Transporting Causal Effects Beyond Common Support

Mind the Gap & Think Like a Scientist

Harsh Parikh

Yale University

Vending machine operation

- *K machines*. Each day, choose what to stock and how to price.

Photo: tsubasuke5, Wikimedia Commons, CC BY 2.0.

Vending machine operation

- K *machines*. Each day, choose what to stock and how to price.
- Restock *locally* (fast, costly) or from a *warehouse* (cheap, 2-day lag).

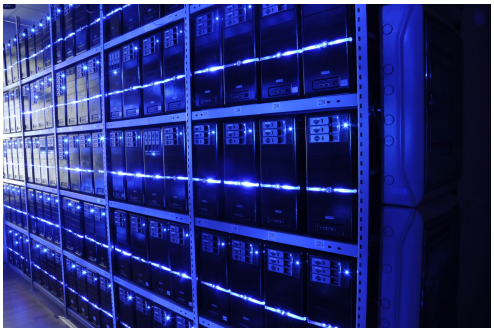
Photo: tsubasuke5, Wikimedia Commons, CC BY 2.0.

Vending machine operation

- K *machines*. Each day, choose what to stock and how to price.
- Restock *locally* (fast, costly) or from a *warehouse* (cheap, 2-day lag).
- *Future demand* is uncertain.

Photo: tsubasuke5, Wikimedia Commons, CC BY 2.0.

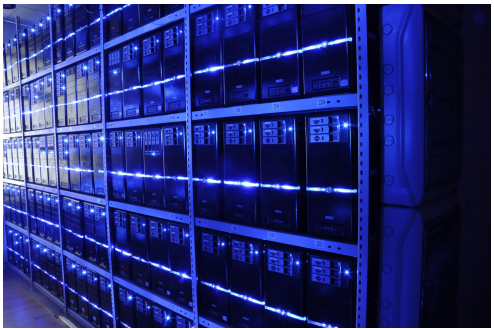
A simulator



- Historically, on-site staff set the stocking.

Photo: BalticServers.com, Wikimedia Commons, CC BY-SA 3.0.

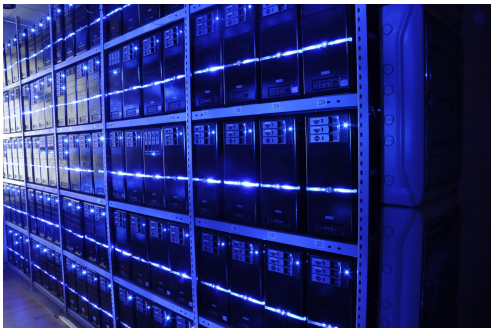
A simulator



- Historically, on-site staff set the stocking.
- *Experiments* find good actions but are *costly*.

Photo: BalticServers.com, Wikimedia Commons, CC BY-SA 3.0.

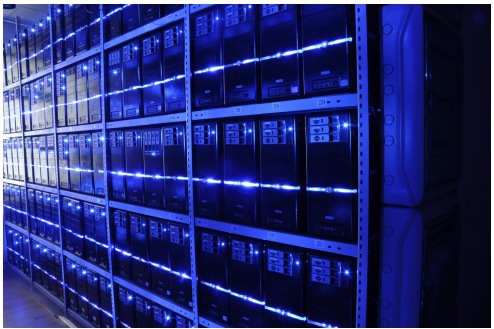
A simulator



- Historically, on-site staff set the stocking.
- *Experiments* find good actions but are *costly*.
- *Observational* and *experimental* data build a *simulator*.

Photo: BalticServers.com, Wikimedia Commons, CC BY-SA 3.0.

A simulator



- Historically, on-site staff set the stocking.
- *Experiments* find good actions but are *costly*.
- *Observational* and *experimental* data build a *simulator*.
- We want an optimal **policy** from it.

Photo: BalticServers.com, Wikimedia Commons, CC BY-SA 3.0.

Two questions

1. What is the *value of experimentation*?

Two questions

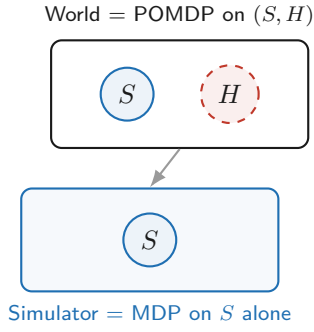
1. What is the *value of experimentation*?

2. What is the *optimal use of a simulator*?

Setup

units i ▪ time t

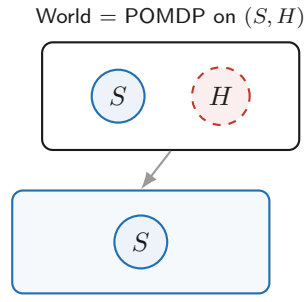
$S_{i,t}$	observed state
$H_{i,t}$	hidden state
$A_{i,t}$	action
$R_{i,t} \in [0, R_{\max}]$	reward
π	policy
$\mathcal{M}$	model (world or simulator)
$\gamma \in [0, 1)$	discount



Setup

units i ▪ time t

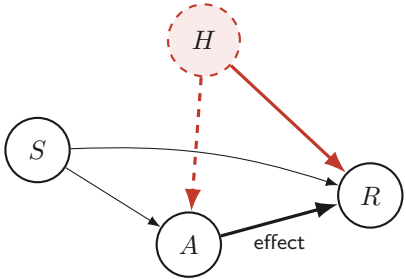
$S_{i,t}$	observed state
$H_{i,t}$	hidden state
$A_{i,t}$	action
$R_{i,t} \in [0, R_{\max}]$	reward
π	policy
$\mathcal{M}$	model (world or simulator)
$\gamma \in [0, 1)$	discount



$$V^{\pi}(\mathcal{M}) \;=\; \mathbb{E}^{\pi,\mathcal{M}}\Big[\sum_{t=0}^{T} \gamma^{t} \sum_{i} R_{i,t}\Big]$$

value of policy π on model $\mathcal{M}$

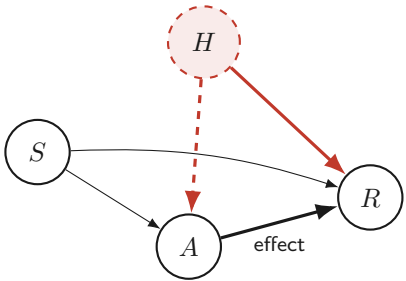
Confounding



H hidden ▪ S state ▪ A action ▪ R reward

- Historical actions followed unobserved demand signals.

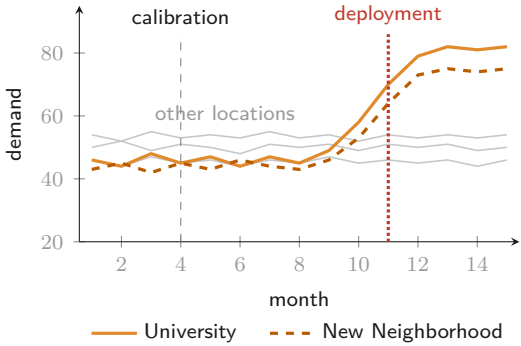
Confounding



H hidden ▪ S state ▪ A action ▪ R reward

- Historical actions followed unobserved demand signals.
- $A \not\perp H \mid S$.

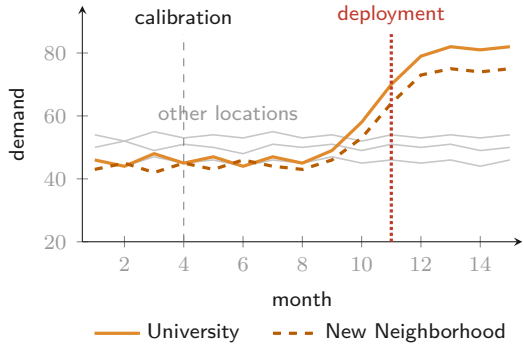
Drift



- *Drift*. The (hidden) state distribution shifts from *calibration* to *deployment*.

Seasonality calibrated from Kaggle vending sales (Singh 2022).

Drift

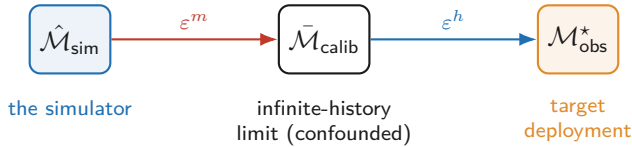


- *Drift*. The (hidden) state distribution shifts from *calibration* to *deployment*.

$$\mathbb{P}_{\text{deploy}}(H \mid S) \neq \mathbb{P}_{\text{calib}}(H \mid S)$$

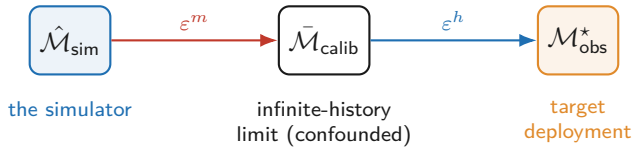
Seasonality calibrated from Kaggle vending sales (Singh 2022).

Errors!



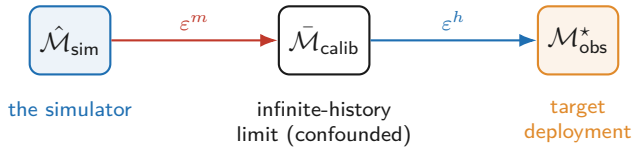
$\hat{\cdot}$ estimated $\bar{\cdot}$ infinite-data limit $\cdot^\star$ ground truth

Errors!



- ε^m: The simulator versus its calibration limit.

Errors!



- ε^m: The simulator versus its calibration limit.
- ε^h: Confounding and drift.

Three Choices

SOP
non-adaptive

Simulator-Optimal Policy.
$\pi^{\mathsf{SOP}} \in \arg\max_{\pi} V^{\pi}(\hat{\mathcal{M}}_{\mathsf{sim}})$, fixed for all t.

Solve in the simulator, deploy, never update.

Three Choices

SOP
non-adaptive

Simulator-Optimal Policy.

$\pi^{\mathsf{SOP}} \in \arg\max_{\pi} V^{\pi}(\hat{\mathcal{M}}_{\mathsf{sim}})$, fixed for all t.

Solve in the simulator, deploy, never update.

A-SOP
passive

Adaptive SOP.

$\pi_t \in \arg\max_{\pi} V^{\pi}(\hat{\mathcal{M}}_t), \qquad \hat{\mathcal{M}}_t = \mathbb{E}[\mathcal{M} \mid \mathcal{D}_t]$

Re-solve as on-path data arrives, with the simulator as a *prior*.

Three Choices

SOP
non-adaptive

Simulator-Optimal Policy.

$\pi^{\text{SOP}} \in \arg\max_{\pi} V^{\pi}(\hat{\mathcal{M}}_{\text{sim}})$, fixed for all t.

Solve in the simulator, deploy, never update.

A-SOP
passive

Adaptive SOP.

$\pi_t \in \arg\max_{\pi} V^{\pi}(\hat{\mathcal{M}}_t), \qquad \hat{\mathcal{M}}_t = \mathbb{E}[\mathcal{M} \mid \mathcal{D}_t]$

Re-solve as on-path data arrives, with the simulator as a *prior*.

SEP
active

Simulation-aided Experimental Policy.

A-SOP plus *deliberate experimentation*.

Learns where the policy would never go on its own.

Question 1

Value of experimentation

The value of experimentation

All values are discounted return in the *deployed world* $\mathcal{M}^{\star}_{\text{obs}}$, written $V^{\pi} := V^{\pi}(\mathcal{M}^{\star}_{\text{obs}})$.

Three *deployable* policies, observed state only.

$$\pi_{\text{SOP}}, \qquad \pi_{\text{A-SOP}}, \qquad \pi_{\text{SEP}}$$

An *oracle* π_H may condition on the hidden state.

$V^{\pi_H} \geq V^{\pi}$ for every observed-state π.

The value of experimentation

All values are discounted return in the *deployed world* $\mathcal{M}^{\star}_{\text{obs}}$, written $V^{\pi} := V^{\pi}(\mathcal{M}^{\star}_{\text{obs}})$.

Three *deployable* policies, observed state only.

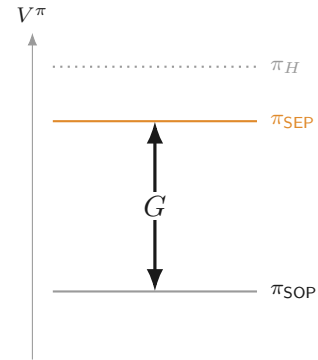
$$\pi_{\text{SOP}}, \qquad \pi_{\text{A-SOP}}, \qquad \pi_{\text{SEP}}$$

An *oracle* π_H may condition on the hidden state.

$$V^{\pi_H} \geq V^{\pi} \quad \text{for every observed-state } \pi.$$

Value of experimentation.

$$\boxed{G := V^{\pi_{\text{SEP}}} - V^{\pi_{\text{SOP}}}}$$



Gaps

Splitting G by adding and subtracting $V^{\pi_{\text{A-SOP}}}$

$$G = \underbrace{V^{\pi_{\text{A-SOP}}} - V^{\pi_{\text{SOP}}}}_{G_{\text{local}}} + \underbrace{V^{\pi_{\text{SEP}}} - V^{\pi_{\text{A-SOP}}}}_{G_{\text{reach}}}$$

G_{local} *within support* ▪ G_{reach} *out of support*

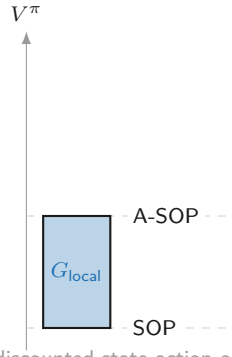
d^{π} is the discounted state-action occupancy.

Gaps

Splitting G by adding and subtracting $V^{\pi_{\text{A-SOP}}}$

$$G = \underbrace{V^{\pi_{\text{A-SOP}}} - V^{\pi_{\text{SOP}}}}_{G_{\text{local}}} + \underbrace{V^{\pi_{\text{SEP}}} - V^{\pi_{\text{A-SOP}}}}_{G_{\text{reach}}}$$

G_{local} *within support* ▪ G_{reach} *out of support*



d^{π} is the discounted state-action occupancy.

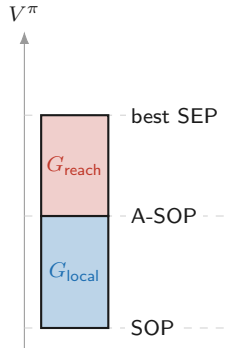
G_{local}. A-SOP's on-path data is *interventional* ($A \perp H \mid S$) and learns the truth within support, with no experiment.

Gaps

Splitting G by adding and subtracting $V^{\pi_{\text{A-SOP}}}$

$$G = \underbrace{V^{\pi_{\text{A-SOP}}} - V^{\pi_{\text{SOP}}}}_{G_{\text{local}}} + \underbrace{V^{\pi_{\text{SEP}}} - V^{\pi_{\text{A-SOP}}}}_{G_{\text{reach}}}$$

G_{local} *within support* ▪ G_{reach} *out of support*



d^{π} is the discounted state-action occupancy.

G_{local}. A-SOP's on-path data is *interventional* $(A \perp H \mid S)$ and learns the truth within support, with no experiment.

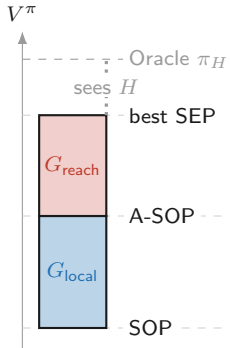
G_{reach}. Reachable only by playing $(s, a) \notin \text{supp}(d^{\pi_{\text{A-SOP}}})$, which needs a pilot.

Gaps

Splitting G by adding and subtracting $V^{\pi_{\text{A-SOP}}}$

$$G = \underbrace{V^{\pi_{\text{A-SOP}}} - V^{\pi_{\text{SOP}}}}_{G_{\text{local}}} + \underbrace{V^{\pi_{\text{SEP}}} - V^{\pi_{\text{A-SOP}}}}_{G_{\text{reach}}}$$

G_{local} *within support* ▪ G_{reach} *out of support*



d^{π} is the discounted state-action occupancy.

G_{local}. A-SOP's on-path data is *interventional* ($A \perp H \mid S$) and learns the truth within support, with no experiment.

G_{reach}. Reachable only by playing $(s, a) \notin \text{supp}(d^{\pi_{\text{A-SOP}}})$, which needs a pilot.

Simulation lemma

Building on classic results from Kearns & Singh (2002).

For any observed-state policy π, split the per-pair error along the chain $\hat{\mathcal{M}}_{\mathsf{sim}} \xrightarrow{\varepsilon^m} \bar{\mathcal{M}}_{\mathsf{calib}} \xrightarrow{\varepsilon^h} \mathcal{M}^\star_{\mathsf{obs}}$.

$$|V^\pi(\mathcal{M}^\star_{\mathsf{obs}}) - V^\pi(\hat{\mathcal{M}}_{\mathsf{sim}})| \;\le\; \underbrace{\frac{2}{1-\gamma}}_{\text{lin. } T_{\mathsf{eff}}} (\varepsilon_r^h + \varepsilon_r^m) \;+\; \underbrace{\frac{2\gamma R_{\max}}{(1-\gamma)^2}}_{\text{quad. } T_{\mathsf{eff}}} (\varepsilon_p^h + \varepsilon_p^m)$$

worst-case per pair. ε_r reward, ε_p transition (TV).

Simulation lemma

Building on classic results from Kearns & Singh (2002).

For any observed-state policy π, split the per-pair error along the chain $\hat{\mathcal{M}}_{\text{sim}} \xrightarrow{\varepsilon^m} \bar{\mathcal{M}}_{\text{calib}} \xrightarrow{\varepsilon^h} \mathcal{M}^\star_{\text{obs}}$.

$$|V^\pi(\mathcal{M}^\star_{\text{obs}}) - V^\pi(\hat{\mathcal{M}}_{\text{sim}})| \;\le\; \underbrace{\frac{2}{1-\gamma}}_{\text{lin. } T_{\text{eff}}} (\varepsilon^h_r + \varepsilon^m_r) \;+\; \underbrace{\frac{2\gamma R_{\max}}{(1-\gamma)^2}}_{\text{quad. } T_{\text{eff}}} (\varepsilon^h_p + \varepsilon^m_p)$$

worst-case per pair. ε_r reward, ε_p transition (TV).

- ε^h: randomized pilots ($A \perp H$) target identifying $\mathcal{M}^\star_{\text{obs}}$.

Simulation lemma

Building on classic results from Kearns & Singh (2002).

For any observed-state policy π, split the per-pair error along the chain $\hat{\mathcal{M}}_{\text{sim}} \xrightarrow{\varepsilon^m} \bar{\mathcal{M}}_{\text{calib}} \xrightarrow{\varepsilon^h} \mathcal{M}^\star_{\text{obs}}$.

$$|V^\pi(\mathcal{M}^\star_{\text{obs}}) - V^\pi(\hat{\mathcal{M}}_{\text{sim}})| \;\le\; \underbrace{\frac{2}{1-\gamma}}_{\text{lin. } T_{\text{eff}}} (\varepsilon_r^h + \varepsilon_r^m) \;+\; \underbrace{\frac{2\gamma R_{\max}}{(1-\gamma)^2}}_{\text{quad. } T_{\text{eff}}} (\varepsilon_p^h + \varepsilon_p^m)$$

worst-case per pair. ε_r reward, ε_p transition (TV).

- ε^h: randomized pilots ($A \perp H$) target identifying $\mathcal{M}^\star_{\text{obs}}$.
- ε^m: shrinks as the data grows.

Bounding the Gap

All values in $\mathcal{M}^{\star}_{\text{obs}}$. Telescope G through the passive policy.

$$G \leq \underbrace{\left|V^{\pi_{\text{A-SOP}}} - V^{\pi_{\text{SOP}}}\right|}_{\textit{local}} + \underbrace{\left|V^{\pi_{\text{SEP}}} - V^{\pi_{\text{A-SOP}}}\right|}_{\textit{reachability}}$$

Bounding the Gap

All values in $\mathcal{M}^{\star}_{\text{obs}}$. Telescope G through the passive policy.

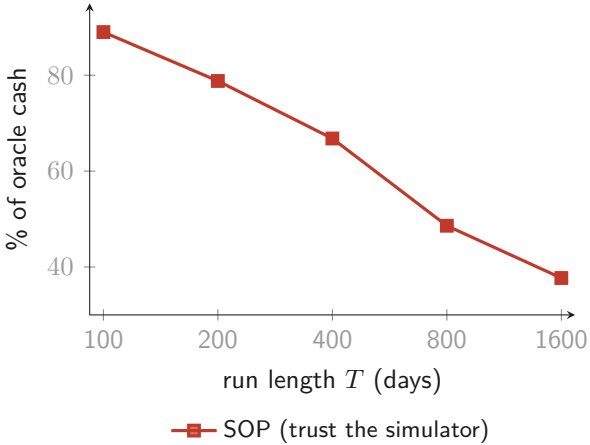
$$G \;\le\; \underbrace{\left|V^{\pi_{\text{A-SOP}}} - V^{\pi_{\text{SOP}}}\right|}_{\textit{local}} + \underbrace{\left|V^{\pi_{\text{SEP}}} - V^{\pi_{\text{A-SOP}}}\right|}_{\textit{reachability}}$$

- Within the simulator, the two coincide, $V^{\pi_{\text{A-SOP}}}(\hat{\mathcal{M}}_{\text{sim}}) = V^{\pi_{\text{SOP}}}(\hat{\mathcal{M}}_{\text{sim}})$, Thus, *local gap* is pure model-transfer error.

$$\left|V^{\pi_{\text{A-SOP}}} - V^{\pi_{\text{SOP}}}\right| \;\le\; 2 \sup_{\pi} \left|V^{\pi}(\mathcal{M}^{\star}_{\text{obs}}) - V^{\pi}(\hat{\mathcal{M}}_{\text{sim}})\right| \;\le\; 2B.$$

$B = \frac{2}{1-\gamma}(\varepsilon_r^h + \varepsilon_r^m) + \frac{2\gamma R_{\max}}{(1-\gamma)^2}(\varepsilon_p^h + \varepsilon_p^m)$, the simulation-lemma bound.

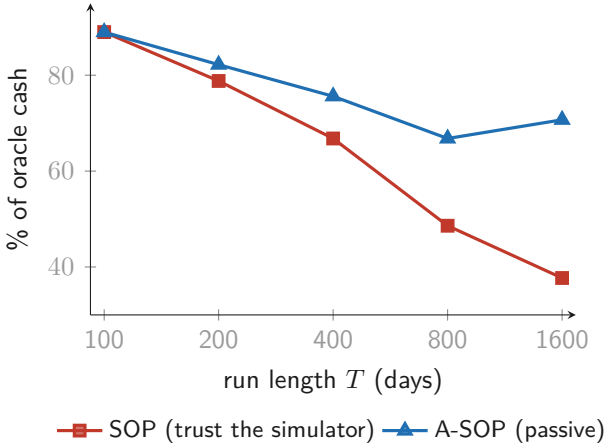
Back to Operating Vending Machines



- *SOP* trusts the simulator. Value *decays* with the horizon.

Multi-horizon table, Parikh et al. (arXiv:2605.21458), 30 common-seed trials.

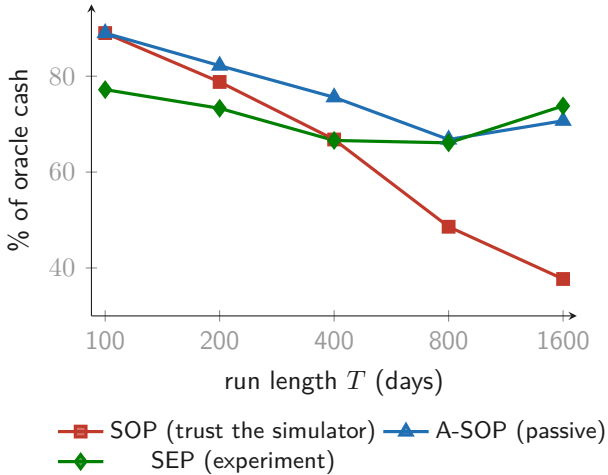
Back to Operating Vending Machines



- *SOP* trusts the simulator. Value *decays* with the horizon.
- *A-SOP* adapts passively and recovers most of the loss, $+33$ pp with no experiment.

Multi-horizon table, Parikh et al. (arXiv:2605.21458), 30 common-seed trials.

Back to Operating Vending Machines



- *SOP* trusts the simulator. Value *decays* with the horizon.
- *A-SOP* adapts passively and recovers most of the loss, $+33$ pp with no experiment.
- **SEP** experiments. It pays an early *cost* and passes passive only at long horizons.

Multi-horizon table, Parikh et al. (arXiv:2605.21458), 30 common-seed trials.

Question 2

Optimal use of a simulator

A better use of the simulator

- *SOP* deploys the simulator's policy unchanged. Value *decays*.

A better use of the simulator

- *SOP* deploys the simulator's policy unchanged. Value *decays*.
- *A-SOP* uses the simulator as a *warm start* and recovers most of the loss.

A better use of the simulator

- *SOP* deploys the simulator's policy unchanged. Value *decays*.
- *A-SOP* uses the simulator as a *warm start* and recovers most of the loss.
- *SEP* adds experimentation and improves further.

A better use of the simulator

- *SOP* deploys the simulator's policy unchanged. Value *decays*.
- *A-SOP* uses the simulator as a *warm start* and recovers most of the loss.
- *SEP* adds experimentation and improves further.

$\Rightarrow$ Can the simulator tell us *where* to experiment?

Idea

- $\mathcal{M}$ is parameterized by unknown parameters $\{\theta_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$

$$\theta_{s,a} = \big(\underbrace{r_{s,a}}_{\text{reward}}, \underbrace{p_{s,a}}_{\text{transition}} \big)$$

Idea

- $\mathcal{M}$ is parameterized by unknown parameters $\{\theta_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$

$$\theta_{s,a} = (\ \underbrace{r_{s,a}}_{\text{reward}},\ \underbrace{p_{s,a}}_{\text{transition}}\ )$$

- Parameter uncertainty $\implies$ uncertainty in predicted *target policy's value* $V^{\pi^{\text{tgt}}}(\hat{\theta})$.

Idea

- $\mathcal{M}$ is parameterized by unknown parameters $\{\theta_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$

$$\theta_{s,a} = \big(\underbrace{r_{s,a}}_{\text{reward}}, \underbrace{p_{s,a}}_{\text{transition}} \big)$$

- Parameter uncertainty $\implies$ uncertainty in predicted *target policy's value* $V^{\pi^{\text{tgt}}}(\hat{\theta})$.
- A good pilot must *reduce that uncertainty the most*.

$$\text{minimize} \quad \mathrm{Var}(\hat{V}^{\pi^{\text{tgt}}} \mid \text{pilot})$$

Ingredients for Criterion

To first order (delta method), variance combines two quantities at each (s', a').

Ingredients for Criterion

To first order (delta method), variance combines two quantities at each (s', a').

- *Value sensitivity* $\nabla_\theta \hat{V}^{\pi^{\text{tgt}}}$: How much predicted value moves when θ shifts.

Ingredients for Criterion

To first order (delta method), variance combines two quantities at each (s', a').

- *Value sensitivity* $\nabla_\theta \hat{V}^{\pi^{\text{tgt}}}$: How much predicted value moves when θ shifts.
- *Posterior uncertainty* $\left[\Sigma_\theta^{-1} + n\,\mathcal{I}_\theta^{\text{int}}\right]^{-1}$: Prior precision + *pilot information*.

The Fisher-SEP criterion

Combine sensitivity and uncertainty over the states π^{tgt} visits.

$$\mathrm{Var}(\pi; \pi^{\text{tgt}}) = \sum_{s} d_{\pi^{\text{tgt}}}(s) \sum_{(s',a')} \left(\nabla_\theta \hat{V}^{\pi^{\text{tgt}}}\right)^\top \left[\Sigma_\theta^{-1} + n_{s',a'}(\pi)\, \mathcal{I}^{\text{int}}\right]^{-1} \nabla_\theta \hat{V}^{\pi^{\text{tgt}}}$$

π enters only through $n_{s',a'}(\pi) = T\, d_\pi(s')\, \pi(a' \mid s')$ where $d_\pi(s')$ is visitation probability.

The Fisher-SEP criterion

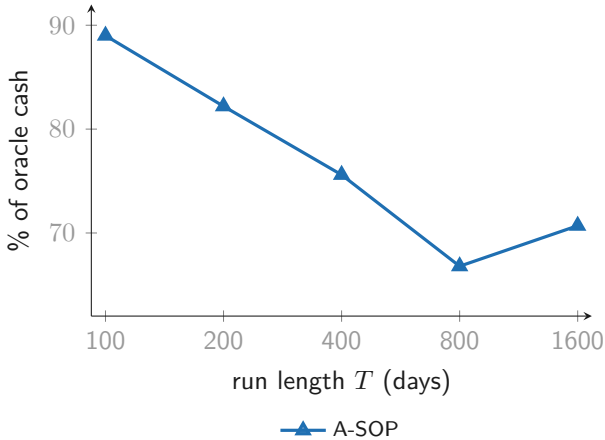
Combine sensitivity and uncertainty over the states π^{tgt} visits.

$$\mathrm{Var}(\pi; \pi^{\text{tgt}}) = \sum_{s} d_{\pi^{\text{tgt}}}(s) \sum_{(s',a')} \left(\nabla_\theta \hat{V}^{\pi^{\text{tgt}}}\right)^\top \left[\Sigma_\theta^{-1} + n_{s',a'}(\pi)\, \mathcal{I}^{\text{int}}\right]^{-1} \nabla_\theta \hat{V}^{\pi^{\text{tgt}}}$$

π enters only through $n_{s',a'}(\pi) = T\, d_\pi(s')\, \pi(a' \mid s')$ where $d_\pi(s')$ is visitation probability.

- **Fisher-SEP** is $\pi^\star = \arg\min_\pi \mathrm{Var}(\pi; \pi^{\text{tgt}})$.
  » Prefer a pilot study where *sensitivity* and *uncertainty* are *both* high.

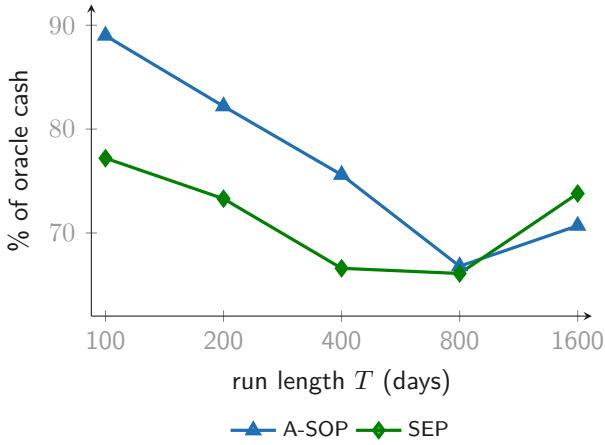
Back to Operating Vending Machines (v2.0)



- *A-SOP* is the bar to beat.
  Passive already recovers most.

Multi-horizon table, Parikh et al. (arXiv:2605.21458), 30 common-seed trials.

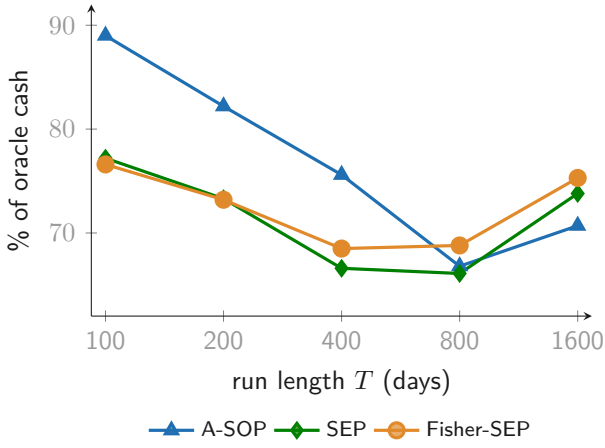
Back to Operating Vending Machines (v2.0)



- *A-SOP* is the bar to beat. Passive already recovers most.
- Undirected **SEP** explores broadly, at higher cost and smaller gain.

Multi-horizon table, Parikh et al. (arXiv:2605.21458), 30 common-seed trials.

Back to Operating Vending Machines (v2.0)



- *A-SOP* is the bar to beat. Passive already recovers most.
- Undirected **SEP** explores broadly, at higher cost and smaller gain.
- *Fisher-SEP* concentrates the pilot where the simulator's error most affects value.

Multi-horizon table, Parikh et al. (arXiv:2605.21458), 30 common-seed trials.

A Few Takeaways

1. *Value of experimentation.*

Set by the **horizon** and by what the gap is made of. Passive learning absorbs *confounding*, and experimentation pays against *drift*, once the horizon amortizes its cost.

A Few Takeaways

1. *Value of experimentation.*

   Set by the **horizon** and by what the gap is made of. Passive learning absorbs *confounding*, and experimentation pays against *drift*, once the horizon amortizes its cost.

2. *Optimal use of the simulator.*

   Transport *where to experiment*, chosen by the *Fisher criterion* (value sensitivity $\times$ uncertainty), along with transporting the optimal policy.

Open questions

- **Networked units.** Interference and spillovers when units sit in a *network*.
- **Cost-aware design.** Price of simulator *(re)training* and *experiment* cost.
- **Simulator structure.** Its *parametric form*, here fully nonparametric.
- **Imperfect recruitment.** Units enroll into the pilot only *with some probability*.
- **Non-stationary population.** The unit pool *grows and churns* over time.

Thank you

Mind the Sim-to-Real Gap & Think Like a Scientist

scan for the paper

Questions & discussion

Joint work with

Gabriel Levin-Konigsberg

Dominique Perrault-Joncas

Alexander Volfovsky

Supported by *Amazon Supply Chain Optimization Technologies (SCOT)*.

Full multi-horizon results

% of oracle cash	T=100	T=200	T=400	T=800	T=1600
SOP	89.0	78.8	66.8	48.6	37.7
ϵ-greedy	75.8	65.0	48.2	39.2	32.8
A-SOP	**89.0**	**82.2**	**75.6**	66.8	70.7
Thompson (PSRL)	88.5	82.6	76.6	68.6	71.2
KG-SEP	80.4	76.9	72.7	67.5	71.8
SEP	77.2	73.3	66.6	66.1	73.8
Fisher-SEP-R	76.6	73.2	68.5	**68.8**	**75.3**

Bold = best non-oracle per horizon. A-SOP $\approx$ Thompson. Crossover near T=400 to 600. ϵ-greedy random exploration, Thompson/PSRL posterior sampling, KG-SEP knowledge-gradient.

Multi-horizon table, Parikh et al. (arXiv:2605.21458), 30 common-seed trials.

Confounding-and-drift ablation

% oracle (T=400)	SOP	A-SOP	KG-SEP	SEP
no confounding, no drift	92	85	78	67
confounding only	73	86	77	66
confounding + drift	67	76	73	67

Passive A-SOP absorbs confounding (86 vs 85). Drift is the residual it cannot fully close (76).

Confounding-and-drift ablation, Parikh et al. (arXiv:2605.21458), 30 trials.

Notation

Symbol	Meaning
$S,\ H,\ A,\ R$	observed state, hidden state, action, reward
$(s, a),\ (s', a')$	a state-action pair, a candidate pilot pair
π	policy (observed history to action)
$\mathrm{do}(A{=}a),\ \not\perp$	set A by intervention, not conditionally independent
$\gamma \in [0, 1),\ R_{\max}$	discount factor, reward bound ($R \in [0, R_{\max}]$)
$T,\ T_{\text{eff}} = \frac{1}{1-\gamma}$	run length (days), effective planning horizon
$V^{\pi}(\mathcal{M})$	value of policy π on model $\mathcal{M}$
$\hat{\mathcal{M}}_{\text{sim}},\ \bar{\mathcal{M}}_{\text{calib}},\ \mathcal{M}^{\star}_{\text{obs}}$	simulator, infinite-data confounded limit, deployment target
$\varepsilon_r,\ \varepsilon_p$	per-pair reward error, transition error (total variation)
$\varepsilon^h,\ \varepsilon^m$	calibration-to-deployment gap, finite-sample residual
$G = G_{\text{local}} + G_{\text{reach}}$	design gap, within-support plus out-of-support
SOP, A-SOP, SEP	simulator-optimal, adaptive (passive), experimental (adaptive)
$\mathrm{Var},\ \mathrm{Var}_r$	posterior variance of target value, reward-only reduction
$d(s)$	discounted state-visitation distribution
$\sigma^0,\ n_{s',a'}$	prior SD, pilot sample count at (s', a')
$\tau^{\text{int}}_{s',a'} = \mathrm{Var}(R \mid s', \mathrm{do}(a'))^{-1}$	inverse interventional reward variance (pilot-identified)
P^{π} (e.g. $P^{\pi^{\text{tgt}}}$)	state-transition matrix induced by policy π