Improved Offline Contextual Bandits with Second-Order Bounds: Betting and Freezing

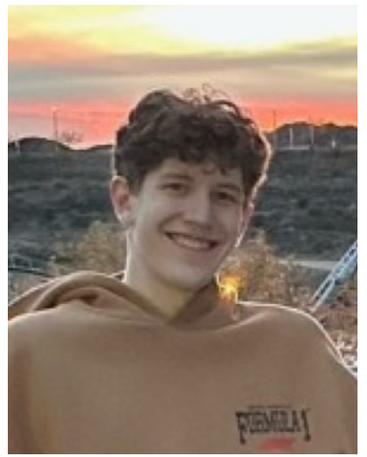
Jongha (Jon) Ryu

MIT EECS

RL Theory Seminars

December 16, 2025

Jeongyeol Kwon
UW-Madison/Meta

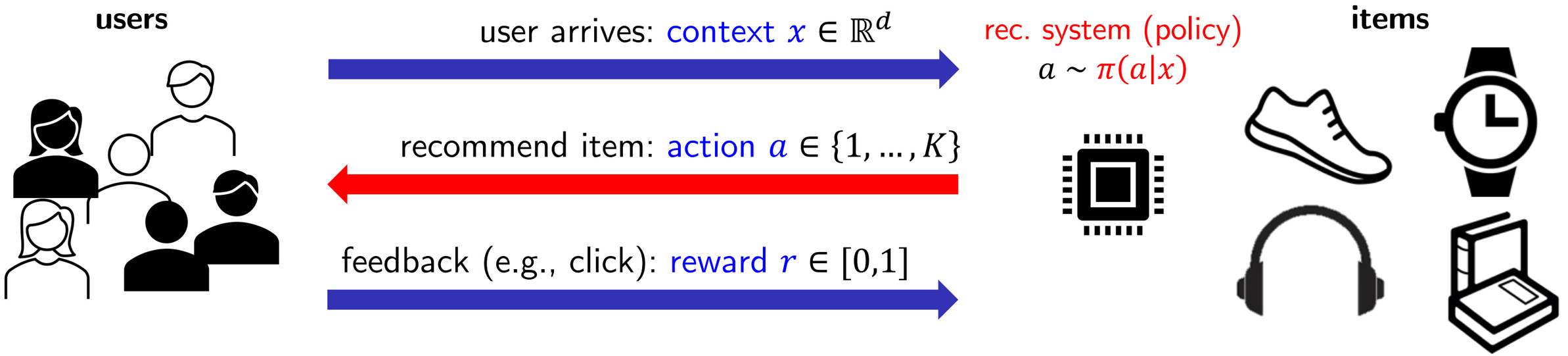


Benjamin Koppe
Cornell

Kwang-Sung Jun
U. of Arizona

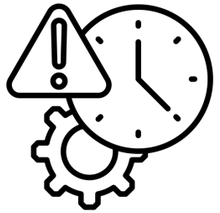
Contextual Bandits

- **Example**: recommendation systems



- **Goal**: find π that maximizes cumulative reward!
- Ideally, we can **try** different policies $\pi \in \Pi$ to find the best (via online interaction)

Online Interaction is Expensive!



- System constraints

- Organizational policies

- Real-time feedback is expensive

Let's go offline!

Problem: **Off-Policy** Contextual Bandit

- Behavior (logging) policy $\pi_0(a|x)$
- Offline data $D_n = \{(x_t, a_t, r_t)\}_{t=1}^n \sim p(x)\pi_0(a|x)p(r|a,x)$
  - For $t = 1, \ldots, n$
    - Observe **context** $x_t \sim p(x)$
    - Choose **action** $a_t \sim \pi_0(a|x_t)$
    - Observe **reward** $r_t \sim p(r|a_t, x_t)$

- **Goal**: Given a policy class Π, find the best policy $\hat{\pi} \in \arg\max_{\pi \in \Pi} v(\pi)$
- **Criterion:** Value $v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$ (expected reward)

Off-Policy Contextual Bandit: **Challenge**

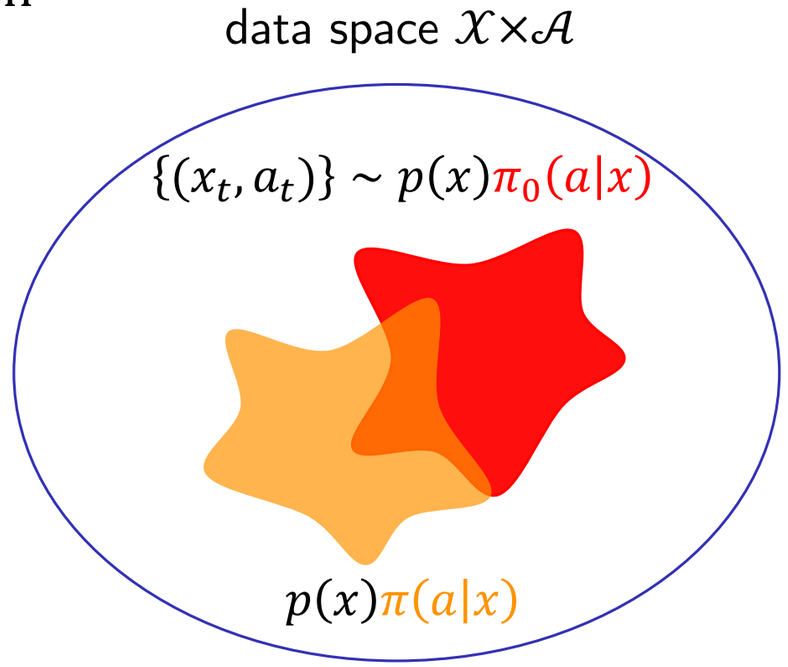
- **Goal**: Given a policy class Π, find the best policy $\hat{\pi} \in \arg\max_{\pi \in \Pi} v(\pi)$

- Difficulty? logging $\pi_0 \neq$ target π!
- At time t, given context x_t

behavior policy π_0 chose action 2

target policy π chooses action 3

	action	reward
observed	1	1.0
	2	0.1
	3	0.5
unobserved!	...	
	K	0.3



Risk of **over-optimization:** a low-value policy π might seem good on $\mathcal{D}$

Off-Policy Problems

~35 min ~10 min

	evaluation	selection	learning (=optimization)
$\lvert\Pi\rvert$ (size of policy class)	1	$< \infty$	∞ (e.g., neural nets)
Goal	estimate $v(\pi)$	find $\operatorname{argmax}_{\pi\in\Pi} v(\pi)$	find $\operatorname{argmax}_{\pi\in\Pi} v(\pi)$
Data split used	test	validation	train
Requirement	none	none	"argmax-oracle-efficiency"

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$

Part 1. Off-Policy Selection

Off-Policy Selection

Off-Policy Selection

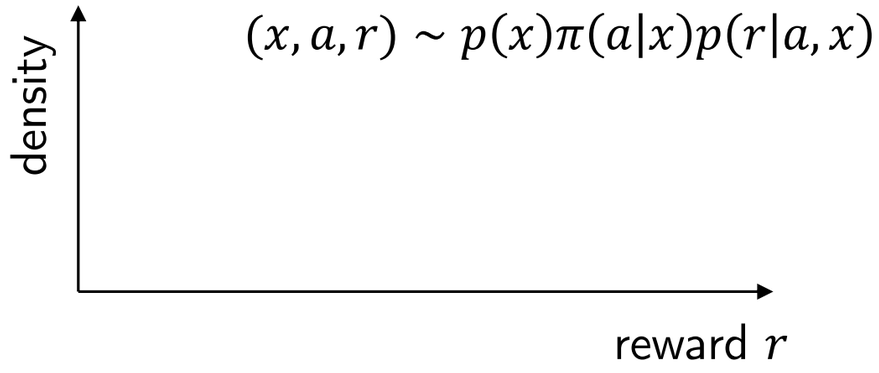
- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$

Off-Policy Selection

- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

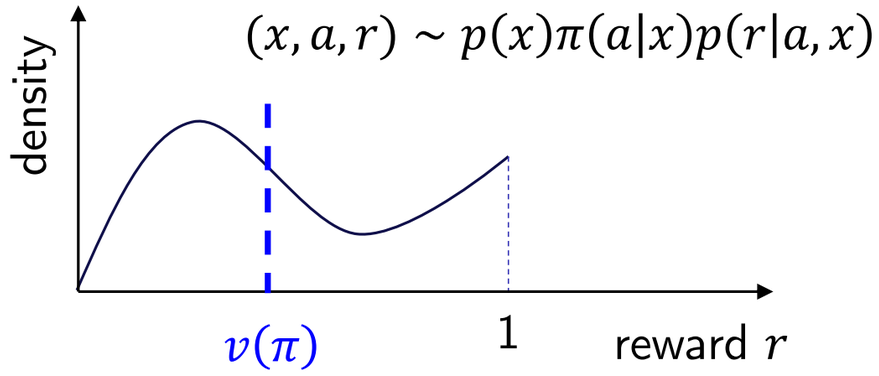
$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$



Off-Policy Selection

- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$



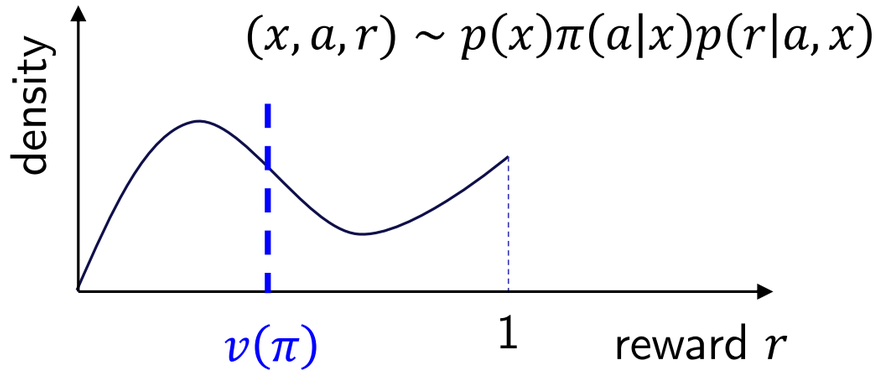
Off-Policy Selection

- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$

$$= \mathbb{E}_{p(x)\pi_0(a|x)p(r|a,x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)} r\right]$$

importance weight



Off-Policy Selection

- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$

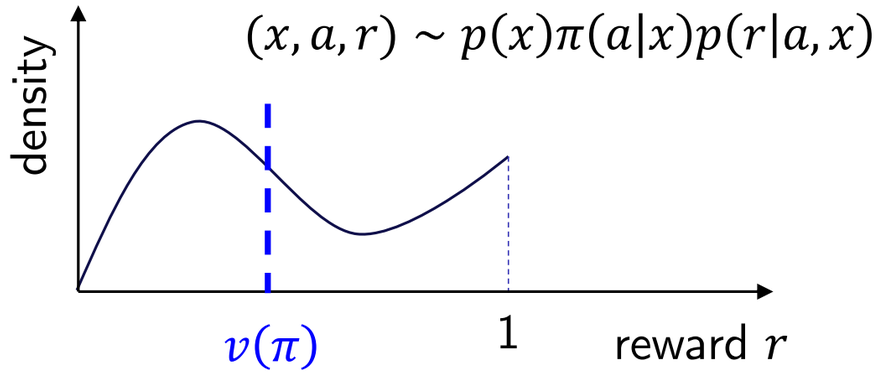
$$= \mathbb{E}_{p(x)\pi_0(a|x)p(r|a,x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)} r\right]$$

importance weight

$$\approx \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t \stackrel{\text{def}}{=} \mathrm{IPS}(\pi)$$

inverse propensity score (IPS)
inverse propensity weighting (IPW)
importance weighted estimator (IW)

unbiased!



Off-Policy Selection

- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$

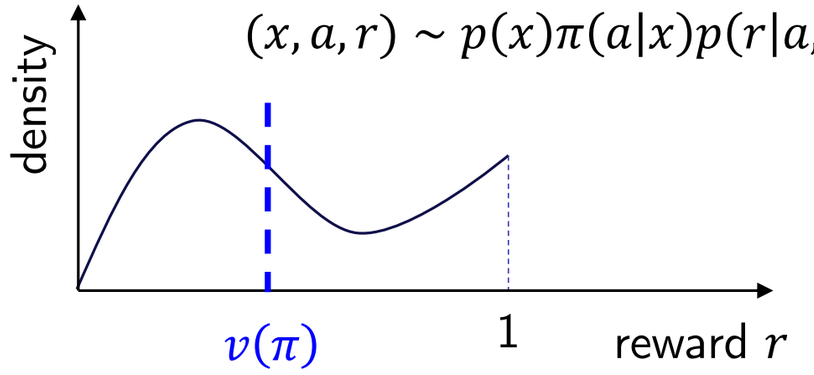
$$= \mathbb{E}_{p(x)\pi_0(a|x)p(r|a,x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)} r\right]$$

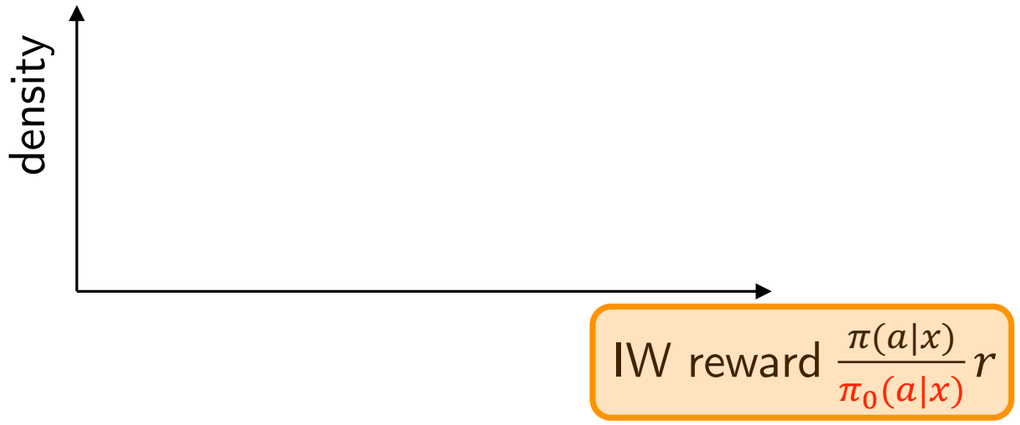
importance weight

$$\approx \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t \stackrel{\text{def}}{=} \mathrm{IPS}(\pi)$$

inverse propensity score (IPS)
inverse propensity weighting (IPW)
importance weighted estimator (IW)

unbiased!





Off-Policy Selection

- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$

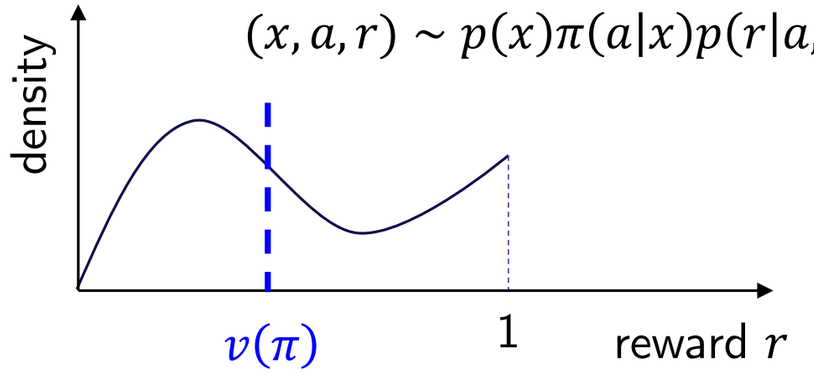
$$= \mathbb{E}_{p(x)\pi_0(a|x)p(r|a,x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)} r\right]$$

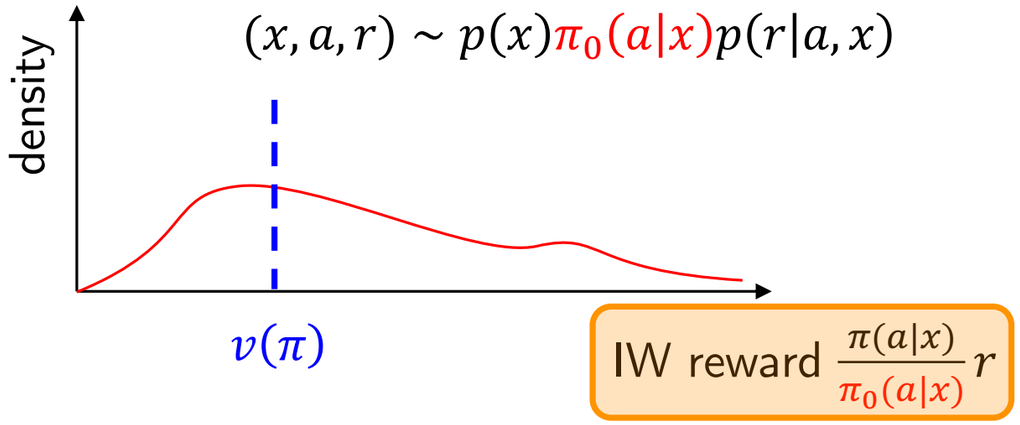
importance weight

$$\approx \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t \stackrel{\text{def}}{=} \mathrm{IPS}(\pi)$$

unbiased!

inverse propensity score (IPS)
inverse propensity weighting (IPW)
importance weighted estimator (IW)





Off-Policy Selection

- Straightforward if we estimate **the value function** $\pi \mapsto v(\pi)$, using $\mathcal{D}$

$$v(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r]$$

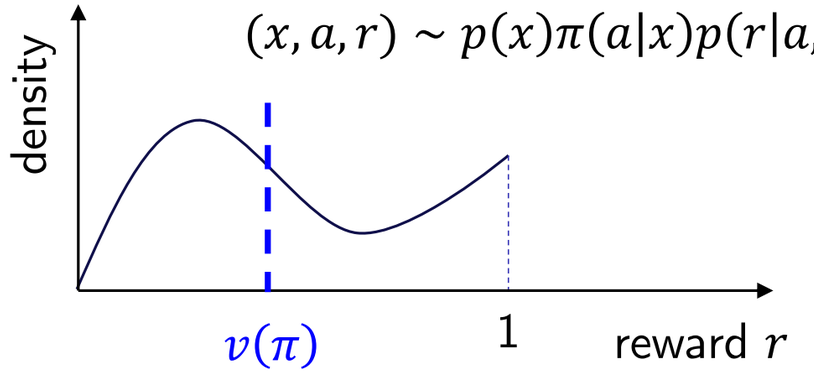
$$= \mathbb{E}_{p(x)\pi_0(a|x)p(r|a,x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)} r\right]$$

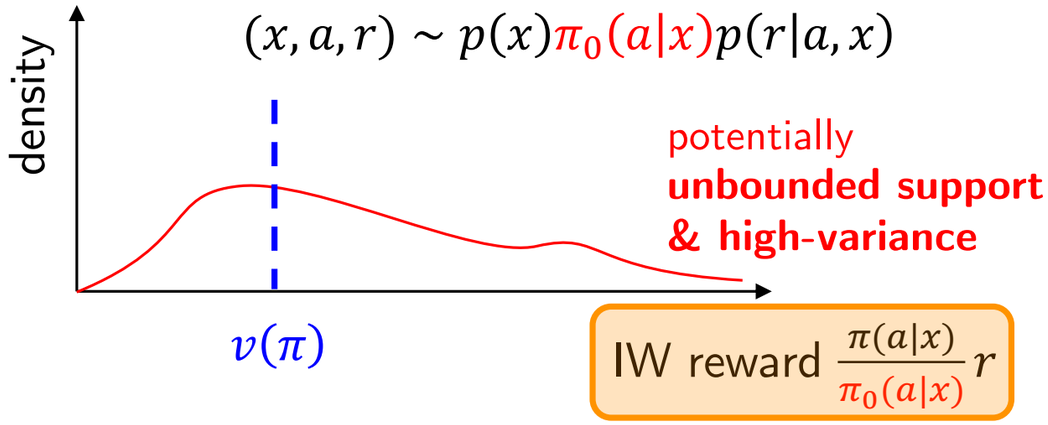
importance weight

$$\approx \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t \stackrel{\text{def}}{=} \mathrm{IPS}(\pi)$$

unbiased!

inverse propensity score (IPS)
inverse propensity weighting (IPW)
importance weighted estimator (IW)



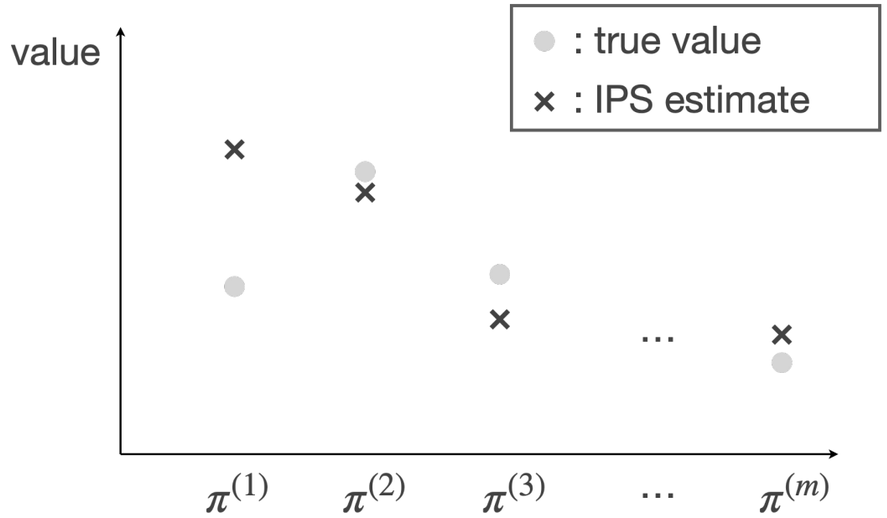


Off-Policy Selection: **IPS** and **Pessimism**

- $\text{IPS}(\pi)$ is unbiased but might be high-variance!

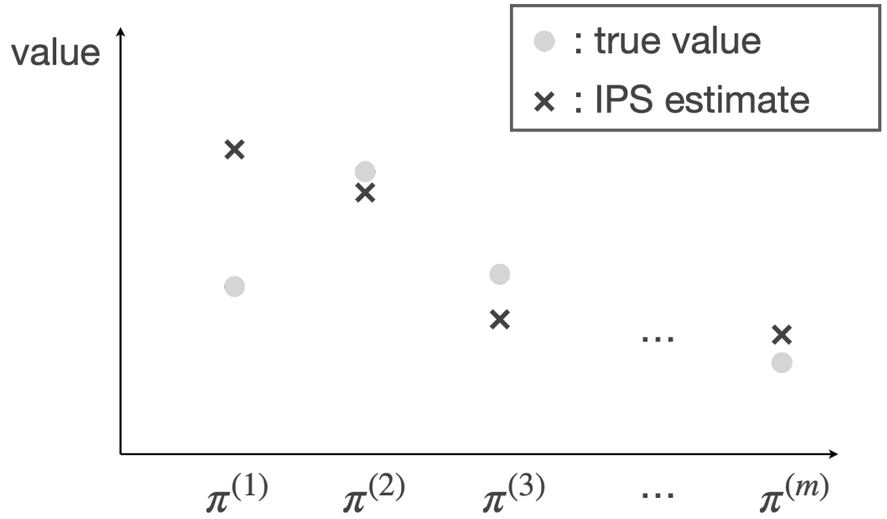
Off-Policy Selection: **IPS** and **Pessimism**

- $\mathrm{IPS}(\pi)$ is unbiased but might be high-variance!



Off-Policy Selection: **IPS** and **Pessimism**

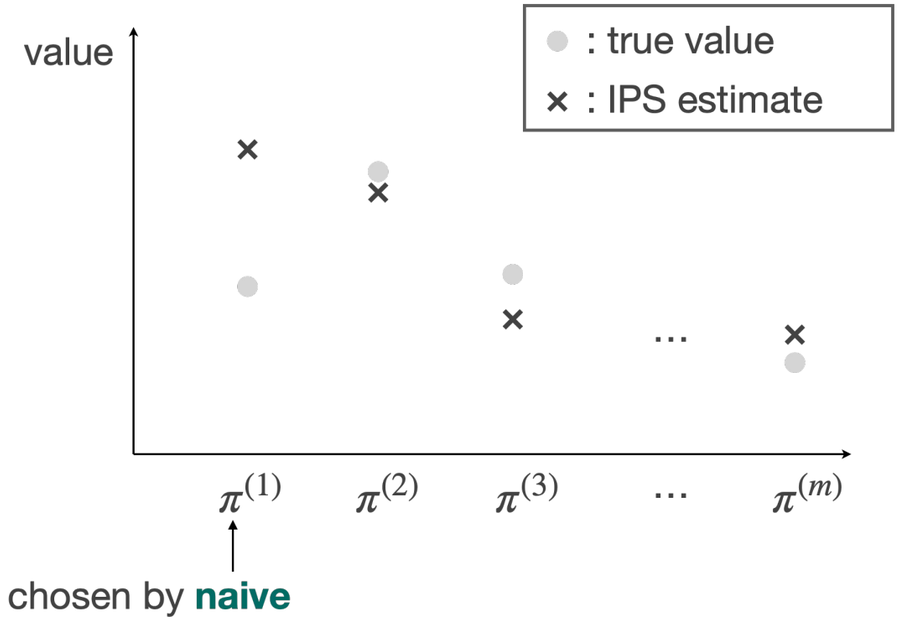
- $\mathrm{IPS}(\pi)$ is unbiased but might be high-variance!
- **Naive selection** with $\mathrm{IPS}(\pi)$ (i.e., $\operatorname{argmax}_{\pi \in \Pi} \mathrm{IPS}(\pi)$)



Off-Policy Selection: **IPS** and **Pessimism**

- $\mathrm{IPS}(\pi)$ is unbiased but might be high-variance!
- **Naive selection** with $\mathrm{IPS}(\pi)$ (i.e., $\operatorname{argmax}_{\pi \in \Pi} \mathrm{IPS}(\pi)$)

  might pick up low-value, high-variance policy

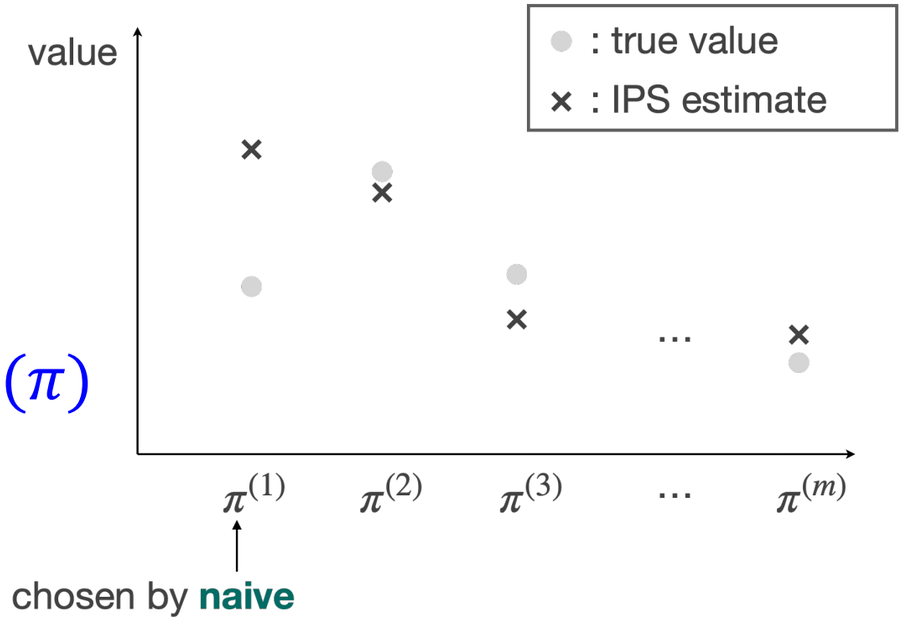


Off-Policy Selection: **IPS** and **Pessimism**

- $\mathrm{IPS}(\pi)$ is unbiased but might be high-variance!
- **Naive selection** with $\mathrm{IPS}(\pi)$ (i.e., $\operatorname{argmax}_{\pi \in \Pi} \mathrm{IPS}(\pi)$)

  might pick up low-value, high-variance policy
- **Pessimism**: Use lower confidence bound (LCB) on $v(\pi)$

$$\operatorname*{argmax}_{\pi \in \Pi} \mathrm{LCB}(\pi)$$

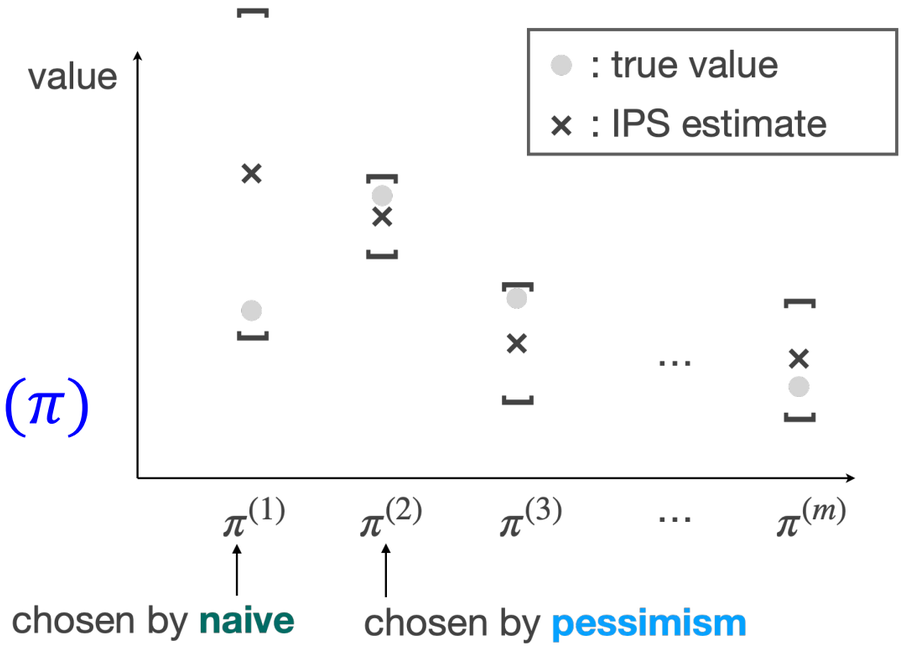


Off-Policy Selection: **IPS** and **Pessimism**

- $\mathrm{IPS}(\pi)$ is unbiased but might be high-variance!
- **Naive selection** with $\mathrm{IPS}(\pi)$ (i.e., $\operatorname{argmax}_{\pi \in \Pi} \mathrm{IPS}(\pi)$)

  might pick up low-value, high-variance policy
- **Pessimism**: Use lower confidence bound (LCB) on $v(\pi)$

$$\operatorname*{argmax}_{\pi \in \Pi} \mathrm{LCB}(\pi)$$



Off-Policy Selection: **IPS** and **Pessimism**

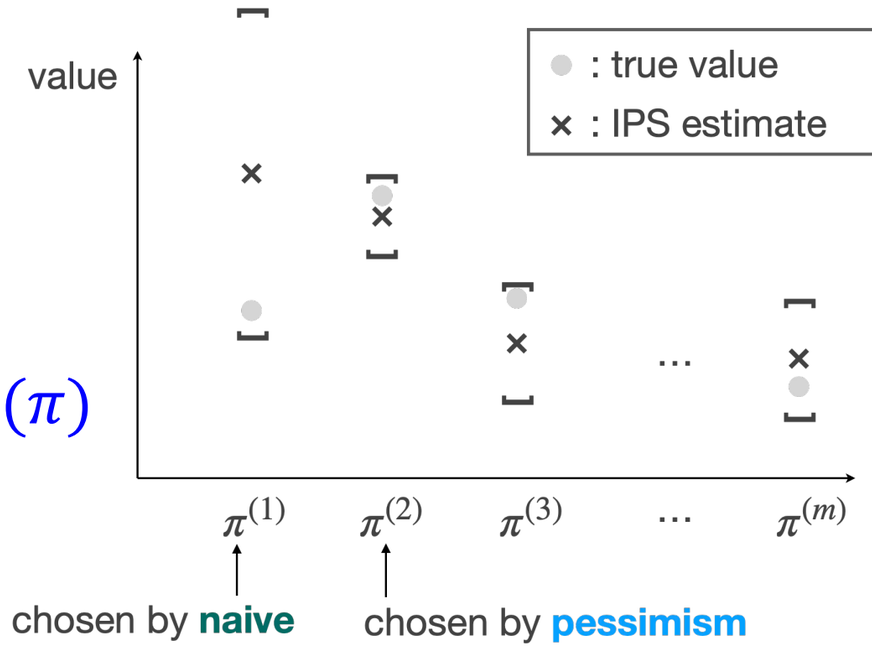
- $\mathrm{IPS}(\pi)$ is unbiased but might be high-variance!
- **Naive selection** with $\mathrm{IPS}(\pi)$ (i.e., $\operatorname{argmax}_{\pi \in \Pi} \mathrm{IPS}(\pi)$)

  might pick up low-value, high-variance policy
- **Pessimism**: Use lower confidence bound (LCB) on $v(\pi)$

  $$\operatorname*{argmax}_{\pi \in \Pi} \mathrm{LCB}(\pi)$$

  Can avoid over-optimization!

  (caveat: high-value, high-variance policy might be still not visible)



Off-Policy Selection: **IPS** and **Pessimism**

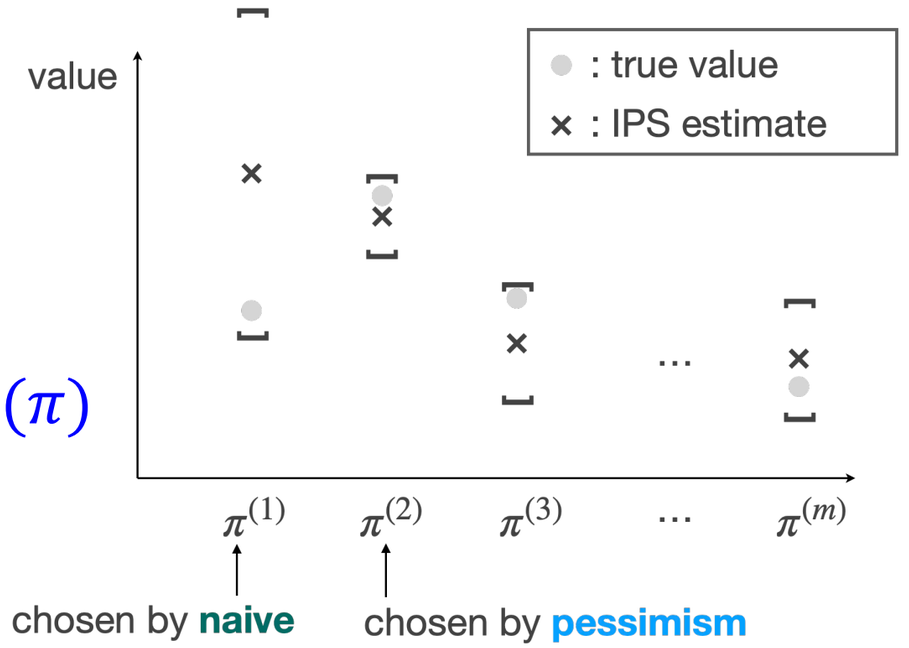
- $\mathrm{IPS}(\pi)$ is unbiased but might be high-variance!
- **Naive selection** with $\mathrm{IPS}(\pi)$ (i.e., $\operatorname{argmax}_{\pi \in \Pi} \mathrm{IPS}(\pi)$)

  might pick up low-value, high-variance policy
- **Pessimism**: Use lower confidence bound (LCB) on $v(\pi)$

  $$\operatorname*{argmax}_{\pi \in \Pi} \mathrm{LCB}(\pi)$$

  Can avoid over-optimization!

  (caveat: high-value, high-variance policy might be still not visible)



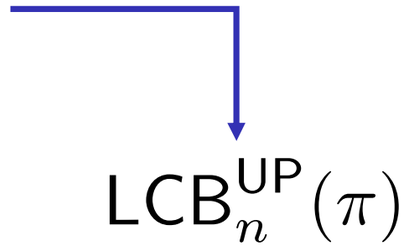
Our contribution

We propose **a "tight" LCB** that yields **provably good** selection via **pessimism**,
(1) with variance-adaptive property, (2) without hyperparameter!

Pessimism by **semi-Unbounded coin Betting** (PUB)

Pessimism by **semi-Unbounded coin Betting** (PUB)

- New LCB based on betting + universal portfolio (will elaborate soon)



Pessimism by **semi-Unbounded coin Betting** (PUB)

- New LCB based on betting + universal portfolio (will elaborate soon)

- **Pessimism** by **semi-Unbounded coin Betting** (PUB)

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \mathsf{LCB}_n^{\mathsf{UP}}(\pi)$$

Pessimism by **semi-Unbounded coin Betting** (PUB)

- New LCB based on betting + universal portfolio (will elaborate soon)
- **Pessimism** by **semi-Unbounded coin Betting** (PUB)

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \mathsf{LCB}_n^{\mathsf{UP}}(\pi)$$

$$\forall \pi^* \in \Pi, \quad \boxed{v(\pi^*) - v(\hat{\pi})} \lesssim \sqrt{\mathsf{Var}(\tilde{r}_1(\pi^*))\frac{1}{n}\ln\frac{|\Pi|}{\delta}}$$

optimality gap

$$\tilde{r}_t(\pi) \triangleq \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t$$

Pessimism by **semi-Unbounded coin Betting** (PUB)

- New LCB based on betting + universal portfolio (will elaborate soon)
- **Pessimism** by **semi-Unbounded coin Betting** (PUB) $\hat{\pi} = \arg\max_{\pi \in \Pi} \mathsf{LCB}_n^{\mathsf{UP}}(\pi)$

$$\forall \pi^* \in \Pi, \quad \boxed{v(\pi^*) - v(\hat{\pi})} \lesssim \sqrt{\underbrace{\mathsf{Var}(\tilde{r}_1(\pi^*))}_{\text{variance-adaptive, w/o hyperparameter tuning!}} \frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

optimality gap

$$\tilde{r}_t(\pi) \triangleq \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t$$

Pessimism by **semi-Unbounded coin Betting** (PUB)

- New LCB based on betting + universal portfolio (will elaborate soon)
- **Pessimism** by **semi-Unbounded coin Betting** (PUB) $\hat{\pi} = \arg\max_{\pi \in \Pi} \mathsf{LCB}_n^{\mathsf{UP}}(\pi)$

$$\forall \pi^* \in \Pi, \quad \boxed{v(\pi^*) - v(\hat{\pi})} \lesssim \sqrt{\underbrace{\mathsf{Var}(\tilde{r}_1(\pi^*))}_{\text{variance-adaptive, w/o hyperparameter tuning!}} \frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

optimality gap

$$\tilde{r}_t(\pi) \triangleq \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t$$

Prior art: Logarithmic smoothing (Sakhi+2024)

$$\lesssim (1 + \mathbb{E}[(\tilde{r}_1(\pi^*))^2]) \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

Pessimism by **semi-Unbounded coin Betting** (PUB)

- New LCB based on betting + universal portfolio (will elaborate soon)
- **Pessimism** by **semi-Unbounded coin Betting** (PUB) $\hat{\pi} = \arg\max_{\pi \in \Pi} \mathrm{LCB}_n^{\mathsf{UP}}(\pi)$

$$\forall \pi^* \in \Pi, \quad \boxed{v(\pi^*) - v(\hat{\pi})} \lesssim \sqrt{\underbrace{\mathsf{Var}(\tilde{r}_1(\pi^*))}_{\text{variance-adaptive, w/o hyperparameter tuning!}} \frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

optimality gap

$$\tilde{r}_t(\pi) \triangleq \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)} r_t$$

Prior art: Logarithmic smoothing (Sakhi+2024)

$$\lesssim (1 + \mathbb{E}[(\tilde{r}_1(\pi^*))^2]) \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

$$\sqrt{\mathsf{Var}(\tilde{r}_1(\pi^*))} \leq 1 + \mathsf{Var}(\tilde{r}_1(\pi^*)) \leq 1 + \mathbb{E}[(\tilde{r}_1(\pi^*))^2]$$

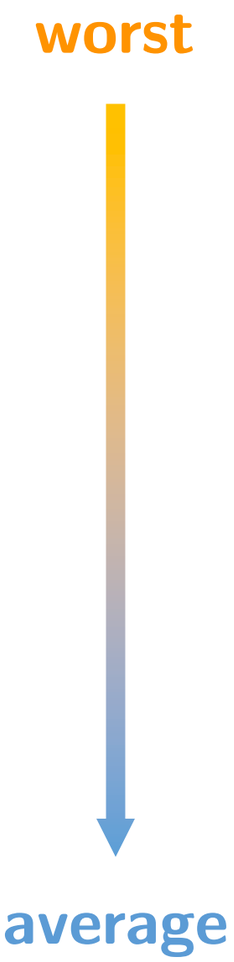
two-step improvement!

Comparison to Prior Arts

	Requires boundedness / moment condition?	Rate guarantee (=confidence width)	Bound is adaptive to
Empirical Bernstein [MaurerP2009]	Y	Y	$\sqrt{\mathrm{Var}(\tilde{r}(\pi^*))}$
Oracle-efficient pessimism [Wang+2024]	Y	Y	("pseudo-loss") × (worst-ratio factor)
[Waudby-Smith+2022]	N	unknown	unknown
Implicit exploration (IX) [Gabbianelli+2024]	N	Y	"smoothed coverage ratio"
Logarithmic smoothing (LS) [Sakhi+2024]	N	Y	$1 + \mathbb{E}\left[\left(\tilde{r}(\pi^*)\right)^2\right]$
PUB (ours)	**N**	**Y**	$\sqrt{\mathrm{Var}(\tilde{r}(\pi^*))}$

Remarks on **Coverage Measures**

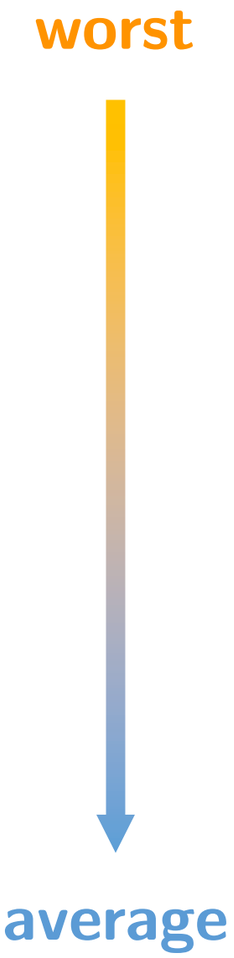
$$v(\pi) - v(\hat{\pi}) \lesssim \boxed{?} \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$



Remarks on **Coverage Measures**

$$v(\pi) - v(\hat{\pi}) \lesssim \boxed{?} \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

- In RL, "worst-case" coverage $\max_{(a,x)} \frac{\pi(a|x)}{\pi_0(a|x)}$



Remarks on **Coverage Measures**

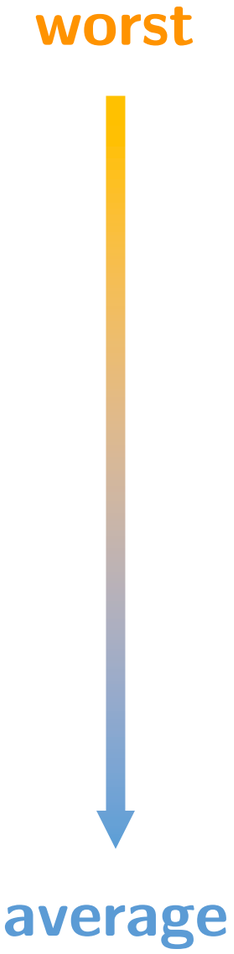
$$v(\pi) - v(\hat{\pi}) \lesssim \boxed{?} \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

- In RL, "worst-case" coverage

$$\max_{(a,x)} \frac{\pi(a|x)}{\pi_0(a|x)}$$

- Pseudo-loss [Wang+2024]

$$\mathbb{E}_{p(x)}\left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)}\right]$$



Remarks on **Coverage Measures**

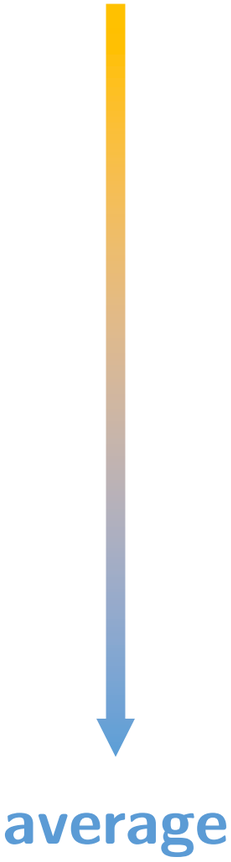
$$v(\pi) - v(\hat{\pi}) \lesssim \boxed{?} \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

- In RL, "worst-case" coverage
- Pseudo-loss [Wang+2024]
- Smoothed coverage ratio (IX) [Gabbianelli+2024]

$$\max_{(a,x)} \frac{\pi(a|x)}{\pi_0(a|x)}$$

$$\mathbb{E}_{p(x)} \left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)} \right]$$

$$\mathbb{E}_{p(x)} \left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x) + \gamma} \mathbb{E}_{p(r|x,a)}[r] \right]$$



Remarks on **Coverage Measures**

$$v(\pi) - v(\hat{\pi}) \lesssim \boxed{?} \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

- In RL, "worst-case" coverage
- Pseudo-loss [Wang+2024]
- Smoothed coverage ratio (IX) [Gabbianelli+2024]
- Second moment (LS) [Sakhi+2024]

$$\max_{(a,x)} \frac{\pi(a|x)}{\pi_0(a|x)}$$

$$\mathbb{E}_{p(x)} \left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)} \right]$$

$$\mathbb{E}_{p(x)} \left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x) + \gamma} \mathbb{E}_{p(r|x,a)}[r] \right]$$

$$\mathbb{E}[(\tilde{r}_1(\pi^*))^2] = \mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)} \left[\left(\frac{\pi(a|x)}{\pi_0(a|x)} r \right)^2 \right]$$

worst

average

Remarks on **Coverage Measures**

$$v(\pi) - v(\hat{\pi}) \lesssim \boxed{?} \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

- In RL, "worst-case" coverage
- Pseudo-loss [Wang+2024]
- Smoothed coverage ratio (IX) [Gabbianelli+2024]
- Second moment (LS) [Sakhi+2024]
- Variance (EB, PUB (ours))

$$\max_{(a,x)} \frac{\pi(a|x)}{\pi_0(a|x)}$$

$$\mathbb{E}_{p(x)}\left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)}\right]$$

$$\mathbb{E}_{p(x)}\left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)+\gamma}\mathbb{E}_{p(r|x,a)}[r]\right]$$

$$\mathbb{E}[(\tilde{r}_1(\pi^*))^2] = \mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}r\right)^2\right]$$

$$\mathrm{Var}(\tilde{r}_1(\pi^*)) = \mathrm{Var}_{p(x)\pi_0(a|x)p(r|x,a)}\left(\frac{\pi(a|x)}{\pi_0(a|x)}r\right)$$

worst

average

Remarks on **Coverage Measures**

$$v(\pi) - v(\hat{\pi}) \lesssim \boxed{?} \sqrt{\frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

- In RL, "worst-case" coverage

$$\max_{(a,x)} \frac{\pi(a|x)}{\pi_0(a|x)}$$

- Pseudo-loss [Wang+2024]

$$\mathbb{E}_{p(x)}\left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)}\right]$$

- Smoothed coverage ratio (IX) [Gabbianelli+2024]

$$\mathbb{E}_{p(x)}\left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)+\gamma}\mathbb{E}_{p(r|x,a)}[r]\right]$$

- Second moment (LS) [Sakhi+2024]

$$\mathbb{E}[(\tilde{r}_1(\pi^*))^2] = \mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}r\right)^2\right]$$

- Variance (EB, PUB (ours))

$$\mathsf{Var}(\tilde{r}_1(\pi^*)) = \mathsf{Var}_{p(x)\pi_0(a|x)p(r|x,a)}\left(\frac{\pi(a|x)}{\pi_0(a|x)}r\right)$$

worst

average

IX vs. LS vs. PUB

- Smoothed coverage ratio (IX) [Gabbianelli+2024]

$$\mathbb{E}_{p(x)}\left[\sum_a \frac{\pi(a|x)}{\pi_0(a|x)+\gamma}\mathbb{E}_{p(r|x,a)}[r]\right]$$

- Second moment (LS) [Sakhi+2024]

$$\mathbb{E}[(\tilde{r}_1(\pi^*))^2] = \mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}r\right)^2\right]$$

- Variance (EB, PUB (ours))

$$\mathrm{Var}(\tilde{r}_1(\pi^*)) = \mathrm{Var}_{p(x)\pi_0(a|x)p(r|x,a)}\left(\frac{\pi(a|x)}{\pi_0(a|x)}r\right)$$

IX vs. LS vs. PUB

- Smoothed coverage ratio (IX) [Gabbianelli+2024]

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{(\pi_0(a|x)+\gamma)\tilde{r}(\pi)}\right]$$

- **Smoothed** second moment (LS) [Sakhi+2024]

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{1+\gamma\tilde{r}(\pi)}\right]$$

- **Smoothed** variance (PUB (ours))

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi)-v(\pi))^2}{1+\gamma(\tilde{r}(\pi)-v(\pi))}\right]$$

IX vs. LS vs. PUB

- Smoothed coverage ratio (IX) [Gabbianelli+2024]

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{(\pi_0(a|x)+\gamma)\tilde{r}(\pi)}\right]$$

VI

- **Smoothed** second moment (LS) [Sakhi+2024]

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{1+\gamma\tilde{r}(\pi)}\right]$$

VI

- **Smoothed** variance (PUB (ours))

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi)-v(\pi))^2}{1+\gamma(\tilde{r}(\pi)-v(\pi))}\right]$$

IX vs. LS vs. PUB

- Smoothed coverage ratio (IX) [Gabbianelli+2024]
- **Smoothed** second moment (LS) [Sakhi+2024]
- **Smoothed** variance (PUB (ours))

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{(\pi_0(a|x)+\gamma)\tilde{r}(\pi)}\right]$$

VI

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{1+\gamma\tilde{r}(\pi)}\right]$$

VI

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi)-v(\pi))^2}{1+\gamma(\tilde{r}(\pi)-v(\pi))}\right]$$

γ needs to be tuned; tuned with oracle, becomes sqrt(...)

IX vs. LS vs. PUB

- Smoothed coverage ratio (IX) [Gabbianelli+2024]
- **Smoothed** second moment (LS) [Sakhi+2024]
- **Smoothed** variance (PUB (ours))

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{(\pi_0(a|x)+\gamma)\tilde{r}(\pi)}\right]$$

VI

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi))^2}{1+\gamma\tilde{r}(\pi)}\right]$$

γ needs to be tuned; tuned with oracle, becomes sqrt(...)

VI

$$\mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{(\tilde{r}(\pi)-v(\pi))^2}{1+\gamma(\tilde{r}(\pi)-v(\pi))}\right]$$

we can achieve sqrt(...) without tuning!
(*universal portfolio* has auto-tuning feature ☺)

Technical Interlude: **On the LCB Construction via Betting**

Lower Confidence Bound

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

Lower Confidence Bound

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

Naive attempts:

Empirical Bernstein?

$$\frac{1}{n}\sum_{t=1}^{n} Y_t - \sqrt{2\widehat{\mathsf{Var}}_n \frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

Lower Confidence Bound

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

Naive attempts:

Empirical Bernstein?

$$\frac{1}{n}\sum_{t=1}^{n} Y_t - \sqrt{2\widehat{\mathsf{Var}}_n \frac{1}{n} \ln \frac{|\Pi|}{\delta}}$$

Issue 1. theoretically, analysis relies on boundedness/tail condition on Y_t

Issue 2. practically, need finite fourth moment at least

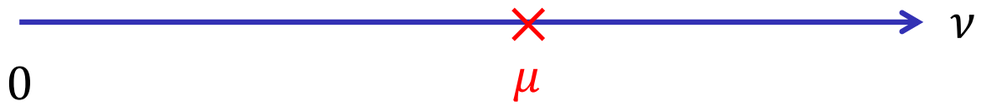
Lower Confidence Bound via Testing

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

Lower Confidence Bound via Testing

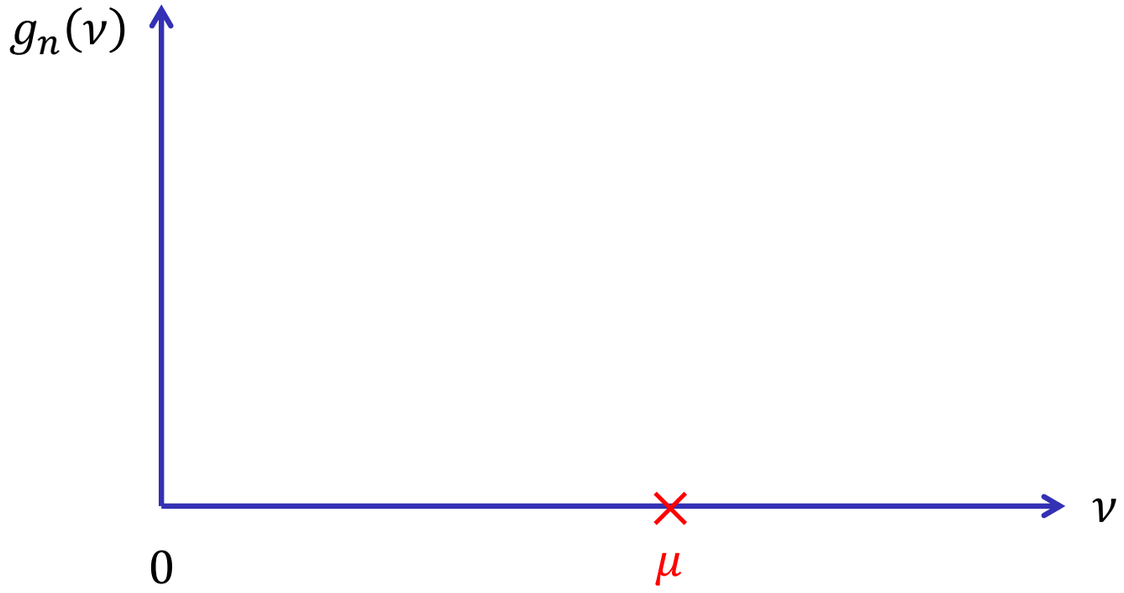
- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?



Lower Confidence Bound via Testing

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

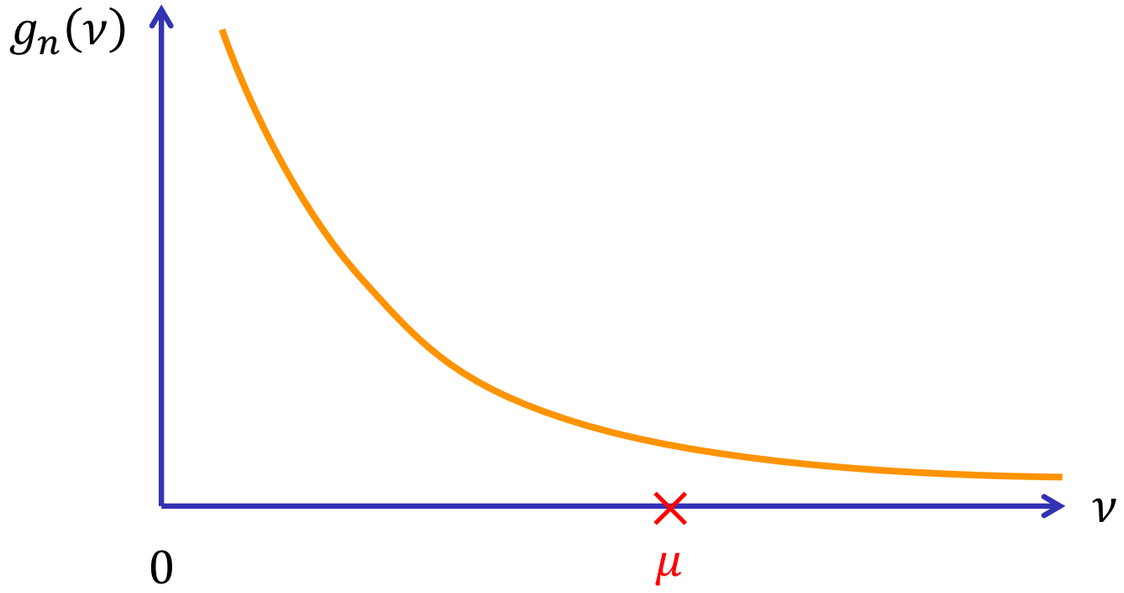


For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:

Lower Confidence Bound via Testing

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

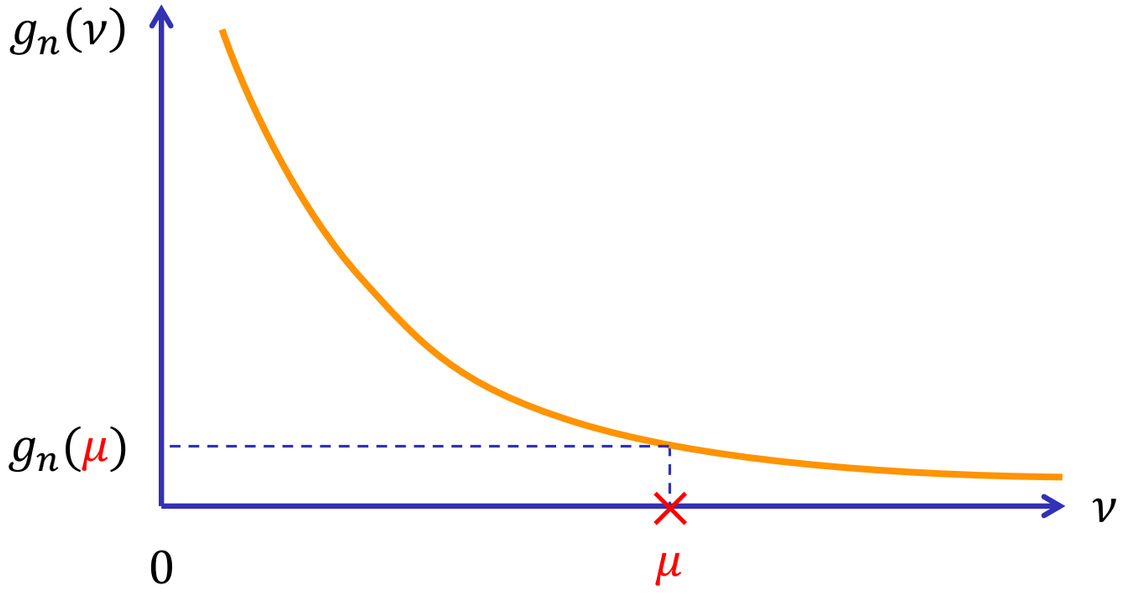


For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:
1) monotonically decreasing;

Lower Confidence Bound via Testing

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$



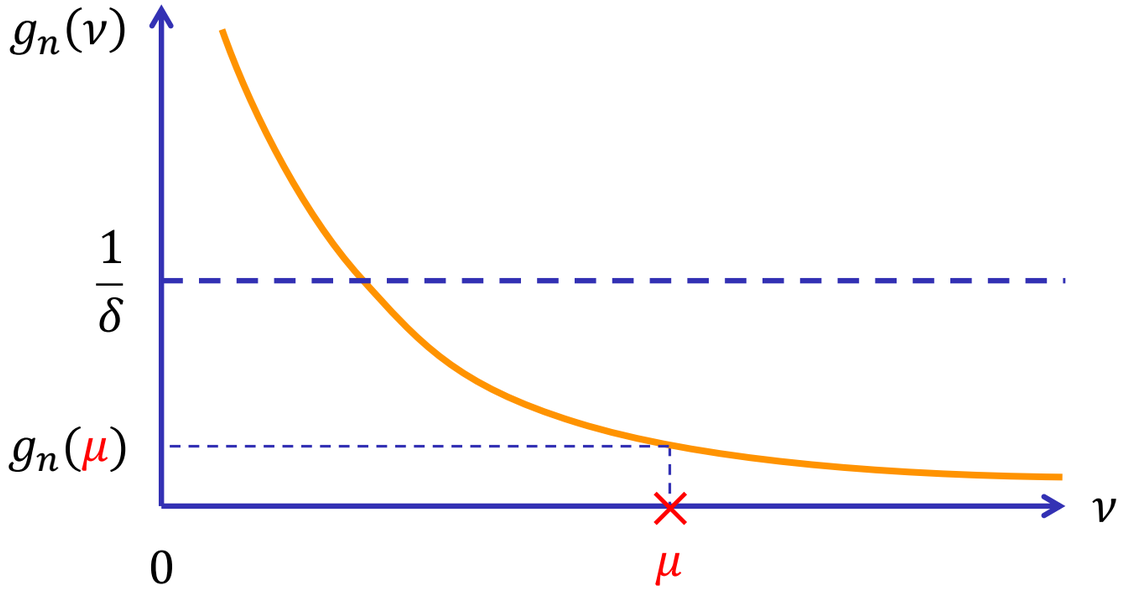
For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:
1) monotonically decreasing;

2) $g_n(\mu)$ must be "small" with high prob.

Lower Confidence Bound via Testing

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$

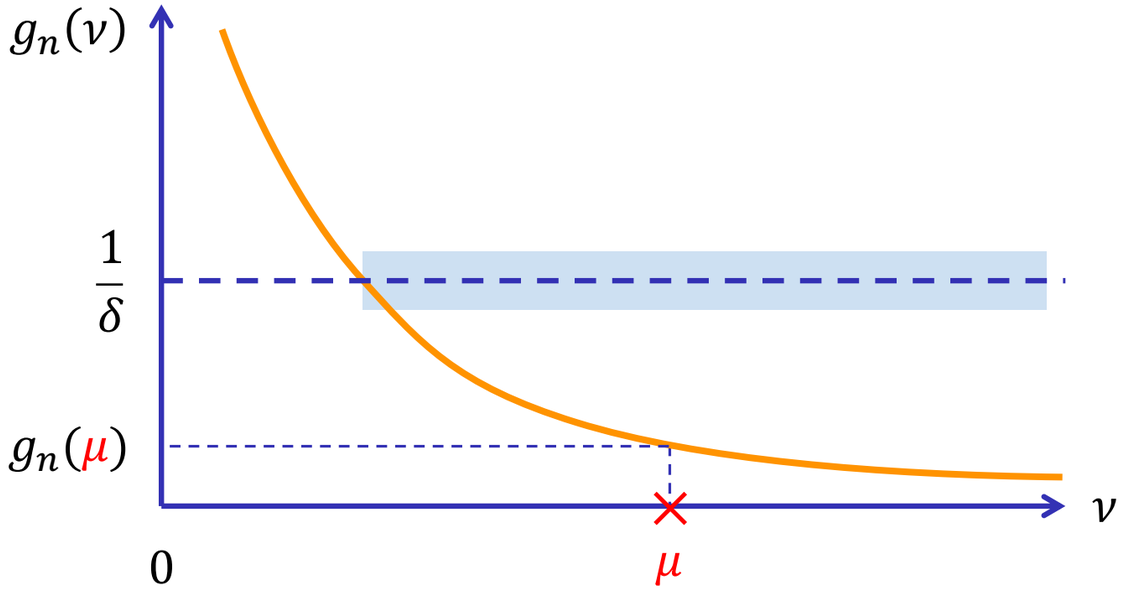


For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:
1) monotonically decreasing;
2) $g_n(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$

Lower Confidence Bound via Testing

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$



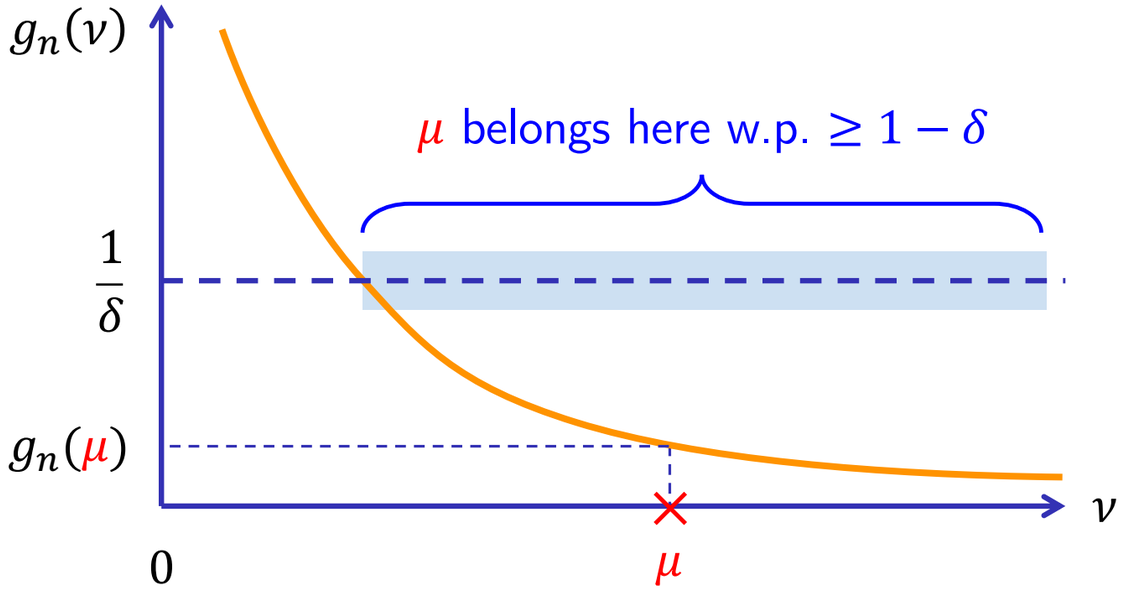
$$\left\{\nu \in (0, \infty) \;\middle|\; g_n(\nu) \leq \frac{1}{\delta}\right\}$$

For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:
1) monotonically decreasing;
2) $g_n(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$

Lower Confidence Bound via Testing

- Observe $Y_1, \ldots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$



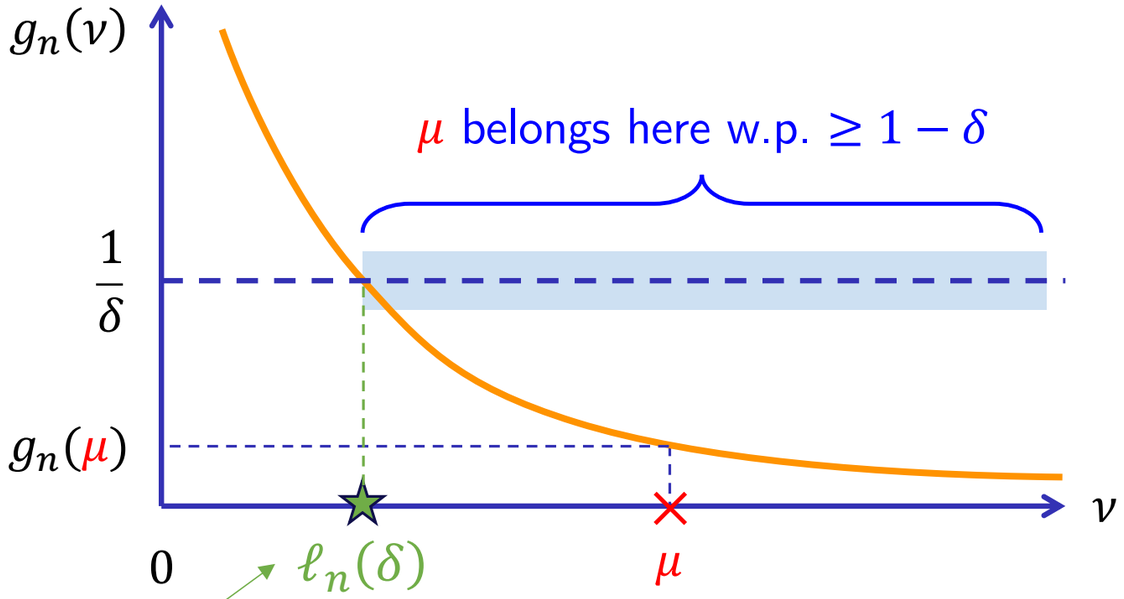
$$\left\{\nu \in (0, \infty) \;\middle|\; g_n(\nu) \leq \frac{1}{\delta}\right\}$$

For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:
1) monotonically decreasing;
2) $g_n(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$

Lower Confidence Bound via Testing

- Observe $Y_1, \dots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$



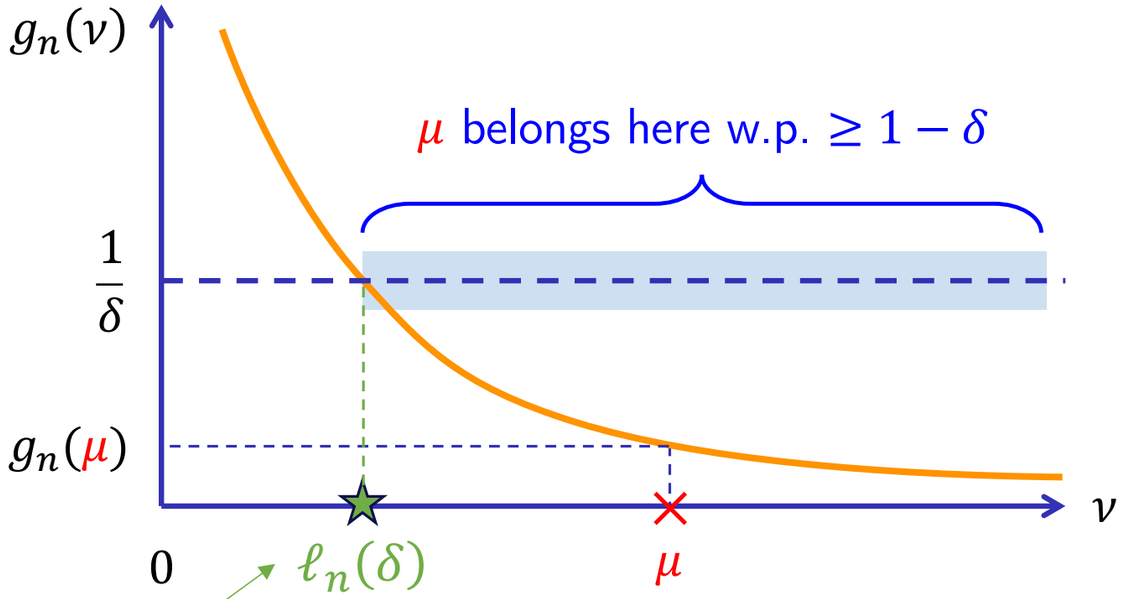
$$\stackrel{\text{def}}{=} \inf\left\{\nu \in (0, \infty) \;\middle|\; g_n(\nu) \leq \frac{1}{\delta}\right\}$$

For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:
1) monotonically decreasing;
2) $g_n(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$

Lower Confidence Bound via Testing

- Observe $Y_1, \dots, Y_n \in [0, \infty)$ with $\mathbb{E}[Y_t] = \mu$ (unknown mean) (e.g., $Y_t = \tilde{r}_t(\pi)$)
- **Want**: construct a LCB $\ell_n(\delta) = \ell_n(Y_{1:n}; \delta)$ such that $\mathbb{P}\big(\mu \geq \ell_n(\delta)\big) \geq 1 - \delta$



For each $\nu \in (0, \infty)$, **test**: is $\nu = \mu$?

Design a "potential function" $\nu \mapsto g_n(\nu)$:
1) monotonically decreasing;
2) $g_n(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$

We can **design** such a potential function as a wealth function from "betting" against a *synthetic* stock market ☺

Potential Function Design by **Betting**

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

cash 

stock 

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor 

cash 

stock 

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor 

$b_t \in [0,1]$

cash 

stock 

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor  $b_t \in [0,1]$

cash  $b_t \times \text{wealth}_{t-1}$

stock  $(1 - b_t) \times \text{wealth}_{t-1}$

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor  $b_t \in [0,1]$ $\frac{\text{price}_t}{\text{price}_{t-1}}$

cash  $b_t \times \text{wealth}_{t-1}$

stock  $(1 - b_t) \times \text{wealth}_{t-1}$

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor		$b_t \in [0,1]$	$\frac{\text{price}_t}{\text{price}_{t-1}}$
cash		$b_t \times \text{wealth}_{t-1}$	1
stock		$(1 - b_t) \times \text{wealth}_{t-1}$	$\frac{Y_t}{v}$

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor		$b_t \in [0,1]$	$\frac{\text{price}_t}{\text{price}_{t-1}}$
cash		$b_t \times \text{wealth}_{t-1}$	1
stock		$(1 - b_t) \times \text{wealth}_{t-1}$	$\frac{Y_t}{v}$

$$\frac{\text{wealth}_t(v)}{\text{wealth}_{t-1}(v)} = b_t + (1 - b_t)\frac{Y_t}{v}$$

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor		$b_t \in [0,1]$	$\frac{\text{price}_t}{\text{price}_{t-1}}$
cash		$b_t \times \text{wealth}_{t-1}$	1
stock		$(1 - b_t) \times \text{wealth}_{t-1}$	$\frac{Y_t}{v}$

$$\frac{\text{wealth}_t(v)}{\text{wealth}_{t-1}(v)} = b_t + (1 - b_t)\frac{Y_t}{v}$$

$$g_n(v) \stackrel{\text{def}}{=} \text{wealth}_n(v)$$

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

investor 	$b_t \in [0,1]$	$\frac{\text{price}_t}{\text{price}_{t-1}}$
cash 	$b_t \times \text{wealth}_{t-1}$	1
stock 	$(1 - b_t) \times \text{wealth}_{t-1}$	$\frac{Y_t}{v}$

"semi-unbounded coin betting"

$$\frac{\text{wealth}_t(v)}{\text{wealth}_{t-1}(v)} = b_t + (1 - b_t) \frac{Y_t}{v}$$

$$g_n(v) \stackrel{\text{def}}{=} \text{wealth}_n(v)$$

Potential Function Design by Betting

- $g_n(v) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b_t + (1 - b_t) \frac{Y_t}{v} \right)$ satisfies **the desiderata**:
  - 1) monotonically decreasing;
  - 2) $\left(g_t(\mu)\right)_{t=1}^{n}$ is a **martingale** $\Rightarrow \sup_{t \geq 1} g_t(\mu) \leq \frac{1}{\delta}$ with prob. $1 - \delta$ (by Ville's inequality)
- This is a "cumulative wealth" from a synthetic two-stock market setup

Historical Bits

- [W-S&R24], [OJ24] studied betting framework for bounded variables.
- [W-S+22] proposed this setup for nonnegative variables in disguise.
- [RyuB24] proposed two-stock market formulation.
- [RyuW24] proposed k-stock market variants for bounded vectors.

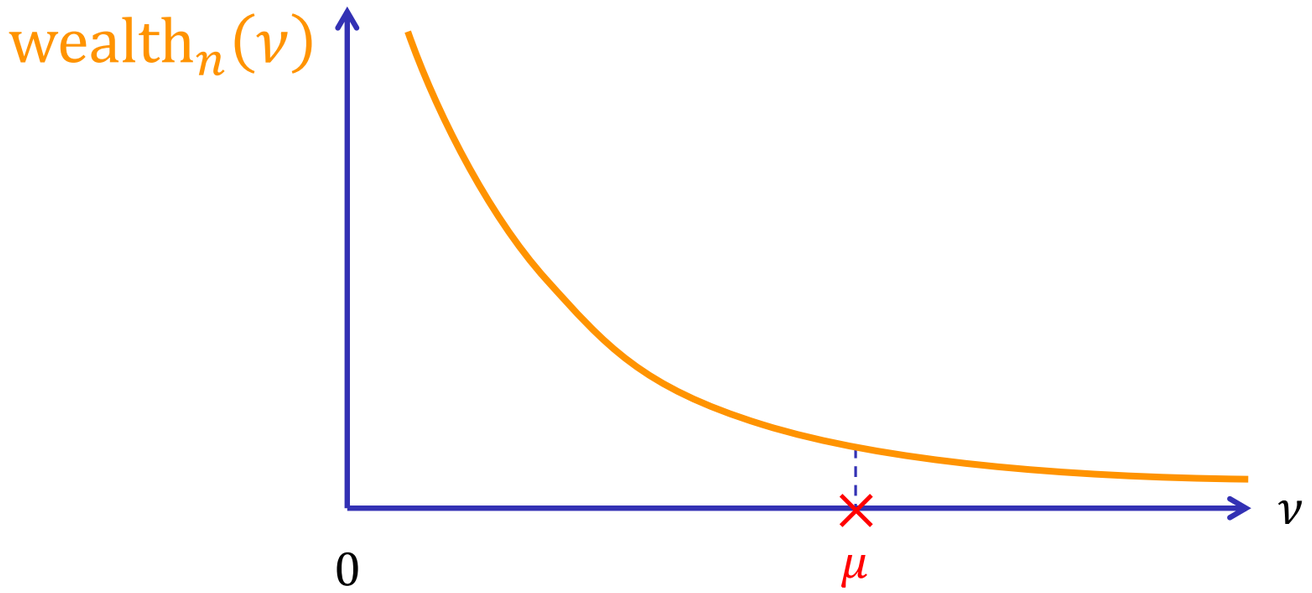
Template: **LCB via Betting**

Template: **LCB via Betting**

- **Choose a (causal) betting strategy**  $\left(b_t(Y_{1:t-1})\right)_{t\geq 1}$

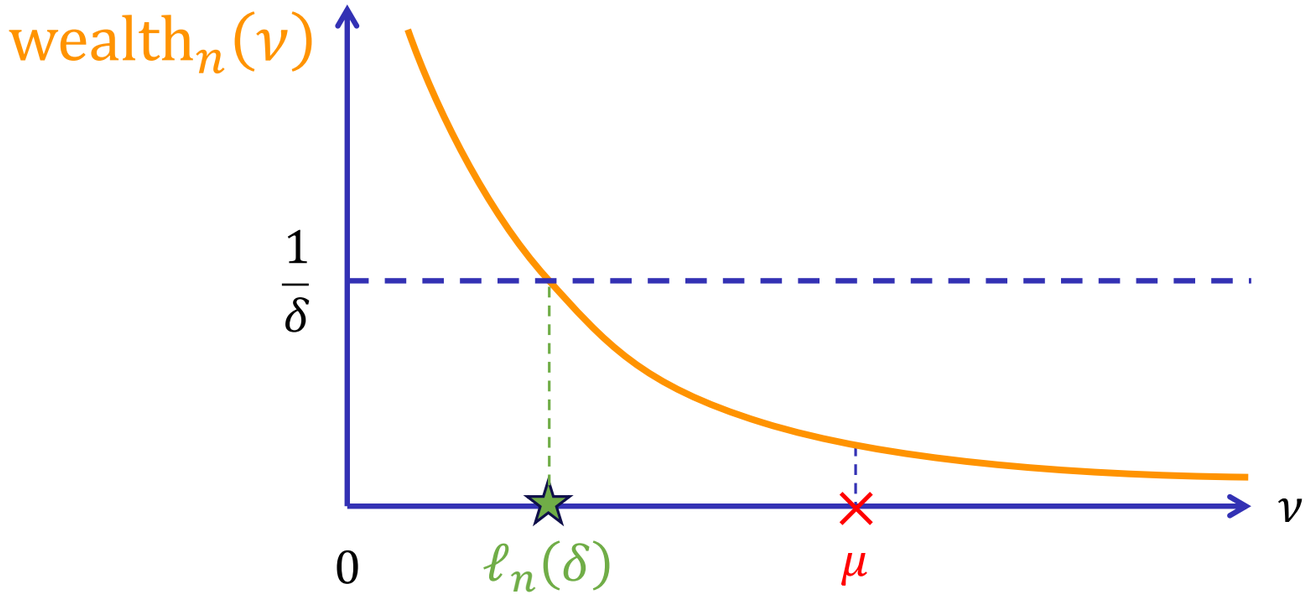
Template: **LCB via Betting**

- **Choose a (causal) betting strategy**  $\big(b_t(Y_{1:t-1})\big)_{t\geq 1}$
- Compute the induced wealth function $\nu \mapsto \mathrm{wealth}_n(\nu)$ from the two-stock market



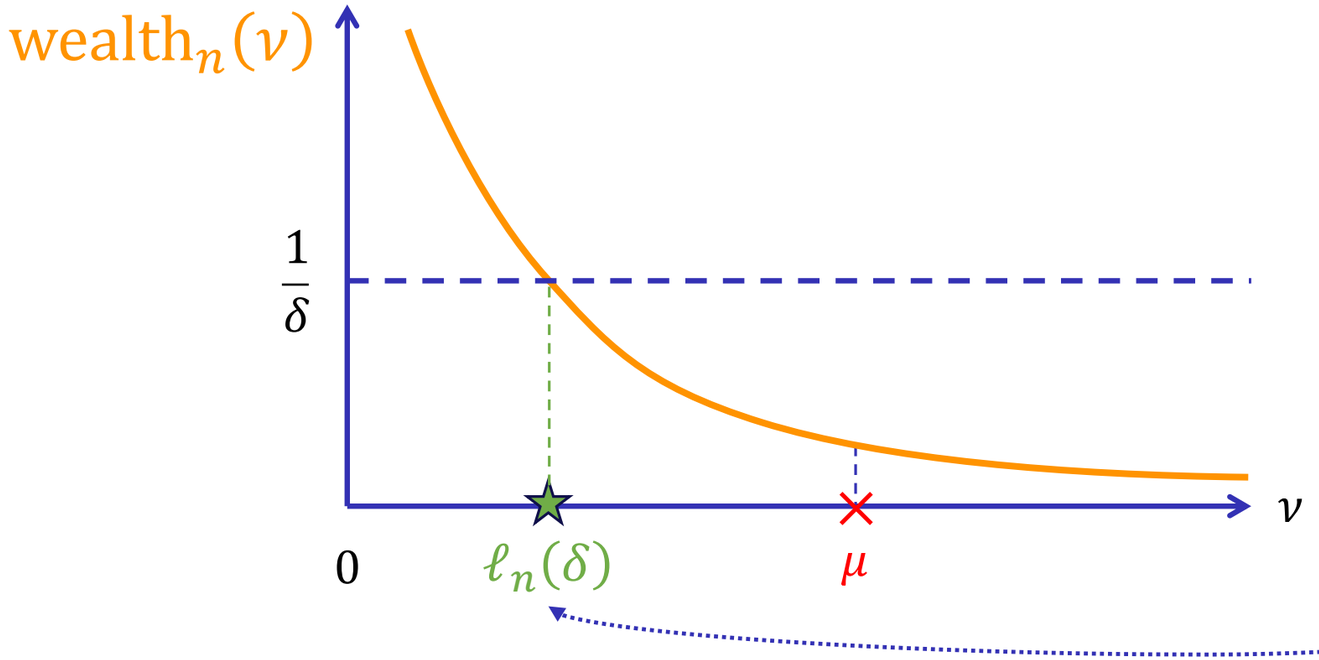
Template: **LCB via Betting**

- **Choose a (causal) betting strategy**  $\big(b_t(Y_{1:t-1})\big)_{t\geq 1}$
- Compute the induced wealth function $\nu \mapsto \mathrm{wealth}_n(\nu)$ from the two-stock market
- Construct LCB via testing with $\nu \mapsto \mathrm{wealth}_n(\nu)$



Template: **LCB via Betting**

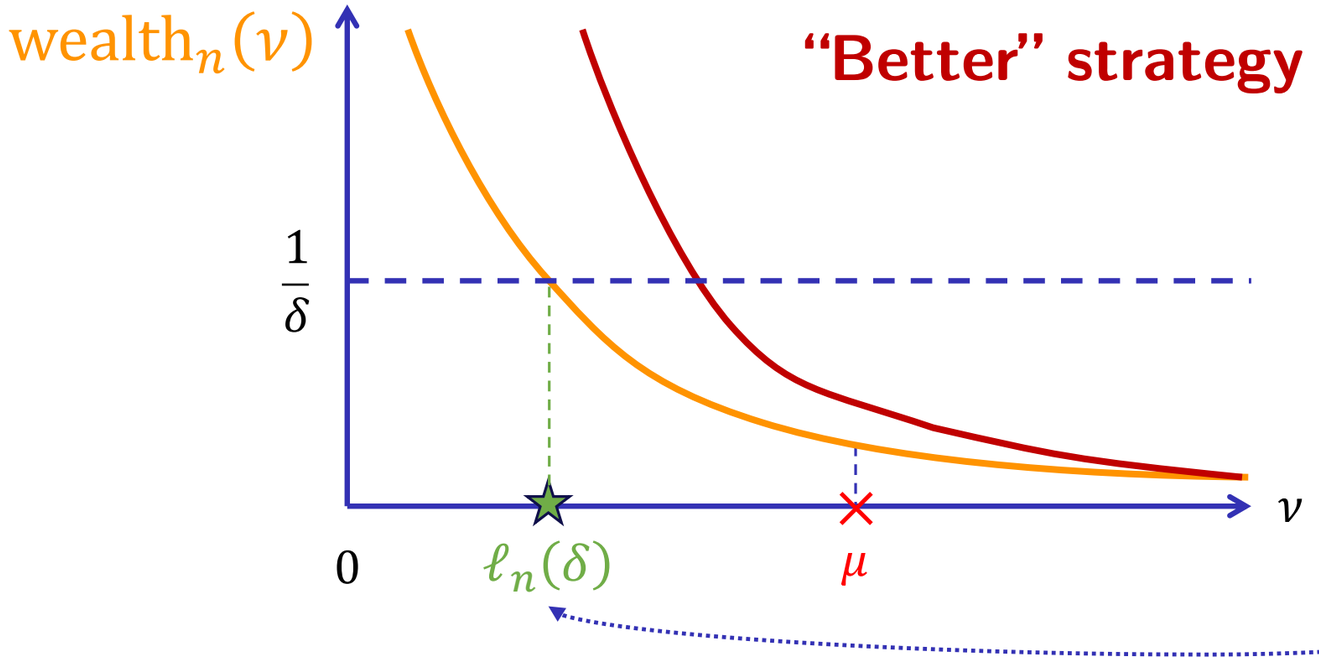
- **Choose a (causal) betting strategy**  $\big(b_t(Y_{1:t-1})\big)_{t\geq 1}$
- Compute the induced wealth function $\nu \mapsto \mathrm{wealth}_n(\nu)$ from the two-stock market
- Construct LCB via testing with $\nu \mapsto \mathrm{wealth}_n(\nu)$



Root finding can be done in $O\left(\log\frac{1}{\epsilon}\right)$
(modulo function evaluation)

Template: **LCB via Betting**

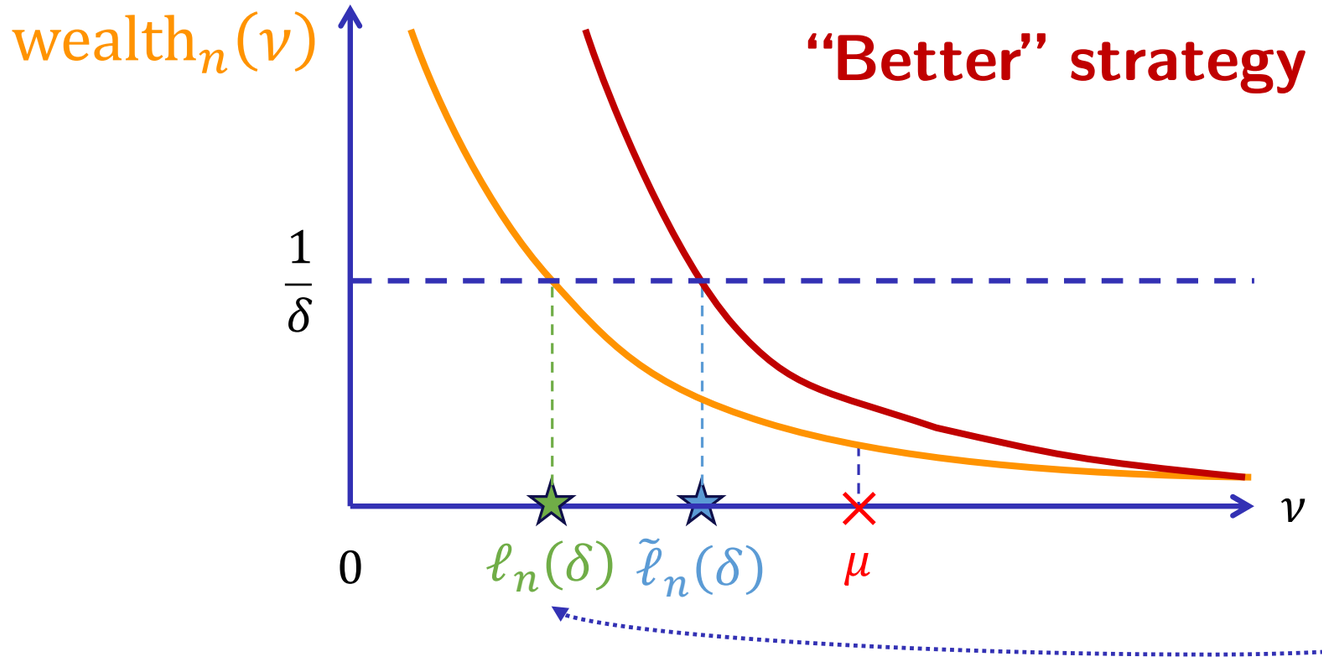
- **Choose a (causal) betting strategy**  $\big(b_t(Y_{1:t-1})\big)_{t\geq 1}$
- Compute the induced wealth function $\nu \mapsto \mathrm{wealth}_n(\nu)$ from the two-stock market
- Construct LCB via testing with $\nu \mapsto \mathrm{wealth}_n(\nu)$



Root finding can be done in $O\left(\log\frac{1}{\epsilon}\right)$ (modulo function evaluation)

Template: **LCB via Betting**

- **Choose a (causal) betting strategy**  $\left(b_t(Y_{1:t-1})\right)_{t\geq 1}$
- Compute the induced wealth function $\nu \mapsto \mathrm{wealth}_n(\nu)$ from the two-stock market
- Construct LCB via testing with $\nu \mapsto \mathrm{wealth}_n(\nu)$



"Better" strategy may lead to "tighter" LCB!

Root finding can be done in $O\left(\log\frac{1}{\epsilon}\right)$ (modulo function evaluation)

What Strategy  To Use?

What Strategy  To Use?

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b $\in$ (0,1)

What Strategy  To Use?

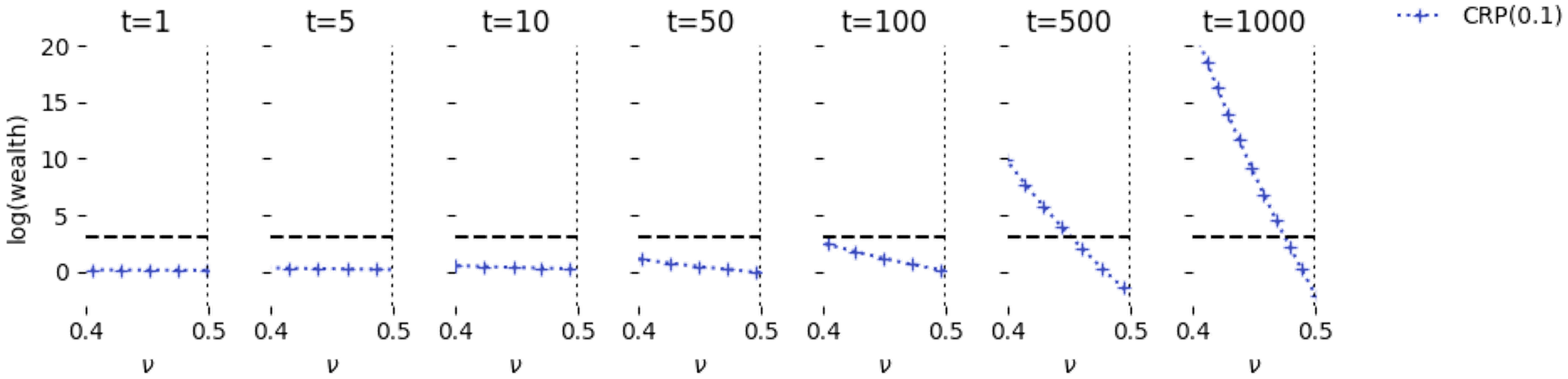
$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)

What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

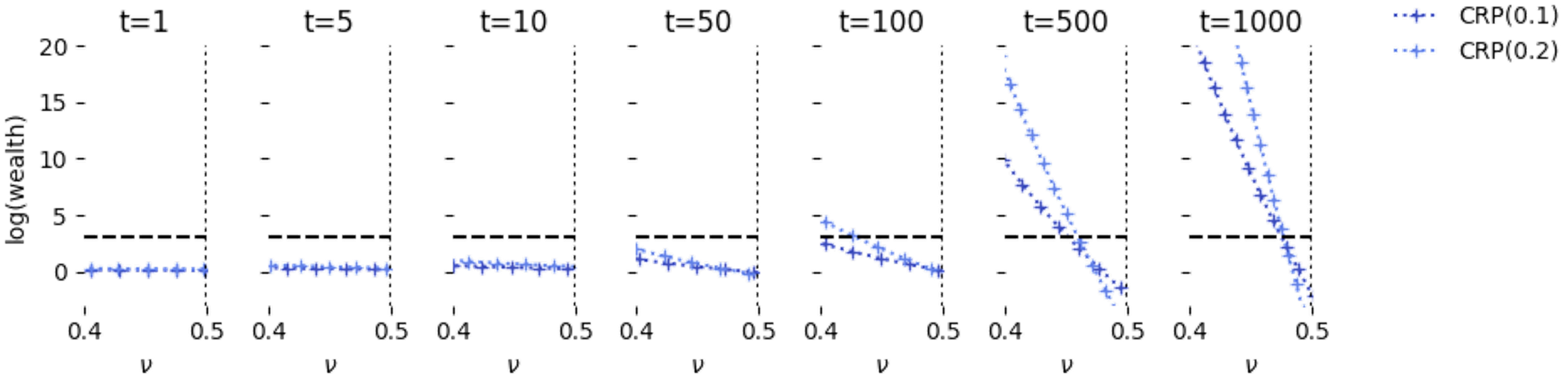
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

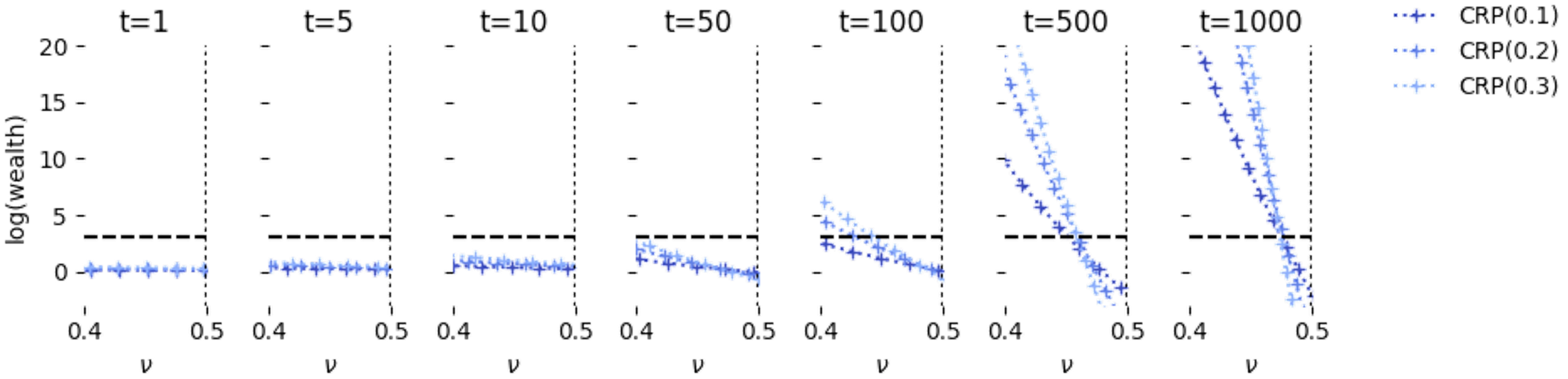
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

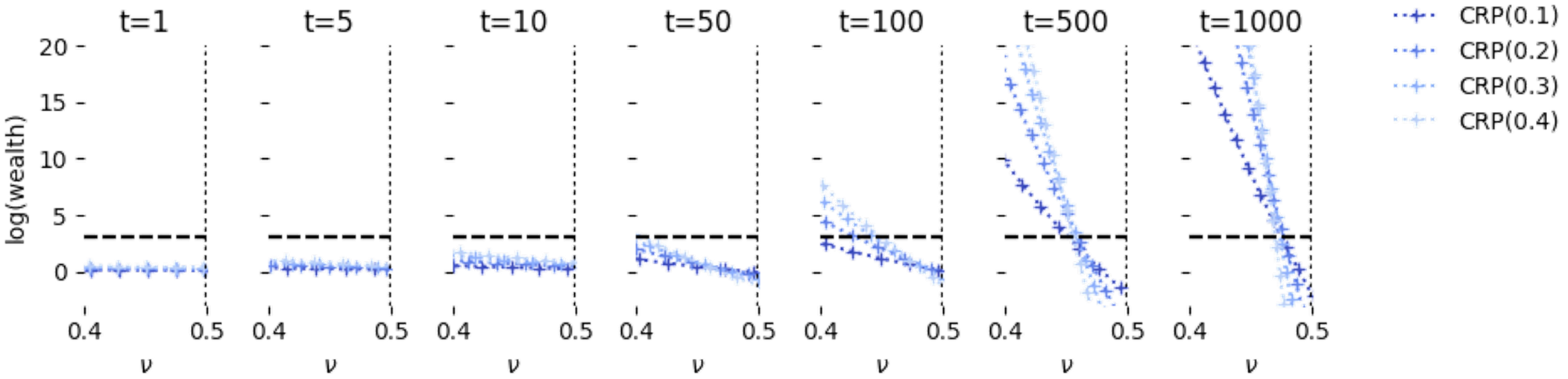
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

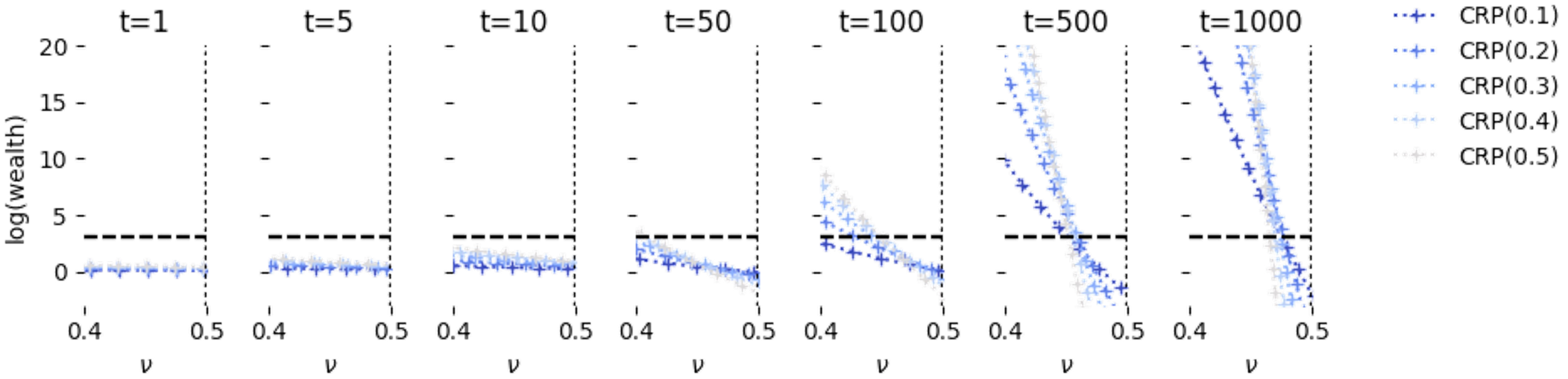
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

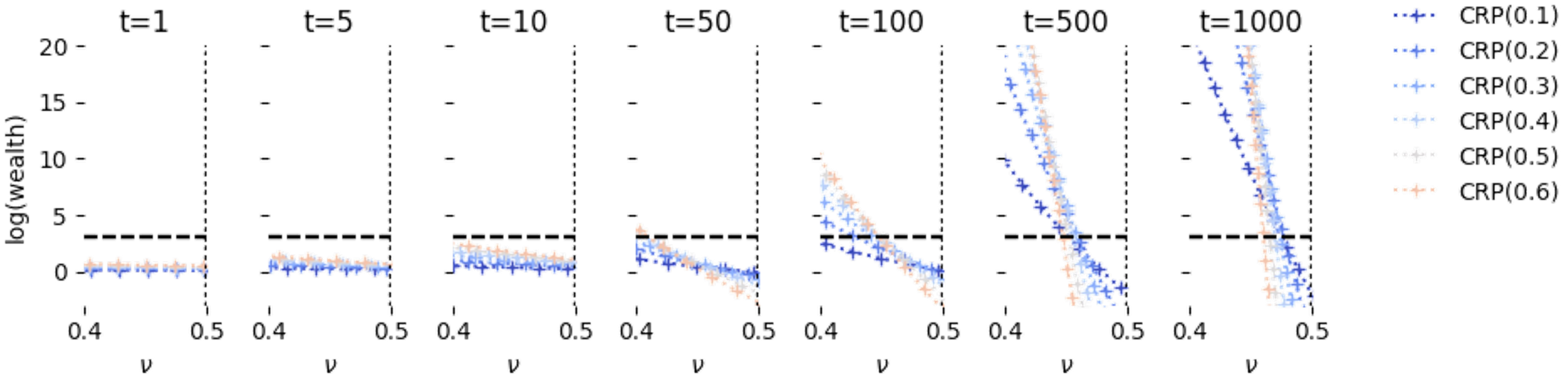
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

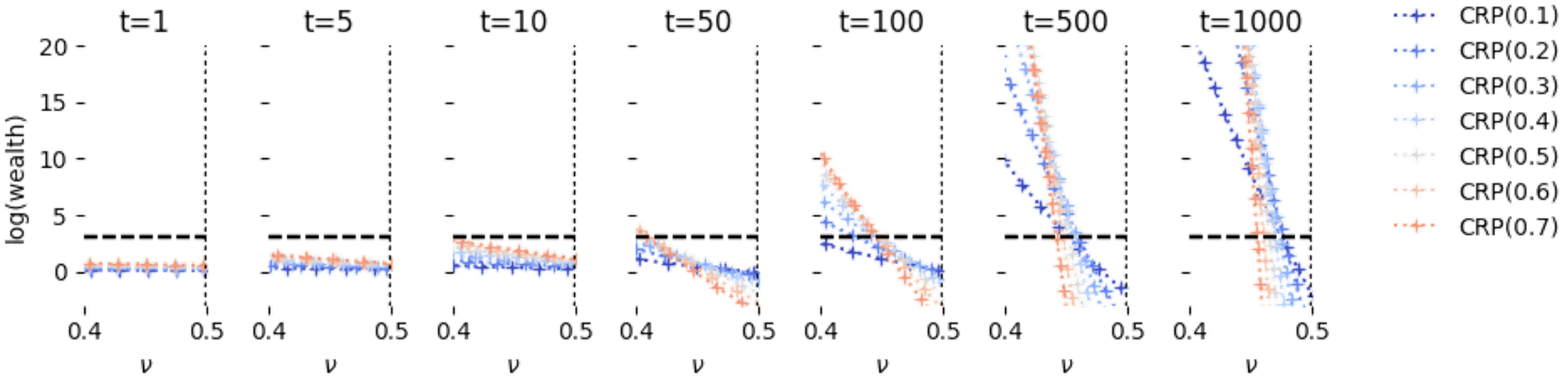
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

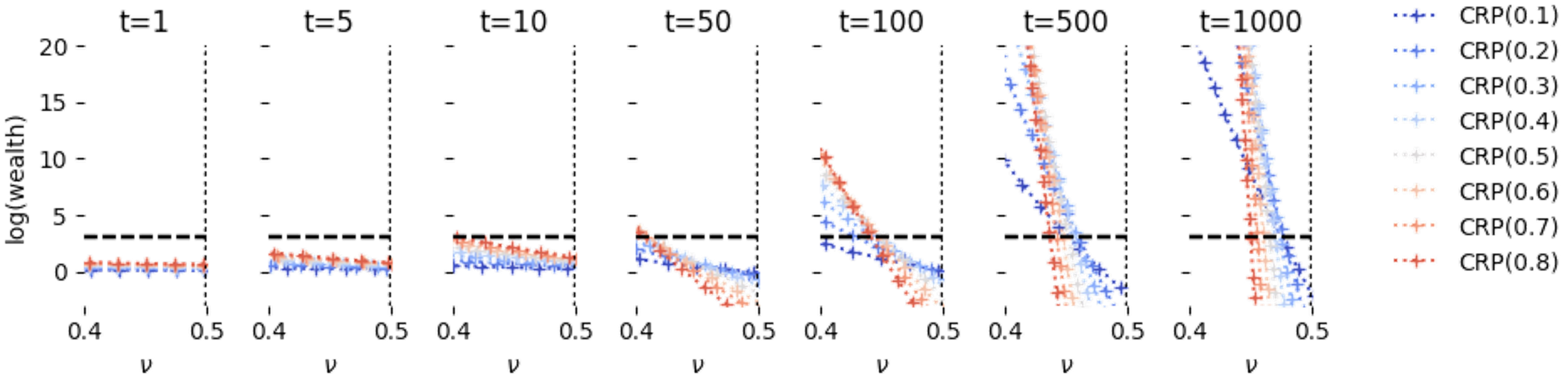
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

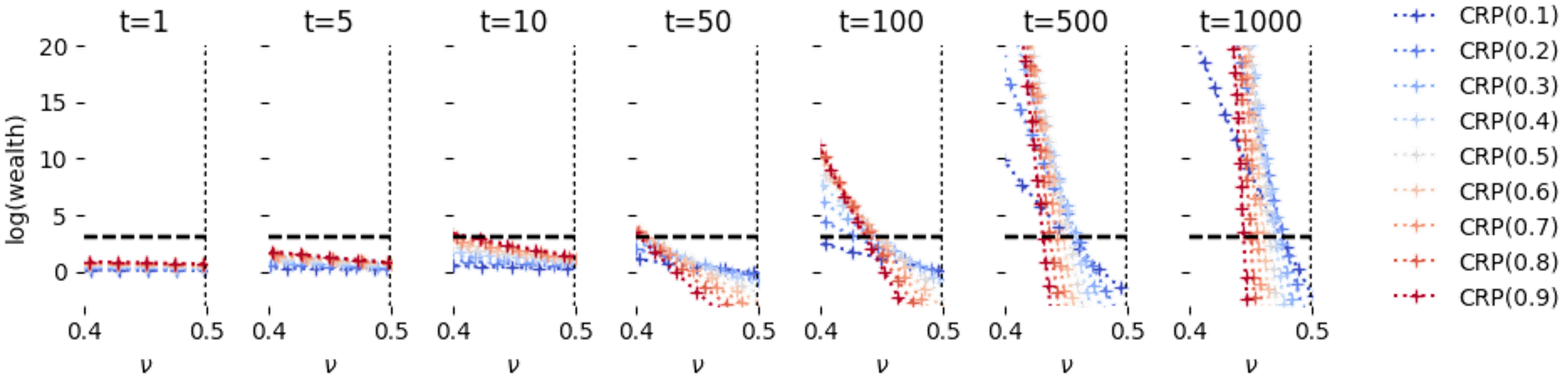
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

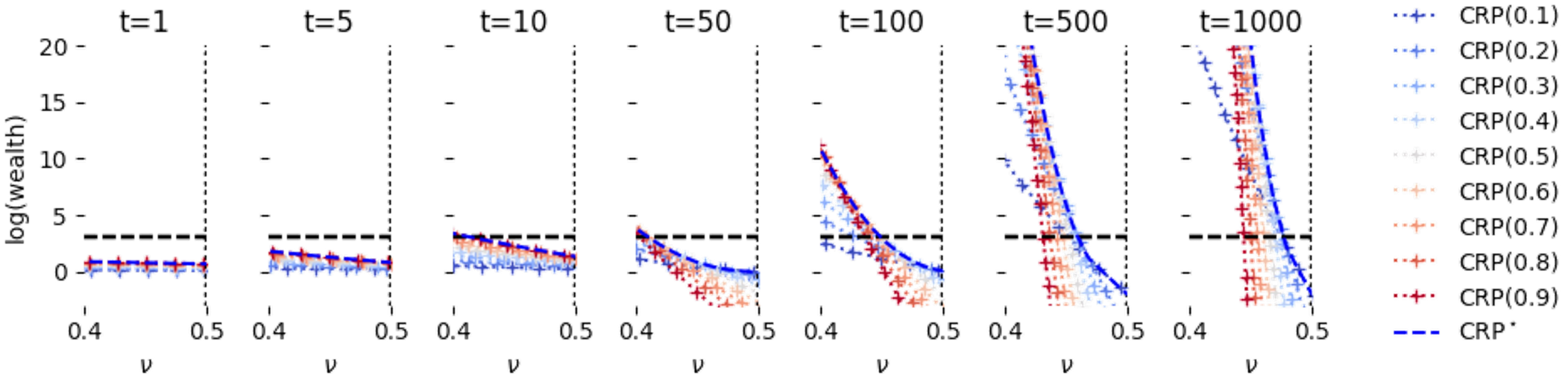
- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)

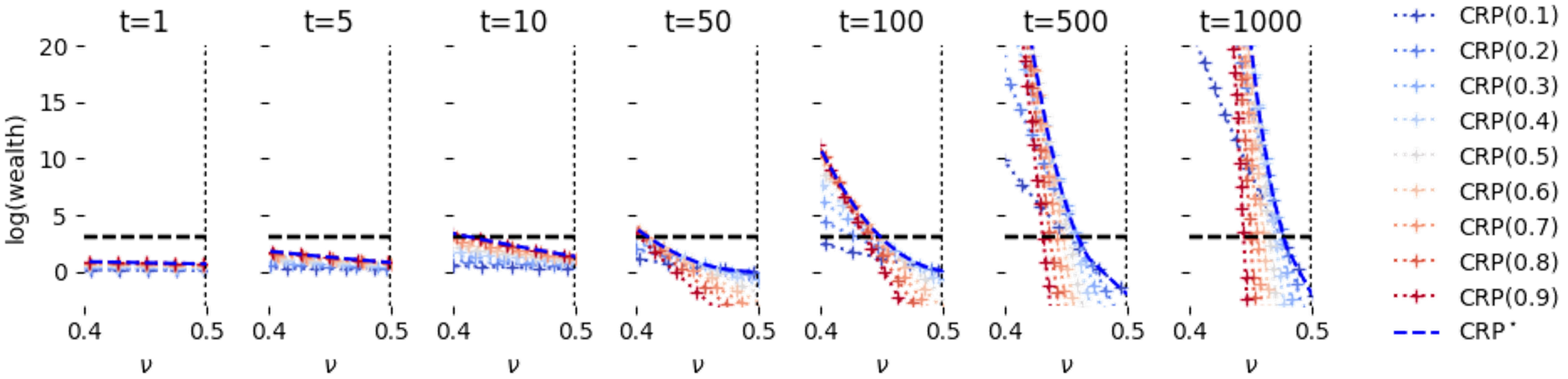


Q. Can we attain the maximum wealth of CRPs **in hindsight**? $\nu \mapsto \max_{b \in (0,1)} \text{wealth}_t^b(\nu)$

What Strategy To Use?

$$\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$$

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)

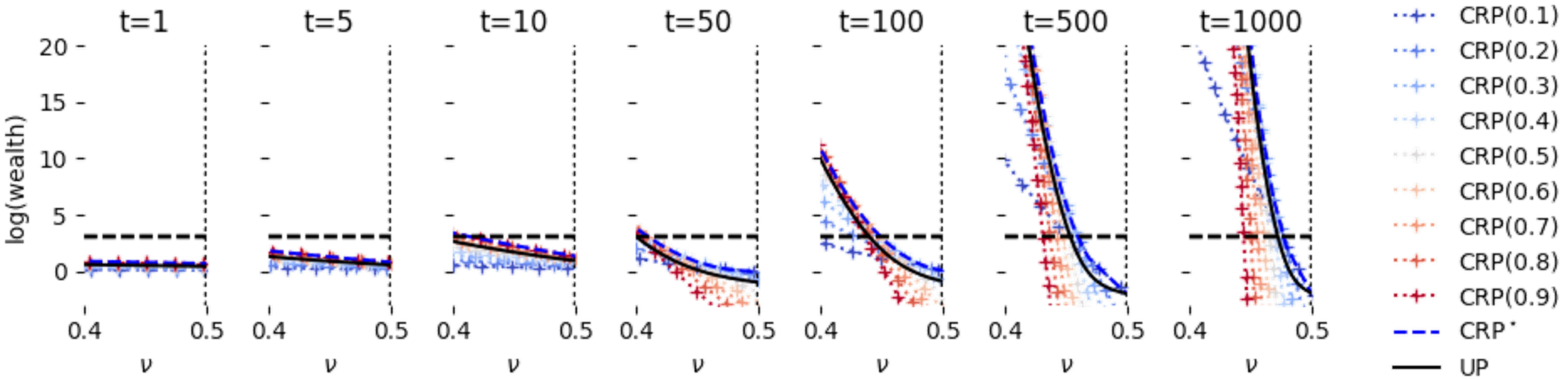


Q. Can we attain the maximum wealth of CRPs **in hindsight**? $\nu \mapsto \max_{b \in (0,1)} \text{wealth}_t^b(\nu)$

What Strategy To Use?

! $\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



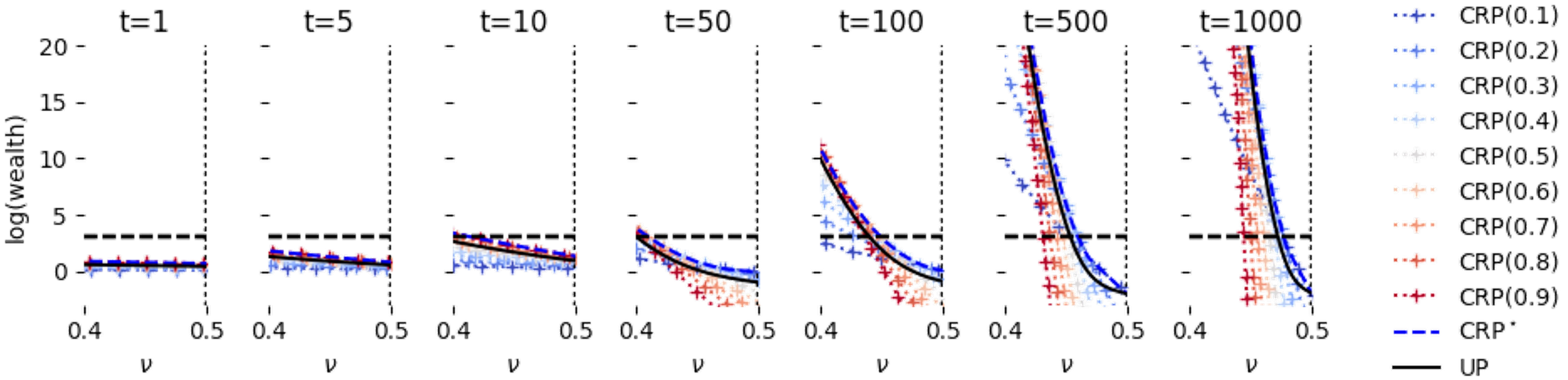
Q. Can we attain the maximum wealth of CRPs **in hindsight**? $\nu \mapsto \max_{b \in (0,1)} \text{wealth}_t^b(\nu)$

A. Yes, by [Cover91]'s **universal portfolio (UP)**, defined as mixture wealth

What Strategy To Use?

! $\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



Q. Can we attain the maximum wealth of CRPs **in hindsight**? $\nu \mapsto \max_{b \in (0,1)} \text{wealth}_t^b(\nu)$

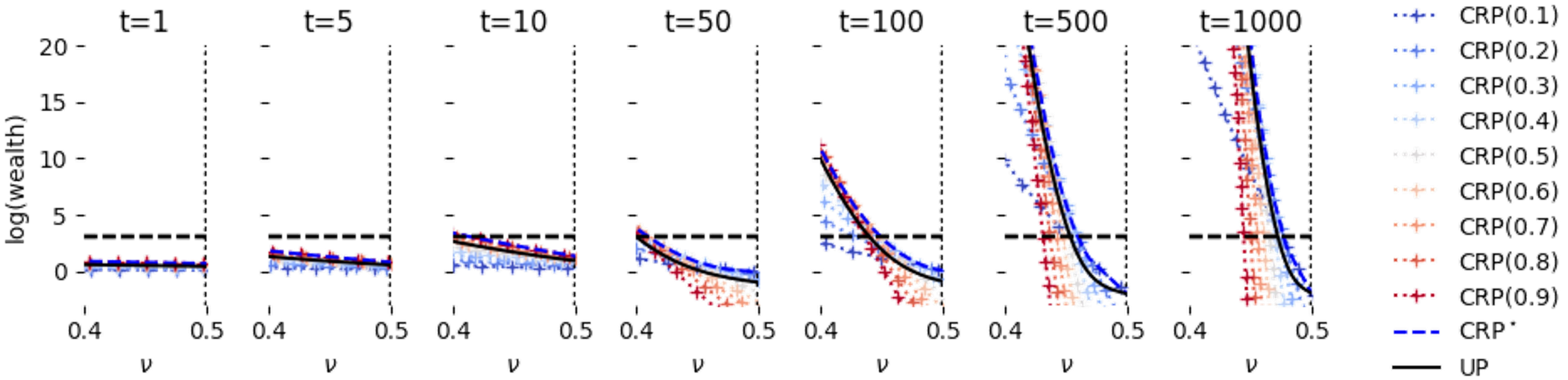
A. Yes, by [Cover91]'s **universal portfolio (UP)**, defined as mixture wealth

$$\text{wealth}_t^{\text{UP}}(\nu) \stackrel{\text{def}}{=} \int_0^1 \text{wealth}_t^b(\nu)\, \pi(b)\, db$$

What Strategy To Use?

! $\text{wealth}_t^b(\nu) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



Q. Can we attain the maximum wealth of CRPs **in hindsight**? $\nu \mapsto \max_{b \in (0,1)} \text{wealth}_t^b(\nu)$

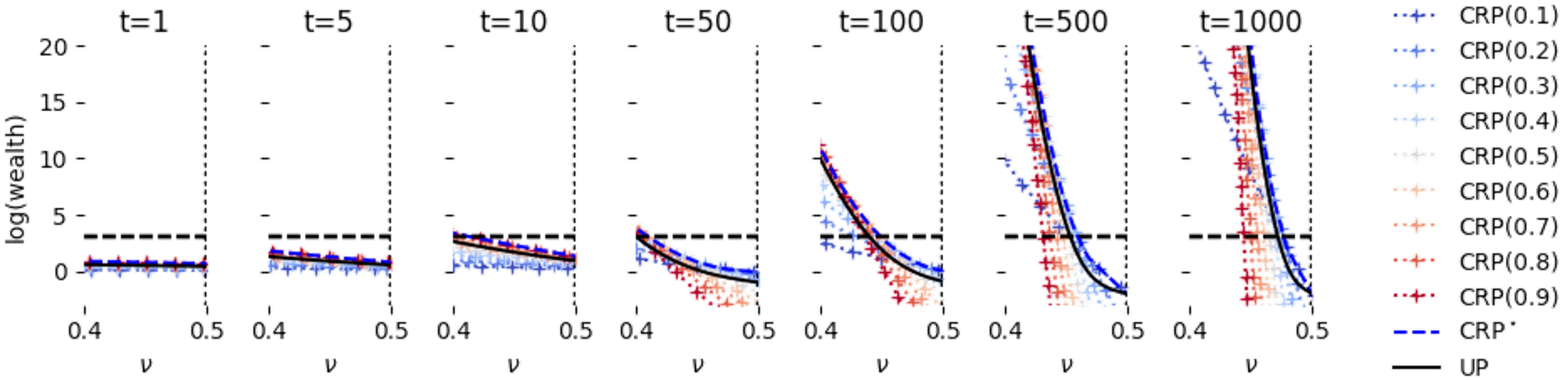
A. Yes, by [Cover91]'s **universal portfolio (UP)**, defined as mixture wealth

$$\text{wealth}_t^{\text{UP}}(\nu) \stackrel{\text{def}}{=} \int_0^1 \text{wealth}_t^b(\nu)\, \pi(b)\, db \stackrel{\text{[CoverO96]}}{\geq} \frac{1}{\sqrt{\pi(t+1)}} \max_{b \in (0,1)} \text{wealth}_t^b(\nu)$$

What Strategy To Use?

! $\text{wealth}_t^b(\nu) \overset{\text{def}}{=} \prod_{t=1}^{n} \left(b + (1-b)\frac{Y_t}{\nu} \right)$

- **Simple strategy**: Constant Rebalanced Portfolio (CRP) (b, 1-b) for b ∈ (0,1)



Q. Can we attain the maximum wealth of CRPs **in hindsight**? $\nu \mapsto \max_{b \in (0,1)} \text{wealth}_t^b(\nu)$

A. Yes, by [Cover91]'s **universal portfolio (UP)**, defined as mixture wealth

$$\text{wealth}_t^{\text{UP}}(\nu) \overset{\text{def}}{=} \int_0^1 \text{wealth}_t^b(\nu)\, \pi(b)\, db \overset{\text{[CoverO96]}}{\geq} \frac{1}{\sqrt{\pi(t+1)}} \max_{b \in (0,1)} \text{wealth}_t^b(\nu) \quad \text{[OJ24]} \quad \text{[RyuB24]} \quad \text{[RyuW24]}$$

UP-LCB

- **UP-LCB** = semi-unbounded coin betting + Cover's universal portfolio

UP-LCB

- **UP-LCB** = semi-unbounded coin betting + Cover's universal portfolio

$$\text{wealth}_n^{\text{UP}}(v) \stackrel{\text{def}}{=} \int_0^1 \text{wealth}_n^b(v)\,\pi(b)\,db$$

$$\geq \text{wealth}_n^{\text{pCRP}^\star}(v) \stackrel{\text{def}}{=} \frac{1}{\sqrt{\pi(n+1)}} \max_{b\in(0,1)} \text{wealth}_n^b(v)$$

UP-LCB

- **UP-LCB** = semi-unbounded coin betting + Cover's universal portfolio

$$\text{wealth}_n^{\text{UP}}(v) \stackrel{\text{def}}{=} \int_0^1 \text{wealth}_n^b(v)\,\pi(b)\,db \longrightarrow O(n^2)$$

$$\geq \text{wealth}_n^{\text{pCRP}^\star}(v) \stackrel{\text{def}}{=} \frac{1}{\sqrt{\pi(n+1)}} \max_{b\in(0,1)} \text{wealth}_n^b(v) \longrightarrow O(n)$$

UP-LCB

- **UP-LCB** = semi-unbounded coin betting + Cover's universal portfolio

$$\text{wealth}_n^{\text{UP}}(v) \stackrel{\text{def}}{=} \int_0^1 \text{wealth}_n^b(v)\,\pi(b)\,db \longrightarrow O(n^2)$$

$$\geq \text{wealth}_n^{\text{pCRP}^\star}(v) \stackrel{\text{def}}{=} \frac{1}{\sqrt{\pi(n+1)}} \max_{b\in(0,1)} \text{wealth}_n^b(v) \longrightarrow O(n)$$

Theorem (informal) With probability $\geq 1-\delta$,

$$0 \leq \mu - \ell_n^{\text{UP}}(\delta) \leq \mu - \ell_n^{\text{pCRP}^\star}(\delta) \lesssim \sqrt{\text{Var}(Y_1)\frac{1}{n}\log\frac{1}{\delta}}$$

UP-LCB

- **UP-LCB** = semi-unbounded coin betting + Cover's universal portfolio

$$\text{wealth}_n^{\text{UP}}(\nu) \stackrel{\text{def}}{=} \int_0^1 \text{wealth}_n^b(\nu)\,\pi(b)\,db \longrightarrow O(n^2)$$

$$\geq \text{wealth}_n^{\text{pCRP}^\star}(\nu) \stackrel{\text{def}}{=} \frac{1}{\sqrt{\pi(n+1)}} \max_{b\in(0,1)} \text{wealth}_n^b(\nu) \longrightarrow O(n)$$

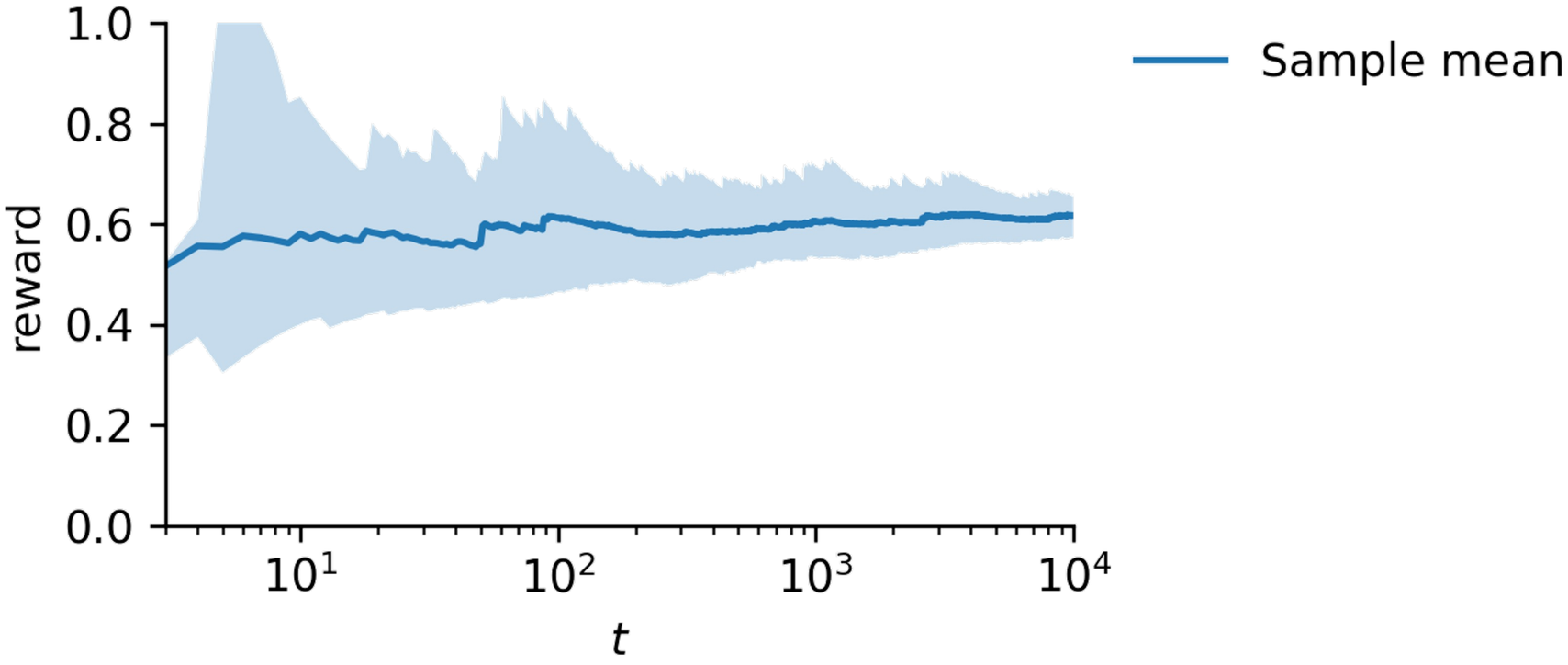
Theorem (informal) With probability $\geq 1-\delta$,

$$0 \leq \mu - \ell_n^{\text{UP}}(\delta) \leq \mu - \ell_n^{\text{pCRP}^\star}(\delta) \lesssim \sqrt{\text{Var}(Y_1)\frac{1}{n}\log\frac{1}{\delta}}$$

The variance-adaptive selection guarantee immediately follows as a corollary!

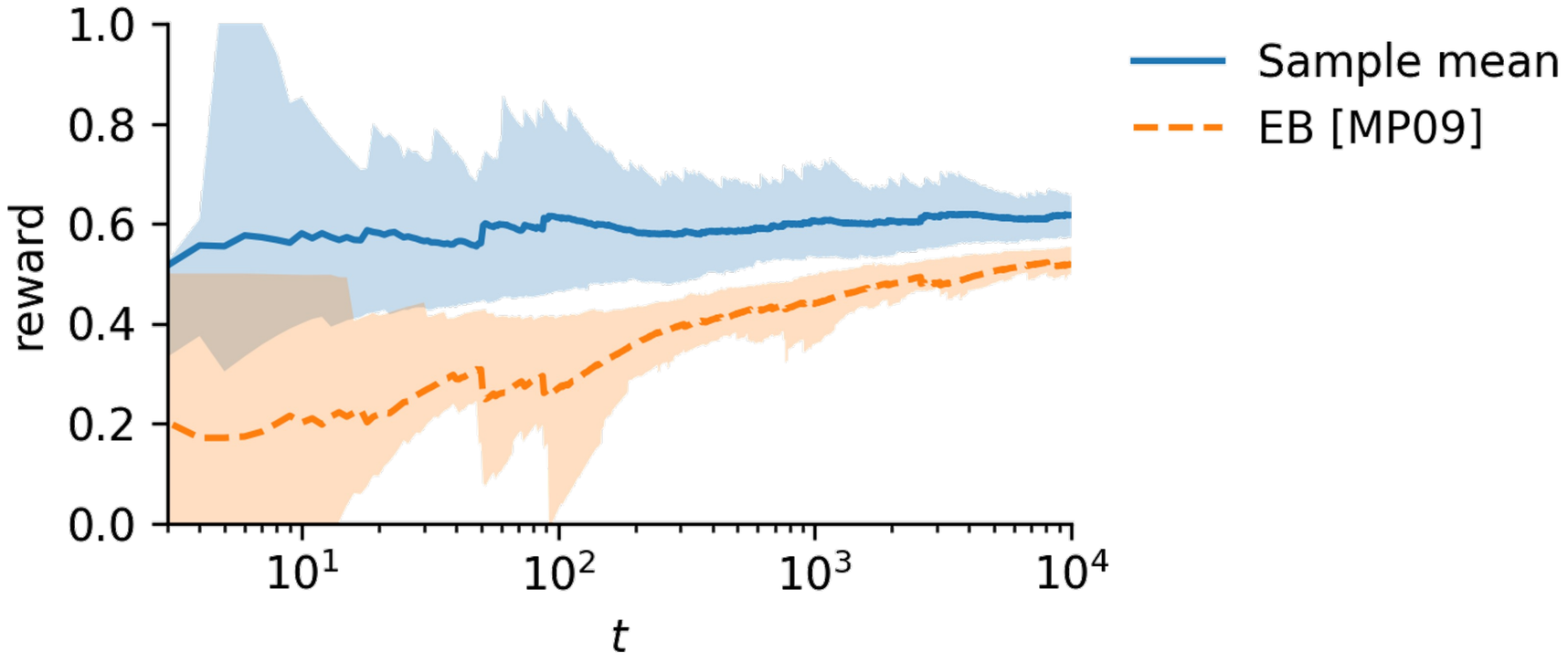
Experiment: UP-LCB

- Synthetic reward processes with **infinite** 4th moment
- Empirical Bernstein (EB) = Plug-in empirical variance (heuristic)



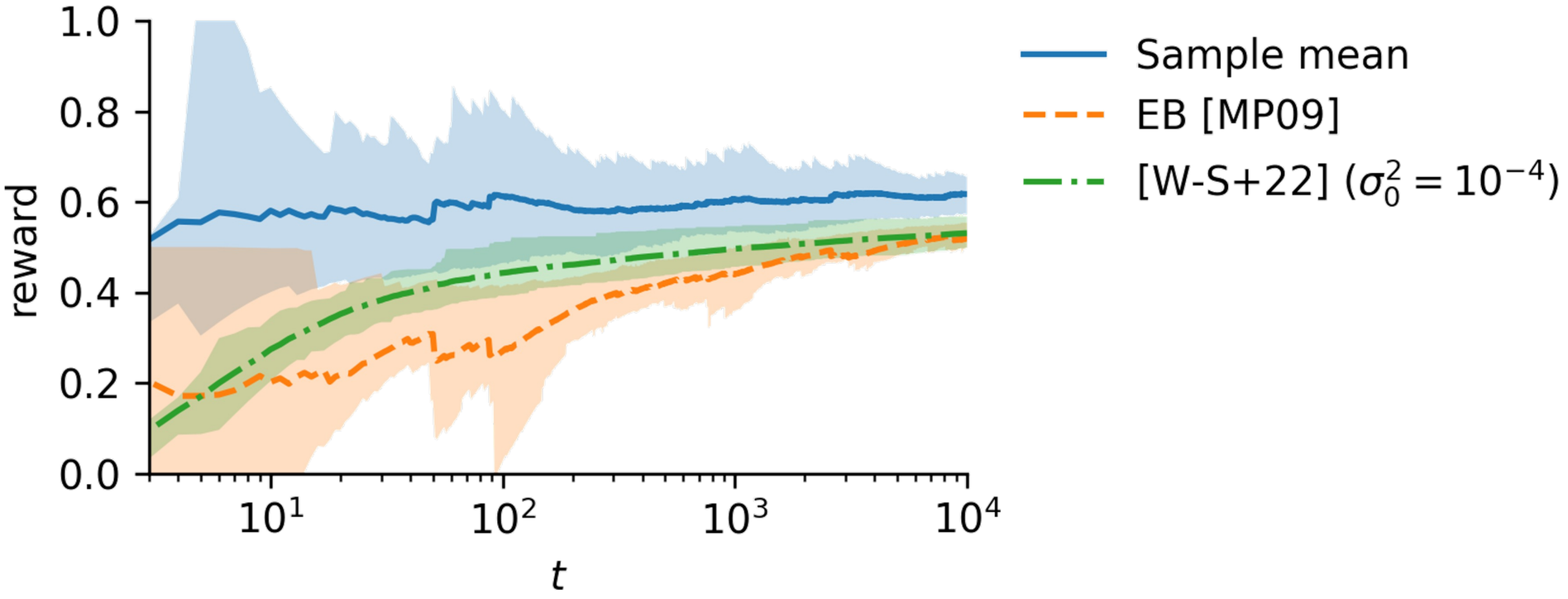
Experiment: UP-LCB

- Synthetic reward processes with **infinite** 4th moment
- Empirical Bernstein (EB) = Plug-in empirical variance (heuristic)



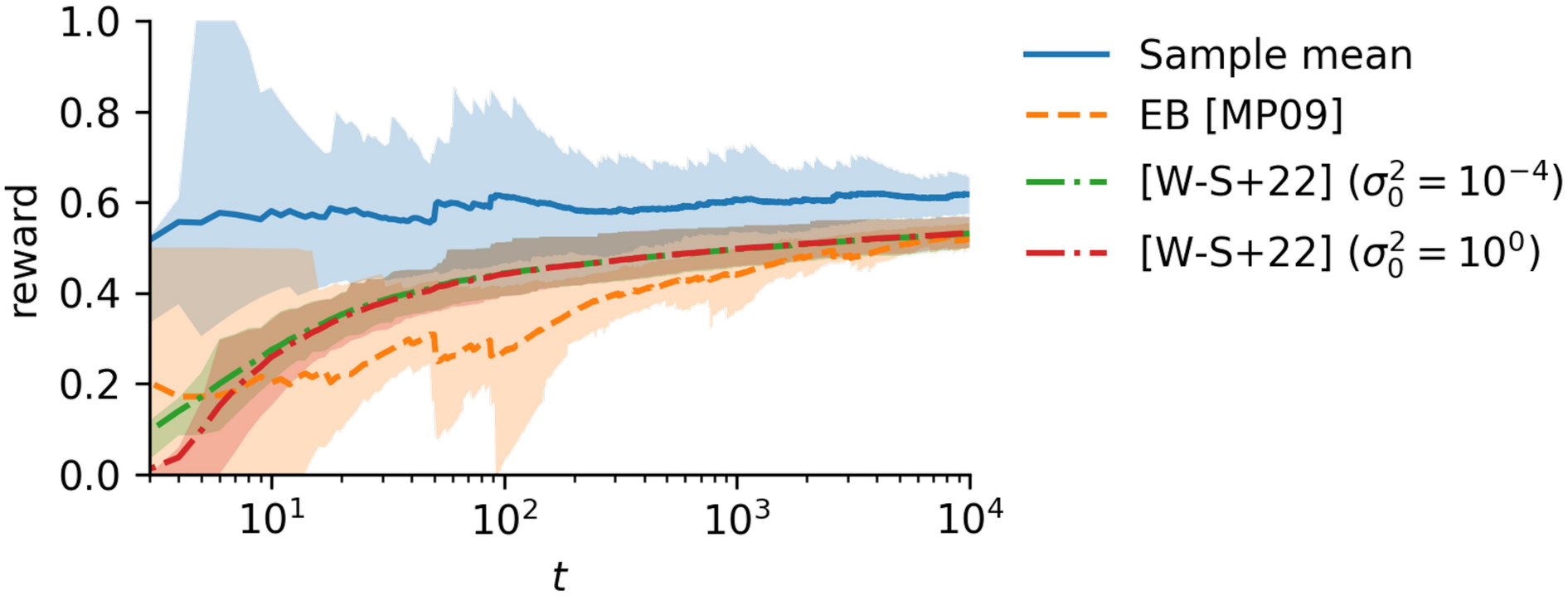
Experiment: UP-LCB

- Synthetic reward processes with **infinite** 4th moment
- Empirical Bernstein (EB) = Plug-in empirical variance (heuristic)



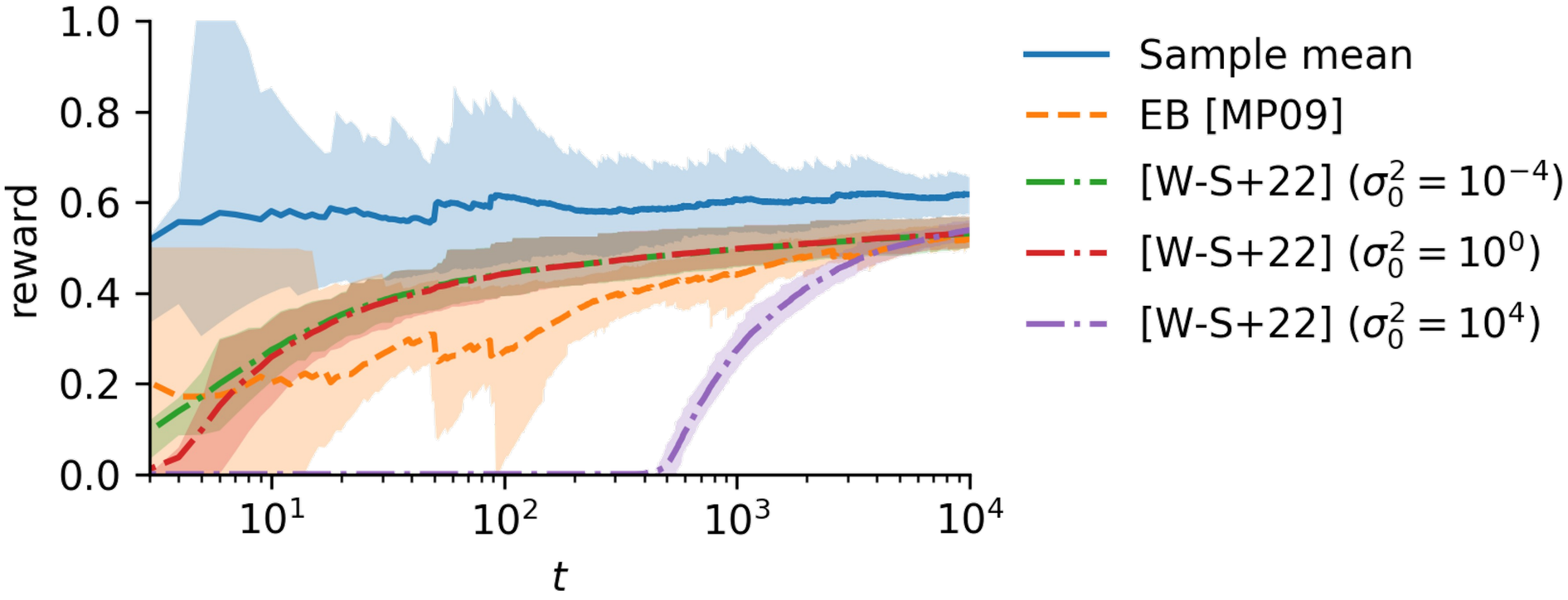
Experiment: UP-LCB

- Synthetic reward processes with **infinite** 4th moment
- Empirical Bernstein (EB) = Plug-in empirical variance (heuristic)



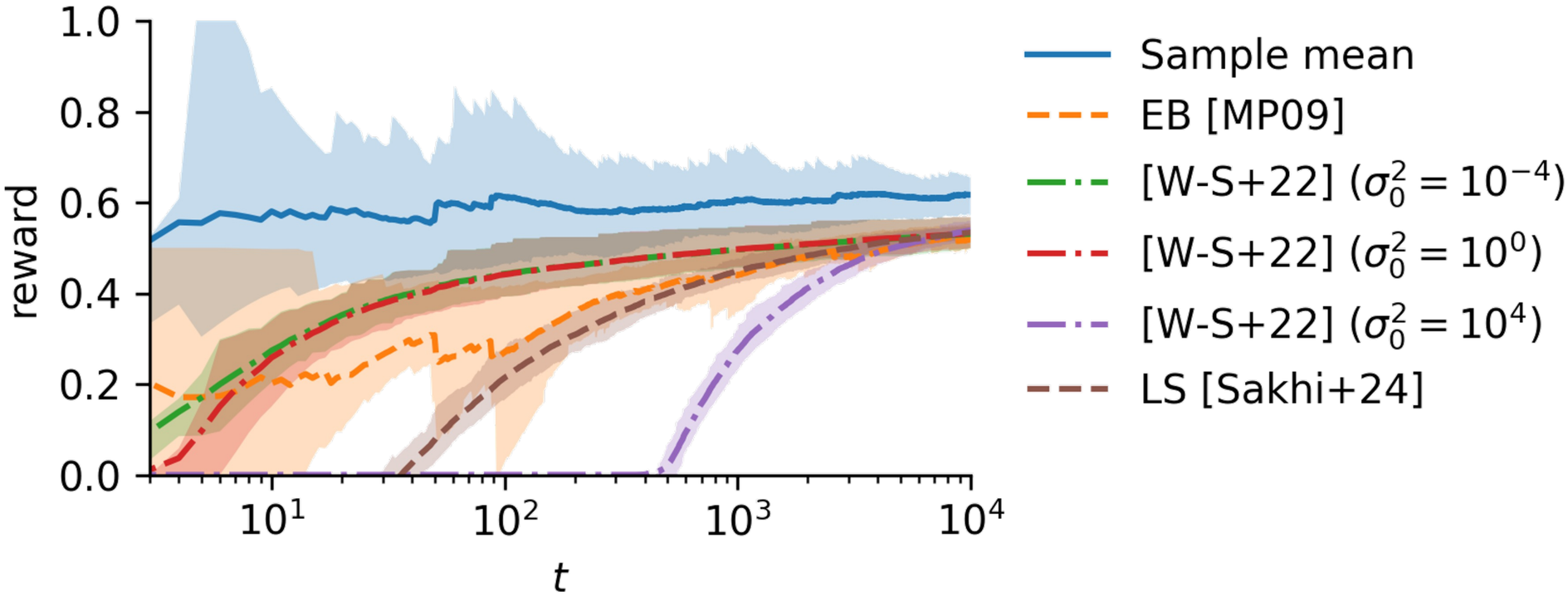
Experiment: UP-LCB

- Synthetic reward processes with **infinite** 4th moment
- Empirical Bernstein (EB) = Plug-in empirical variance (heuristic)



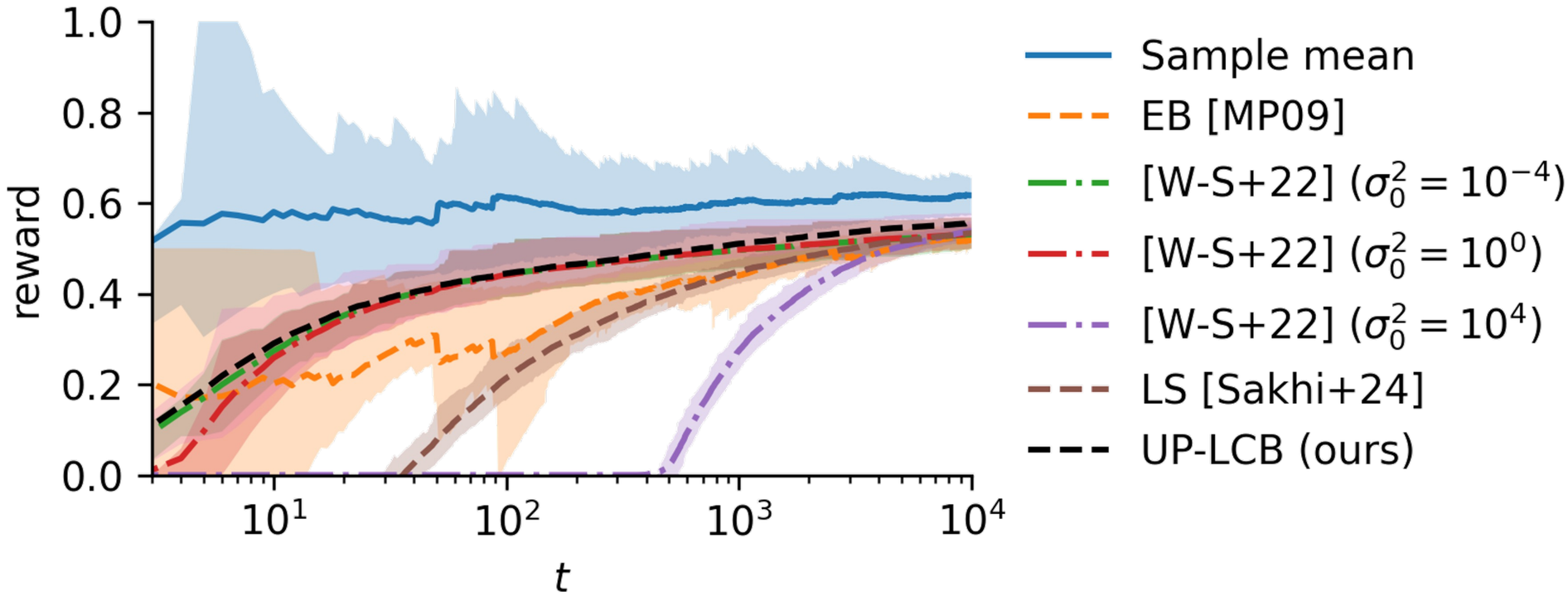
Experiment: UP-LCB

- Synthetic reward processes with **infinite** 4th moment
- Empirical Bernstein (EB) = Plug-in empirical variance (heuristic)



Experiment: UP-LCB

- Synthetic reward processes with **infinite** 4th moment
- Empirical Bernstein (EB) = Plug-in empirical variance (heuristic)



Experiment: UP-LCB + Pessimism

- Experiment setup from [WangKS2024]

Relative improvement from 'Naive' (higher = better)

Dataset	PenDigits			SatImage			JPVowel		
Size	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
LS	34.93	**23.60**	3.02	30.88	26.26	**17.84**	40.60	**27.30**	**0.87**
EB	41.34	**22.67**	**3.77**	33.76	26.79	**17.68**	**44.58**	**27.84**	**1.59**
PUB (ours)	**44.02**	**22.74**	**4.01**	**36.14**	**28.16**	**16.35**	**44.51**	**28.59**	**1.40**

[Sakhi+24] → LS

→ EB

Empirical Bernstein: A heuristic based on the central limit theorem.

Experiment: UP-LCB + Pessimism

- Experiment setup from [WangKS2024]

Relative improvement from 'Naive' (higher = better)

Dataset	PenDigits			SatImage			JPVowel		
Size	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
LS	34.93	**23.60**	3.02	30.88	26.26	**17.84**	40.60	**27.30**	**0.87**
EB	41.34	**22.67**	**3.77**	33.76	26.79	**17.68**	**44.58**	**27.84**	**1.59**
PUB (ours)	**44.02**	**22.74**	**4.01**	**36.14**	**28.16**	**16.35**	**44.51**	**28.59**	**1.40**

[Sakhi+24] → LS

→ EB

Empirical Bernstein: A heuristic based on the central limit theorem.

Always either **the best** or **statistically indistinguishable from the best**!

Part 2. Off-Policy Learning

Off-Policy Learning (=Optimization)

- Policy class is infinitely large: $|\Pi| = \infty$ (like neural networks)

Off-Policy Learning (=Optimization)

- Policy class is infinitely large: $|\Pi| = \infty$ (like neural networks)
- Using LCB-based pessimism? Root-finding via optimization is possible, but...

Off-Policy Learning (=Optimization)

- Policy class is infinitely large: $|\Pi| = \infty$ (like neural networks)
- Using LCB-based pessimism? Root-finding via optimization is possible, but...

$$\max_{\pi\in\Pi}\max_{\alpha\geq 0}\min_{\nu}\left\{\nu+\alpha\left(\frac{1}{\sqrt{\pi(n+1)}}\max_{b\in(0,1)}\log \mathsf{wealth}_t^b(\nu;\pi)-\log\frac{|\Pi|}{\delta}\right)\right\}$$

Off-Policy Learning (=Optimization)

- Policy class is infinitely large: $|\Pi| = \infty$ (like neural networks)
- Using LCB-based pessimism? Root-finding via optimization is possible, but...

$$\max_{\pi\in\Pi}\max_{\alpha\geq 0}\min_{\nu}\left\{\nu+\alpha\left(\frac{1}{\sqrt{\pi(n+1)}}\max_{b\in(0,1)}\log \mathsf{wealth}_t^b(\nu;\pi)-\log\frac{|\Pi|}{\delta}\right)\right\}$$

DIFFICULT!

Off-Policy Learning (=Optimization)

- Policy class is infinitely large: $|\Pi| = \infty$ (like neural networks)
- Using LCB-based pessimism? Root-finding via optimization is possible, but...

$$\max_{\pi\in\Pi}\max_{\alpha\geq 0}\min_{\nu}\left\{\nu+\alpha\left(\frac{1}{\sqrt{\pi(n+1)}}\max_{b\in(0,1)}\log \mathsf{wealth}_t^b(\nu;\pi)-\log\frac{|\Pi|}{\delta}\right)\right\}$$

DIFFICULT!

- **Oracle-efficiency**: objective is **optimizable** efficiently with an optimization oracle

Off-Policy Learning (=Optimization)

- **Our proposal**: A family of objectives achieving the SOTA guarantee

$$\max_{\pi \in \Pi} \sum_{t=1}^{n} \phi(\beta \tilde{r}_t(\pi))$$

hyperparameter

Off-Policy Learning (=Optimization)

- **Our proposal**: A family of objectives achieving the SOTA guarantee

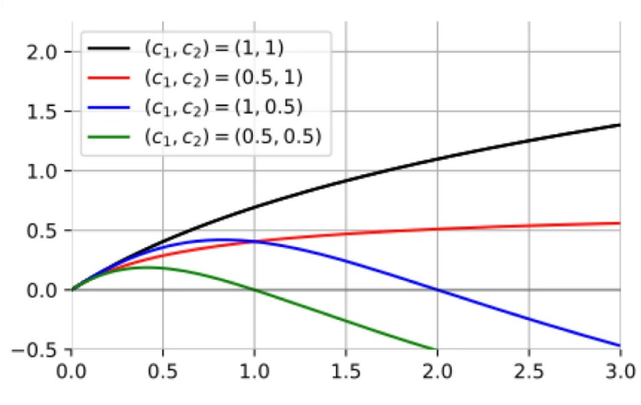
$$\max_{\pi \in \Pi} \sum_{t=1}^{n} \phi(\beta \tilde{r}_t(\pi))$$

hyperparameter

Observation: any $\phi(\cdot)$ satisfying

$$-\log\left(1 - x + \frac{x^2}{c_1 + c_2 x}\right) \leq \phi(x) \leq \log(1+x)$$

implements pessimism ($\approx$ underestimation)



Off-Policy Learning (=Optimization)

- **Our proposal**: A family of objectives achieving the SOTA guarantee

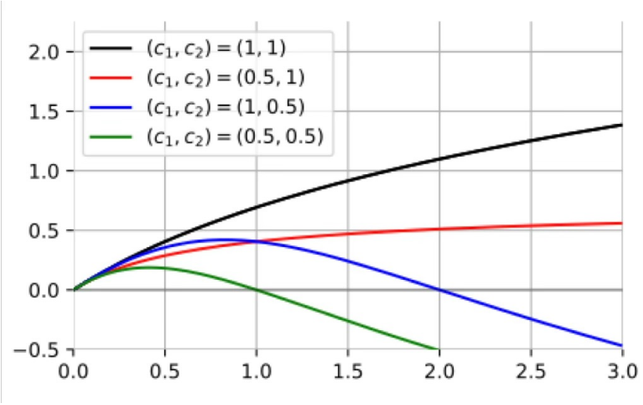
$$\max_{\pi\in\Pi} \sum_{t=1}^{n} \phi(\beta \tilde{r}_t(\pi))$$

hyperparameter

Observation: any $\phi(\cdot)$ satisfying

$$-\log\left(1 - x + \frac{x^2}{c_1 + c_2 x}\right) \le \phi(x) \le \log(1+x)$$

implements pessimism ($\approx$ underestimation)



Logarithmic smoothing: [Sakhi+24] $\quad \phi(x) = \log(1+x)$

Freezing: (Ours) $\quad \phi(x) = \log(1 + \mathbb{I}\{x \le 1\} x)$

more aggressively reduces variance;
advantageous with small sample size

Off-Policy Learning (=Optimization)

- **Our proposal**: A family of objectives achieving the SOTA guarantee

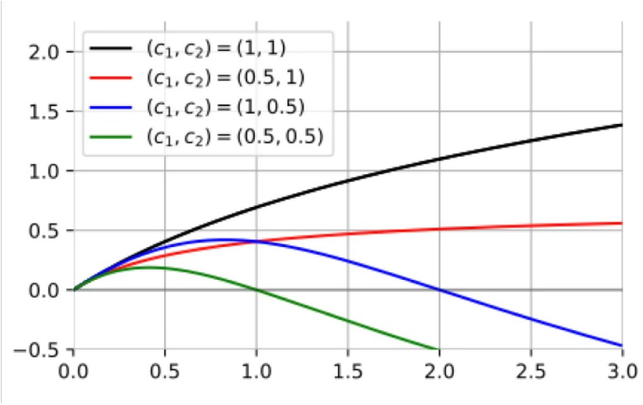
$$\max_{\pi\in\Pi} \sum_{t=1}^{n} \phi(\beta \tilde{r}_t(\pi))$$

hyperparameter

Observation: any $\phi(\cdot)$ satisfying

$$-\log\left(1 - x + \frac{x^2}{c_1 + c_2 x}\right) \le \phi(x) \le \log(1+x)$$

implements pessimism (≈ underestimation)



Logarithmic smoothing: [Sakhi+24] $\quad \phi(x) = \log(1+x)$

Freezing: (Ours) $\quad \phi(x) = \log(1 + \mathbb{I}\{x \le 1\} x)$

more aggressively reduces variance; advantageous with small sample size

Theorem

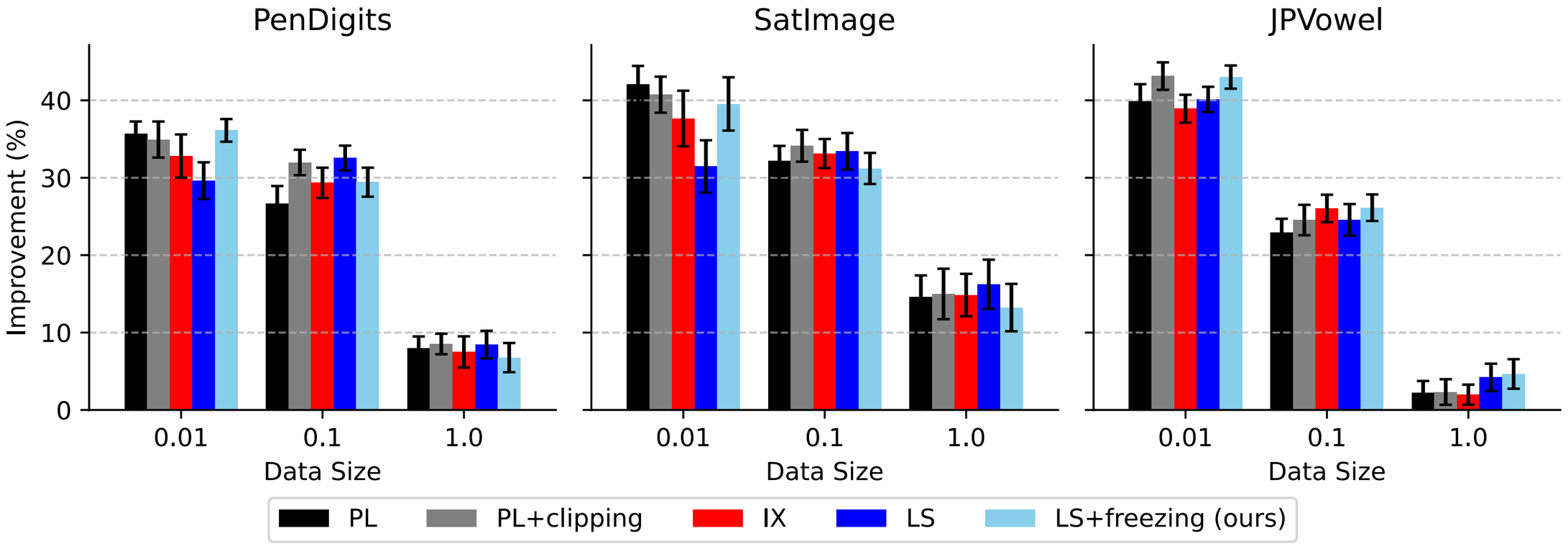
$$\mu(\pi^\star) - \mu(\hat{\pi}_n) \le \beta\, \mathbb{E}\left[\frac{(\tilde{r}^{\pi^*})^2}{c_1 + c_2 \tilde{r}^{\pi^*}}\right] + \frac{2}{\beta n} \ln \frac{|\Pi|}{\delta} - F_\beta(\phi)$$

"negative influence"

$$F_\beta(\phi^{\text{freezing}}) \ge F_\beta(\phi^{\text{clipping}}) \ge F_\beta(\phi^{\text{LS}})$$

Experiment: Off-Policy Learning

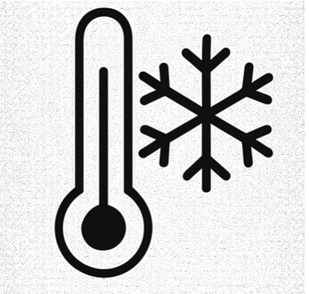
- Experiment setup from [WangKS2024] (value regression setup)
- Again, best or almost the best!



Concluding Remarks

- **Betting** for off-policy selection (and evaluation too ☺)
- **Freezing** for off-policy learning

⇒ Improved second-order bounds for optimality gap!



- **Efficient offline learning** will become increasingly important

- **Future research directions**
  - Betting-type idea for $|\Pi| = \infty$?
  - Similarly, more sophisticated techniques for RL?

Thank You!