Universal Off-Policy Evaluation






Yash Chandak

Scott Niekum

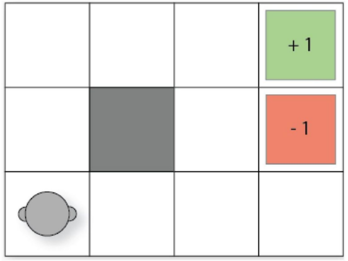
Bruno C. da Silva

Erik Learned-Miller

Emma Brunskill

Philip Thomas

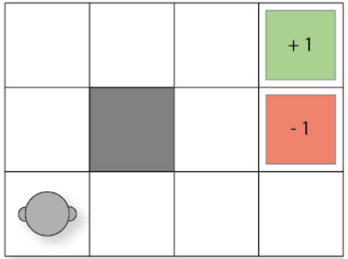
Motivating Examples



Games
(**Expected** performance)

Mean

Motivating Examples



Games
(**Expected**
performance)

Mean

Automated
healthcare
(Mitigate **risk**)

VaR, CVaR

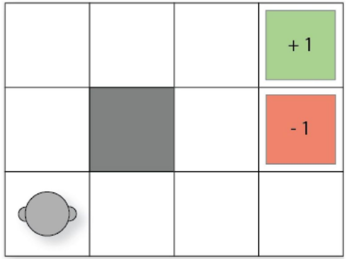
Mechanical control
(Mitigate
uncertainty)

Variance, Entropy

Online recommendation
(Robust to **noisy**
data-collection)

Median, Inter-quantiles

Motivating Examples



Games
(**Expected**
performance)

Mean

Markovian

Automated
healthcare
(Mitigate **risk**)

VaR, CVaR

Partial Observation

Mechanical control
(Mitigate
uncertainty)

Variance, Entropy

Non-Markovian

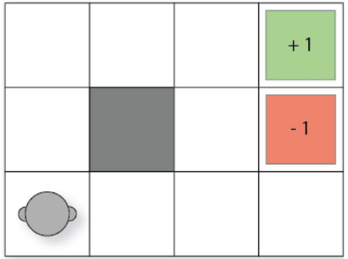
Online recommendation
(Robust to **noisy**
data-collection)

Median, Inter-quantiles

Non-stationary

Motivating Examples

Real-world problems are often high-stakes.
Evaluate a policy's performance ***before*** deployment (**off-policy**).



Games (**Expected** performance)

Mean

Markovian

Automated healthcare (Mitigate **risk**)

VaR, **CVaR**

Partial Observation

Mechanical control (Mitigate **uncertainty**)

Variance, **Entropy**

Non-Markovian

Online recommendation (Robust to **noisy** data-collection)

Median, **Inter-quantiles**

Non-stationary

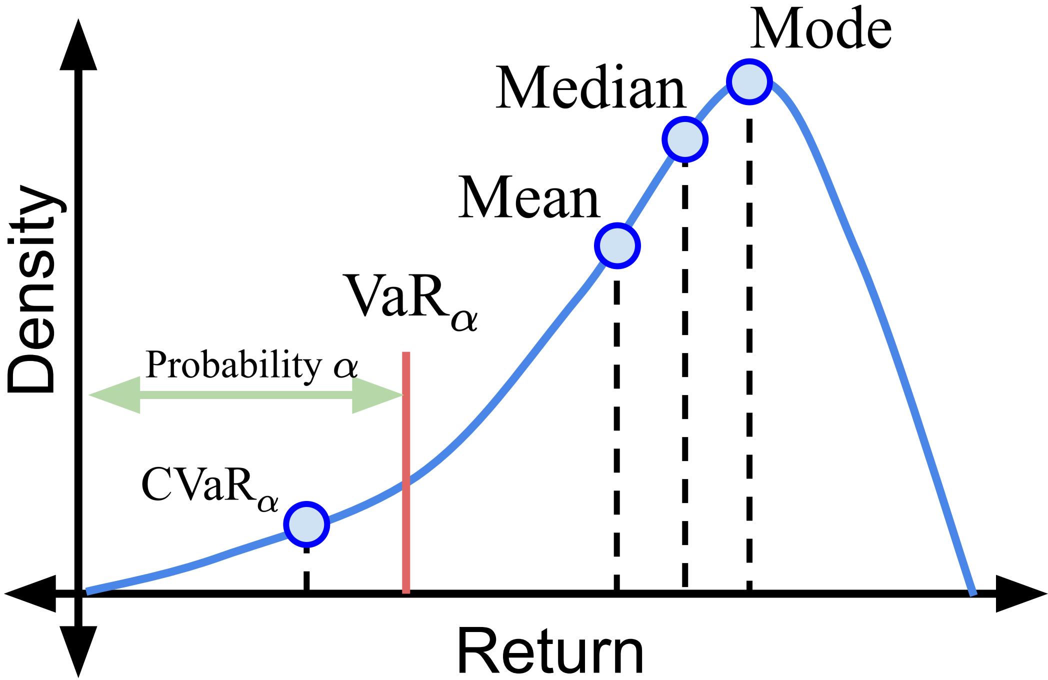
Goal

Given: Trajectories collected using one or multiple ***past*** (behavior) policies.

Goal

Given: Trajectories collected using one or multiple ***past*** (behavior) policies.

Aim: Evaluate and bound the desired performance metric (mean, variance, CVaR, etc.) of the return distribution under a ***new*** policy, using the given trajectories.

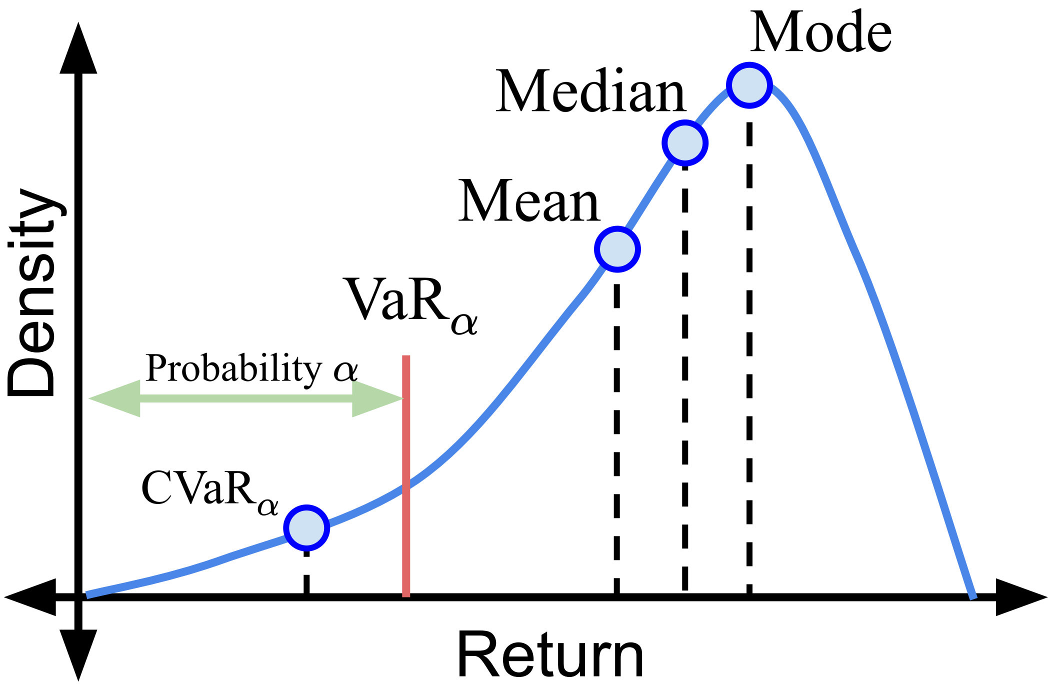


Goal

Assumptions

- Any outcome under the evaluation policy is possible under the behavior policy (**support assumption**)

- Knowledge of **action probabilities** under the behavior policy



Prior work

- **Model-based**
  - Additional requirement for estimating **reward distribution** for **each state-action pair**
  - Hard to estimate accurate models in **non-tabular** settings
- **Typical IS based estimators**
  - Only corrects for the **mean**
- **Distributional RL**
  - **On-policy**

A Universal Evaluation Procedure

- **Off-policy**
  - **Model-free**

- Different **performance metrics** (Estimation + High-confidence bounds)
  - Mean**,**
  - **VaR, CVaR,**
  - Variance**, Entropy,**
  - **Median, Inter-quantiles**
  - **Etc**.

- Different **domain settings**
  - **Markovian, Non-Markovian**
  - **Fully observable, Partially-observable**
  - **Smoothly non-stationary, discrete distribution shifts**

Core Idea

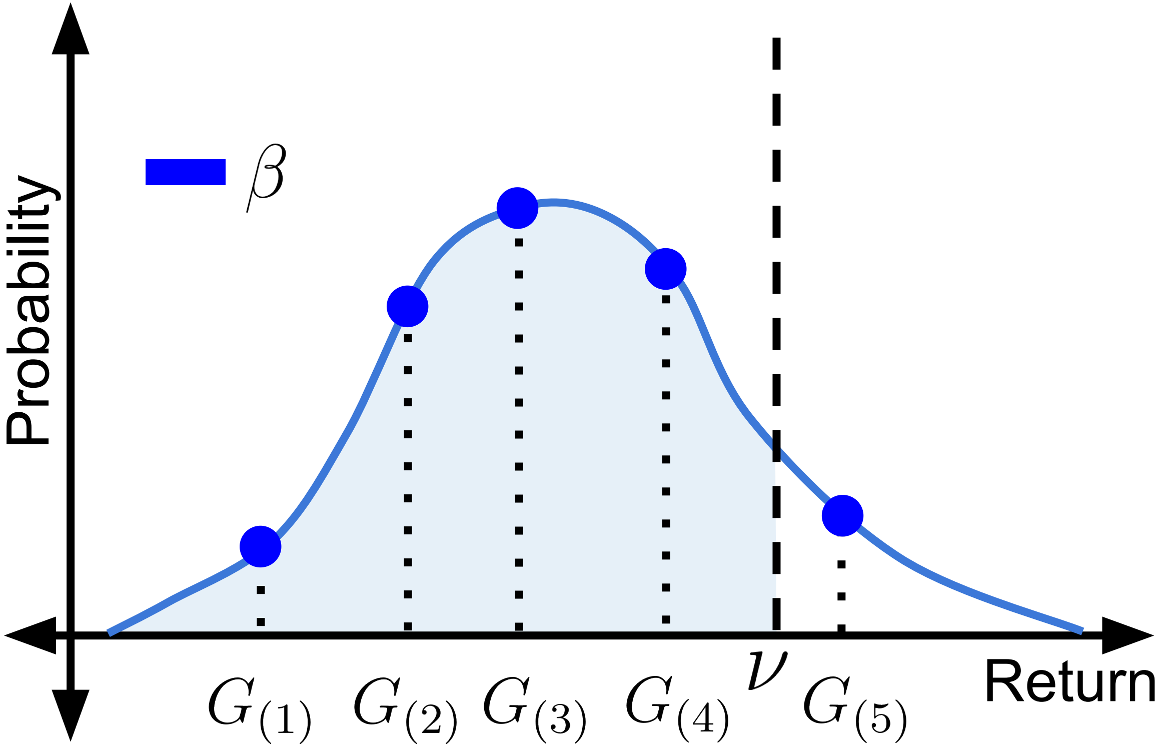
- If we have an **estimator for the CDF** then we can obtain an estimator for any of its parameters

$$F_\pi(\nu) := \Pr\left(G_\pi \leq \nu\right), \qquad \forall \nu \in \mathbb{R}$$

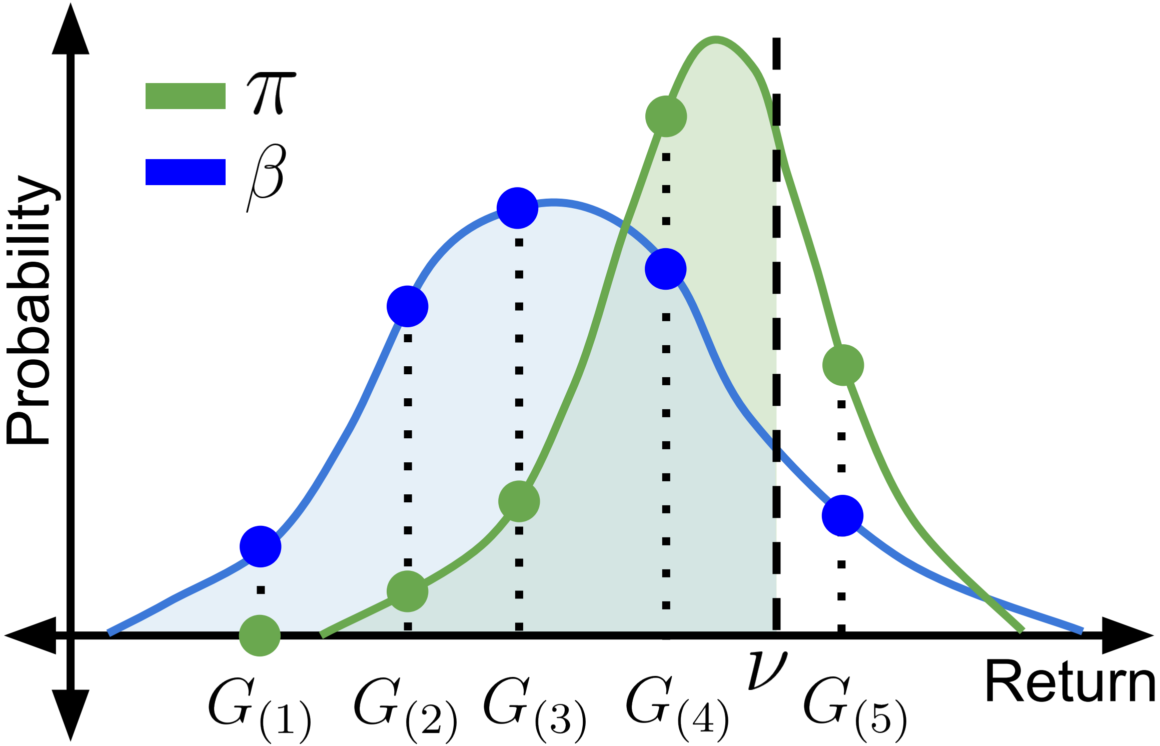
- **Bounds for the CDF** can directly be used to obtain bounds on its parameters

$$\Pr(\forall \nu \in \mathbb{R}, F_\pi(\nu) \in \mathcal{F}(\nu)) \geq 1 - \delta$$

Intuition for CDF Estimator



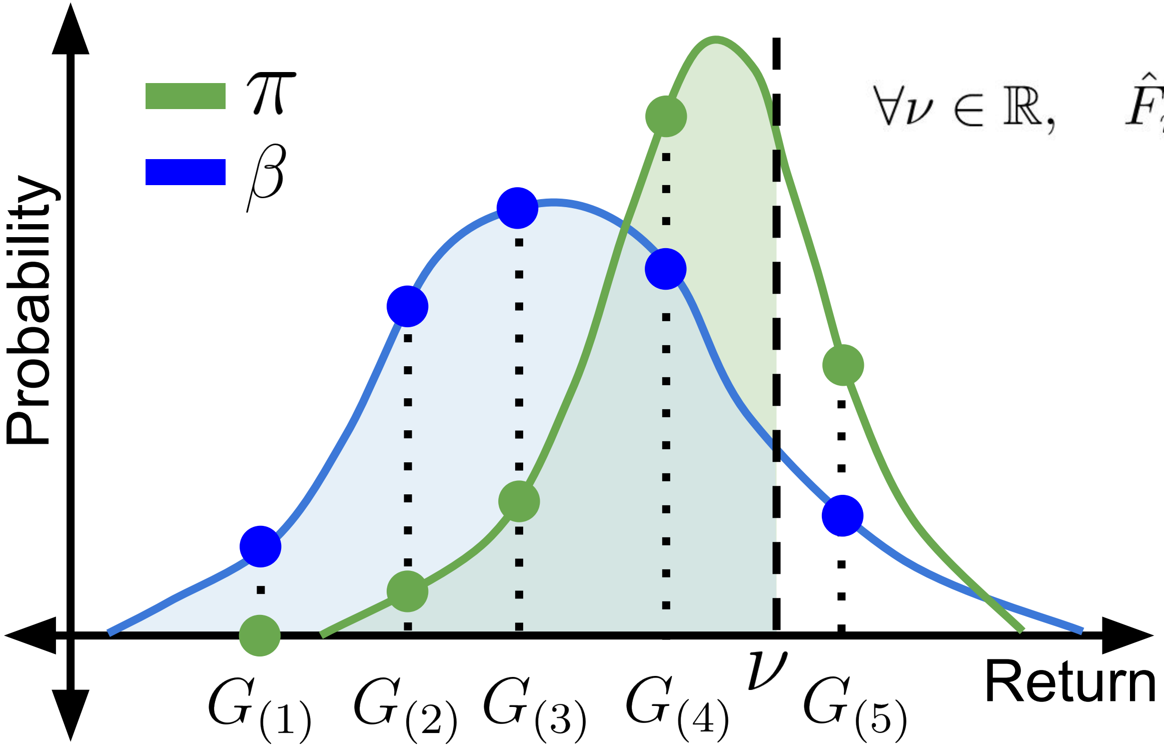
Intuition for CDF Estimator



Intuition for CDF Estimator

Let $\rho_i := \prod_{j=0}^{T} \frac{\pi(A_j|S_j)}{\beta_i(A_j|S_j)}$,

$$\forall \nu \in \mathbb{R}, \quad \hat{F}_n(\nu) := \frac{1}{n}\sum_{i=1}^{n} \rho_i \Big(1_{\{G_i \le \nu\}} \Big).$$



CDF Estimator

Theorem 1. *Under Assumption 1, $\hat{F}_n$ is an unbiased and uniformly consistent estimator of F_π. That is,*

$$\forall \nu \in \mathbb{R}, \quad \mathbb{E}_{\mathcal{D}}\left[\hat{F}_n(\nu)\right] = F_\pi(\nu),$$

$$\sup_{\nu \in \mathbb{R}} \left|\hat{F}_n(\nu) - F_\pi(\nu)\right| \xrightarrow{a.s.} 0.$$

Estimates for Different Parameters

$$\hat{F}_n^{-1}(\alpha) := \min\left\{g \in (G_{(i)})_{i=1}^{n} \,\middle|\, \hat{F}_n(g) \geq \alpha\right\},$$

CDF Inverse

$$\mathrm{d}\hat{F}_n(G_{(i)}) := \hat{F}_n(G_{(i)}) - \hat{F}_n(G_{(i-1)}),$$

PDF

Estimates for Different Parameters

$$\hat{F}_n^{-1}(\alpha) := \min\left\{g \in (G_{(i)})_{i=1}^n \middle| \hat{F}_n(g) \geq \alpha\right\},$$

CDF Inverse

$$\mathrm{d}\hat{F}_n(G_{(i)}) := \hat{F}_n(G_{(i)}) - \hat{F}_n(G_{(i-1)}),$$

PDF

$$\mu_\pi(\hat{F}_n) := \sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)})G_{(i)},$$

$$\sigma_\pi^2(\hat{F}_n) := \sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)})\Big(G_{(i)} - \mu_\pi(\hat{F}_n)\Big)^2,$$

$$\mathcal{H}_\pi(\hat{F}_n) := -\sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)}) \log \mathrm{d}\hat{F}_n(G_{(i)}).$$

$$Q_\pi^\alpha(\hat{F}_n) := \hat{F}_n^{-1}(\alpha),$$

$$\mathrm{IQR}_\pi^{\alpha_1,\alpha_2}(\hat{F}_n) := Q_\pi^{\alpha_2}(\hat{F}_n) - Q_\pi^{\alpha_1}(\hat{F}_n),$$

$$\mathrm{CVaR}_\pi^\alpha(\hat{F}_n) := \frac{1}{\alpha}\sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)})G_{(i)}\mathbb{1}_{\{G_{(i)} \leq Q_\pi^\alpha(\hat{F}_n)\}}.$$

Estimates for Different Parameters

$$\hat{F}_n^{-1}(\alpha) := \min\left\{g \in (G_{(i)})_{i=1}^n \,\middle|\, \hat{F}_n(g) \geq \alpha\right\}, \qquad \mathrm{d}\hat{F}_n(G_{(i)}) := \hat{F}_n(G_{(i)}) - \hat{F}_n(G_{(i-1)}),$$

Mean estimate ***exactly*** equal to the common (trajectory-based) IS estimate.

$$\mu_\pi(\hat{F}_n) := \sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)})G_{(i)},$$

$$\sigma^2_\pi(\hat{F}_n) := \sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)})\Big(G_{(i)} - \mu_\pi(\hat{F}_n)\Big)^2,$$

$$\mathcal{H}_\pi(\hat{F}_n) := -\sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)}) \log \mathrm{d}\hat{F}_n(G_{(i)}).$$

$$Q^\alpha_\pi(\hat{F}_n) := \hat{F}_n^{-1}(\alpha),$$

$$\mathrm{IQR}^{\alpha_1,\alpha_2}_\pi(\hat{F}_n) := Q^{\alpha_2}_\pi(\hat{F}_n) - Q^{\alpha_1}_\pi(\hat{F}_n),$$

$$\mathrm{CVaR}^\alpha_\pi(\hat{F}_n) := \frac{1}{\alpha}\sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)})G_{(i)}\mathbb{1}_{\{G_{(i)} \leq Q^\alpha_\pi(\hat{F}_n)\}}.$$

Bounds for Different Parameters

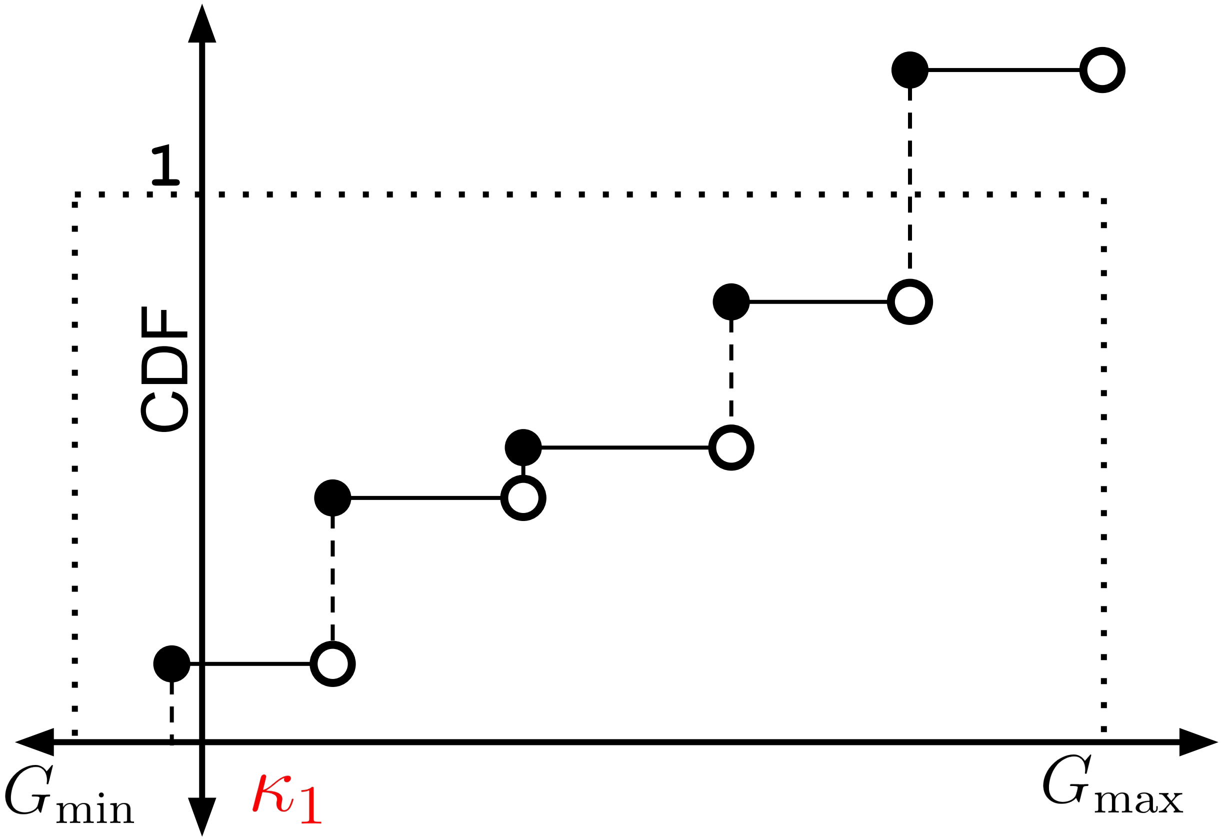
- Mean
- Quantile
- CVaR
- Inter-quantile
- Entropy
- Variance
- ….

Bounds for Different Parameters

- Mean
- **Quantile**
- **CVaR**
- **Inter-quantile**
- **Entropy**
- **Variance**
- ….

- Estimates for different parameters might be **biased**
- Importance sampling results in high **variance**
- Need to obtain **high-confidence bounds with guaranteed coverage** for reliability.

Bounds

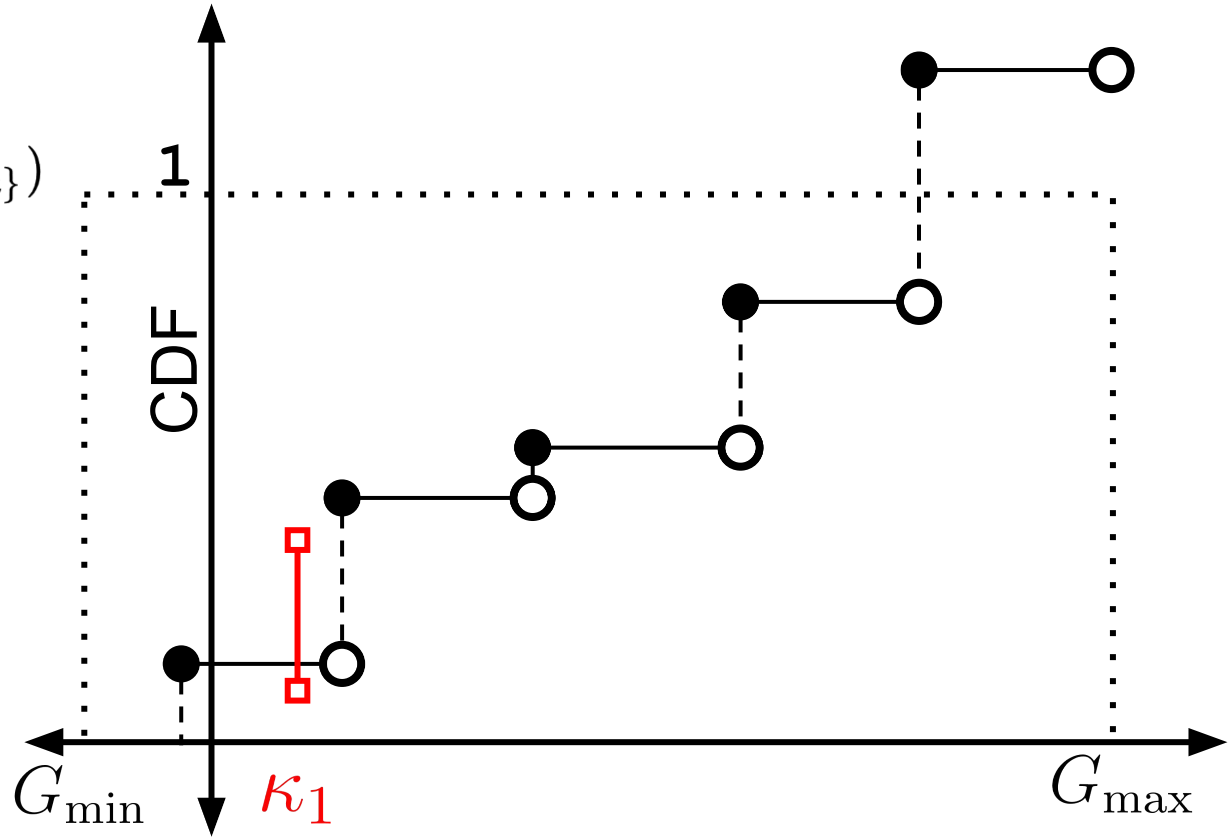


Bounds

$$\hat{F}_n(\kappa) := \frac{1}{n}\sum_{i=1}^{n} \rho_i(1_{\{G_i \le \kappa\}})$$

Mean Estimation!

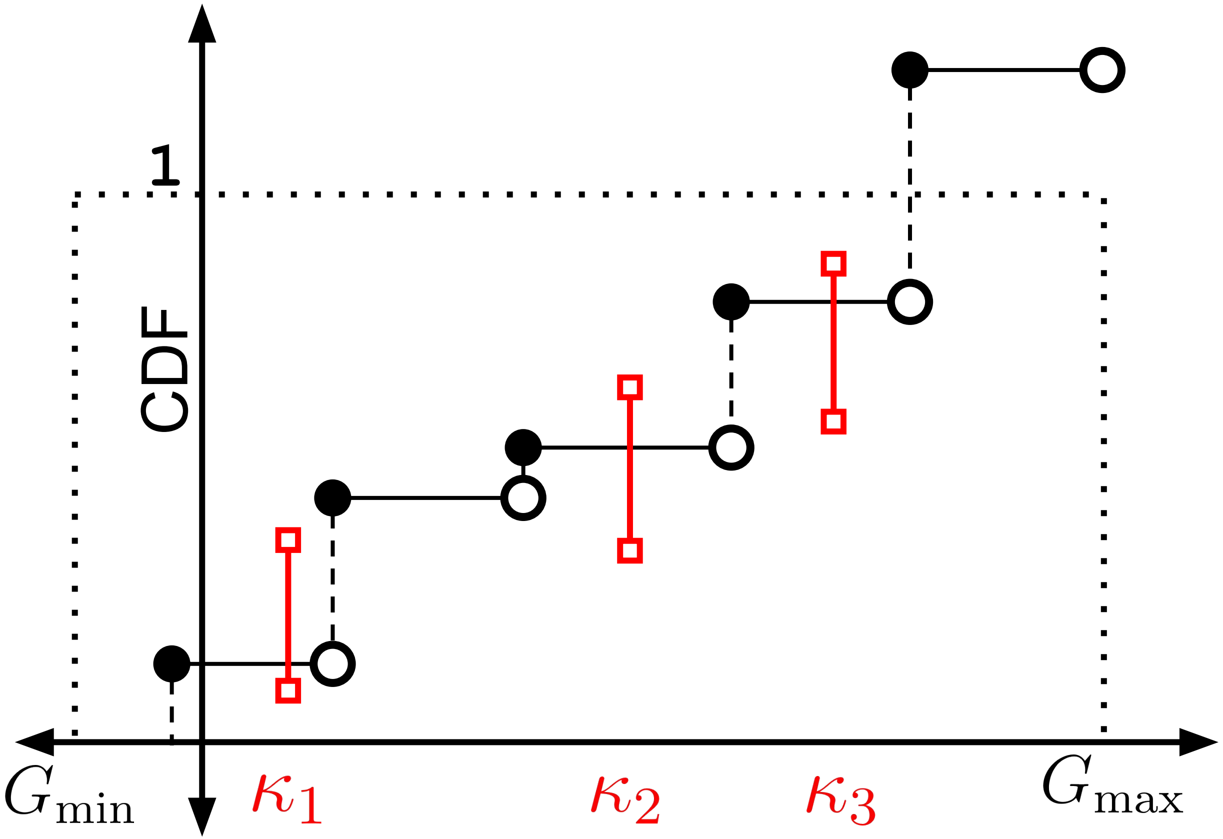
Let $X := \rho(1_{\{G \le \kappa\}})$.
$\mathbb{E}_{\mathcal{D}}[X] = F_\pi(\kappa)$.



Bounds

Mean Estimation!

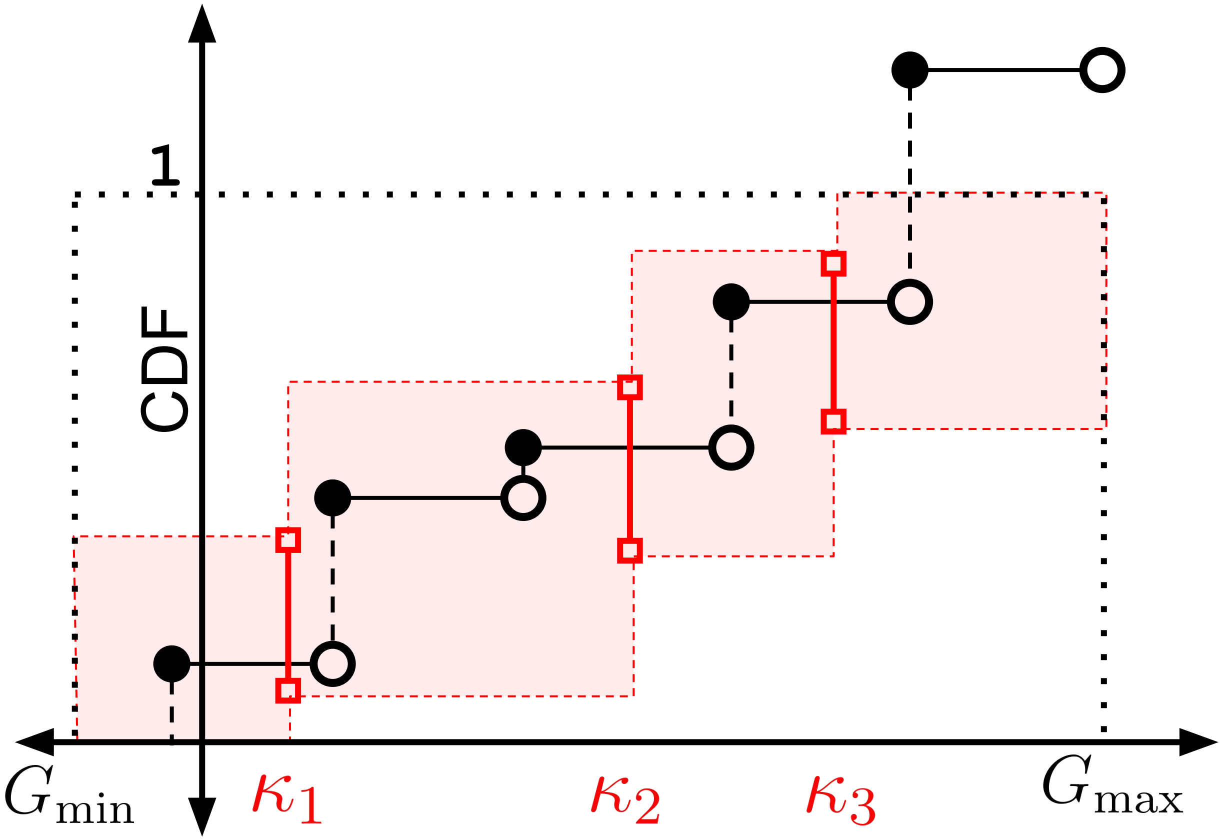
Let $X := \rho(1_{\{G \leq \kappa\}})$.
$\mathbb{E}_{\mathcal{D}}[X] = F_{\pi}(\kappa)$.



Bounds

Mean Estimation!

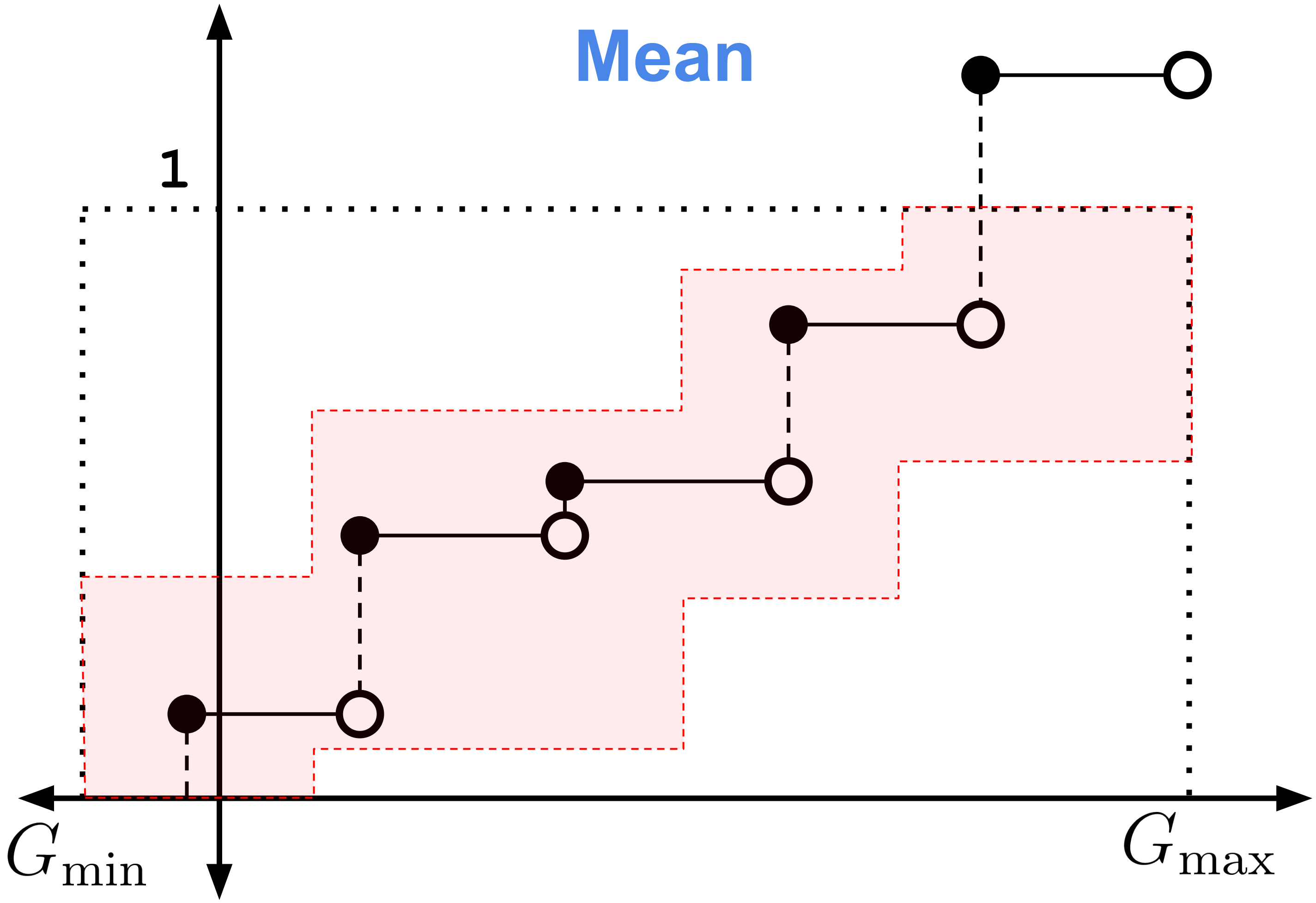
Let $X := \rho(1_{\{G \leq \kappa\}})$.
$\mathbb{E}_{\mathcal{D}}[X] = F_\pi(\kappa)$.

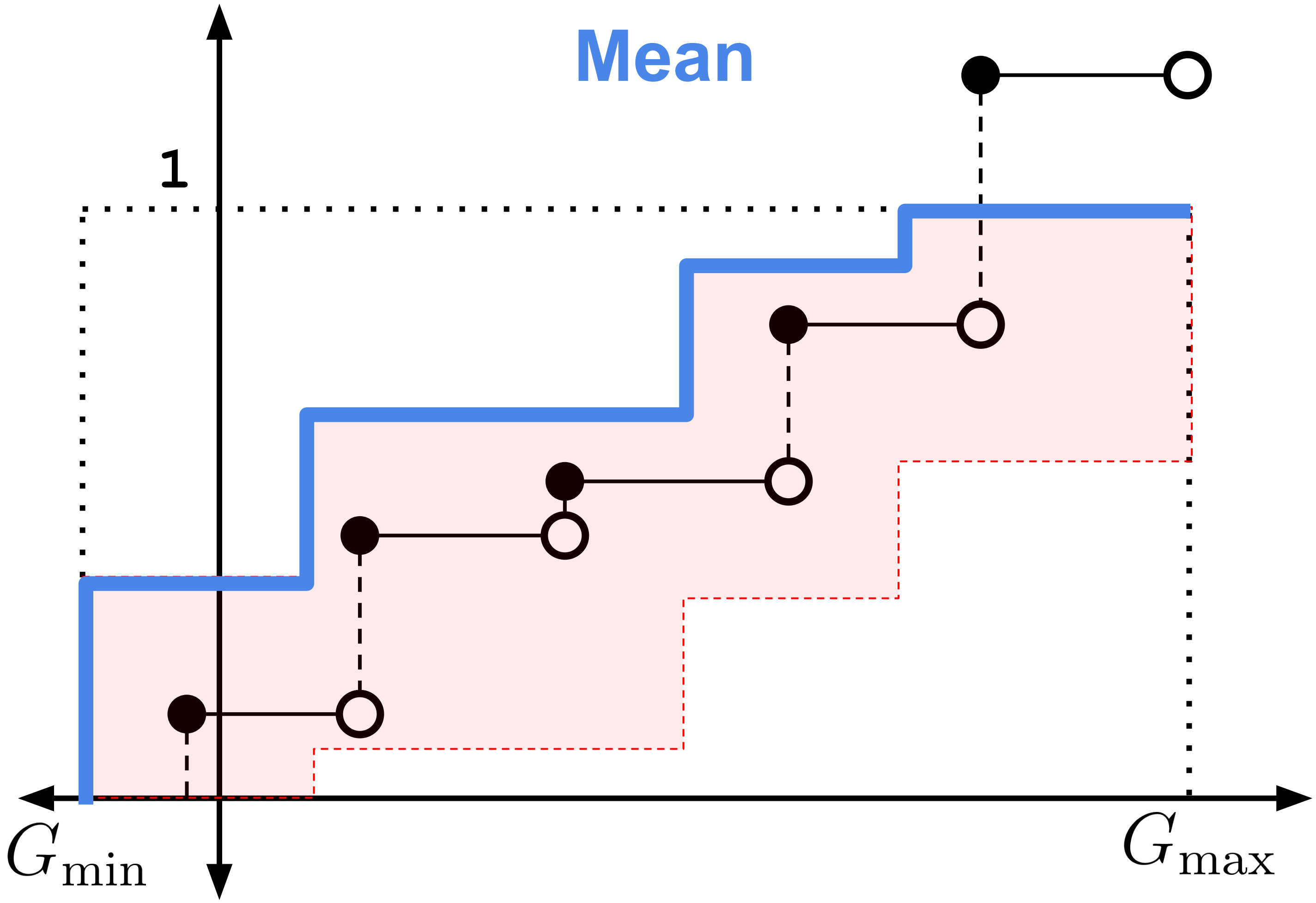


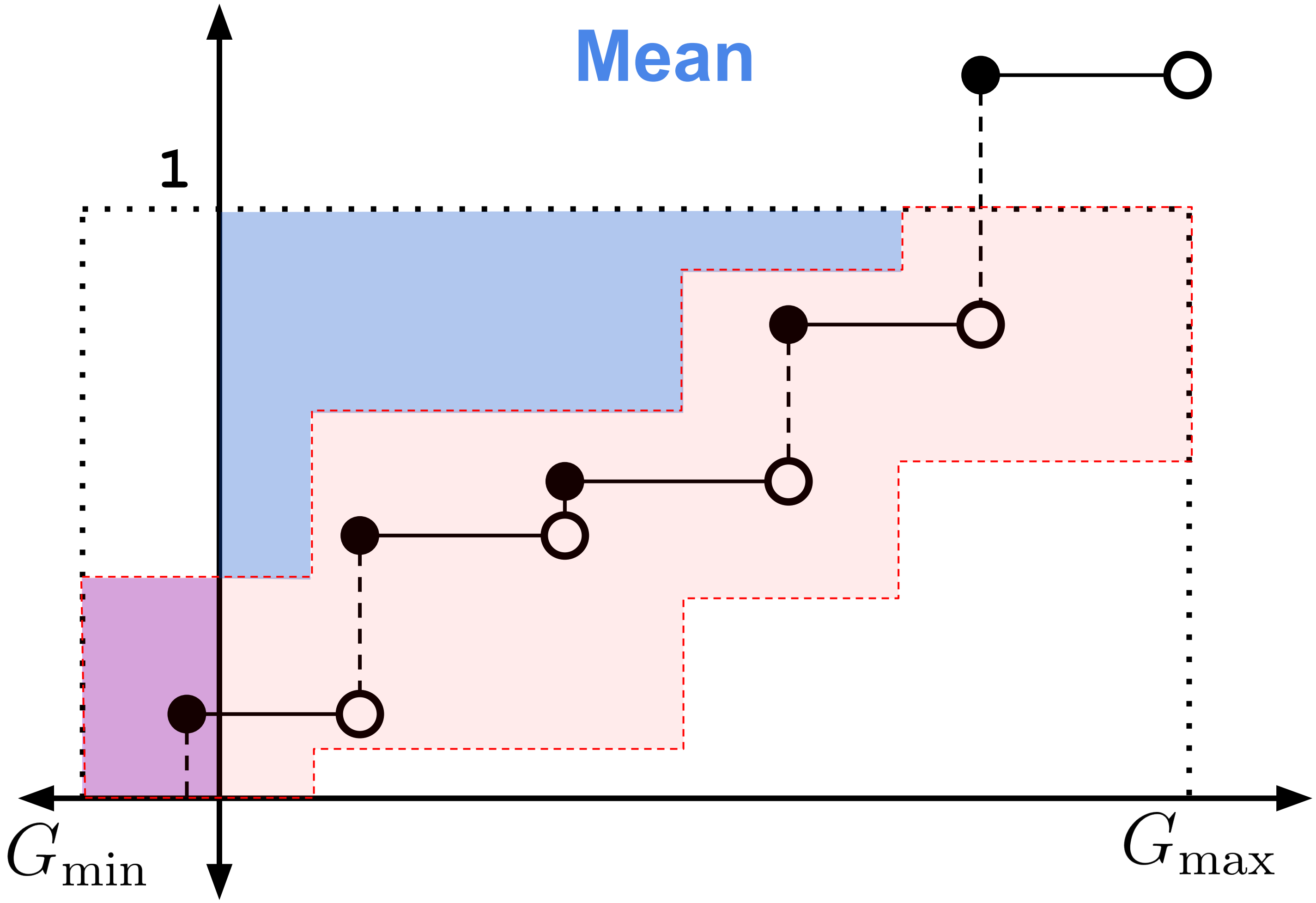
Bounds

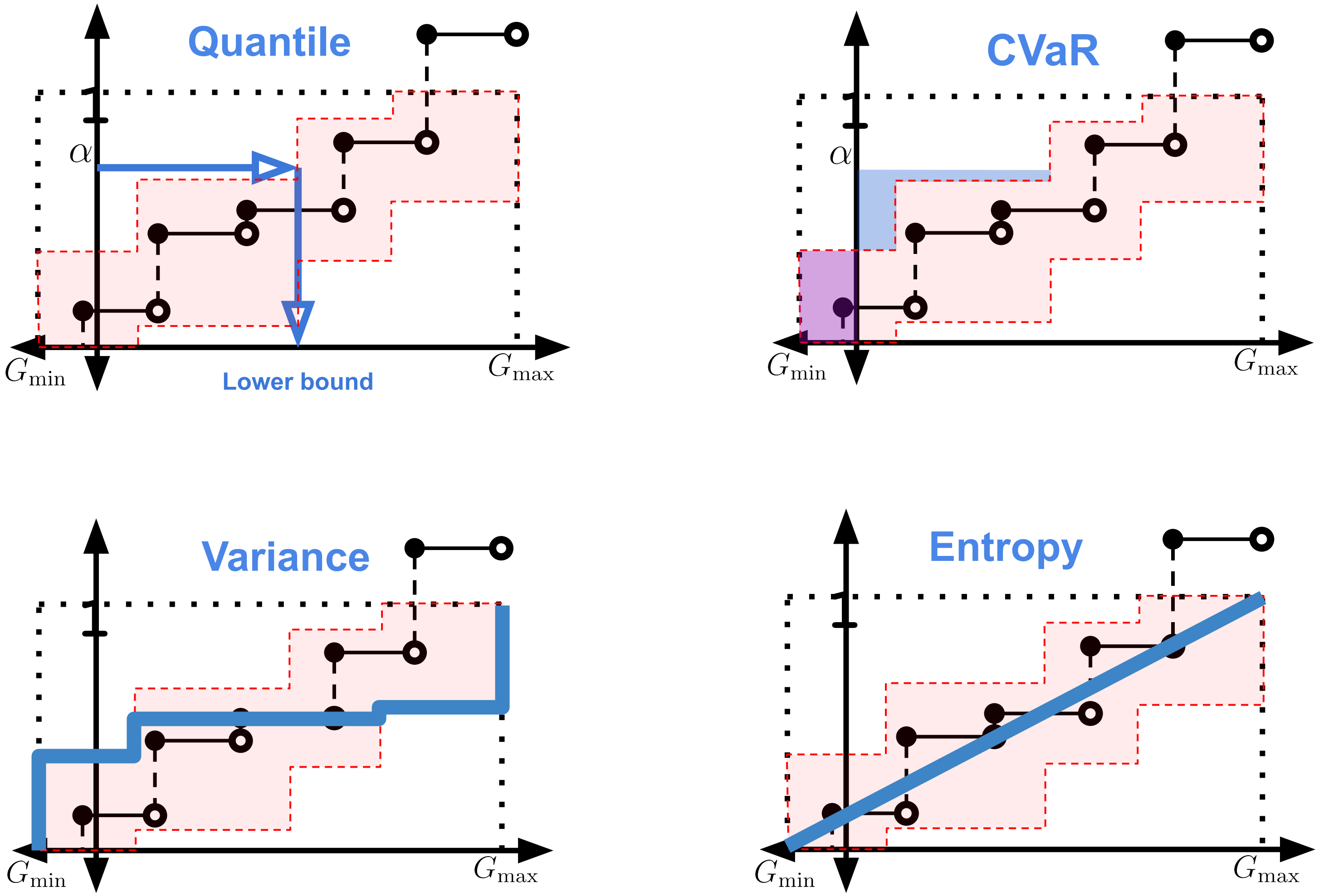
Theorem 3. *Under Assumption 1, for any* $\delta \in (0, 1]$, *if* $\sum_{i=1}^{K} \delta_i \leq \delta$, *then the confidence band defined by* F_- *and* F_+ *provides guaranteed coverage for* F_π. *That is,*

$$\Pr\left(\forall \nu,\ F_-(\nu) \leq F_\pi(\nu) \leq F_+(\nu)\right) \geq 1 - \delta.$$


Mean
1
$G_{\min}$
$G_{\max}$


Mean
1
$G_{\min}$
$G_{\max}$


Mean
1
$G_{\min}$
$G_{\max}$



Quantile
α
1
$G_{\min}$
Lower bound
$G_{\max}$
CVaR
α
1
$G_{\min}$
$G_{\max}$
Variance
1
$G_{\min}$
$G_{\max}$
Entropy
1
$G_{\min}$
$G_{\max}$

Bootstrap

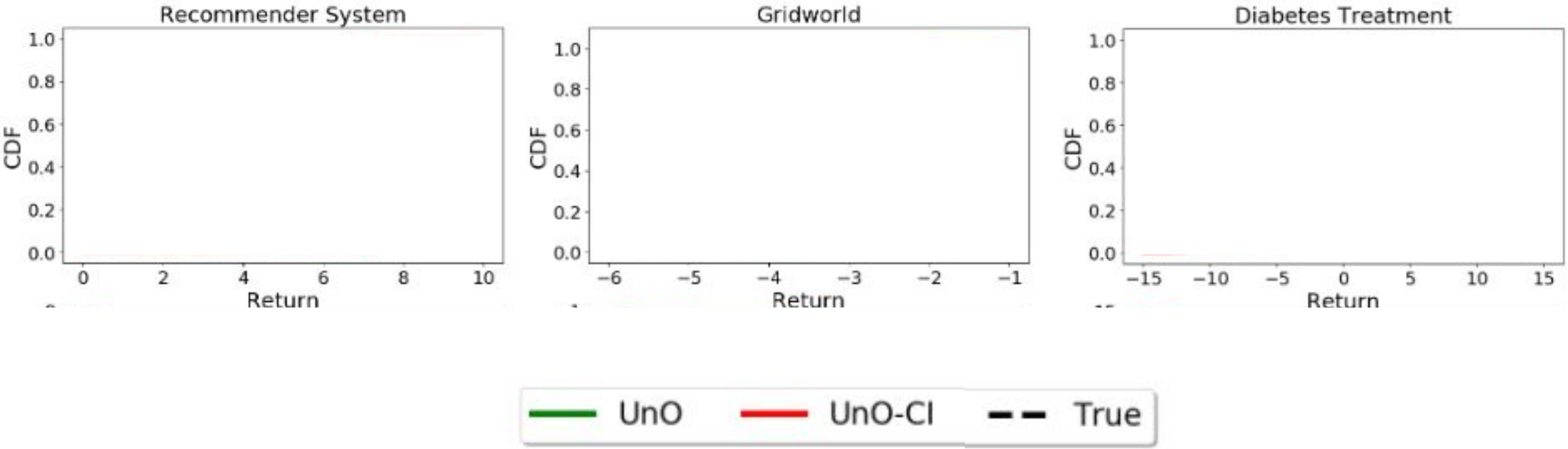
Algorithm 1: Bootstrap Bounds for $\psi(F_\pi)$

Input: Dataset $\mathcal{D}$, Confidence level $1-\delta$
Bootstrap B datasets $\{\mathcal{D}_i^*\}_{i=1}^B$ and create $\{\bar{F}_{n,i}^*\}_{i=1}^B$
Bootstrap estimates $\{\psi(\bar{F}_{n,i}^*)\}_{i=1}^B$ using $\{\bar{F}_{n,i}^*\}_{i=1}^B$.
Compute $\{\psi_-, \psi_+\}$ using BCa($\{\psi(\bar{F}_{n,i}^*)\}_{i=1}^B, \delta$) [1]
Return $\{\psi_-, \psi_+\}$

- Approximate
- Significantly Tighter

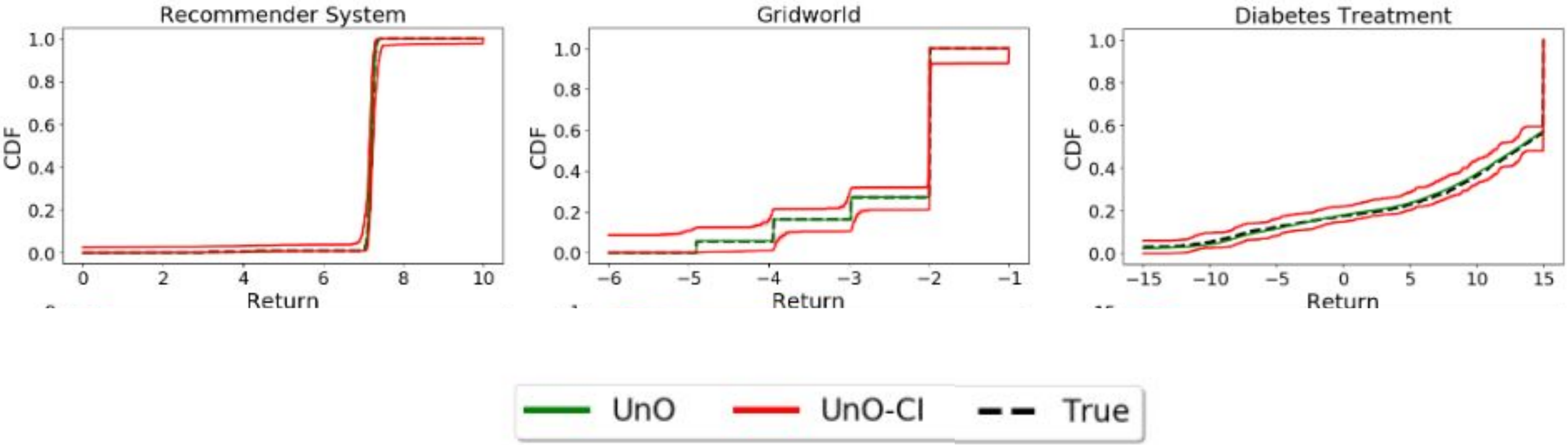
[1] DiCiccio, Thomas J., and Bradley Efron. "Bootstrap confidence intervals." *Statistical science* 11.3 (1996)

Empirical Results



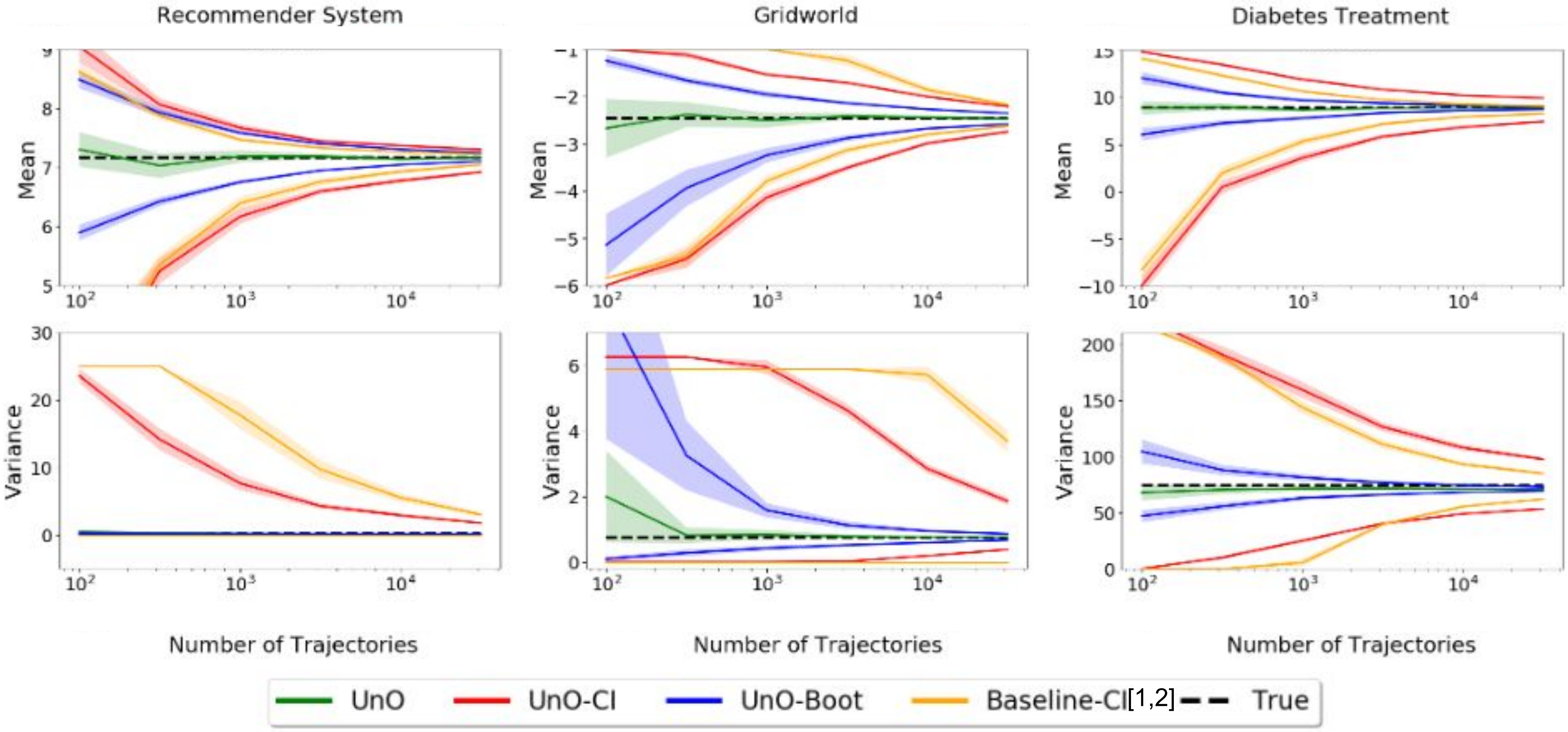
Empirical Results

30k samples
30 Trials

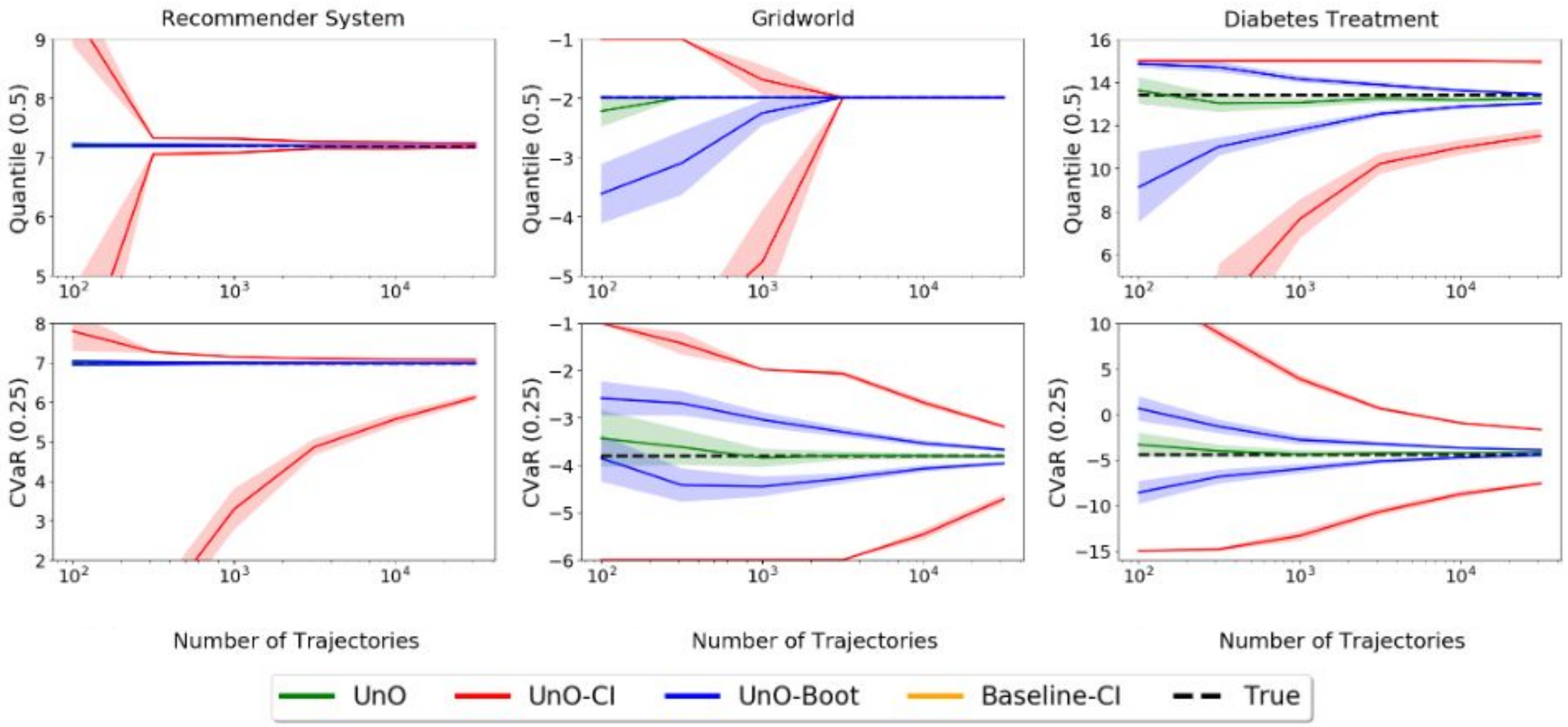


Empirical Results

[1] Thomas, Philip, Georgios Theocharous, and Mohammad Ghavamzadeh. "High-confidence off-policy evaluation." AAAI 2015.
[2] Chandak, Yash, Shiv Shankar, and Philip S. Thomas. "High-Confidence Off-Policy (or Counterfactual) Variance Estimation." AAAI 2021.



Empirical Results



Extensions

- **Weighted IS** based UnO for variance reduction*
- UnO for **partially observable MDPs***
- UnO for **discrete distributional shifts***
- UnO for **smooth non-stationarities***

- Parallel work at NeurIPS'21 by Audrey et al. [1] provides uniform **convergence rates** for off-policy CDF and Lipschitz risk functionals.

*see our paper for more details.

[1] Huang, Audrey, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. "Off-Policy Risk Assessment in Contextual Bandits." NeurIPS 2021.

thank you!