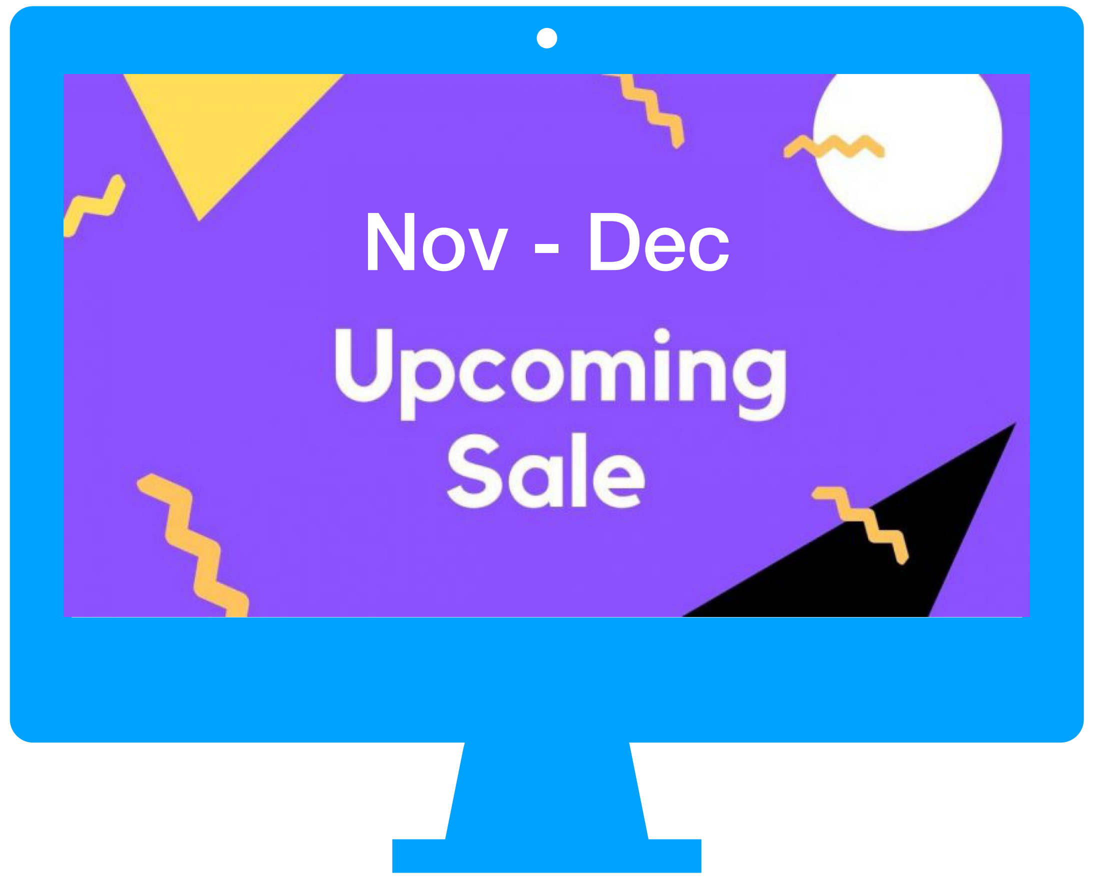
Instance-Optimal PAC Contextual Bandits

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

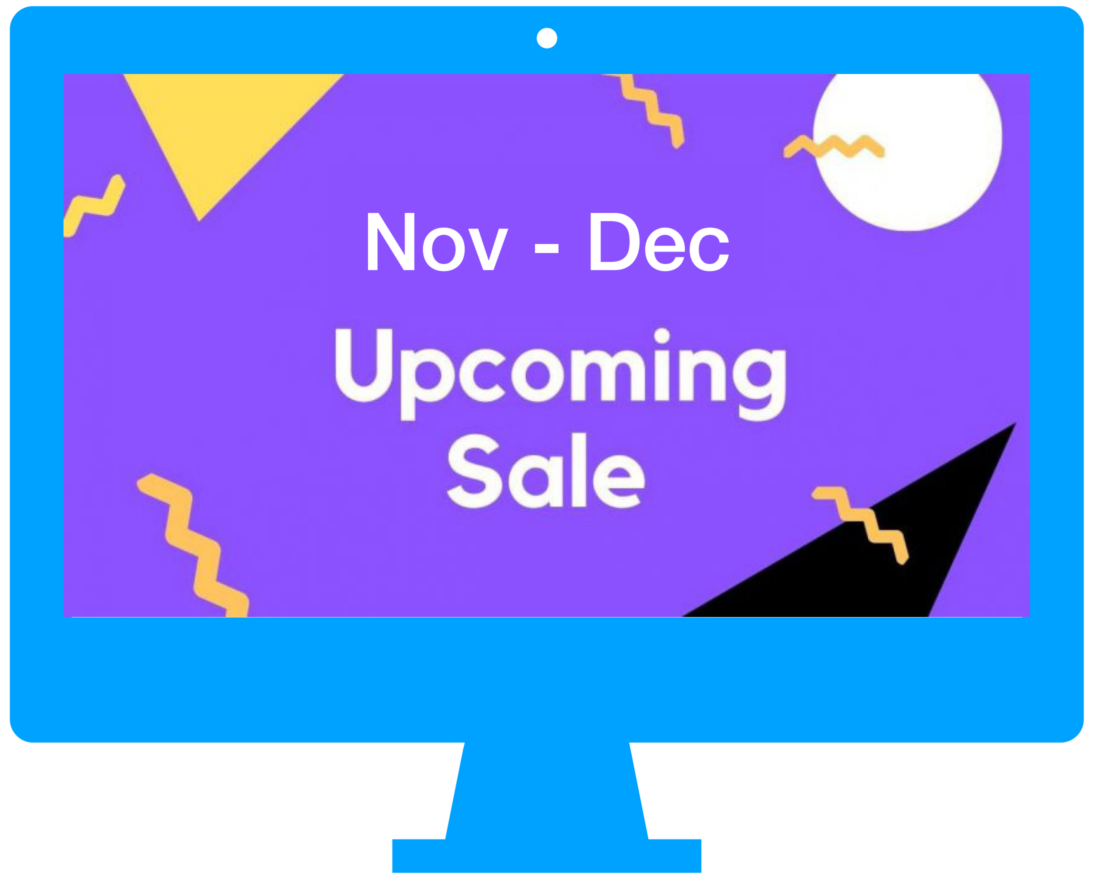
*University of Washington

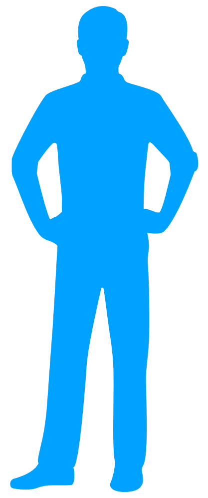
†Amazon

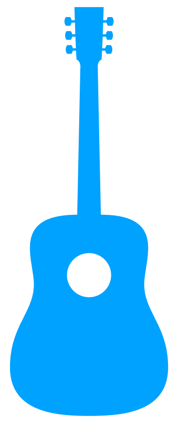
Motivation



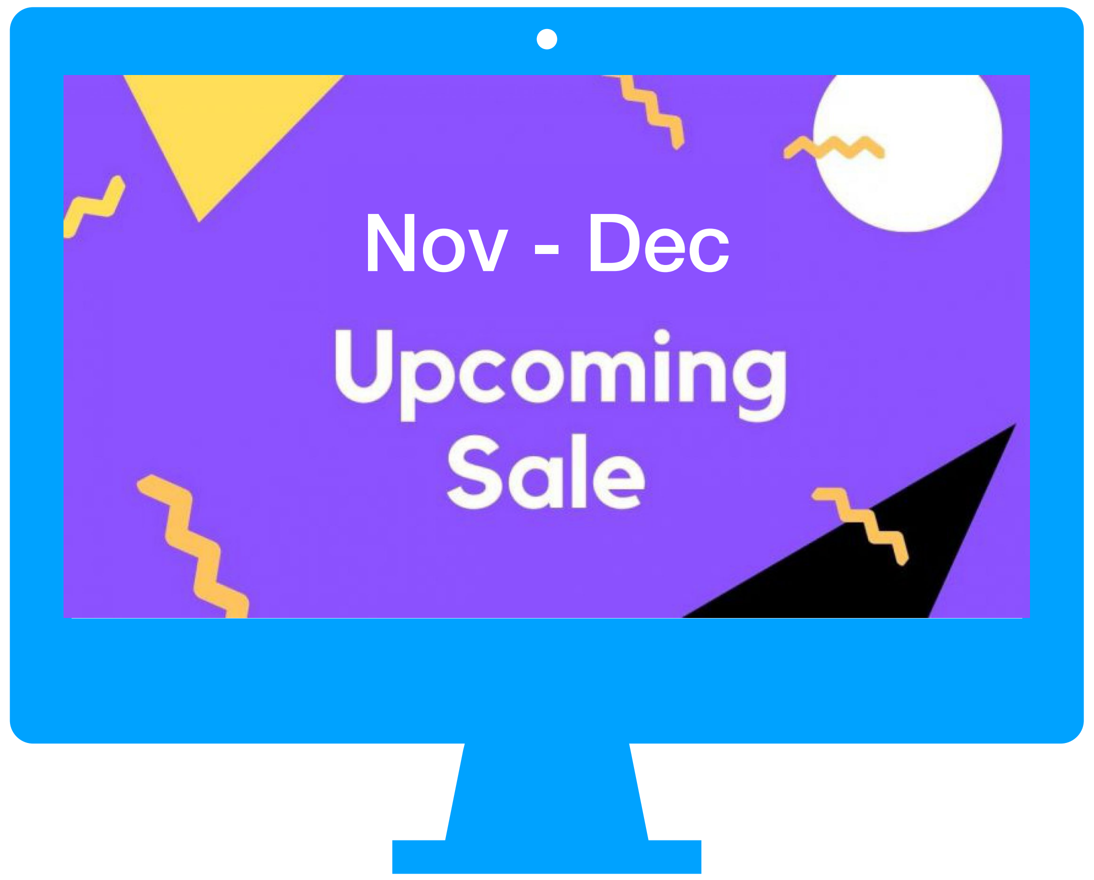
Motivation

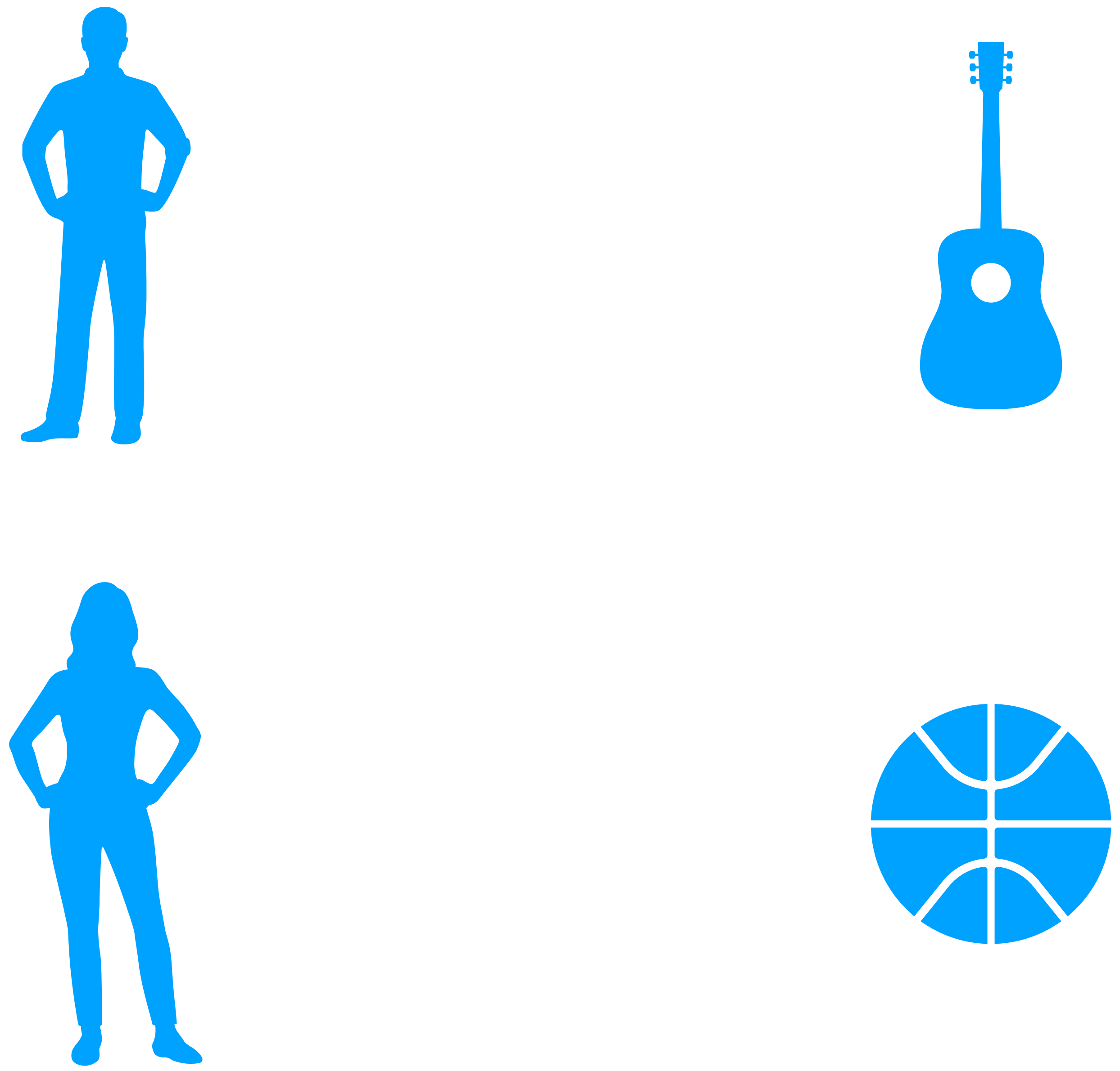




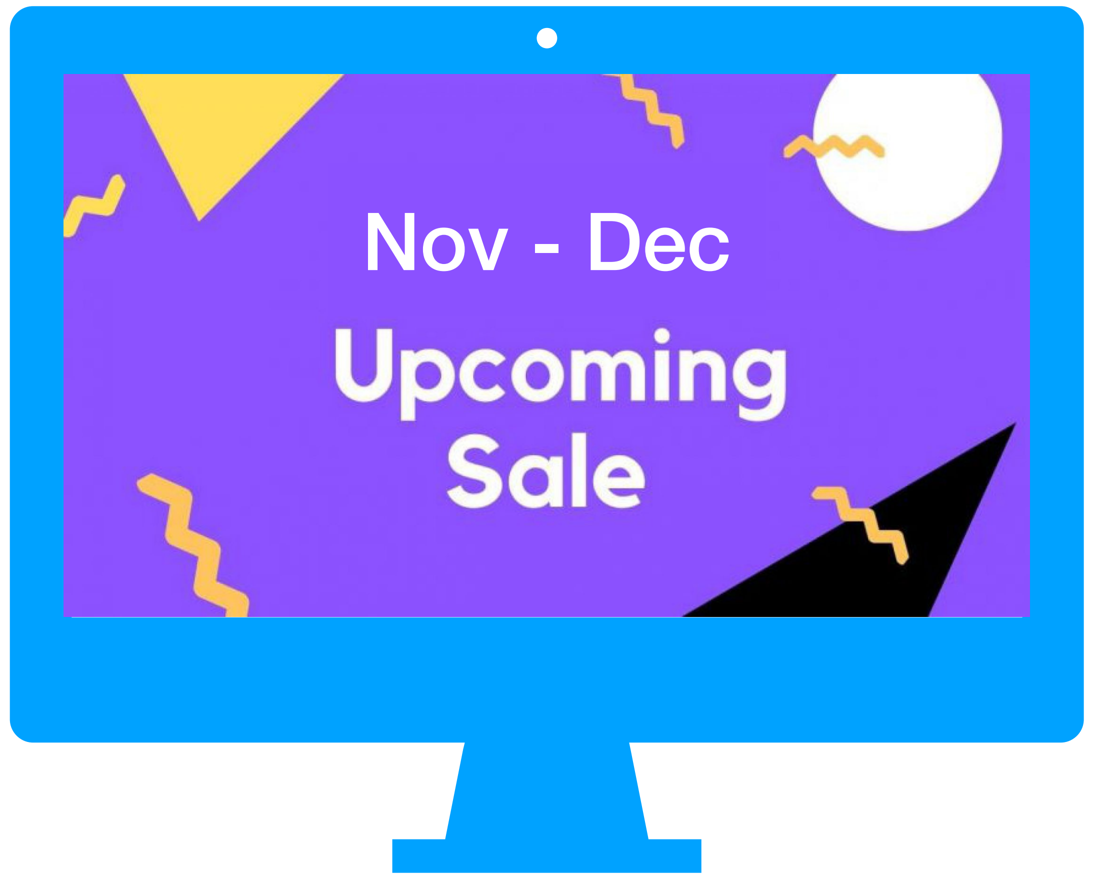


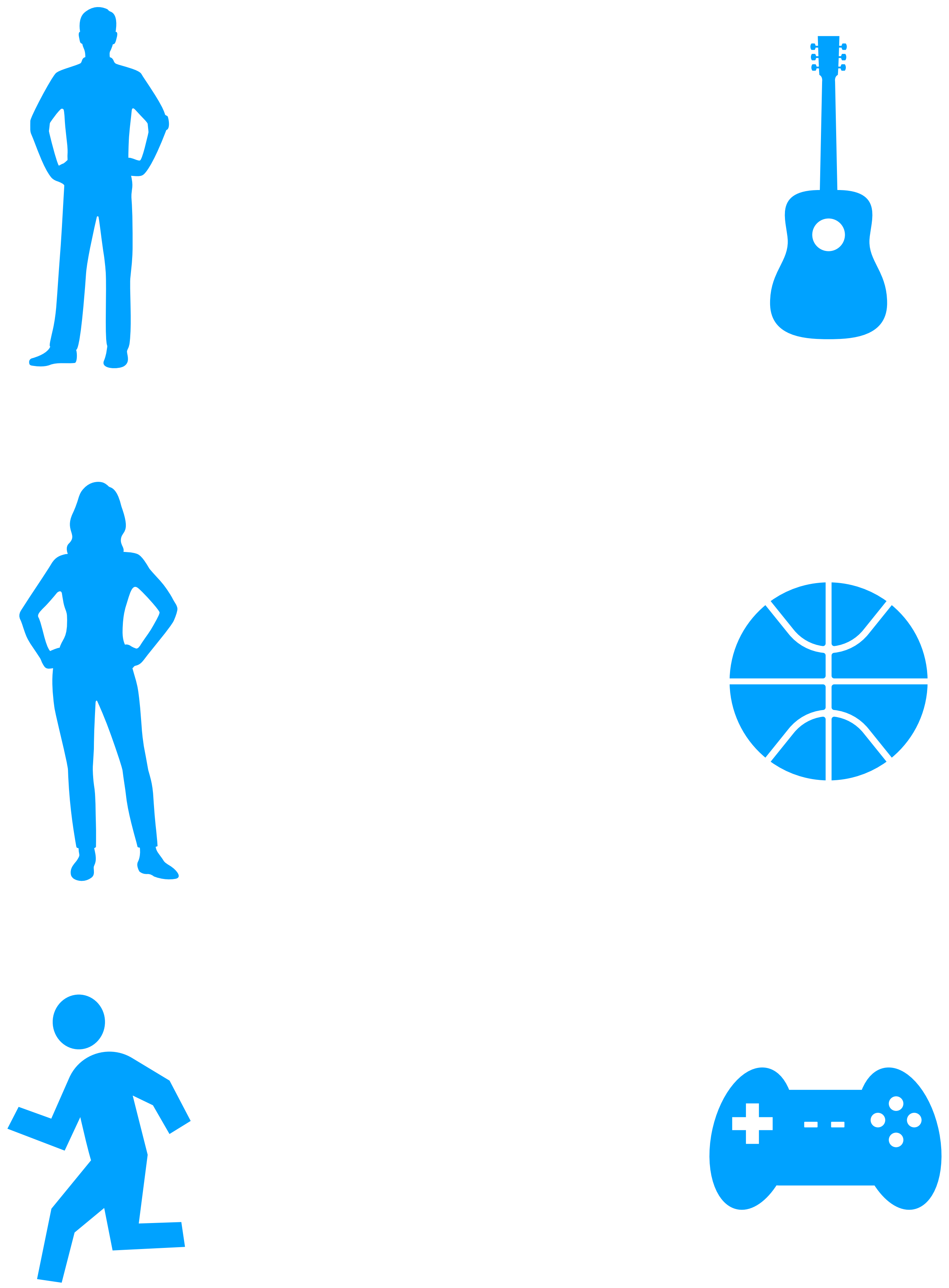
Motivation



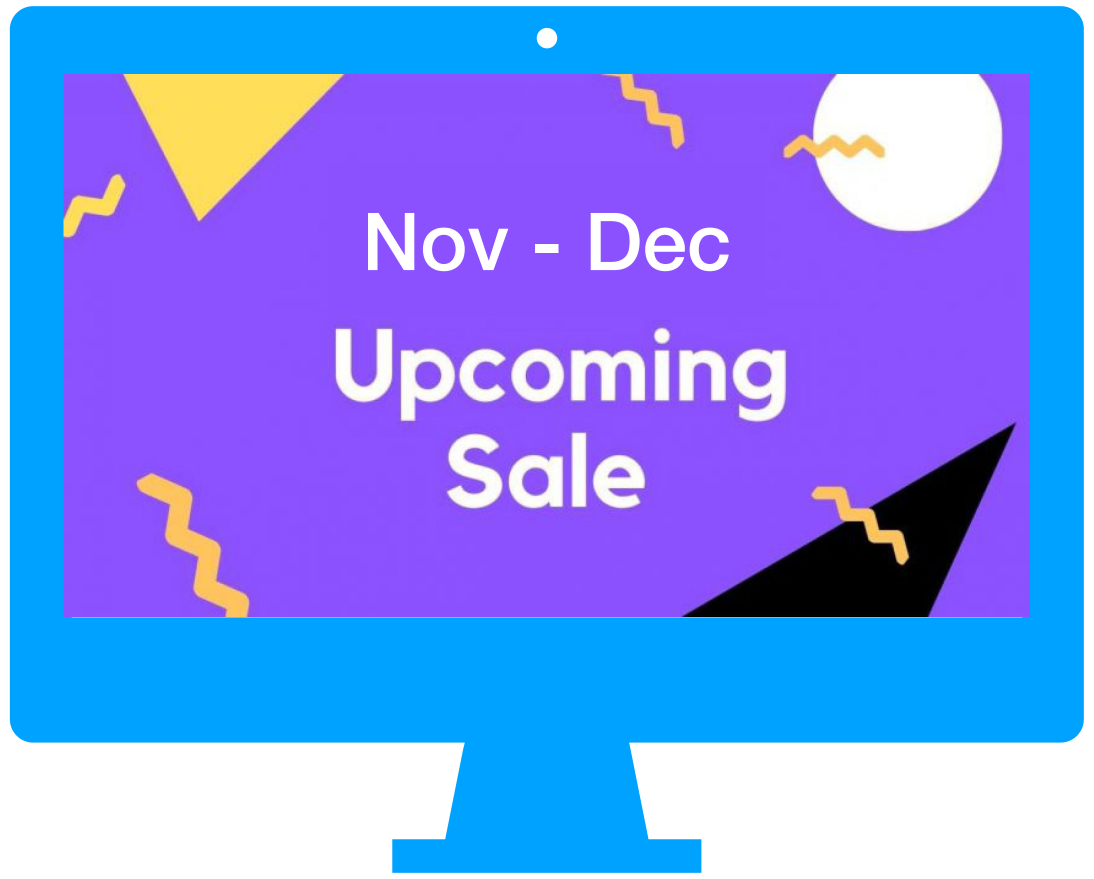


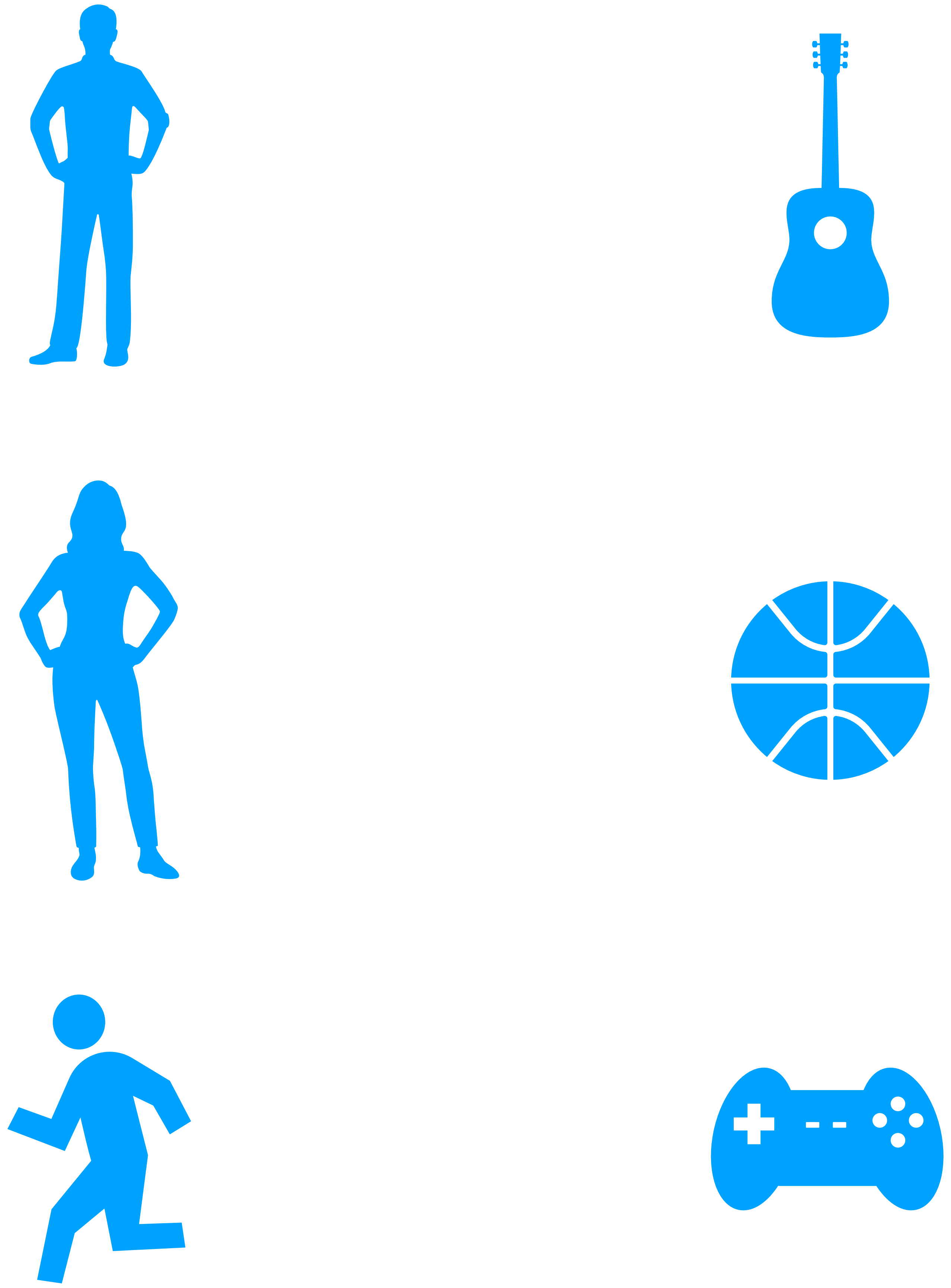
Motivation





Motivation



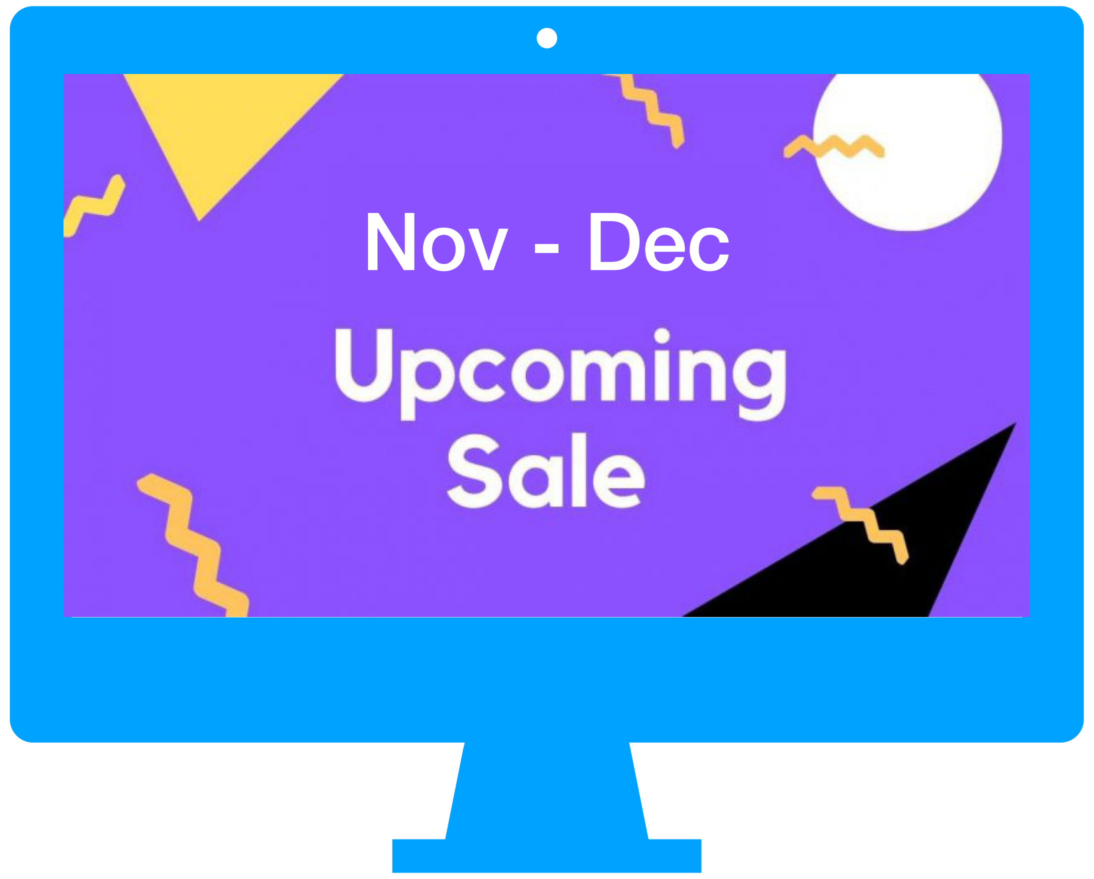


Question: What is the best way to give personalized recommendations?

Motivation

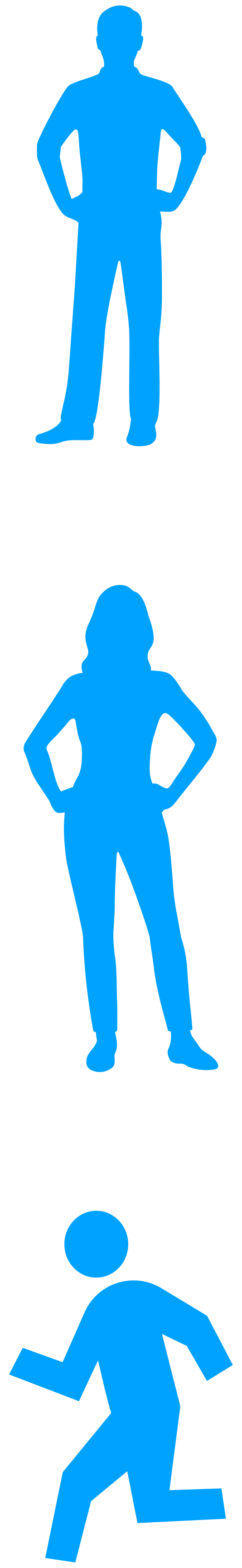
policy

π



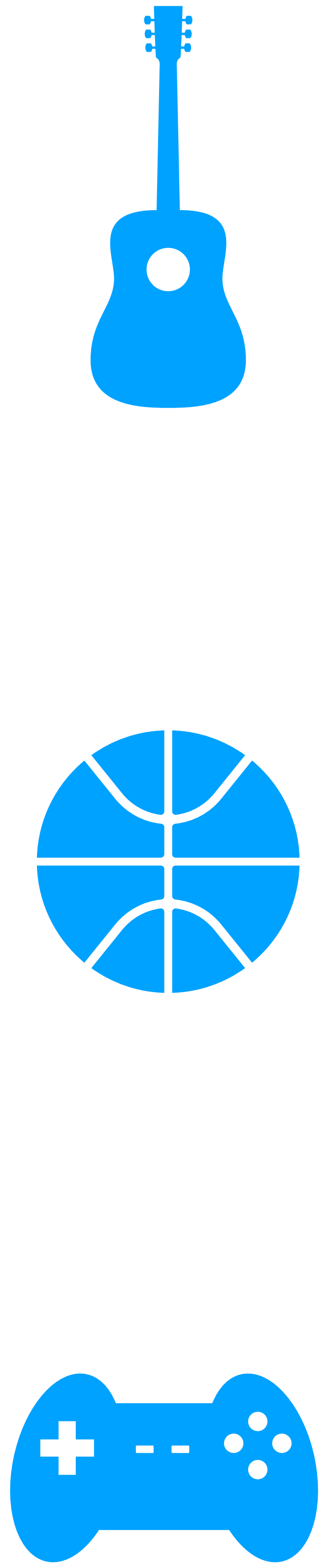
context

c



action

a



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

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 $\pi \in \Pi, \pi : \mathrm{C} \to \mathrm{A}$
- Average reward: $V(\pi) := \mathbb{E}_{c \sim \nu}[r(c, \pi(c))]$
- Optimal policy: $\pi_\star := \arg\max_{\pi \in \Pi} V(\pi)$

<u>$(\epsilon, \delta) -$ **PAC Guarantee**</u>

Return $\hat{\pi}$ satisfying, $V(\hat{\pi}) \geq V(\pi_*) - \varepsilon$ 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:

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

Regret Minimization vs. Policy Identification

- Regret heavily studied:

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

- ILOVETOCONBANDITS [Agarwal et al. 2014] achieves $R_T = O(\sqrt{|\mathsf{A}|T\log(\Pi)})$, computationally efficient

Regret Minimization vs. Policy Identification

- Regret heavily studied:

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

- ILOVETOCONBANDITS [Agarwal et al. 2014] achieves $R_T = O(\sqrt{|\mathsf{A}| T \log(\Pi)})$, computationally efficient

- Modification gives (ϵ, δ)- PAC algorithm w/ sample complexity $O(|\mathsf{A}| \log(\Pi/\delta)/\epsilon^2)$, also see [Zanette et al. 2021]

Regret Minimization vs. Policy Identification

- Regret heavily studied:

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

- ILOVETOCONBANDITS [Agarwal et al. 2014] achieves $R_T = O(\sqrt{|\mathsf{A}| T \log(\Pi)})$, computationally efficient

- Modification gives (ϵ, δ)- PAC algorithm w/ sample complexity $O(|\mathsf{A}| \log(\Pi/\delta)/\epsilon^2)$, also see [Zanette et al. 2021]

<u>Two Problems</u>

a) Minimax Result! Does not adapt to hardness of instance.

Regret Minimization vs. Policy Identification

- Regret heavily studied:

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

- ILOVETOCONBANDITS [Agarwal et al. 2014] achieves $R_T = O(\sqrt{|A| T \log(\Pi)})$, computationally efficient

- Modification gives (ϵ, δ)- PAC algorithm w/ sample complexity $O(|A| \log(\Pi/\delta)/\epsilon^2)$, also see [Zanette et al. 2021]

True for any policy class! Not capturing difficulty for learning π_*

<u>Two Problems</u>

a) Minimax Result! Does not adapt to hardness of instance.

Regret Minimization vs. Policy Identification

- Regret heavily studied:

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

- ILOVETOCONBANDITS [Agarwal et al. 2014] achieves $R_T = O(\sqrt{|A| T \log(\Pi)})$, computationally efficient

- Modification gives (ϵ, δ)- PAC algorithm w/ sample complexity $O(|A| \log(\Pi/\delta)/\epsilon^2)$, also see [Zanette et al. 2021]

True for any policy class! Not capturing difficulty for learning π_*

<u>Two Problems</u>

a) Minimax Result! Does not adapt to hardness of instance.

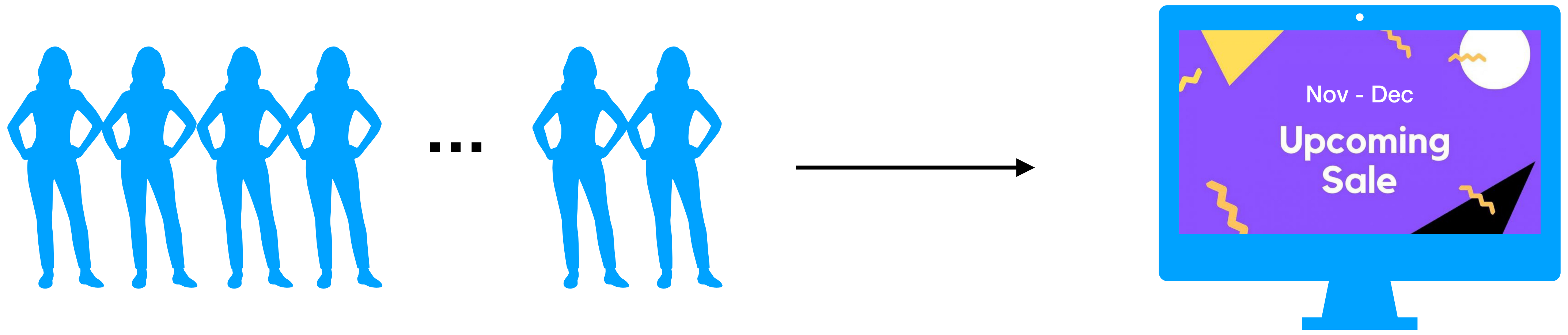
b) Can construct an example, where any optimal regret algorithm won't be instance optimal!

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 $\mathcal{C}$ and policy space Π could be **infinite**!

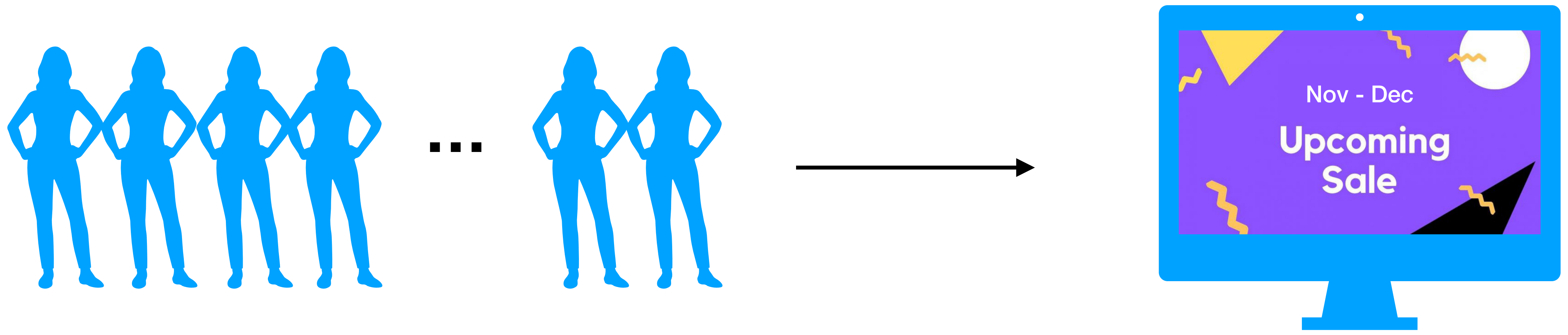
Challenges

- 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 $\mathcal{C}$ and policy space Π could be **infinite**!



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 $\mathcal{C}$ and policy space Π could be **infinite**!



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

- 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 $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

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

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 $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

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

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 $a_t \sim p_{c_t} \in \triangle_{\mathsf{A}}$,

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

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

IPS estimate!

A Lower Bound

A Lower Bound

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

A Lower Bound

- 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 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

Our algorithm

Our algorithm

Our algorithm

Input: Π

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1, 2, \cdots$

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}$, $\forall c \in \mathsf{C}$ and n_l such that

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}$, $\forall c \in \mathsf{C}$ and n_l such that

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}$, $\forall c \in \mathsf{C}$ and n_l such that

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and n_l such that

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPS estimate $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}$ and n_l such that

$$\min_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPS estimate $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
3. Update

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and n_l such that

$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPS estimate $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$

3. Update

$$\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}(\pi, \hat{\pi}_{l-1})$$

Our algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}$ and n_l such that

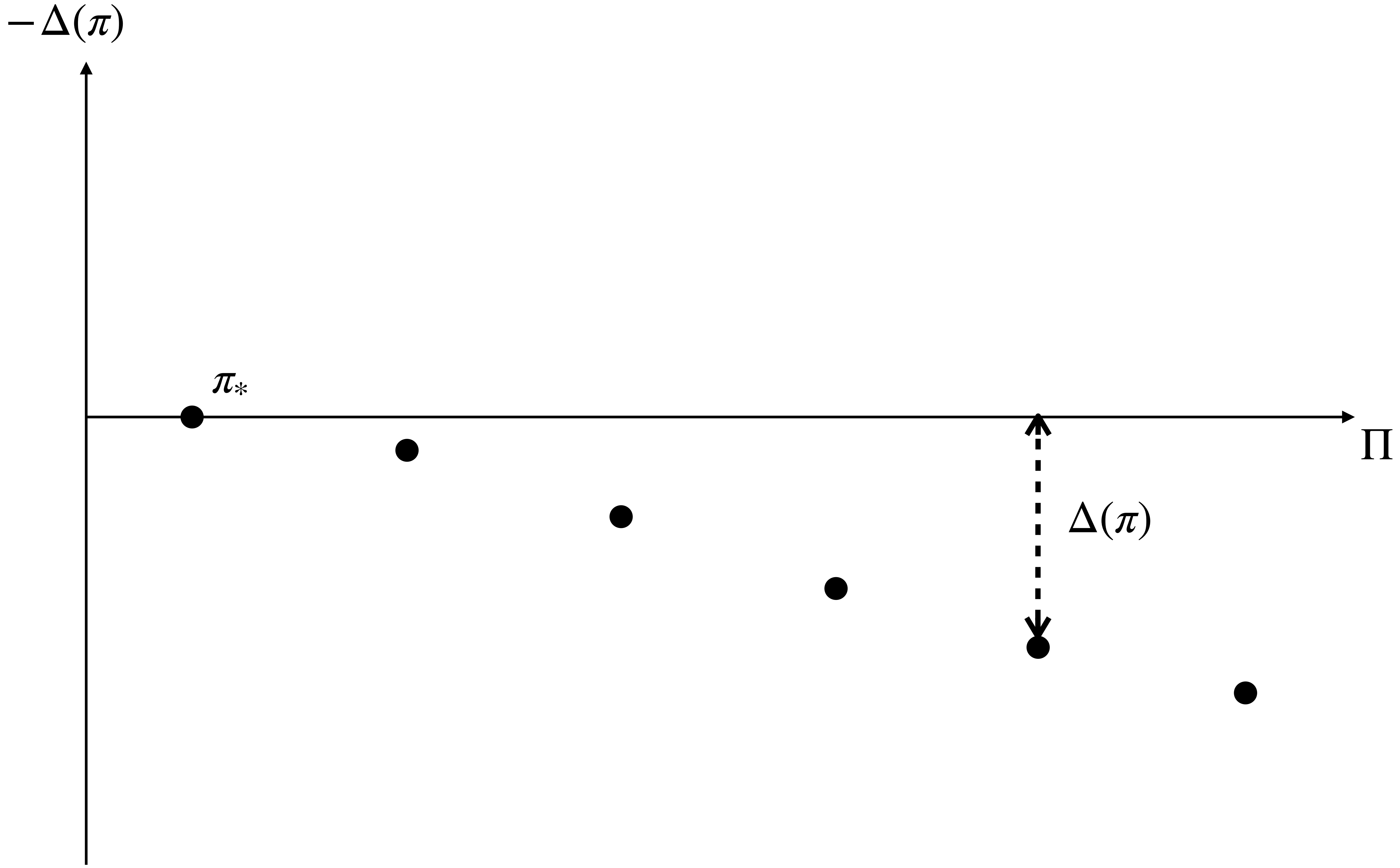
$$\min_{p_c \in \triangle_{\mathsf{A}}, \forall c \in \mathsf{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPS estimate $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$
3. Update

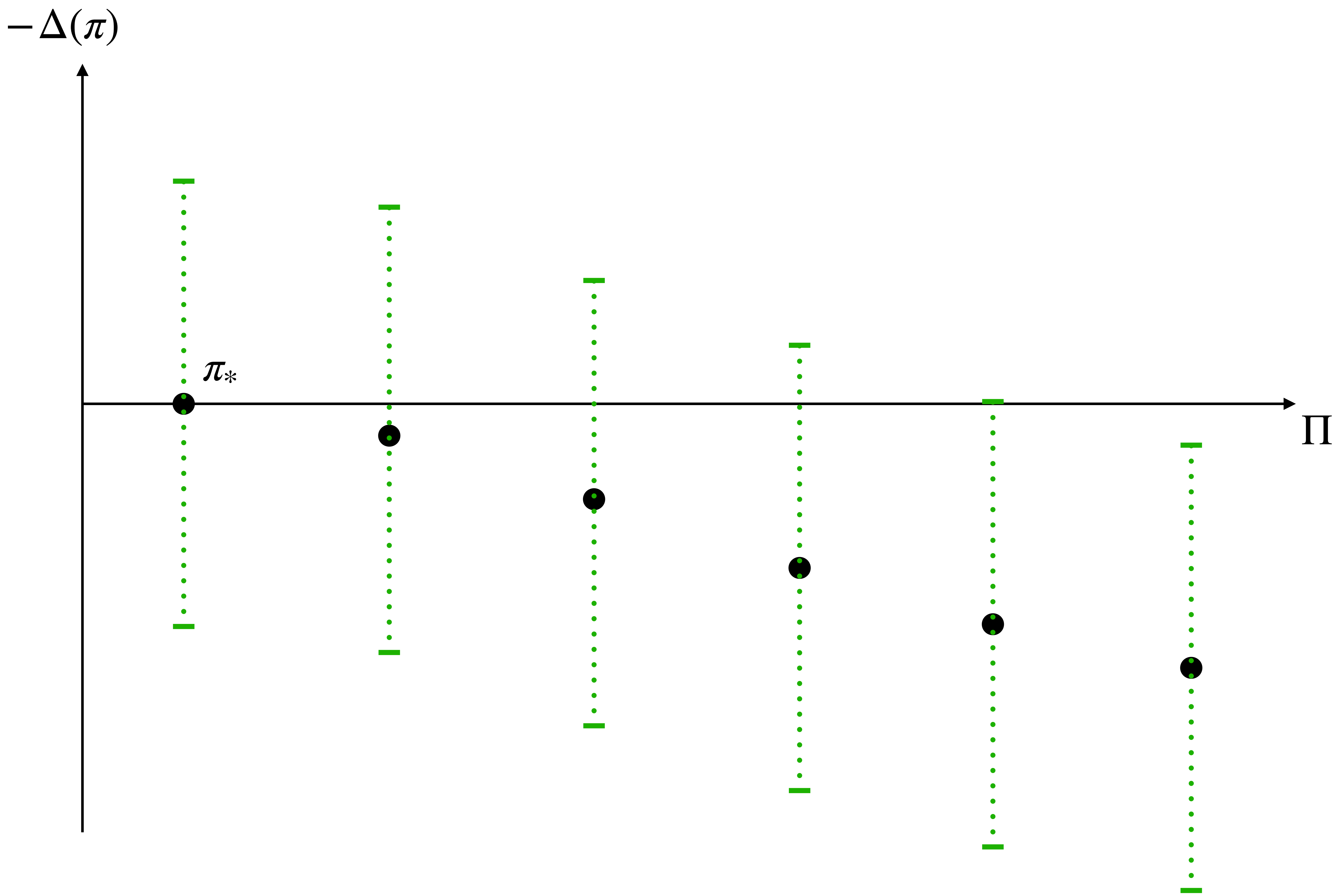
$$\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}(\pi, \hat{\pi}_{l-1})$$

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.

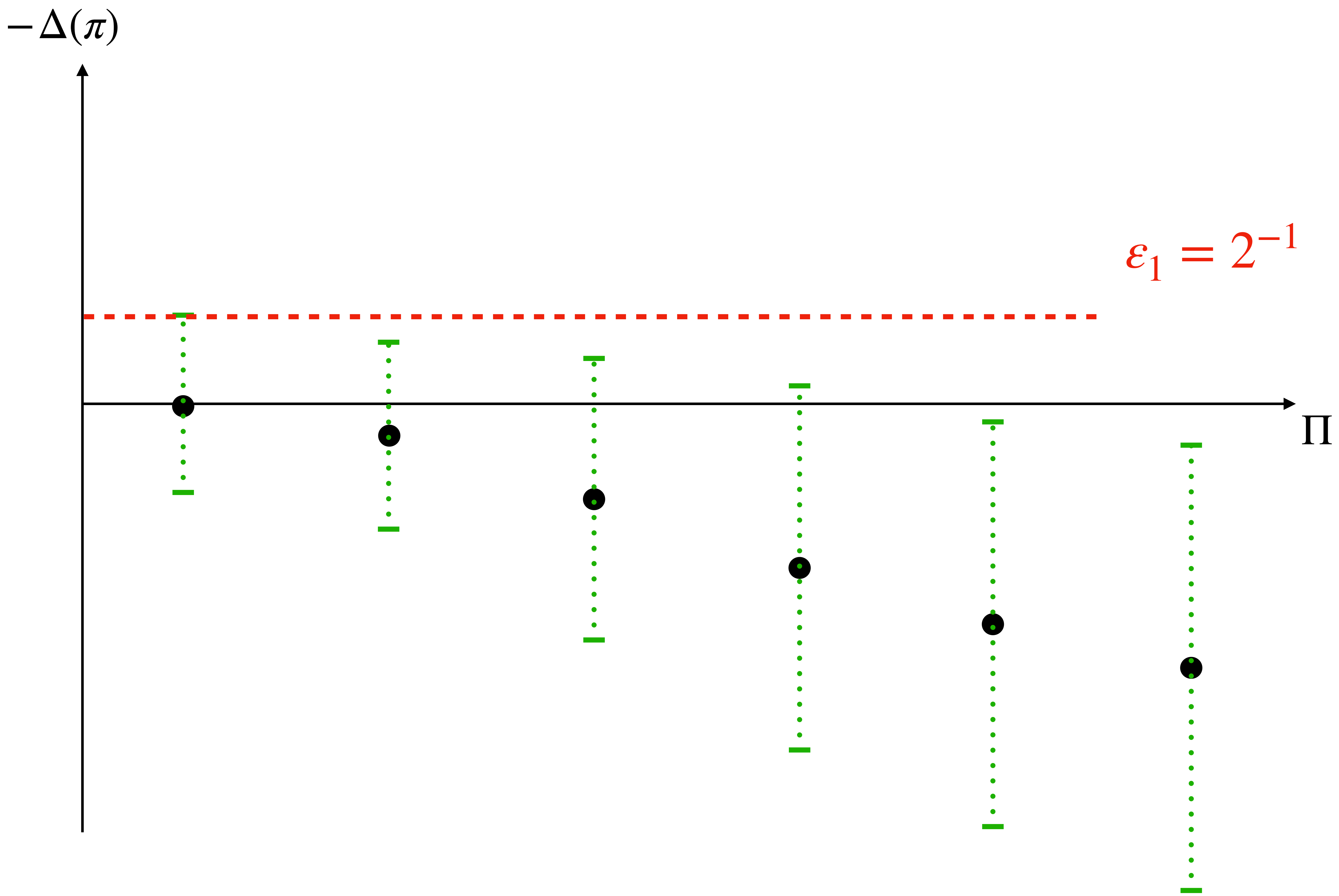
Our algorithm



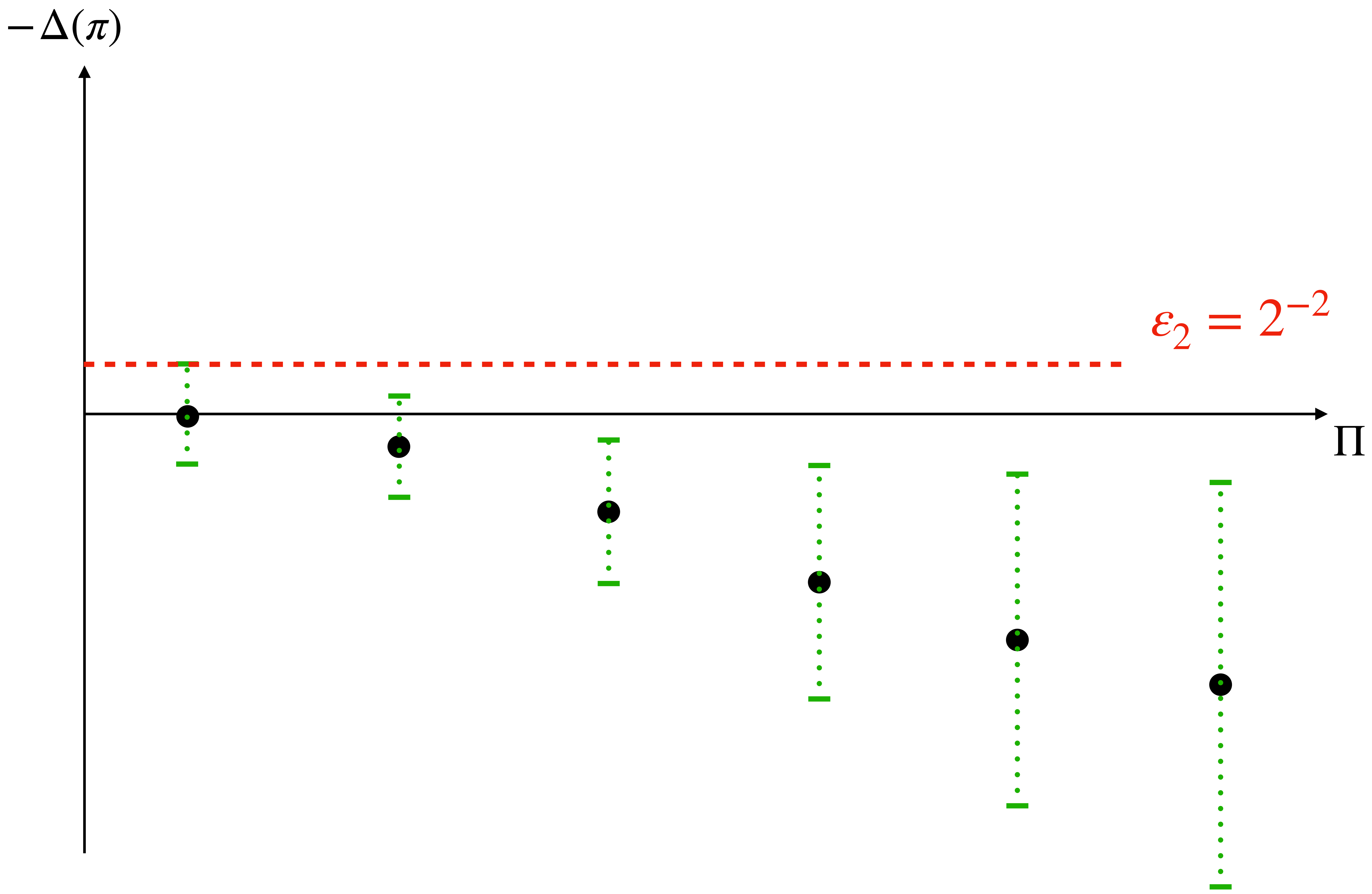
Our algorithm



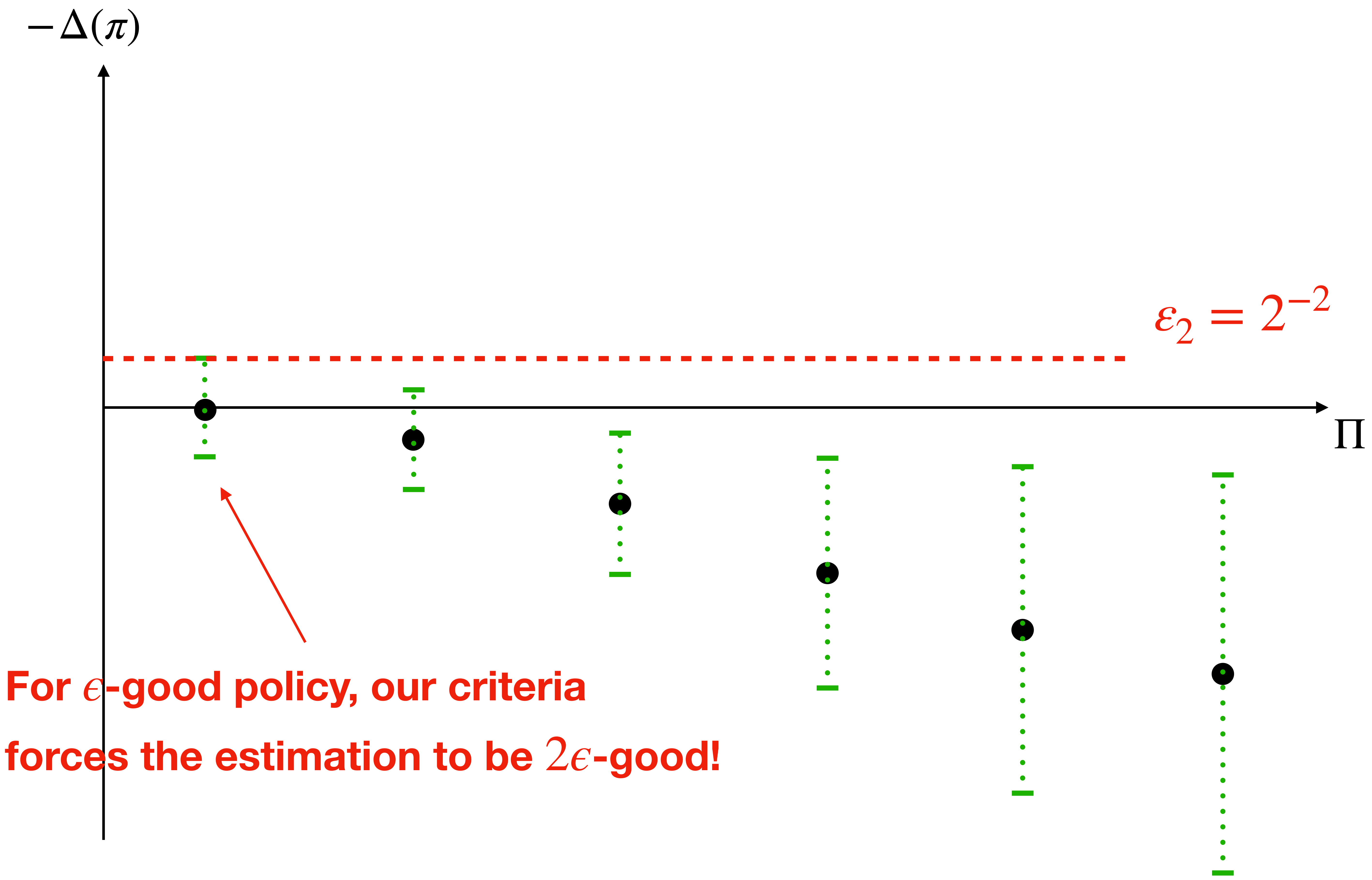
Our algorithm



Our algorithm

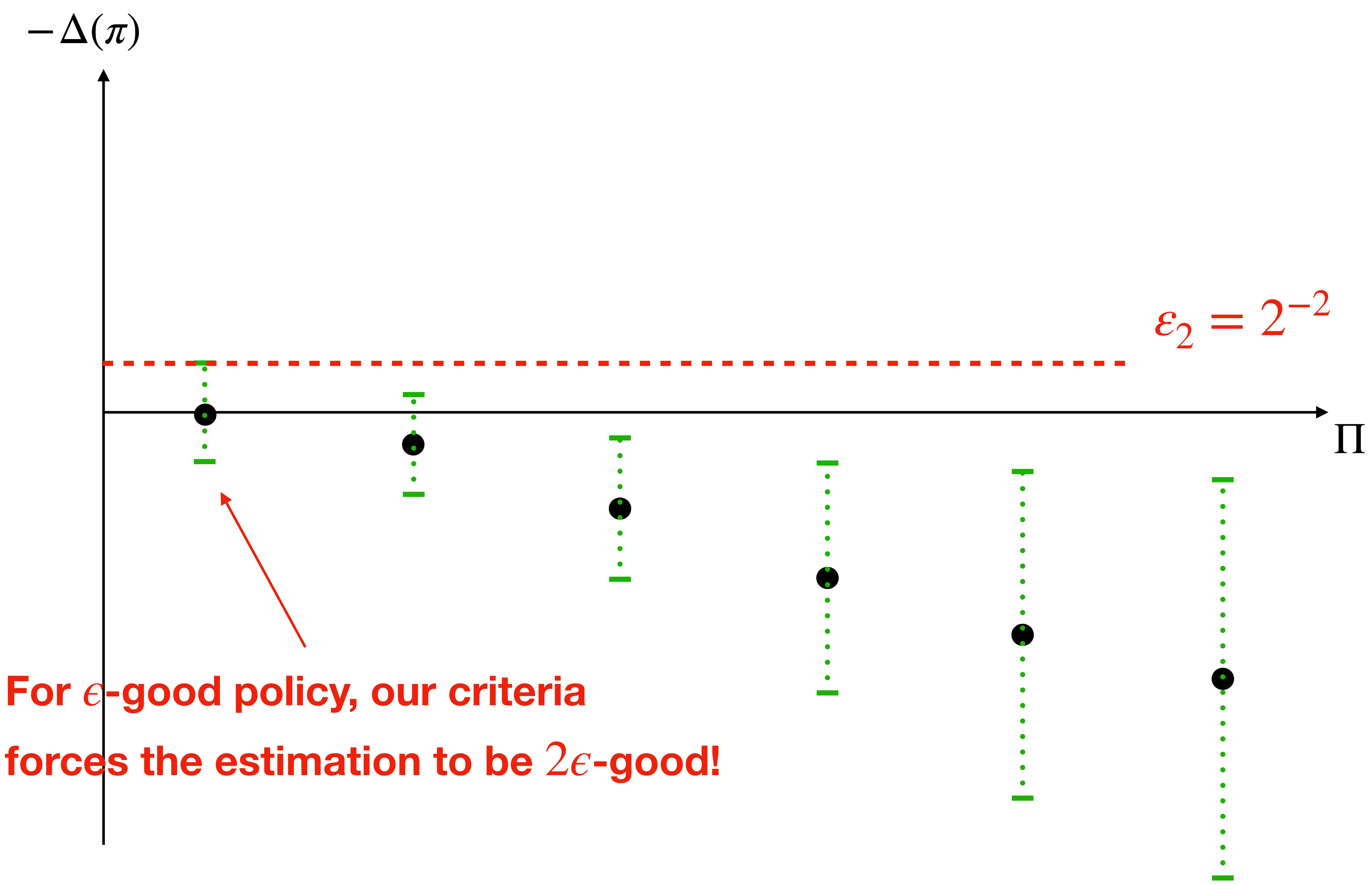


Our algorithm



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

Our algorithm



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

Returning the empirical best policy at the end $\Rightarrow$ at least 2ϵ-good

Towards an efficient algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)}$ and n_l such that

$$\min_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \le 2^{-l}$$

2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPS estimate $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$

3. Update

$$\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}(\pi, \hat{\pi}_{l-1})$$

Towards an efficient algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)}$ and n_l such that

$$\min_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPS estimate $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$

3. Update

$$\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}(\pi, \hat{\pi}_{l-1})$$

Towards an efficient algorithm

Input: Π

Initialize $\Pi_1 = \Pi$

for $l = 1,2,\cdots$

not efficient since cannot hold on to p_c for all c!

1. Choose $p_c^{(l)}$ and n_l such that

$$\min_{p_c \in \triangle_A, \forall c \in C} \max_{\pi \in \Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\hat{\pi}_{l-1}}\|^2_{A(p)^{-1}} \log(1/\delta)}{n_l}} \right) \leq 2^{-l}$$

2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPS estimate $\hat{\Delta}(\pi, \hat{\pi}_{l-1})$ for each $\pi \in \Pi$

3. Update

$$\hat{\pi}_l = \arg\min_{\pi \in \Pi} \hat{\Delta}(\pi, \hat{\pi}_{l-1})$$

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

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

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:

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

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

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

- 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:

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

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

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!

Frank-Wolfe

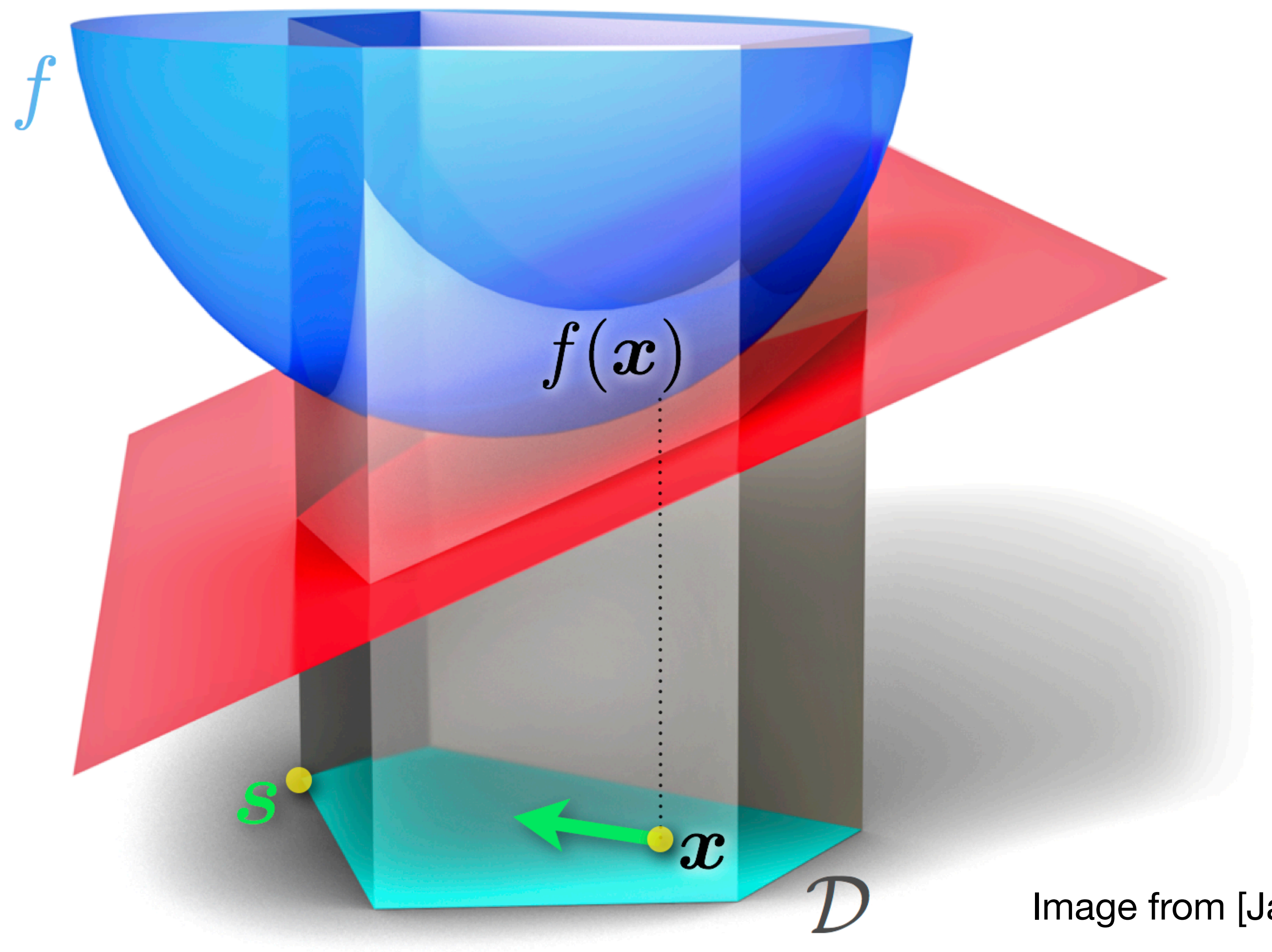


Image from [Jaggi]

- Gives us a sparse yet good enough solution λ
- Plug in solution λ in the closed-form gives us $p_c \in \triangle_A$

Thanks!

Towards an efficient algorithm

- ***argmax*** oracle: given $(c_1, s_1), \cdots, (c_n, s_n) \in \mathscr{C} \times \mathbb{R}^{|\mathscr{A}|}$, returns $\arg\max_{\pi \in \Pi} \sum_{t=1}^{n} s_t(\pi(c_t))$
- Can be computed using cost-sensitive classification

A Lower Bound

- Choose action distribution p such that:

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

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 $[\theta^*]_{c,a} = r(c,a)$

$$\begin{array}{cc} & a \\ c & \begin{pmatrix} & \vdots & \\ \cdots & r(c,a) & \\ & & \end{pmatrix} \end{array} \xrightarrow{\textbf{vectorize}} \theta^*$$

Agnostic Setting Reduces to Linear

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

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

Agnostic Setting Reduces to Linear

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

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

Variance

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

Gap