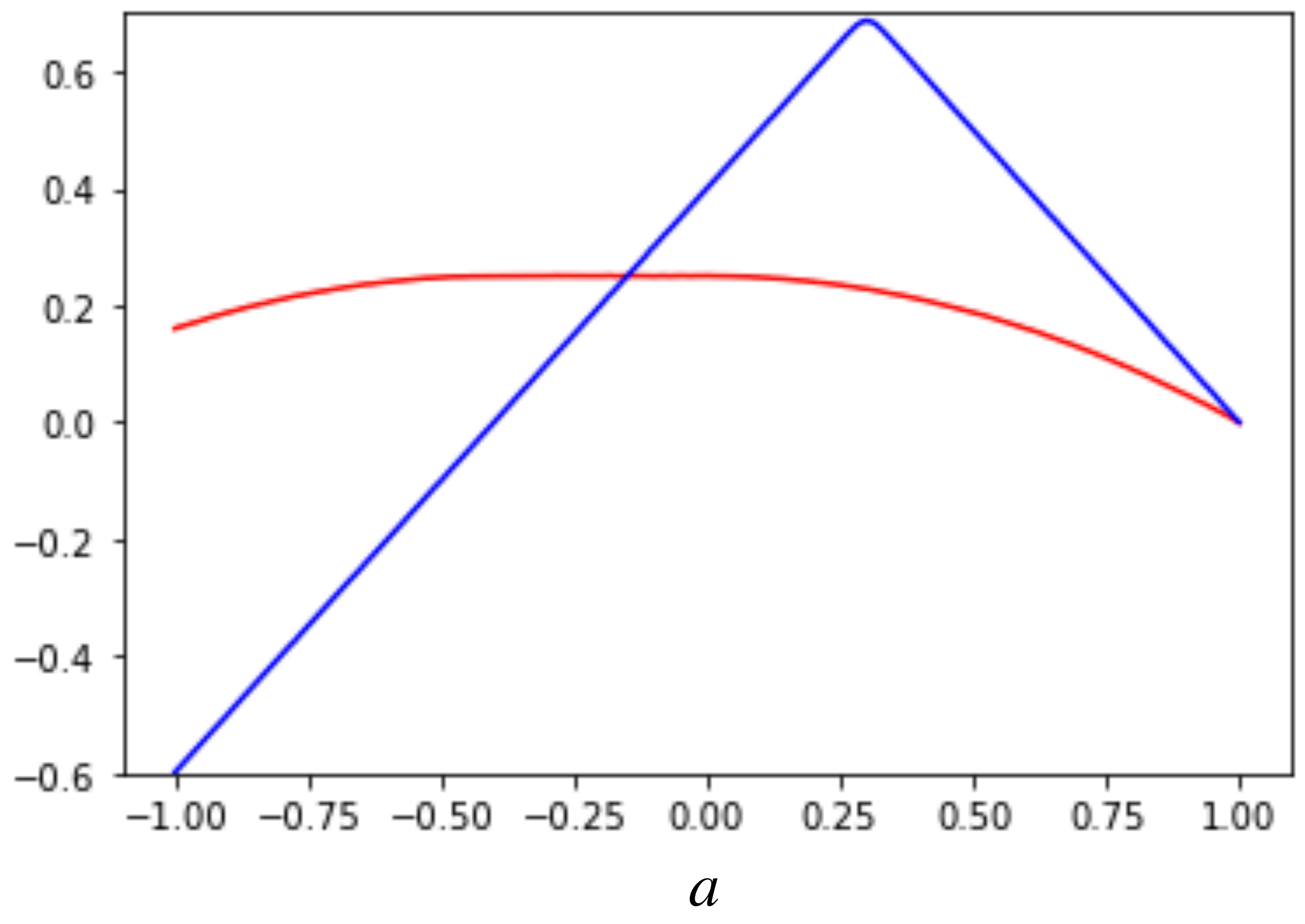


joint approach selects (t, u) which provides the tightest confidence interval

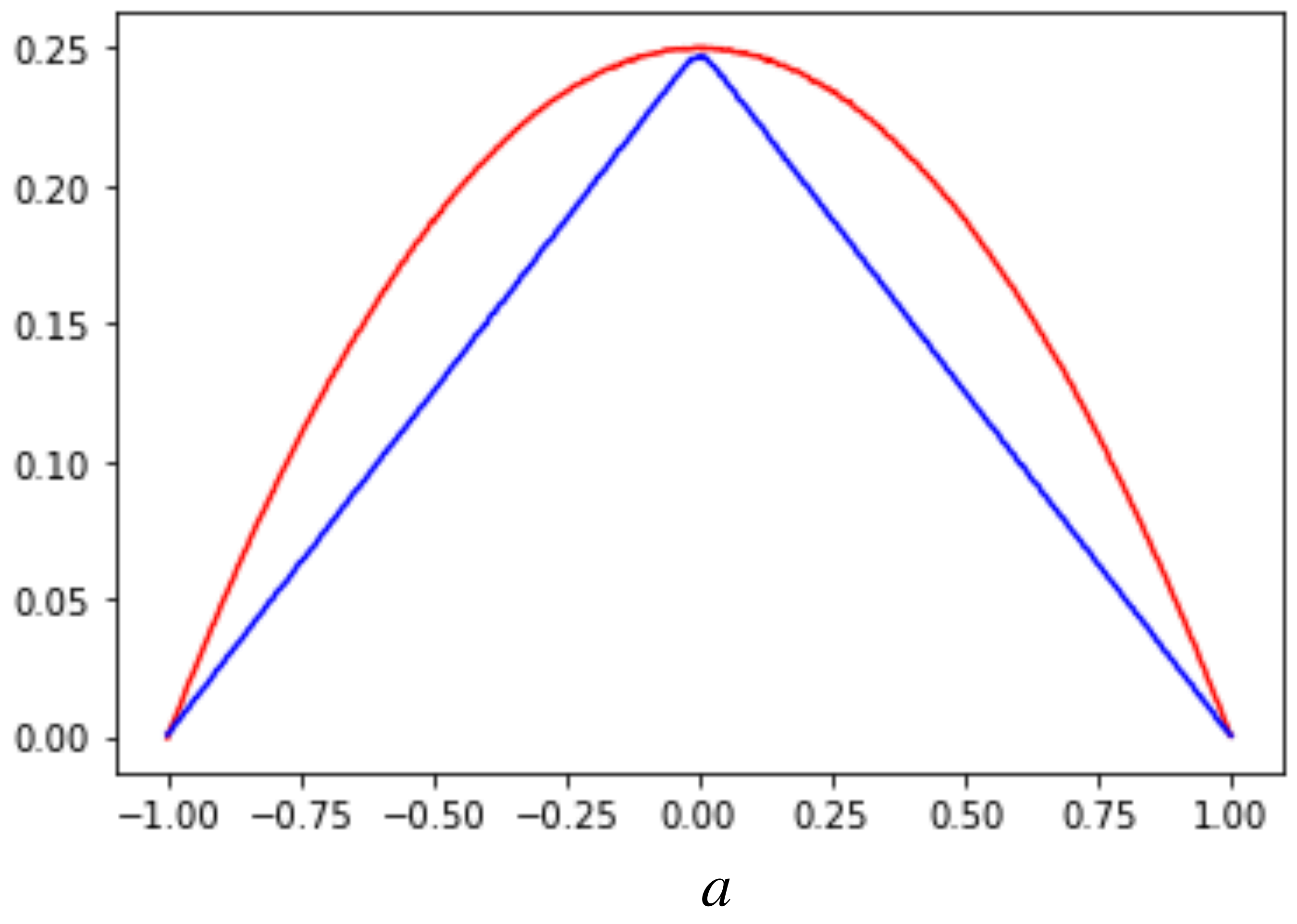
Simulation Setting

- Consider 1D case, threshold policy class $\Pi = \{\mathbf{1}\{x \geq a\} : a \in \mathbb{R}\}$

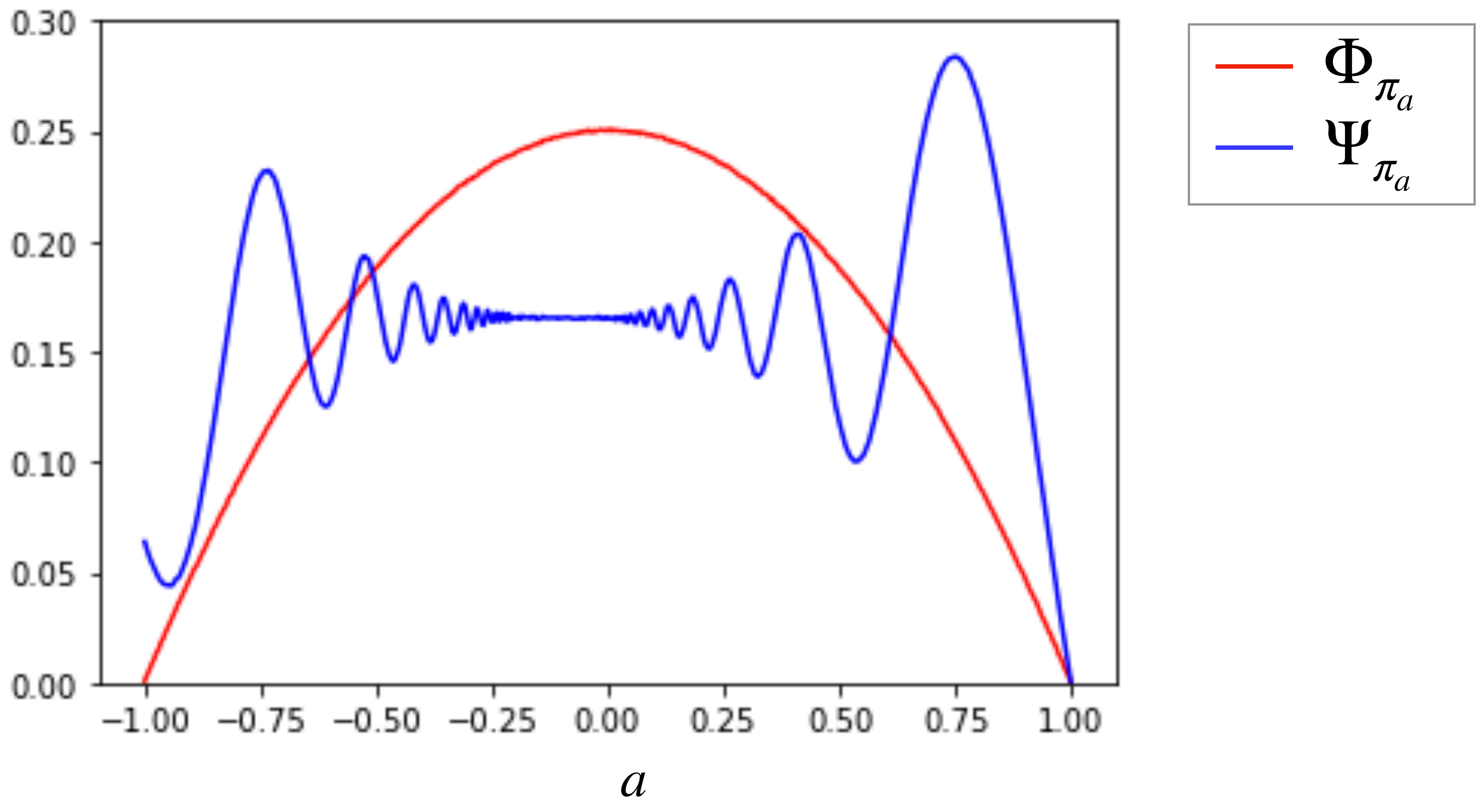
- Consider three scenarios:



π_* is non-unique



π_* is unique, Y^* and $Y^\dagger$ correlated



π_* is unique, Y^* and $Y^\dagger$ not correlated

Detailed Setting

- For the uniform confidence band method and the joint method, estimate the supremum via multiplier bootstrap

- In each setting, we simulate X with a sample size of 8000 for 1000 iterations

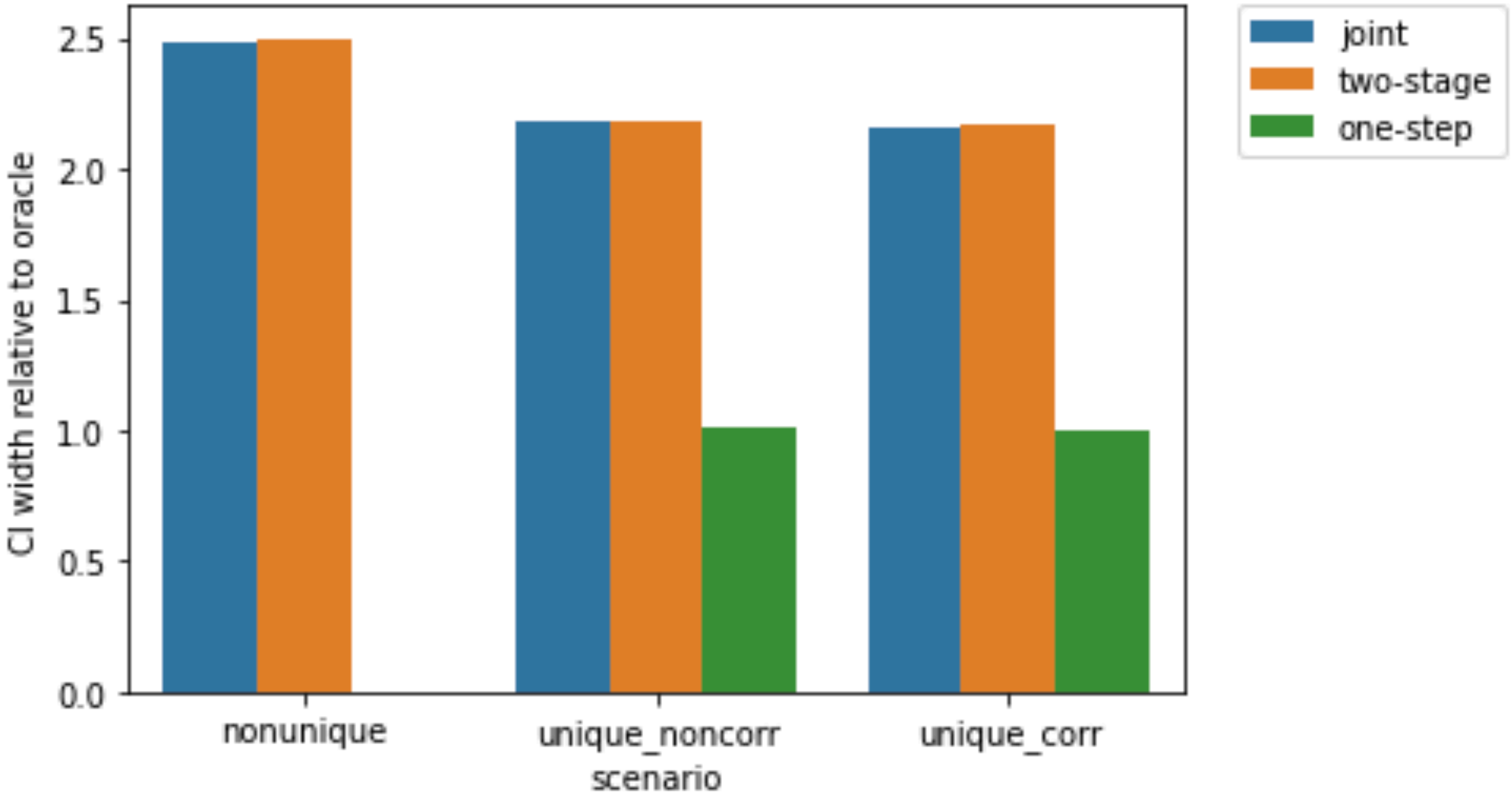
- Use bootstrap sample size of 1000, a confidence level of $\alpha = 0.05$

Simulation Results

Table: coverages for various scenarios

	two-stage	joint	one-step
non-unique	1.0	1.0	0.0
unique non-correlated	0.98	0.98	0.812
unique correlated	0.978	0.981	0.949

Figure: confidence interval width for various scenarios



Summary

- Propose a margin condition and construct an efficient estimator for $\{\Psi_\pi : \pi \in \Pi^*\}$
- Present a two-stage and a joint approach to make inference on $\{\Psi_\pi : \pi \in \Pi^*\}$ without the margin condition
- Run numerical experiments to show the desirable properties of the methods

Outline

- Project 1: Instance-optimal PAC Contextual bandits
- Project 2: Estimation of the mean of subsidiary outcome
- Future Work

Plans for Third Project

- Policy learning when the action space is large
- Application to pricing problem
  - At time t, a customer arrives, the learner plays price p_t and receive revenue $R(p_t)$
  - Assume $p_\star := \arg\max_{p \in \mathbb{R}} R(p)$, one objective is to identify $p_\star$
- Can still use the algorithm before, but will not be computationally efficient

Related Work and Objectives

Related Work and Objectives

- Existing methods:
    - discretizing the action space [Krishnamurthy et al. 2020]: minimax results
- Efficient computation: posterior sampling method

Related Work and Objectives

- Existing methods:
  - discretizing the action space [Krishnamurthy et al. 2020]: minimax results
- Efficient computation: posterior sampling method

Question:

- What is an instance-dependent PAC lower bounds when action space is large?
- Is there a computationally efficient algorithm in this setting?

Thanks!

Inefficiency of low-regret algorithms

Inefficiency of low-regret algorithms

Theorem [Li et al. 2022] There exists an instance μ such that for any α-minimax regret algorithm that is $(0,\delta)$-PAC, the stopping time satisfies

$$\mathbb{E}_\mu[\tau] \geq |\Pi|^2 \Delta^{-2} \log^2(1/2.4\delta)/4\alpha.$$

Posterior Sampling

- Assume $R(p_t)$ has a linear form $R(p_t) = \left\langle \phi_{p_t}, \theta* \right\rangle$, a framework is as follows:

```
Input: Prior $\Pi_0$ for $\theta*$
for $t = 1,2,\cdots$
    1. sample $\tilde{\theta} \sim \Pi_{t-1}$
    2. compute $p_t = \arg\max_p R(p, \tilde{\theta})$
    3. Update posterior $\Pi_t$
```

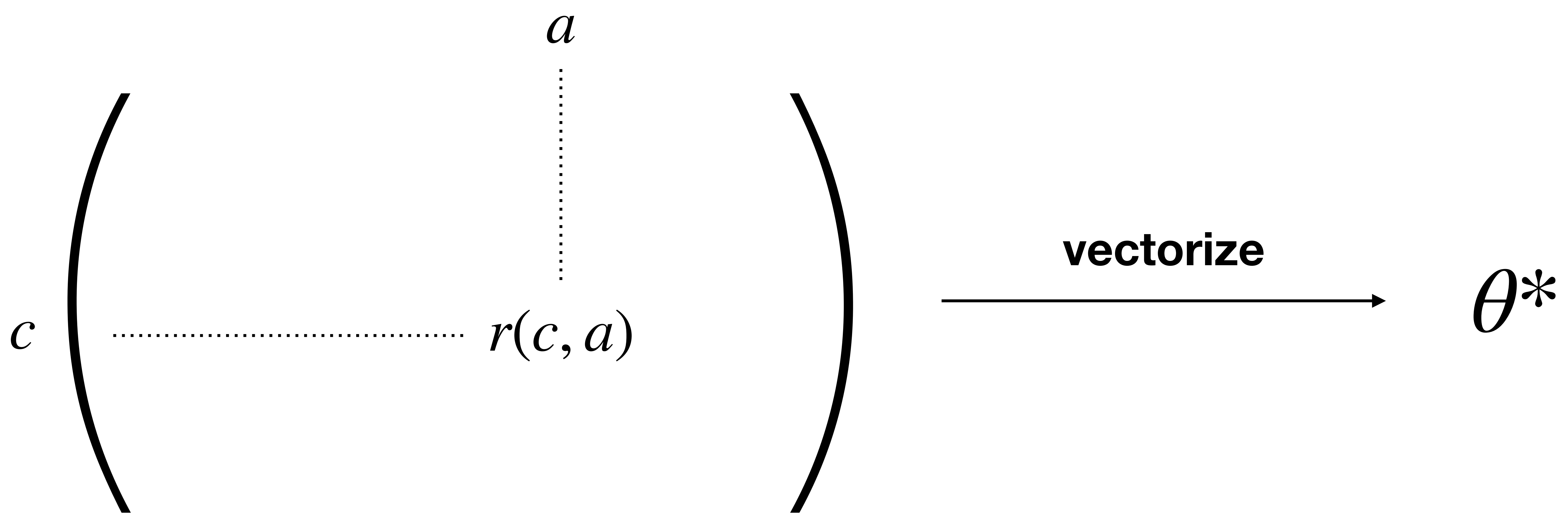
- Can we show that posterior sampling works in this setting? If not, what is the computational limit of posterior sampling methods, i.e. a lower bound?

Agnostic Setting Reduces to Linear

- What if we do not assume linear structure of reward function?

We can reduce it to the previous setting by constructing ϕ!

- Let $\theta^* \in \mathbb{R}^{|C|\times|A|}$ where $[\theta^*]_{c,a} = r(c, a)$



Agnostic Setting Reduces to Linear

$$r(c,a) = \left\langle \textbf{vec}(e_c e_a^\top), \theta^* \right\rangle$$

$$\phi(c,a)$$

$$\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}} = \sum_c \nu_c \sum_a \frac{1}{p_{c,a}} (\mathbf{1}\{\pi(c) = a\} - \mathbf{1}\{\pi_*(c) = a\})^2 = \mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}} + \frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c) \neq \pi(c)\}\right].$$

Agnostic Setting Reduces to Linear

$$r(c,a) = \left\langle \mathbf{vec}(e_c e_a^\top), \theta^* \right\rangle$$

$\phi(c,a)$

$$\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}} = \sum_c \nu_c \sum_a \frac{1}{p_{c,a}}\left(\mathbf{1}\{\pi(c)=a\} - \mathbf{1}\{\pi_*(c)=a\}\right)^2 = \mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}} + \frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq\pi(c)\}\right].$$

Variance

$$\rho_{\Pi,\epsilon} := \min_{p_c \in \triangle_A,\ \forall c \in C}\ \max_{\pi\in\Pi\setminus\pi_*} \frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}} + \frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq\pi(c)\}\right]}{\left(\mathbb{E}_{c\sim\nu}[\, r(c,\pi_*(c)) - r(c,\pi(c))\,] \vee \epsilon\right)^2}.$$

Gap

Uniform Confidence Band

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\{\mathbb{G}f : f \in \mathscr{F}\}$ be some Gaussian process characterizing the behavior of $\hat{\phi}_\pi$

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\{\mathbb{G}f : f \in \mathscr{F}\}$ be some Gaussian process characterizing the behavior of $\hat{\phi}_\pi$
- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\{\mathbb{G}f : f \in \mathscr{F}\}$ be some Gaussian process characterizing the behavior of $\hat{\phi}_\pi$
- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies
- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \frac{\sigma_\pi t_{1-\beta/2}}{n^{1/2}} \right\}.$

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\{\mathbb{G}f : f \in \mathscr{F}\}$ be some Gaussian process characterizing the behavior of $\hat{\phi}_\pi$
- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

$1-\beta/2$ **quantile of** $\sup_{f\in\mathscr{F}} \mathbb{G}f$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \frac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\{\mathbb{G}f : f \in \mathscr{F}\}$ be some Gaussian process characterizing the behavior of $\hat{\phi}_\pi$
- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

$1-\beta/2$ **quantile of** $\sup_{f\in\mathscr{F}} \mathbb{G}f$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi'\in\Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \frac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$

$$\left[\inf_{\pi\in\hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi z_{1-(\alpha-\beta)/2}}{n^{1/2}}\right), \sup_{\pi\in\hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi z_{1-(\alpha-\beta)/2}}{n^{1/2}}\right) \right]$$

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation

- Let $\{\mathbb{G}f : f \in \mathscr{F}\}$ be some Gaussian process characterizing the behavior of $\hat{\phi}_\pi$

- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

$1 - \beta/2$ **quantile of** $\sup_{f \in \mathscr{F}} \mathbb{G}f$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \le \hat{\phi}_\pi + \frac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi z_{1-(\alpha-\beta)/2}}{n^{1/2}}\right), \sup_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi \boxed{z_{1-(\alpha-\beta)/2}}}{n^{1/2}}\right) \right]$$

$1 - (\alpha - \beta)/2$ **quantile of the normal distribution**

Uniform Confidence Band

- Suppose $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\{\mathbb{G}f : f \in \mathcal{F}\}$ be some Gaussian process characterizing the behavior of $\hat{\phi}_\pi$
- We spend $\beta < \alpha$ of the confidence level in the first-stage to eliminate policies

$1 - \beta/2$ **quantile of** $\sup_{f \in \mathcal{F}} \mathbb{G}f$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \frac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \frac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$
- Second stage: construct a uniform confidence interval for the remaining policies

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\beta}} (\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi z_{1-(\alpha-\beta)/2}}{n^{1/2}}), \sup_{\pi \in \hat{\Pi}_{1-\beta}} (\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi \boxed{z_{1-(\alpha-\beta)/2}}}{n^{1/2}}) \right]$$

$1 - (\alpha - \beta)/2$ **quantile of the**
normal distribution

A Joint Approach

A Joint Approach

- Replace the quantiles $(t_{1-\beta/2}, u_{1-(\alpha-\beta)/2})$ by $(t_{1-\alpha/2}, u_{1-\alpha/2})$

A Joint Approach

- Replace the quantiles $(t_{1-\beta/2}, u_{1-(\alpha-\beta)/2})$ by $(t_{1-\alpha/2}, u_{1-\alpha/2})$

A Joint Approach

- Replace the quantiles $(t_{1-\beta/2}, u_{1-(\alpha-\beta)/2})$ by $(t_{1-\alpha/2}, u_{1-\alpha/2})$

remove the union bound argument!

A Joint Approach

- Replace the quantiles $(t_{1-\beta/2}, u_{1-(\alpha-\beta)/2})$ by $(t_{1-\alpha/2}, u_{1-\alpha/2})$

remove the union bound argument!

- More specifically, choose $(t_{1-\alpha/2}, u_{1-\alpha/2})$ such that

$$\inf_{\pi\in\Pi} \Pr\left\{\sup_{f\in\mathscr{F}} |\mathbb{G}f| \leq t_{1-\alpha/2}, \mathbb{G}\tilde{f}_\pi \leq u_{1-\alpha/2}\right\} \geq 1-\alpha/2.$$

A Joint Approach

- Replace the quantiles $(t_{1-\beta/2}, u_{1-(\alpha-\beta)/2})$ by $(t_{1-\alpha/2}, u_{1-\alpha/2})$

remove the union bound argument!

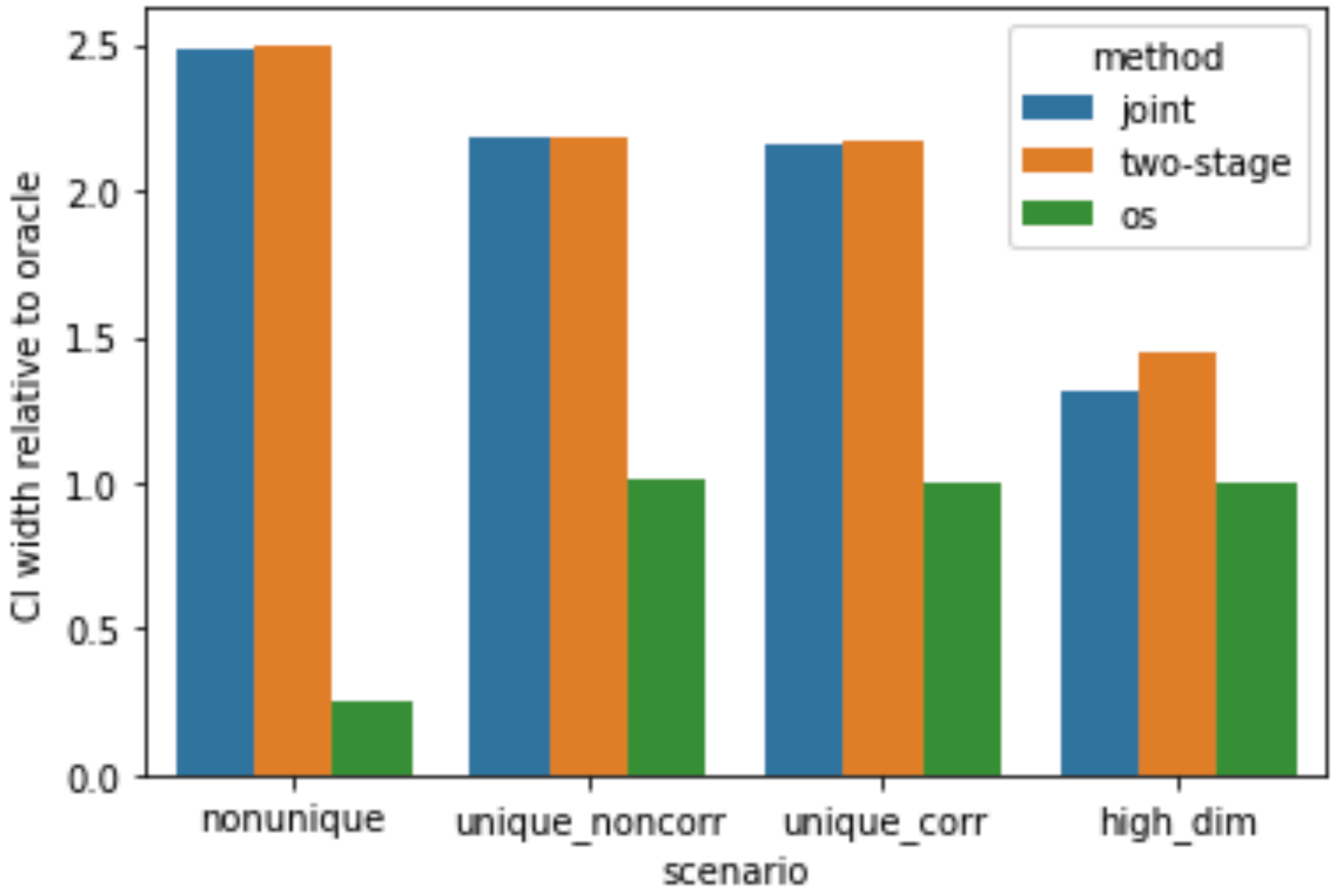
- More specifically, choose $(t_{1-\alpha/2}, u_{1-\alpha/2})$ such that

$$\inf_{\pi \in \Pi} \Pr\left\{ \sup_{f \in \mathcal{F}} |\mathbb{G} f| \leq t_{1-\alpha/2}, \mathbb{G} \tilde{f}_\pi \leq u_{1-\alpha/2} \right\} \geq 1 - \alpha/2.$$

Theorem (confidence interval for Ψ_π). The following confidence interval contains Ψ_π with probability at least $1-\alpha$ asymptotically:

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\alpha}} \left[\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi(P) u_{1-\alpha/2}}{n^{1/2}} \right], \sup_{\pi \in \hat{\Pi}_{1-\alpha}} \left[\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi(P) u_{1-\alpha/2}}{n^{1/2}} \right] \right].$$

3D Simulation



Inefficiency of Anytime CI and Robust Mean Estimator

- Anytime confidence interval scales like $\sqrt{t \log(1/\delta)}$, which is vacuous as $t \to \infty$
- Let $\Psi(P) := \Psi_{\pi_P^*}(P)$. Without the margin condition 2, Ψ will not be pathwise differentiable around some P_0, i.e. the limit $\lim_{\epsilon \to 0} \frac{\Psi(P_\epsilon) - \Psi(P_0)}{\epsilon}$ does not converge.
- Also, $\Psi_{\pi_n^*}(P_0) - \Psi_{\pi_0^*}(P_0)$ is likely not $o_{P_0}(n^{-1/2})$ without the margin condition, so the CI constructed by any robust mean estimator will suffer this as well, which means that it is necessarily worse than the uniform confidence band approach, which has the $n^{-1/2}$ scaling in the confidence interval

Hard Instance

- Fix $m \in \mathbb{N}$, $\Delta \in (0,1]$ and let $\mathsf{C} = [m]$ with uniform distribution, $\mathsf{A} = \{0,1\}$.
- For $i = 1,\ldots,m$, let $\pi_i(j) = \mathbf{1}\{i = j\}$ and define $r(i,j) = \Delta\mathbf{1}\{j = \pi_1(i)\}$.
- Then $V(\pi_1) = \Delta$ and $V(\pi_i) = \Delta(1 - 2/m)$ for all $i \in \mathsf{C}\backslash\{1\}$.
- In this case, $m = |\Pi|$ and $\rho_{\Pi,0} = \dfrac{4/m}{(2\Delta/m)^2} = m\Delta^{-2}$.

Towards Lower Bound: Estimators

- Linear contextual bandit setting:
  - feature map: ϕ : C × A $\rightarrow \mathbb{R}^d$ such that $r(c,a) = \langle \phi(c,a), \theta* \rangle$ for $\theta* \in \Theta \subset \mathbb{R}^d$
- Given dataset D $= \{(c_t, a_t, r_t)\}_{t=1}^{n}$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c,a)\phi(c,a)^\top \theta*] = \sum_c \nu_c \sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top \theta*$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting:
    - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta^*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c, a)\phi(c, a)^\top}}_{A(p)} \theta^*$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting:
  - feature map: ϕ : C × A $\to \mathbb{R}^d$ such that $r(c,a) = \langle \phi(c,a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset D $= \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c,a)\phi(c,a)^\top \theta^*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top}}_{A(p)} \theta^*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} A(p)^{-1} \sum_{t=1}^n \phi(c_t, a_t) r_t$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting:
  - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta^*] = \underset{\color{orange}{A(p)}}{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c, a)\phi(c, a)^\top}}\, \theta^*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} A(p)^{-1} \sum_{t=1}^{n} \phi(c_t, a_t) r_t$$

IPW estimate!

Estimate the Gap

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$

- Let $\phi_\pi := \mathbb{E}_{c\sim\nu}[\phi(c,\pi(c))]$, an estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_\pi)^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_\pi) = \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{n}$$

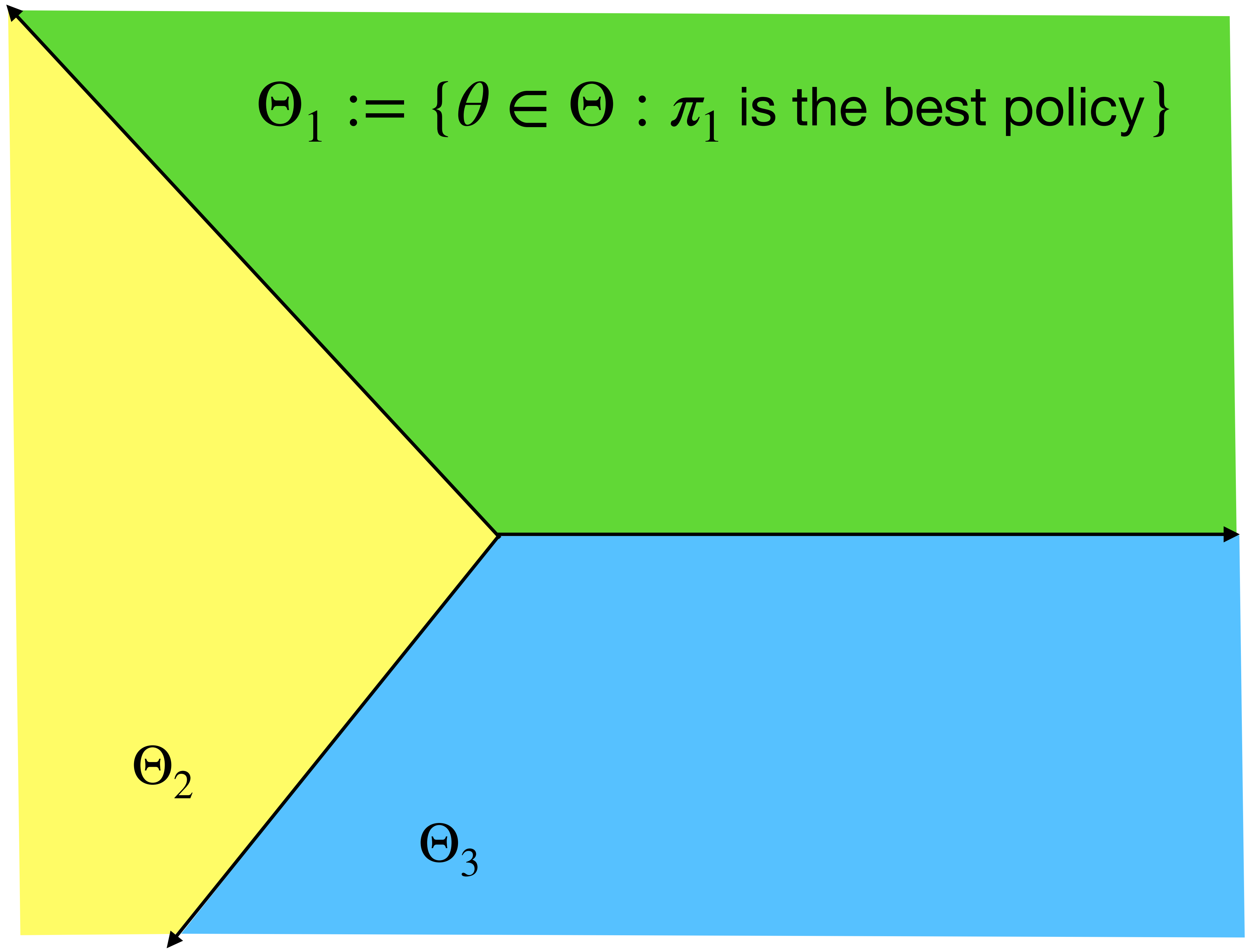
- Assuming Gaussian noise, with probability at least $1 - \delta$,

$$|\hat{\Delta}(\pi) - \Delta(\pi)| \leq \sqrt{\frac{2\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}$$

Towards Lower Bound

- Linear contextual bandit setting:
  - feature map: ϕ : C × A $\rightarrow \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Let $\phi_\pi := \mathbb{E}_{c \sim \nu}[\phi(c, \pi(c))]$, so for any $\pi \in \Pi$, $V(\pi) = \left\langle \phi_\pi, \theta^* \right\rangle$

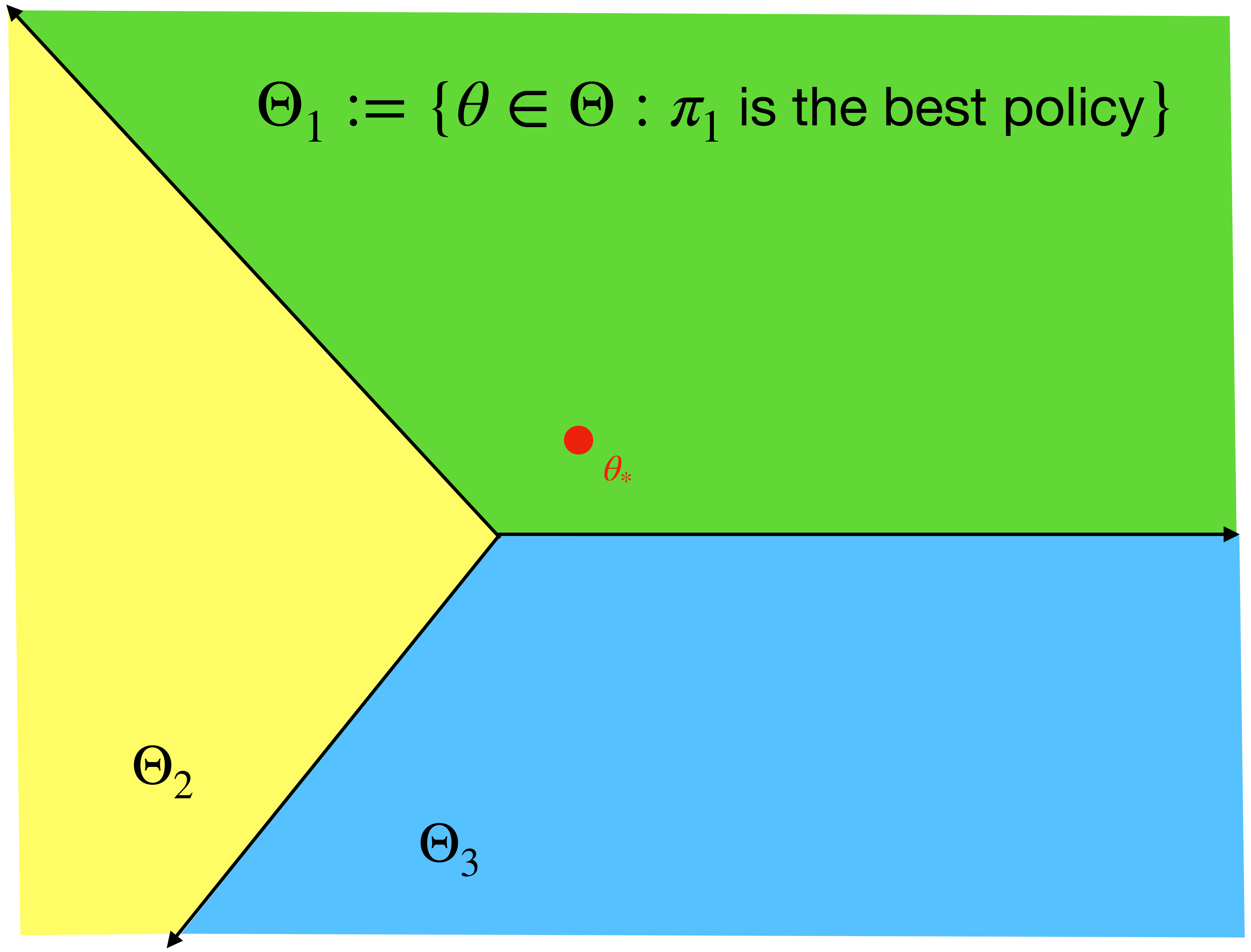
Lower Bound in Linear Contextual Bandits


$\Theta_1 := \{\theta \in \Theta : \pi_1$ is the best policy$\}$

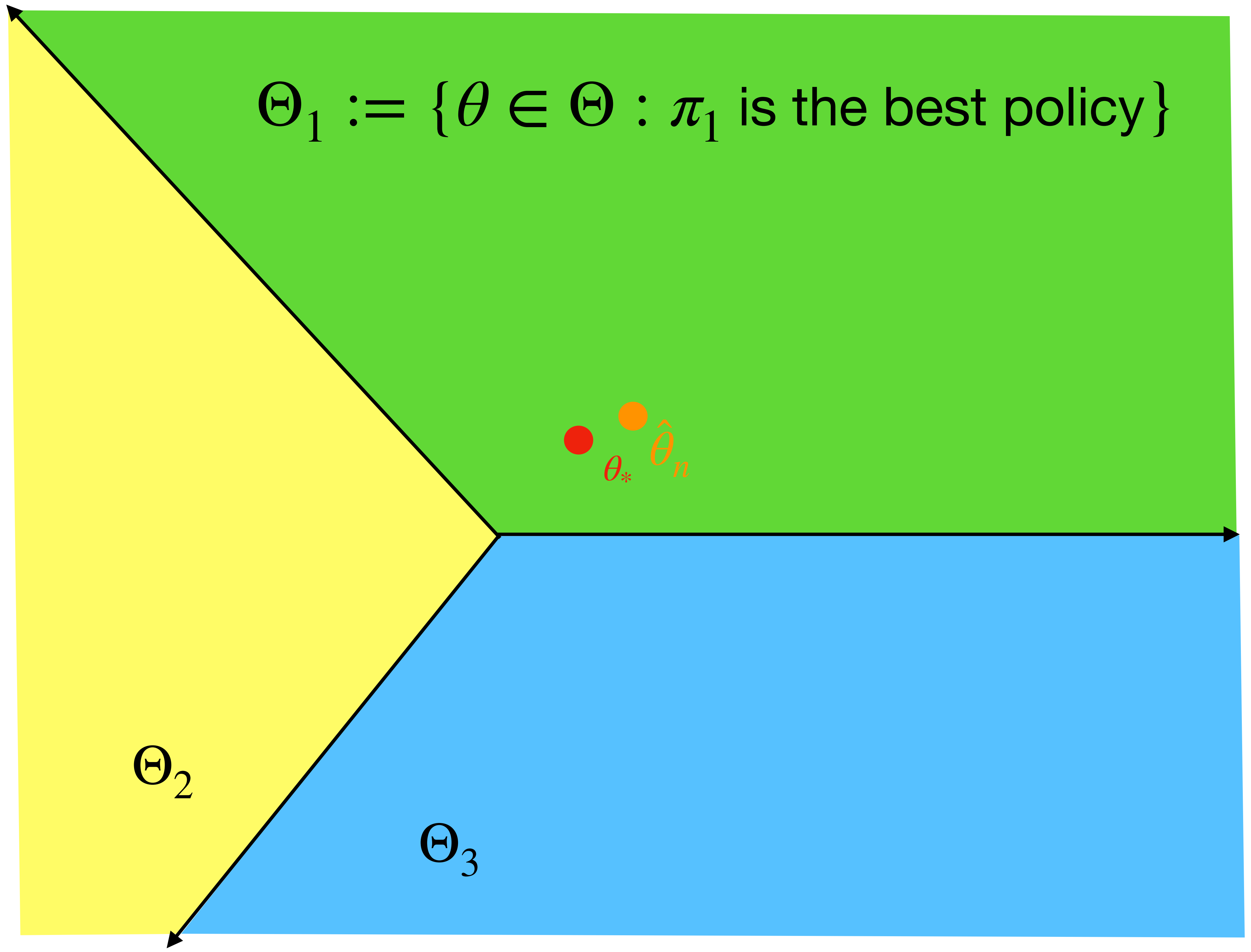
Θ_2

Θ_3

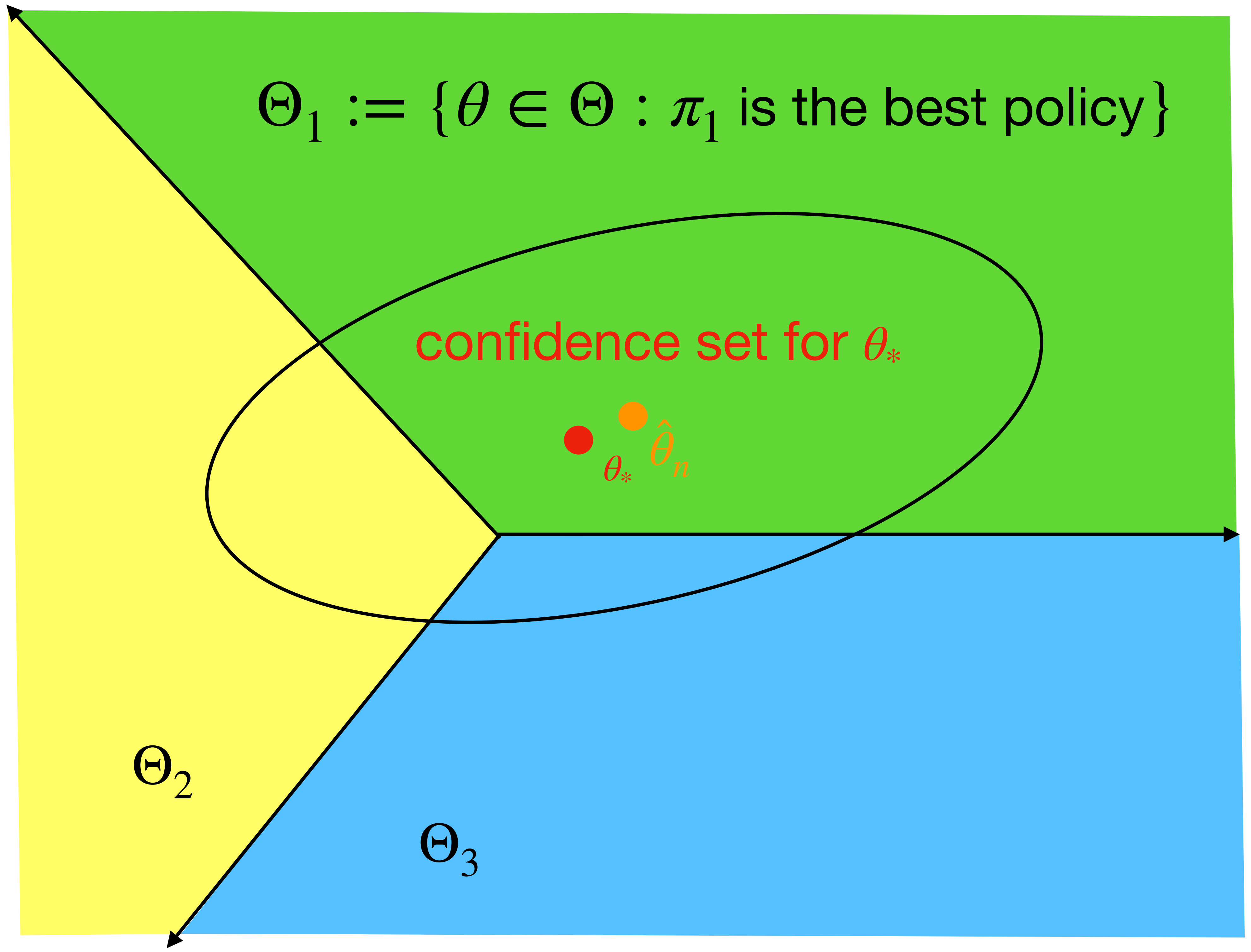
Lower Bound in Linear Contextual Bandits



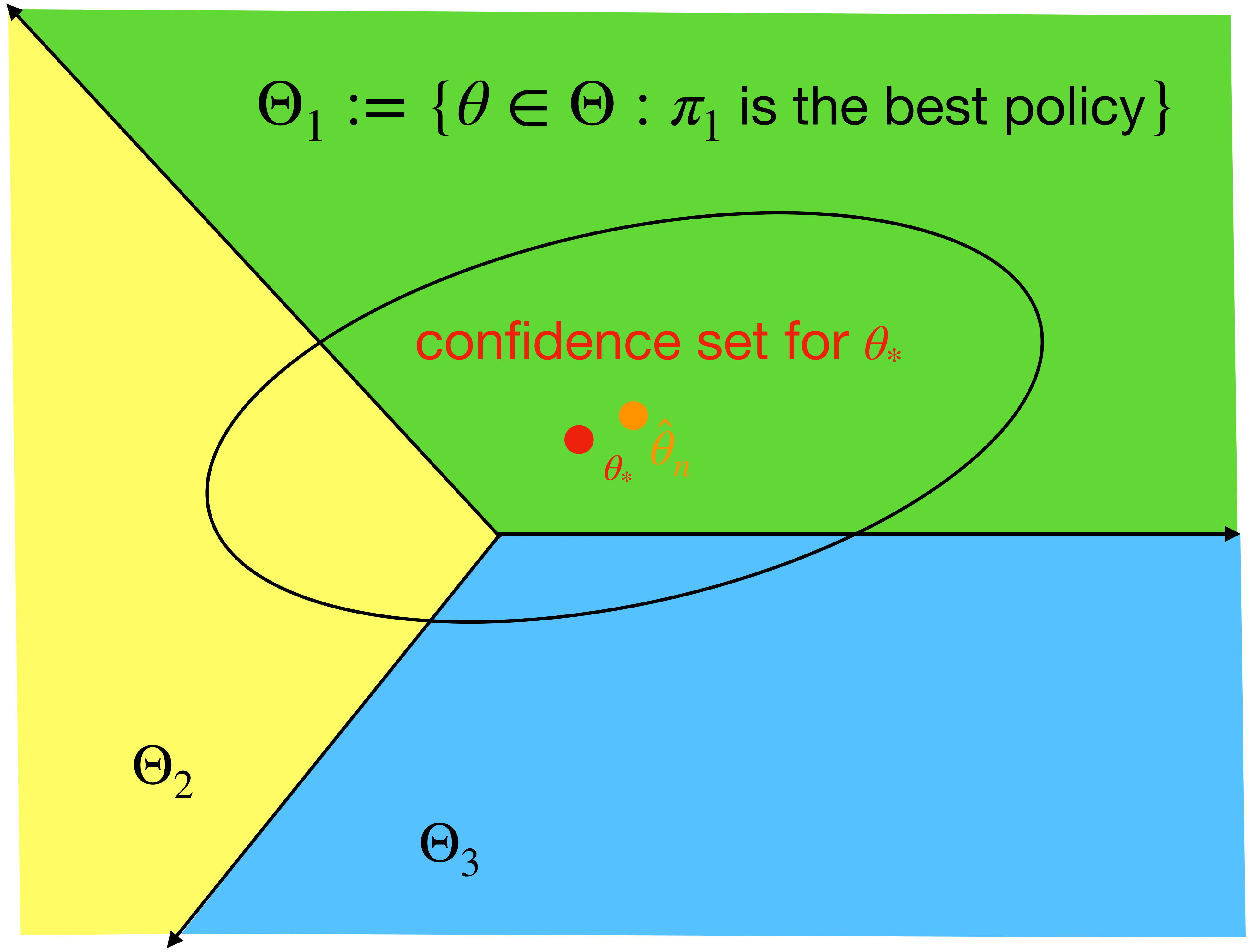
Lower Bound in Linear Contextual Bandits



Lower Bound in Linear Contextual Bandits



Lower Bound in Linear Contextual Bandits



Want confidence set to shrink to Θ_1 as quickly as possible!

A Lower Bound

- Let S_n denote the confidence set

- $\mathsf{S}_n \subset \Theta_1 \Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, V(\pi_*) - V(\pi) \geq 0$
  $\Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, (\phi_{\pi_*} - \phi_\pi)^\top \theta \geq 0$
  $\Leftrightarrow \forall \theta \in \mathsf{S}_n, (\phi_{\pi_*} - \phi_\pi)^\top \theta_* \geq (\phi_{\pi_*} - \phi_\pi)^\top (\theta_* - \theta)$

A Lower Bound

- Let S_n denote the confidence set

- $\mathsf{S}_n \subset \Theta_1 \Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, V(\pi_*) - V(\pi) \geq 0$
  $\Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, (\phi_{\pi_*} - \phi_\pi)^\top \theta \geq 0$
  $\Leftrightarrow \forall \theta \in \mathsf{S}_n, \underset{\textbf{gap}}{\boxed{(\phi_{\pi_*} - \phi_\pi)^\top \theta_*}} \geq (\phi_{\pi_*} - \phi_\pi)^\top (\theta_* - \theta)$

A Lower Bound

- Let S_n denote the confidence set

- $\mathsf{S}_n \subset \Theta_1 \Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, V(\pi_*) - V(\pi) \geq 0$
  $\Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, (\phi_{\pi_*} - \phi_\pi)^\top \theta \geq 0$
  $\Leftrightarrow \forall \theta \in \mathsf{S}_n, \underset{\text{gap}}{\boxed{(\phi_{\pi_*} - \phi_\pi)^\top \theta_*}} \geq \underset{\text{estimation error of the gap}}{\boxed{(\phi_{\pi_*} - \phi_\pi)^\top (\theta_* - \theta)}}$

A Lower Bound

- Let S_n denote the confidence set

- $\mathsf{S}_n \subset \Theta_1 \Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, V(\pi_*) - V(\pi) \geq 0$

$$\Leftrightarrow \forall \pi \in \Pi, \forall \theta \in \mathsf{S}_n, (\phi_{\pi_*} - \phi_\pi)^\top \theta \geq 0$$

$$\Leftrightarrow \forall \theta \in \mathsf{S}_n, \underset{\text{gap}}{\boxed{(\phi_{\pi_*} - \phi_\pi)^\top \theta_*}} \geq \underset{\text{estimation error of the gap}}{\boxed{(\phi_{\pi_*} - \phi_\pi)^\top (\theta_* - \theta)}}$$

Need estimates for θ^* and the gap!

Estimators for $\theta*$

- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^{n}$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t)r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta*] = \sum_{c} \nu_c \sum_{a} p_{c,a}\phi(c, a)\phi(c, a)^\top \theta*$$

Estimators for $\theta*$

- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^{n}$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c,a)\phi(c,a)^\top \theta*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top}}_{A(p)} \theta*$$

Estimators for $\theta*$

- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^{n}$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t)r_t] = \mathbb{E}_{c,a}[\phi(c,a)\phi(c,a)^\top \theta*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a}\phi(c,a)\phi(c,a)^\top}}_{A(p)}\theta*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n}A(p)^{-1}\sum_{t=1}^{n}\phi(c_t, a_t)r_t$$

Estimators for θ^*

- Given dataset $\mathrm{D} = \{(c_t, a_t, r_t)\}_{t=1}^{n}$ where $a_t \sim p_{c_t} \in \triangle_{\mathrm{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c,a)\phi(c,a)^\top \theta^*] = \underbrace{\sum_c \nu_c \sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top}_{A(p)} \theta^*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} A(p)^{-1} \sum_{t=1}^{n} \phi(c_t, a_t) r_t$$

IPW estimate!

Estimate the Gap

Estimate the Gap

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$

Estimate the Gap

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$
- An estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

Estimate the Gap

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$
- An estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_\pi)^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_\pi) = \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{n}$$

Estimate the Gap

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$

- An estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_{\pi}, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_{\pi})^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_{\pi}) = \frac{\|\phi_{\pi_*} - \phi_{\pi}\|^2_{A(p)^{-1}}}{n}$$

- Assuming Gaussian noise, with probability at least $1 - \delta$,

$$|\hat{\Delta}(\pi) - \Delta(\pi)| \leq \sqrt{\frac{2\|\phi_{\pi_*} - \phi_{\pi}\|^2_{A(p)^{-1}} \log(1/\delta)}{n}}$$

A Lower Bound

- Plugging in the guarantee: $\forall \theta \in \mathsf{S}_n, (\phi_{\pi_*} - \phi_\pi)^\top(\theta_* - \theta) \leq (\phi_{\pi_*} - \phi_\pi)^\top \theta_*$

$$\Leftrightarrow \sqrt{\frac{2\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}} \log(1/\delta)}{n}} \leq (\phi_{\pi_*} - \phi_\pi)^\top \theta_*$$

- Choose action distribution p such that:

$$\max_{\pi \in \Pi \backslash \pi_*} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{\Delta(\pi)^2} \leq \frac{n}{2\log(1/\delta)}$$

A Lower Bound

- Plugging in the guarantee: $\forall \theta \in \mathsf{S}_n, (\phi_{\pi_*} - \phi_\pi)^\top(\theta_* - \theta) \leq (\phi_{\pi_*} - \phi_\pi)^\top \theta_*$

$$\Leftrightarrow \sqrt{\frac{2\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}} \log(1/\delta)}{n}} \leq (\phi_{\pi_*} - \phi_\pi)^\top \theta_*$$

- Choose action distribution p such that:

$$\max_{\pi \in \Pi \backslash \pi_*} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{\Delta(\pi)^2} \leq \frac{n}{2\log(1/\delta)}$$

Theorem [Li et al. 2022] Let τ be the stopping time of the algorithm. Any $(0,\delta)$-PAC algorithm satisfies $\tau \geq \rho_{\Pi,0} \log(1/2.4\delta)$ with high probability where

$$\rho_{\Pi,0} = \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi \backslash \pi_*} \frac{\overset{\text{variance}}{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}}{\underset{\text{gap}}{\Delta(\pi)^2}}.$$

Uniform Confidence Band

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P
- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P
- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\sqrt{n}\frac{\hat{\phi}_\pi - \Phi_\pi}{\sigma_\pi} \to \mathbb{G}f$ and $\sqrt{n}\frac{\hat{\psi}_\pi - \Psi_\pi}{\tilde{\sigma}_\pi} \to \mathbb{G}\tilde{f}$

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P
- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\sqrt{n}\frac{\hat{\phi}_\pi - \Phi_\pi}{\sigma_\pi} \to \mathbb{G}f$ and $\sqrt{n}\frac{\hat{\psi}_\pi - \Psi_\pi}{\tilde{\sigma}_\pi} \to \mathbb{G}\tilde{f}$

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P
- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\sqrt{n}\dfrac{\hat{\phi}_\pi - \Phi_\pi}{\sigma_\pi} \to \mathbb{G}f$ and $\sqrt{n}\dfrac{\hat{\psi}_\pi - \Psi_\pi}{\tilde{\sigma}_\pi} \to \mathbb{G}\tilde{f}$
- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \dfrac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \dfrac{\sigma_\pi t_{1-\beta/2}}{n^{1/2}} \right\}.$

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P
- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\sqrt{n}\dfrac{\hat{\phi}_\pi - \Phi_\pi}{\sigma_\pi} \to \mathbb{G}f$ and $\sqrt{n}\dfrac{\hat{\psi}_\pi - \Psi_\pi}{\tilde{\sigma}_\pi} \to \mathbb{G}\tilde{f}$

$1 - \beta/2$ **quantile of** $\sup_{f \in \mathscr{F}} \mathbb{G}f$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \dfrac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \dfrac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P

- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation

- Let $\sqrt{n}\dfrac{\hat{\phi}_\pi - \Phi_\pi}{\sigma_\pi} \to \mathbb{G}f$ and $\sqrt{n}\dfrac{\hat{\psi}_\pi - \Psi_\pi}{\tilde{\sigma}_\pi} \to \mathbb{G}\tilde{f}$

$1 - \beta/2$ **quantile of** $\sup_{f \in \mathscr{F}} \mathbb{G}f$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \dfrac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \dfrac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$

- Second stage: construct a uniform confidence interval for the remaining policies

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P
- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\sqrt{n}\dfrac{\hat{\phi}_\pi - \Phi_\pi}{\sigma_\pi} \to \mathbb{G}f$ and $\sqrt{n}\dfrac{\hat{\psi}_\pi - \Psi_\pi}{\tilde{\sigma}_\pi} \to \mathbb{G}\tilde{f}$

$1-\beta/2$ **quantile of** $\sup_{f\in\mathscr{F}} \mathbb{G}f$

- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \dfrac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \dfrac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$
- Second stage: construct a uniform confidence interval for the remaining policies

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi z_{1-(\alpha-\beta)/2}}{n^{1/2}}\right), \sup_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi z_{1-(\alpha-\beta)/2}}{n^{1/2}}\right) \right]$$

Uniform Confidence Band

- Suppose D_π is a gradient of Φ_π at P, $\tilde{D}_\pi$ is a gradient of Ψ_π at P
- $\sigma_\pi := (PD_\pi^2)^{1/2}$, $\tilde{\sigma}_\pi := (P\tilde{D}_\pi^2)^{1/2}$, standard deviation
- Let $\sqrt{n}\dfrac{\hat{\phi}_\pi - \Phi_\pi}{\sigma_\pi} \to \mathbb{G}f$ and $\sqrt{n}\dfrac{\hat{\psi}_\pi - \Psi_\pi}{\tilde{\sigma}_\pi} \to \mathbb{G}\tilde{f}$
- First stage: $\hat{\Pi}_{1-\beta} := \left\{ \pi \in \Pi : \sup_{\pi' \in \Pi} \left[\hat{\phi}_{\pi'} - \dfrac{\sigma_{\pi'} t_{1-\beta/2}}{n^{1/2}} \right] \leq \hat{\phi}_\pi + \dfrac{\sigma_\pi \boxed{t_{1-\beta/2}}}{n^{1/2}} \right\}.$

$1-\beta/2$ quantile of $\sup_{f\in\mathcal{F}} \mathbb{G}f$

- Second stage: construct a uniform confidence interval for the remaining policies

$$\left[\inf_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi - \frac{\tilde{\sigma}_\pi z_{1-(\alpha-\beta)/2}}{n^{1/2}}\right), \sup_{\pi \in \hat{\Pi}_{1-\beta}} \left(\hat{\psi}_\pi + \frac{\tilde{\sigma}_\pi \boxed{z_{1-(\alpha-\beta)/2}}}{n^{1/2}}\right) \right]$$

$1-(\alpha-\beta)/2$ quantile of the normal distribution

A Lower Bound

A Lower Bound

Theorem [Li et al. 2022] Let τ be the stopping time of the algorithm. Any $(0,\delta)$-PAC algorithm satisfies $\tau \geq \rho_{\Pi,0} \log(1/2.4\delta)$ with high probability where

variance

$$\rho_{\Pi,0} = \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \; \max_{\pi \in \Pi \setminus \pi_*} \frac{\|\phi_{\pi_*} - \phi_{\pi}\|^2_{A(p)^{-1}}}{\Delta(\pi)^2}.$$

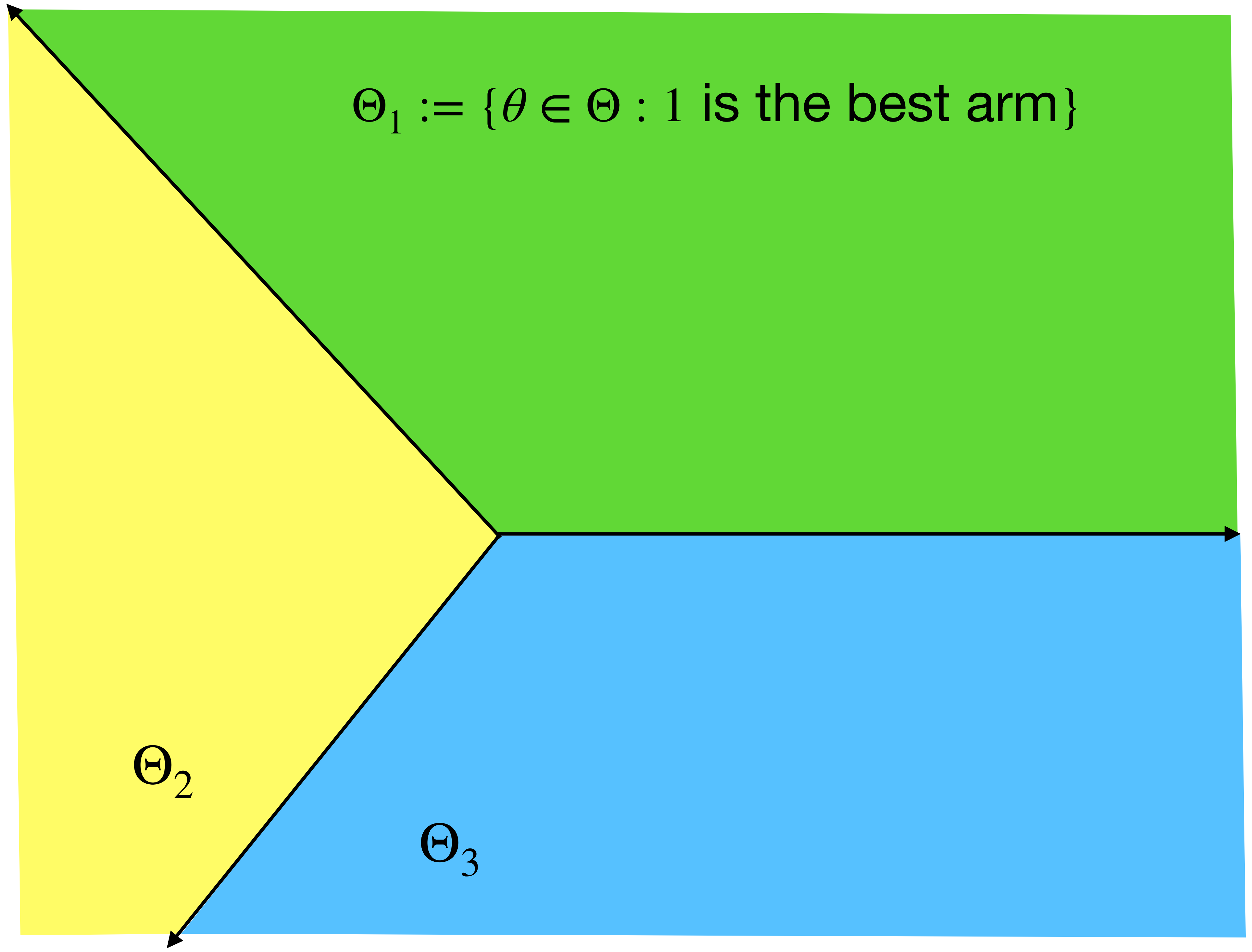
gap

A Lower Bound in Linear Bandits

- Set of features $x \in \mathscr{X}$, some unknown parameter $\theta^* \in \Theta \subset \mathbb{R}^d$
- At each time $t = 1,2,\cdots$:
    - Choose action $a_t \in \mathsf{A}$
    - Receive reward $r_t = \left\langle x_{a_t}, \theta^* \right\rangle + \epsilon$
- Goal: identify $a_* = \arg\max_{a \in \mathsf{A}} \left\langle x_a, \theta_* \right\rangle$

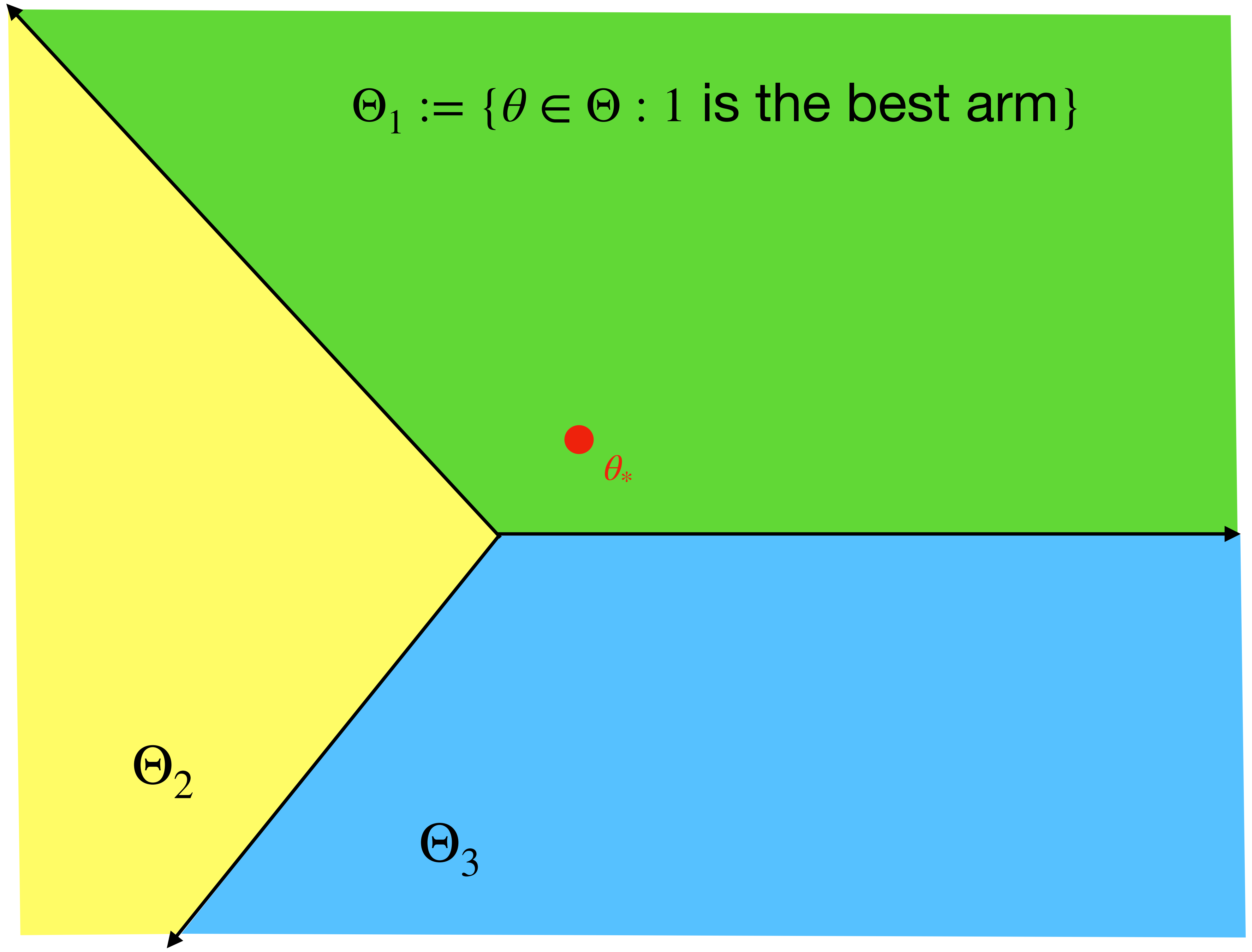
A Lower Bound in Linear Bandits

- Identify $a_* = \arg\max_{a \in \mathcal{A}} \langle x_a, \theta_* \rangle$



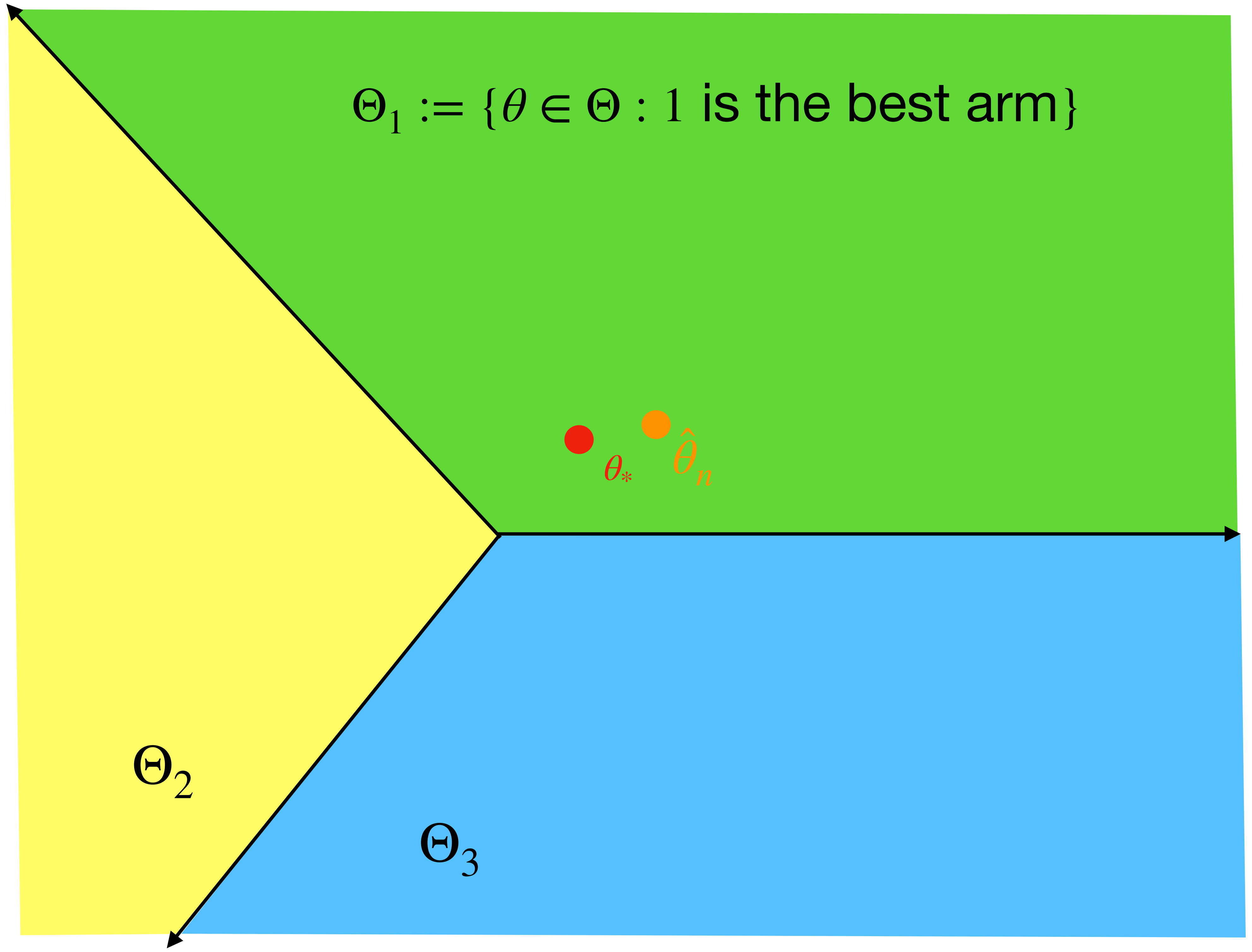
A Lower Bound in Linear Bandits

- Identify $a_* = \arg\max_{a \in \mathcal{A}} \langle x_a, \theta_* \rangle$



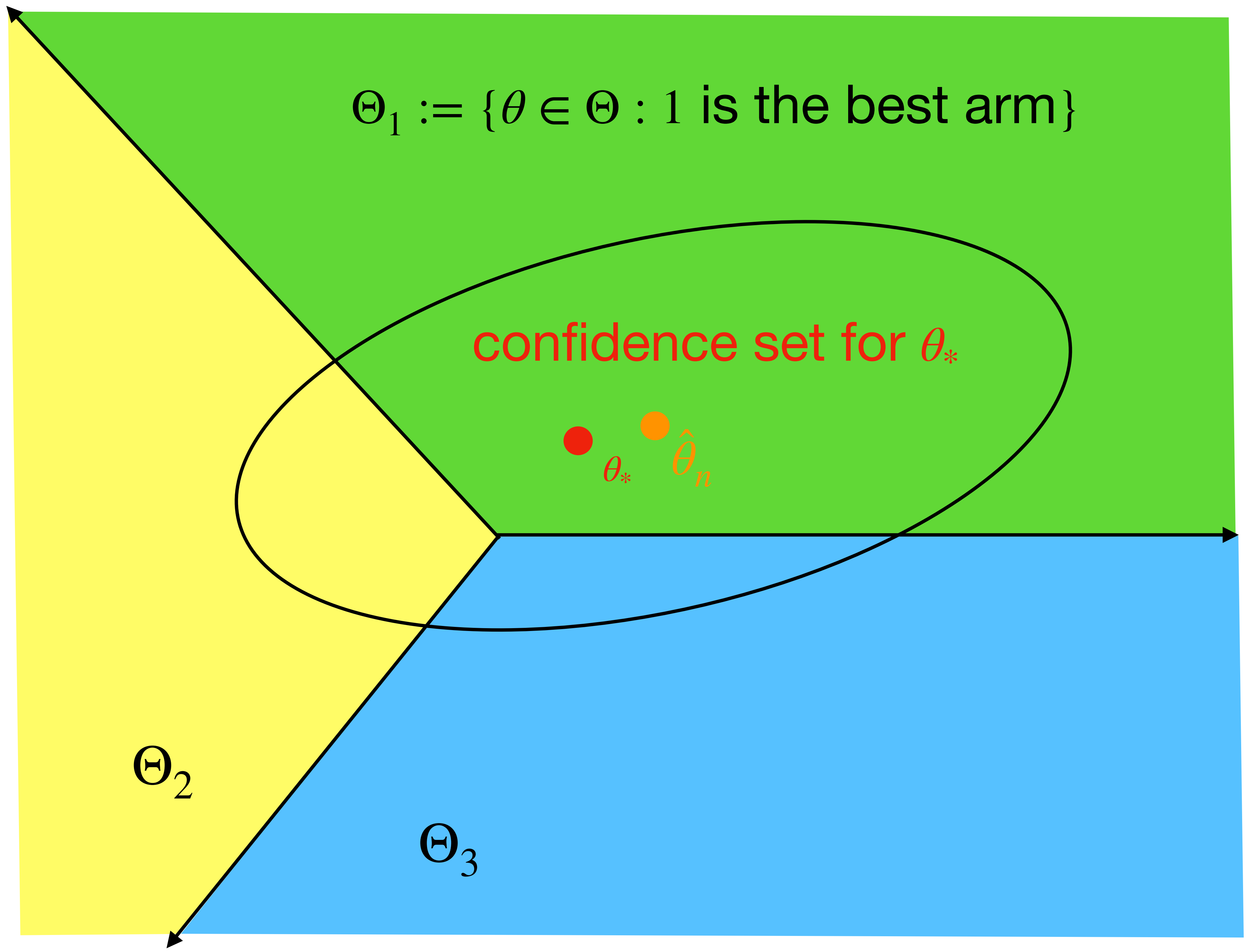
A Lower Bound in Linear Bandits

- Identify $a_* = \arg\max_{a \in \mathcal{A}} \langle x_a, \theta_* \rangle$



A Lower Bound in Linear Bandits

- Identify $a_* = \arg\max_{a \in \mathscr{A}} \left\langle x_a, \theta_* \right\rangle$



A Lower Bound in Linear Bandits

A Lower Bound in Linear Bandits

- Given dataset $\{(a_t, r_t)\}_{t=1}^{n}$, consider the least-squares estimate

$$\hat{\theta}_n = \left(\sum_{t=1}^{n} x_{a_t} x_{a_t}^\top\right)^{-1} \left(\sum_{t=1}^{n} x_{a_t} r_t\right),$$

A Lower Bound in Linear Bandits

- Given dataset $\{(a_t, r_t)\}_{t=1}^{n}$, consider the least-squares estimate

$$\hat{\theta}_n = \Bigg(\underbrace{\sum_{t=1}^{n} x_{a_t} x_{a_t}^\top}_{A_n}\Bigg)^{-1} \Bigg(\sum_{t=1}^{n} x_{a_t} r_t\Bigg),$$

A Lower Bound in Linear Bandits

- Given dataset $\{(a_t, r_t)\}_{t=1}^{n}$, consider the least-squares estimate

$$\hat{\theta}_n = \Bigg(\underbrace{\sum_{t=1}^{n} x_{a_t} x_{a_t}^{\top}}_{A_n}\Bigg)^{-1} \Bigg(\sum_{t=1}^{n} x_{a_t} r_t\Bigg),$$

A Lower Bound in Linear Bandits

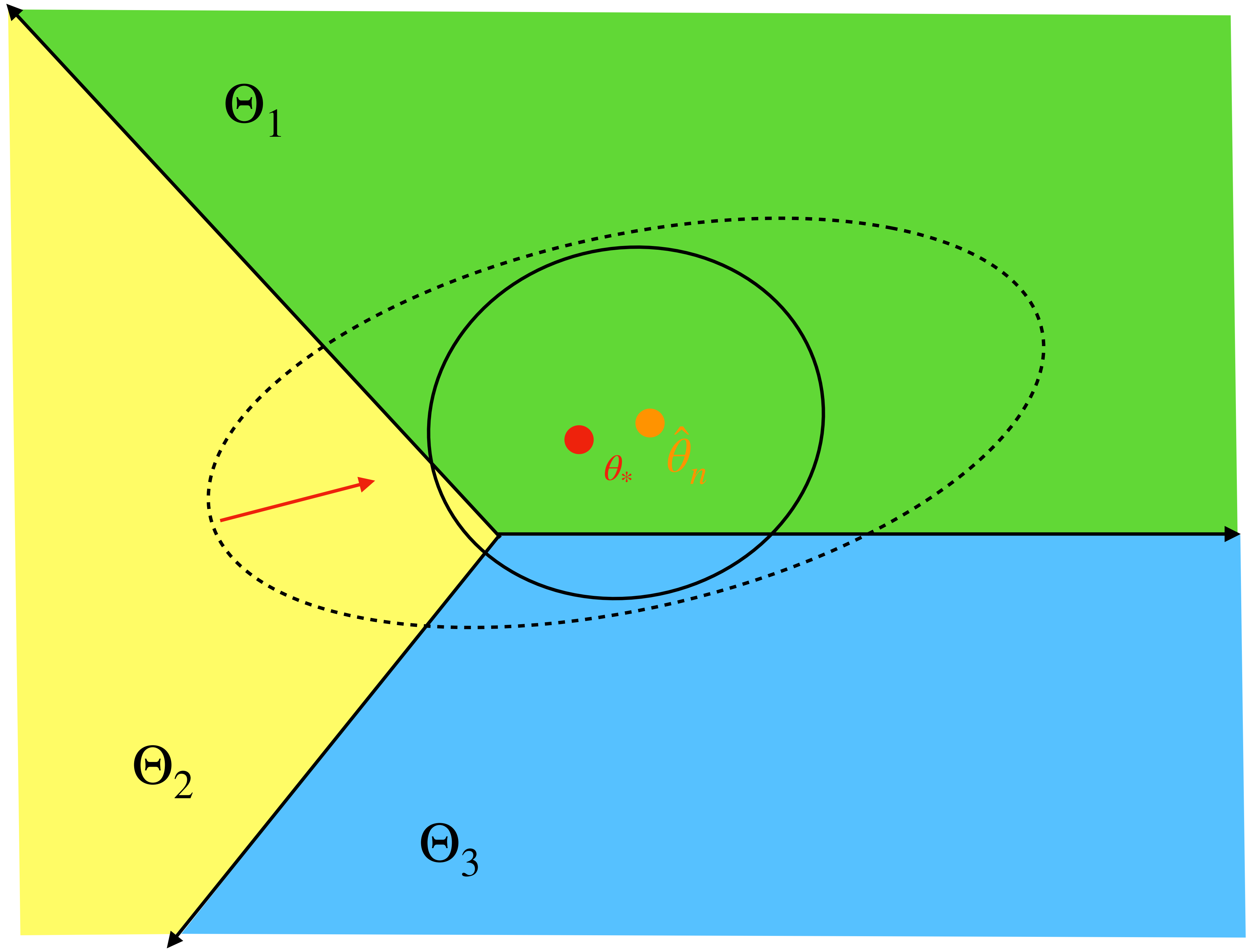
- Given dataset $\{(a_t, r_t)\}_{t=1}^{n}$, consider the least-squares estimate

$$\hat{\theta}_n = \Bigg(\underbrace{\sum_{t=1}^{n} x_{a_t} x_{a_t}^{\top}}_{A_n}\Bigg)^{-1} \Bigg(\sum_{t=1}^{n} x_{a_t} r_t\Bigg),$$

- Can get $|x^{\top}(\theta_* - \hat{\theta}_n)| \leq c\|x\|_{A_n^{-1}}\sqrt{\log(|\mathsf{A}|/\delta)}$ with probability at least $1 - \delta$

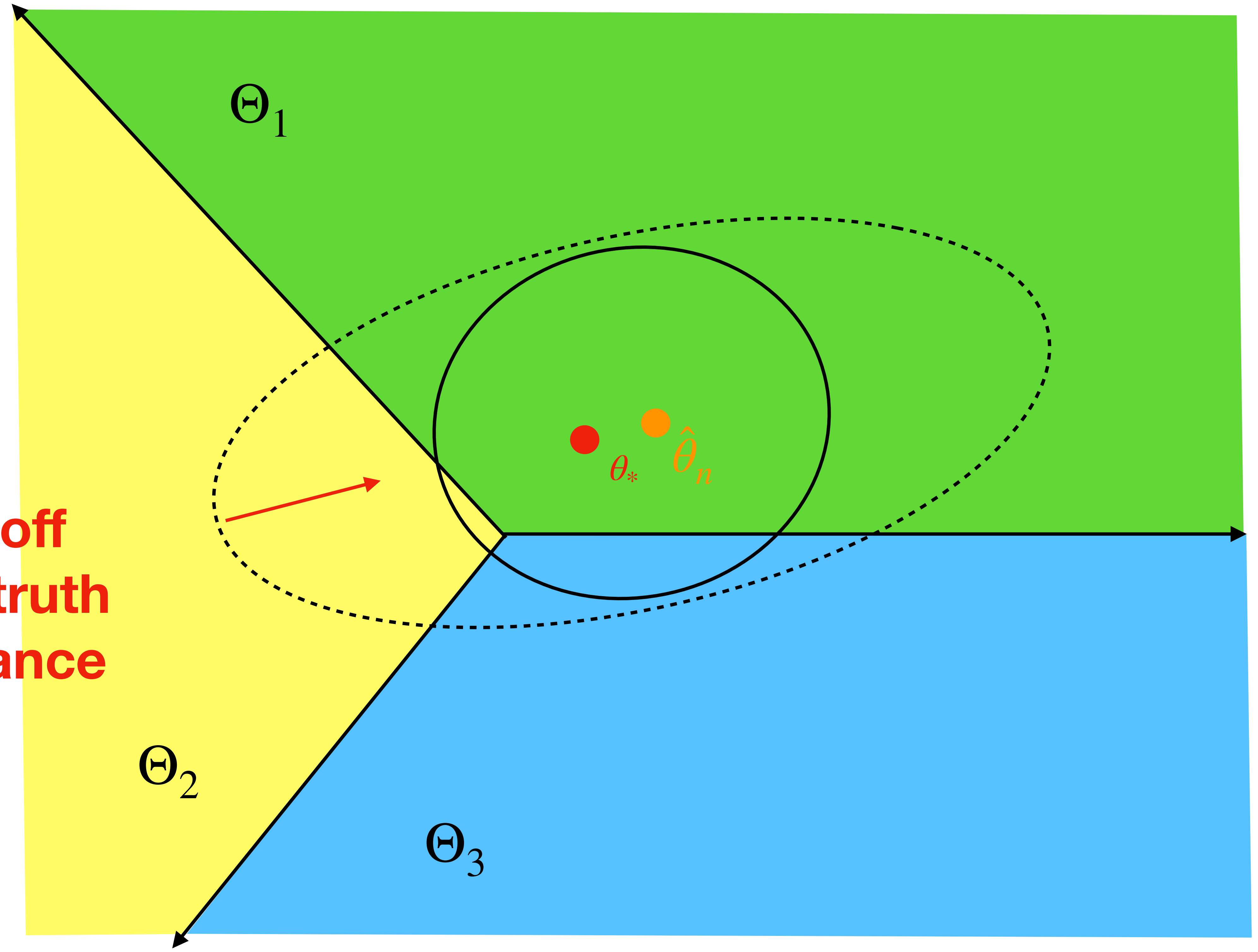
A Lower Bound in Linear Bandits

- $|x^\top(\theta_* - \hat{\theta}_n)| \leq c\|x\|_{A_n^{-1}}\sqrt{\log(|\mathsf{A}|/\delta)}$



A Lower Bound in Linear Bandits

- $|x^\top(\theta_* - \hat{\theta}_n)| \leq c\|x\|_{A_n^{-1}}\sqrt{\log(|\mathsf{A}|/\delta)}$



This direction provides a tradeoff between the deviation from the truth and the uncertainty, i.e. the variance

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: ϕ : C × A $\rightarrow \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset D $= \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta^*] = \sum_c \nu_c \sum_a p_{c,a} \phi(c, a)\phi(c, a)^\top \theta^*$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c,a) = \langle \phi(c,a), \theta* \rangle$ for $\theta* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c,a)\phi(c,a)^\top \theta*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top}}_{A(p)} \theta*$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^{n}$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t) r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta^*] = \underbrace{\sum_c \nu_c \sum_a p_{c,a} \phi(c, a)\phi(c, a)^\top}_{A(p)} \theta^*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} A(p)^{-1} \sum_{t=1}^{n} \phi(c_t, a_t) r_t$$

Towards Lower Bound: Estimators

- Linear contextual bandit setting (agnostic setting could be reduced to linear setting):
  - feature map: $\phi : \mathsf{C} \times \mathsf{A} \to \mathbb{R}^d$ such that $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Given dataset $\mathsf{D} = \{(c_t, a_t, r_t)\}_{t=1}^n$ where $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

$$\mathbb{E}[\phi(c_t, a_t)r_t] = \mathbb{E}_{c,a}[\phi(c, a)\phi(c, a)^\top \theta^*] = \underbrace{\boxed{\sum_c \nu_c \sum_a p_{c,a} \phi(c, a)\phi(c, a)^\top} }_{A(p)}\theta^*$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} A(p)^{-1} \sum_{t=1}^{n} \phi(c_t, a_t) r_t$$

IPW estimate!

A Lower Bound for Contextual Bandits

A Lower Bound for Contextual Bandits

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$

A Lower Bound for Contextual Bandits

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$
- Let $\phi_\pi := \mathbb{E}_{c \sim \nu}[\phi(c, \pi(c))]$, an estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_\pi)^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_\pi) = \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{n}$$

A Lower Bound for Contextual Bandits

- For each $\pi \in \Pi$, define the gap $\Delta(\pi) := V(\pi_*) - V(\pi)$

- Let $\phi_\pi := \mathbb{E}_{c\sim\nu}[\phi(c, \pi(c))]$, an estimate $\hat{\Delta}(\pi) = \hat{V}(\pi_*) - \hat{V}(\pi) = \left\langle \phi_{\pi_*} - \phi_\pi, \hat{\theta} \right\rangle$

$$Var(\hat{\Delta}(\pi)) = (\phi_{\pi_*} - \phi_\pi)^\top Var(\hat{\theta})(\phi_{\pi_*} - \phi_\pi) = \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{n}$$

Theorem [Li et al. 2022] Let τ be the stopping time of the algorithm. Any $(0,\delta)$-PAC algorithm satisfies $\tau \geq \rho_{\Pi,0} \log(1/2.4\delta)$ with high probability where

variance

$$\rho_{\Pi,0} = \min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \; \max_{\pi \in \Pi \setminus \pi_*} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(p)^{-1}}}{\Delta(\pi)^2}.$$

gap