Title
Kasra Fallah(kf2779)
December 16, 2025

1 Introduction

Offline reinforcement learning (RL) seeks to optimize decision-making policies using a fixed dataset collected by a *behavior policy* π_b, without further interaction with the environment. A fundamental primitive in offline RL—and a prerequisite for safe policy improvement—is *off-policy evaluation* (OPE): estimating the performance

$$J(\pi_e) \;=\; \mathbb{E}_{\pi_e}\left[\frac{1}{H}\sum_{t=0}^{H-1}\ell(x_t,u_t)\right]$$

of a candidate *evaluation policy* π_e using data generated under π_b. OPE has been extensively studied in both reinforcement learning and causal inference, yet remains statistically challenging in sequential decision-making problems due to distribution shift between π_e and π_b (1; 2; 4).

A classical approach to OPE is importance sampling (IS), which reweights trajectories by likelihood ratios between the evaluation and behavior policies. While IS is unbiased, its variance typically grows exponentially with the horizon in Markov decision processes (1; 5). Self-normalized variants such as step-wise or per-decision importance sampling (PDIS) reduce variance at the cost of bias (1; 3), but remain fragile when the induced trajectory distributions differ substantially. This fragility is often quantified via the *effective sample size* (ESS), which collapses rapidly under policy mismatch (6; 7).

The difficulty of OPE is exacerbated in *dynamical systems and control*, where even small deviations in feedback gains may compound over time. In linear–quadratic control and related settings, this phenomenon has been observed both empirically and theoretically (10; 11). In partially observed systems, where actions depend on noisy outputs rather than latent states, the induced trajectory distribution depends jointly on the closed-loop dynamics, the observation channel, and the policy. As a result, likelihood-ratio weights behave like a multiplicative stochastic process whose logarithm exhibits a drift when $\pi_e \neq \pi_b$, leading to exponential decay of ESS with horizon.

To mitigate the variance of pure importance sampling, *doubly robust* (DR) estimators combine IS with a model-based control variate (2; 8). DR estimators enjoy a key robustness property: they are unbiased if either the importance weights are correct or the model is correctly specified. Recent works have extended DR ideas to sequential decision processes and reinforcement learning (8; 9), and have shown substantial empirical gains in finite-horizon settings. However, relatively little is understood about the behavior of DR estimators in *closed-loop dynamical systems*, particularly under output feedback and long horizons.

Contributions. In this work, we study off-policy evaluation for linear–Gaussian partially observed dynamical systems under linear–Gaussian output-feedback policies. This setting captures

a canonical abstraction in control (LQG-style evaluation) while remaining rich enough to expose fundamental statistical pathologies of sequential importance sampling. Our contributions are as follows:

- **Importance-weight degeneracy in output-feedback systems.** We show that when $\pi_e \neq \pi_b$, the cumulative log-importance weights exhibit a negative drift, causing ESS to decay exponentially with the horizon and with policy mismatch.
- **Implications for SN-PDIS.** We characterize how ESS collapse induces both variance explosion and systematic negative bias in self-normalized trajectory-wise estimators, explaining their empirical failure under moderate policy shift.
- **Doubly robust OPE for linear–Gaussian control.** We construct a DR estimator using a learned linear–Gaussian model (fit via EM) and establish a double-robustness guarantee. When the model prediction error is small, DR substantially reduces variance relative to SN-PDIS.
- **Empirical validation.** Through controlled experiments, we verify the predicted ESS collapse, bias–variance trade-offs, and the stabilizing effect of DR estimators as the policy shift increases.

Organization. To respect the 5-page main-paper constraint, we present the problem setup, estimators, and main theoretical results in the main text, while deferring proofs and extended diagnostics to the appendix.

2 Problem Formulation

We consider off-policy evaluation (OPE) for a partially observed linear–Gaussian dynamical system

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1}$$

$$y_t = Cx_t + v_t, \tag{2}$$

where $x_t \in \mathbb{R}^n$ is the latent state, $y_t \in \mathbb{R}^p$ is the observed output, and $u_t \in \mathbb{R}^m$ is the control input. The process noise $w_t \sim \mathcal{N}(0, W)$ and observation noise $v_t \sim \mathcal{N}(0, V)$ are i.i.d. and mutually independent. The system (A, B, C) is assumed stabilizable and detectable.

Control is exerted through stationary linear–Gaussian output-feedback policies

$$u_t = Ky_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \Sigma_u), \tag{3}$$

with feedback gain $K \in \mathbb{R}^{m\times p}$. A behavior policy π_b with gain K_b generates a fixed dataset $\mathcal{D} = \{(y_t, u_t, \ell_t)\}_{t=0}^{T-1}$, while an evaluation policy π_e with gain K_e is to be evaluated using $\mathcal{D}$ alone. We quantify policy mismatch by $\delta := \|K_e - K_b\|_F$.

The performance criterion is the finite-horizon average cost

$$J(\pi) := \mathbb{E}_\pi\left[\frac{1}{H}\sum_{t=0}^{H-1} \ell(x_t, u_t)\right], \qquad \ell(x, u) = x^\top Qx + u^\top Ru, \tag{4}$$

where $Q \succeq 0$ and $R \succ 0$. The expectation is taken with respect to the closed-loop distribution induced by (1)–(3).

The OPE task is to estimate $J(\pi_e)$ using data generated by π_b. We study three classes of estimators: (i) model-based (MB) estimators obtained by simulating π_e on a learned linear–Gaussian model; (ii) self-normalized per-decision importance sampling (SN-PDIS), based on likelihood ratios $\rho_t := \pi_e(u_t \mid y_t)/\pi_b(u_t \mid y_t)$; and (iii) doubly robust (DR) estimators combining importance weighting with a model-based control variate. Formal definitions are given in Section 3.

A key diagnostic quantity throughout the paper is the effective sample size (ESS),

$$\mathrm{ESS} := \frac{\left(\sum_{t=0}^{H-1} w_t\right)^2}{\sum_{t=0}^{H-1} w_t^2}, \qquad w_t := \prod_{k=0}^{t} \rho_k, \tag{5}$$

which measures concentration of the normalized importance weights and governs the statistical stability of IS-based estimators.

Our objective is to characterize how partial observability, horizon length H, and policy shift δ jointly affect (i) importance-weight degeneracy, (ii) the bias–variance behavior of SN-PDIS, and (iii) the variance-reduction properties of DR estimators. Theoretical results are presented in Section **??**, with proofs deferred to the appendix.

3 Estimators and Main Results

We describe the off-policy estimators considered in this work and summarize the main theoretical results. All proofs and explicit constants are deferred to the appendix.

Model-based estimator. A linear–Gaussian model $\widehat{\mathcal{M}} = (\widehat{A}, \widehat{B}, \widehat{C}, \widehat{W}, \widehat{V})$ is learned from $\mathcal{D}$ by maximum likelihood via the EM algorithm. The model-based (MB) estimator evaluates π_e by Monte Carlo simulation under $\widehat{\mathcal{M}}$ and returns

$$\widehat{J}_{\mathrm{MB}} := J_{\widehat{\mathcal{M}}}(\pi_e).$$

When $\widehat{\mathcal{M}}$ is accurate, $\widehat{J}_{\mathrm{MB}}$ has low variance; however, it may be biased under model misspecification.

Self-normalized per-decision importance sampling. Define the per-step likelihood ratio

$$\rho_t := \frac{\pi_e(u_t \mid y_t)}{\pi_b(u_t \mid y_t)}, \qquad w_t := \prod_{k=0}^{t} \rho_k.$$

The self-normalized per-decision importance sampling (SN-PDIS) estimator is

$$\widehat{J}_{\mathrm{SN}} = \sum_{t=0}^{H-1} \frac{w_t}{\sum_{s=0}^{H-1} w_s}\, \ell_t. \tag{6}$$

This estimator is biased but commonly used due to its reduced variance relative to ordinary importance sampling.

Doubly robust estimator. Let $\widehat{\ell}_t^{\mathrm{MB}}$ denote a model-based predictor of $\mathbb{E}[\ell_t \mid y_t]$ under $\widehat{\mathcal{M}}$. The doubly robust (DR) estimator is defined as

$$\widehat{J}_{\mathrm{DR}} = \frac{1}{H}\sum_{t=0}^{H-1}\widehat{\ell}_t^{\mathrm{MB}} + \sum_{t=0}^{H-1}\frac{w_t}{\sum_{s=0}^{H-1} w_s}\big(\ell_t - \widehat{\ell}_t^{\mathrm{MB}}\big). \tag{7}$$

3.1 Main Theoretical Results

We now state the main theoretical results characterizing the behavior of these estimators. All results hold under the assumptions of Section 2. Precise constants and proofs are given in the appendix.

Lemma 1 (Importance-weight drift)**.** *Assume $\pi_e \neq \pi_b$. Then there exists $\gamma > 0$ such that*

$$\mathbb{E}[\log \rho_t \mid \mathcal{F}_{t-1}] \leq -\gamma \quad a.s.$$

Consequently, the effective sample size satisfies

$$\mathbb{E}[\mathrm{ESS}] \leq C\exp(-\gamma H),$$

where $C > 0$ is independent of H.

Lemma 2 (Bias and variance of SN-PDIS)**.** *Assume $\ell_t \geq 0$ almost surely. Then*

$$\mathbb{E}[\widehat{J}_{\mathrm{SN}}] \leq J(\pi_e),$$

and

$$\mathrm{Var}(\widehat{J}_{\mathrm{SN}}) \;\geq\; c\,\mathrm{ESS}^{-1},$$

for some constant $c > 0$.

Lemma 3 (Double robustness)**.** *If either*

1. *$\rho_t = \pi_e(u_t \mid y_t)/\pi_b(u_t \mid y_t)$ almost surely, or*
2. *$\widehat{\ell}_t^{\mathrm{MB}} = \mathbb{E}[\ell_t \mid y_t]$ almost surely,*

then

$$\mathbb{E}[\widehat{J}_{\mathrm{DR}}] = J(\pi_e).$$

[Variance reduction] If

$$\mathbb{E}\Big[(\ell_t - \widehat{\ell}_t^{\mathrm{MB}})^2\Big] \leq \varepsilon,$$

then

$$\mathrm{Var}(\widehat{J}_{\mathrm{DR}}) \;\leq\; \varepsilon\,\mathrm{Var}(\widehat{J}_{\mathrm{SN}}).$$

Theorem 1 (Estimator comparison)**.** *Under the assumptions of Section 2,*

1. *SN-PDIS exhibits bias and variance explosion as $\mathrm{ESS} \to 0$;*
2. *MB estimation has low variance but incurs bias proportional to model error;*
3. *DR achieves strictly smaller mean-squared error than SN-PDIS and MB whenever the model estimation error is sufficiently small.*

4 Results and Analysis

This section evaluates the empirical behavior of the proposed off-policy evaluation (OPE) estimators for partially observed linear dynamical systems. We compare three approaches: model-based simulation (MB), self-normalized per-decision importance sampling (SN-PDIS), and a doubly robust estimator (DR). The experiments are designed to isolate the effects of policy mismatch, importance-weight degeneracy, and model error.

All results are obtained using the experimental pipeline described in Section **??**, with algorithmic details deferred to Appendix B and additional figures provided in Appendix A.

4.1 Model learning and validation

We first assess the quality of the learned linear Gaussian state-space model obtained via the EM procedure. As shown in Figure 5a, the innovation energy decreases monotonically across EM iterations, indicating stable convergence. Throughout training, the estimated system matrix remains strictly stable due to the imposed spectral projection, ensuring that the learned dynamics are well-posed for simulation-based evaluation.

Predictive performance is evaluated through one-step-ahead output prediction. Figure 5b demonstrates that the learned model closely tracks the observed output trajectories, while the innovation process remains approximately zero-mean and largely contained within the $\pm 2\sigma$ confidence bands (Figure 5c). These diagnostics suggest that the learned model provides an accurate local approximation of the closed-loop dynamics induced by the behavior policy.

4.2 Single-run off-policy evaluation

We next compare OPE estimates for a fixed policy shift $\delta = 0.10$. Figure 2b reports the true evaluation cost alongside MB, SN-PDIS, and DR estimates for a representative run.

The model-based estimator achieves near-perfect agreement with the ground truth, reflecting the high fidelity of the learned dynamics in this regime. In contrast, SN-PDIS significantly underestimates the true cost and exhibits high variability. The doubly robust estimator remains accurate, closely matching the true evaluation cost despite relying on the same importance weights as SN-PDIS.

This behavior highlights a key phenomenon: although SN-PDIS is unbiased in expectation, its finite-sample performance degrades severely under moderate policy mismatch, whereas the doubly robust estimator leverages the learned model as a control variate to reduce variance.

4.3 Importance-weight degeneracy and effective sample size

To explain the failure mode of SN-PDIS, we examine the evolution of importance weights. Figure 1a shows that the cumulative log-weights rapidly diverge, indicating severe weight imbalance. Correspondingly, the effective sample size (ESS) collapses as the policy shift δ increases (Figure 1b).

This collapse implies that the SN-PDIS estimate is effectively supported by only a small subset of timesteps, rendering it highly sensitive to noise and outliers. The mechanism is further illustrated in Figure 3, where the normalized contribution mass concentrates on a small number of events despite relatively smooth instantaneous costs.

4.4 Sensitivity to policy mismatch

Figure 2 summarizes estimator performance as a function of policy mismatch. As δ increases, the absolute error of SN-PDIS grows rapidly (Figure 2a), closely tracking the decline in ESS. In contrast, the doubly robust estimator remains accurate across the same range of policy shifts, while the model-based estimator exhibits only mild bias attributable to residual model error.

These results empirically confirm the theoretical intuition that importance-sampling-based estimators suffer from exponential variance growth under policy mismatch, whereas doubly robust methods can maintain stability provided that the learned model is sufficiently accurate.

4.5 Robustness across random seeds

Finally, we assess estimator variability across independent runs. Figure 4 reports the distribution and mean-squared error (MSE) of each estimator across multiple random seeds at a fixed policy shift.

SN-PDIS exhibits substantially higher variance and MSE than both MB and DR, consistent with its low effective sample size. The doubly robust estimator achieves the lowest overall MSE, combining the low variance of model-based simulation with the bias correction afforded by importance weighting. These findings indicate that the empirical advantages of DR persist beyond single-run evaluations.

4.6 Summary of empirical findings

Across all experiments, we observe the following consistent trends:

- Importance-sampling-based estimators degrade rapidly under policy mismatch due to weight degeneracy and effective sample size collapse.
- Model-based evaluation performs well when the learned dynamics accurately capture the closed-loop system but may incur bias under model misspecification.
- Doubly robust estimation achieves the most favorable bias–variance tradeoff, remaining accurate and stable across a wide range of policy shifts.

Together, these results demonstrate that doubly robust methods provide a reliable approach to off-policy evaluation in partially observed control systems, particularly in regimes where policy mismatch is unavoidable.

References

[1] D. Precup, R. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. *ICML*, 2000.

[2] M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *ICML*, 2011.

[3] P. Thomas and E. Brunskill. High-confidence off-policy evaluation. *AAAI*, 2016.

[4] P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *ICML*, 2016.

[5] R. Munos et al. Safe and efficient off-policy reinforcement learning. *NeurIPS*, 2016.

[6] A. Kong, J. Liu, and W. Wong. Sequential imputations and Bayesian missing data problems. *JASA*, 1994.

[7] E. Ionides. Truncated importance sampling. *JCGS*, 2008.

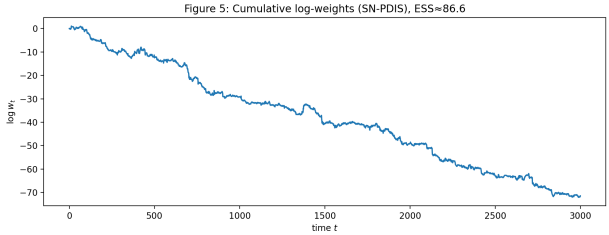
[8] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *ICML*, 2016.

[9] M. Farajtabar et al. More robust doubly robust off-policy evaluation. *ICML*, 2018.

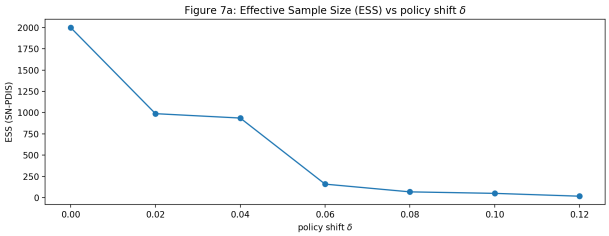
[10] M. Fazel et al. Global convergence of policy gradient methods for the linear quadratic regulator. *ICML*, 2018.

[11] S. Dean et al. Sample complexity of the linear quadratic regulator. *Foundations and Trends in Systems and Control*, 2019.

A Additional Figures

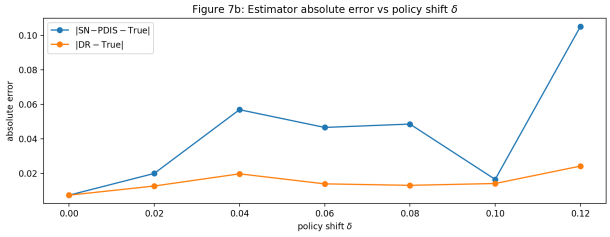


(a) Cumulative log-weights for SN-PDIS (representative run).

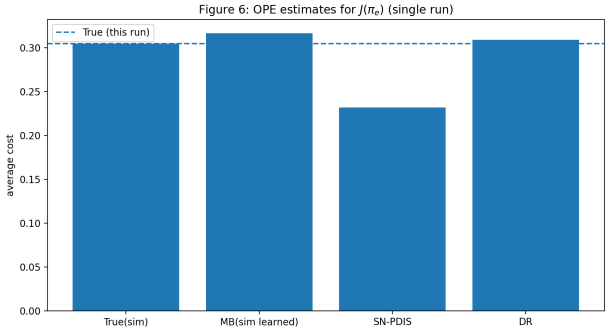


(b) ESS versus policy shift δ.

Figure 1: Importance-weight degeneracy: as δ increases, the effective sample size collapses, explaining the instability of pure IS-based estimators.

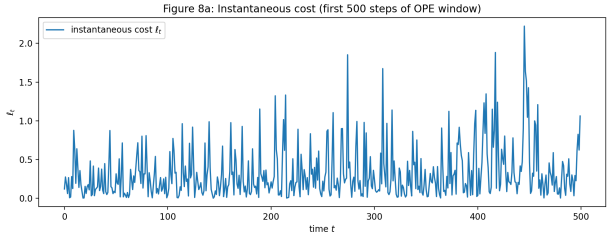


(a) Absolute estimation error versus δ.

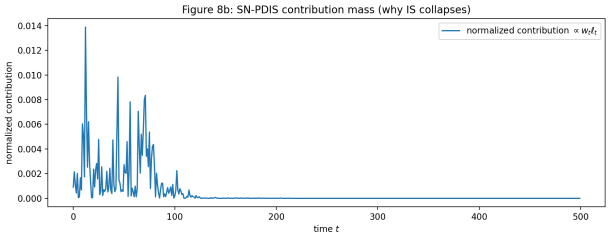


(b) OPE estimates for $J(\pi_e)$ (single run).

Figure 2: Estimator performance summary: DR remains accurate across shifts where SN-PDIS degrades due to weight collapse.

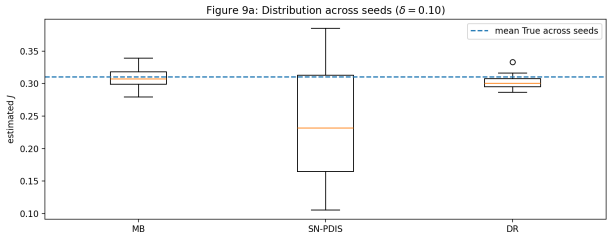


(a) Instantaneous cost ℓ_t over the OPE window.

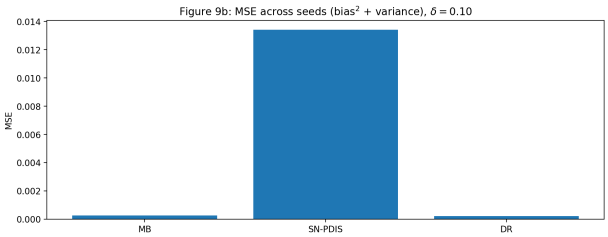


(b) Normalized contribution mass $\propto w_t \ell_t$.

Figure 3: Mechanism behind SN-PDIS failure: contribution concentrates on a small set of timesteps/trajectories when weights degenerate.

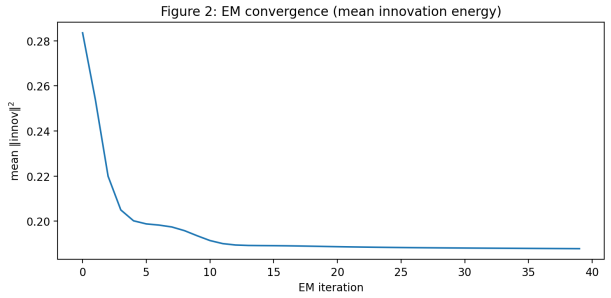


(a) Distribution across random seeds (fixed δ).

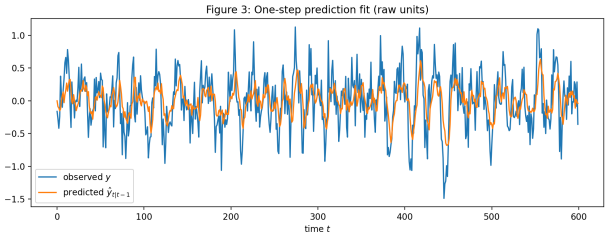


(b) MSE across seeds (fixed δ).

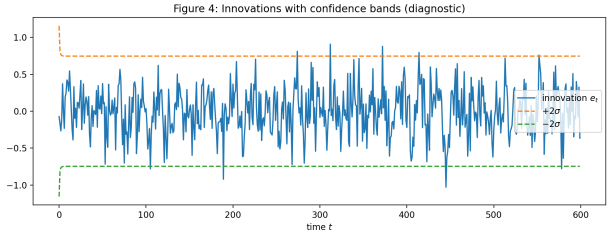
Figure 4: Across-seed variability: SN-PDIS exhibits substantially higher variance/MSE than MB and DR at nontrivial shifts.



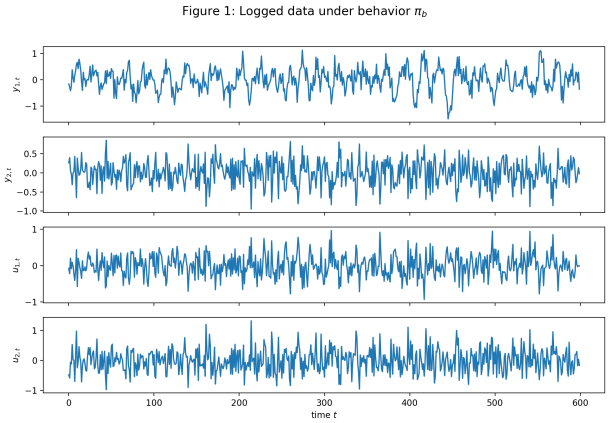
(a) EM convergence.



(b) One-step prediction fit.



(c) Innovation bands.



(d) Logged trajectories (optional).

Figure 5: Model-learning diagnostics (supplementary).

B Algorithms

This appendix summarizes the algorithms used throughout the paper. All procedures correspond exactly to the experimental pipeline described in the main text.

B.1 Data collection under the behavior policy

Algorithm 1 Data collection under behavior policy π_b

Require: Stable linear system (A, B, C), noise covariances (W, V), behavior gain K_b, exploration covariance Σ_u, horizon T

Ensure: Logged trajectory $\{(y_t, u_t)\}_{t=0}^{T-1}$

Initialize $x_0 = 0$
for $t = 0$ to $T - 1$ **do**
    Observe $y_t = Cx_t + v_t, \quad v_t \sim \mathcal{N}(0, V)$
    Sample $u_t = K_b y_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_u)$
    Update $x_{t+1} = Ax_t + Bu_t + w_t, \quad w_t \sim \mathcal{N}(0, W)$
end for

B.2 Linear Gaussian system identification (EM)

Algorithm 2 EM for a linear Gaussian state-space model

Require: Logged data $\{(y_t, u_t)\}_{t=0}^{T-1}$, state dimension n_x, iterations N_{EM}

Ensure: Estimates $(\hat{A}, \hat{B}, \hat{C}, \hat{W}, \hat{V})$

Initialize (A, B, C, W, V)
for $k = 1$ to N_{EM} **do**
    E-step: run Kalman smoother; compute $\mathbb{E}[x_t]$, $\mathbb{E}[x_t x_t^\top]$, $\mathbb{E}[x_{t+1} x_t^\top]$
    M-step: update (A, B, C, W, V) using moment / least-squares updates
    (Optional) stabilize $\hat{A}$ via spectral projection
end for

B.3 Model-based policy evaluation

Algorithm 3 Model-based policy evaluation (MB-OPE)

Require: Learned model $(\hat{A}, \hat{B}, \hat{C}, \hat{W}, \hat{V})$, evaluation gain K_e, horizon H

Ensure: $\widehat{J}_{\mathrm{MB}}(\pi_e)$

Simulate the learned model under π_e over horizon H and compute $\widehat{J}_{\mathrm{MB}}(\pi_e) = \frac{1}{H} \sum_{t=0}^{H-1} \ell_t$

B.4 Self-normalized PDIS

Algorithm 4 Self-normalized PDIS

Require: Logged data $\{(y_t, u_t, \ell_t)\}_{t=0}^{H-1}$, policies π_b, π_e
Ensure: $\widehat{J}_{\mathrm{SN}}(\pi_e)$
1: Compute $\log \rho_t = \log \pi_e(u_t|y_t) - \log \pi_b(u_t|y_t)$
2: Compute cumulative $\log w_t = \sum_{s=0}^{t} \log \rho_s$
3: Normalize $w_t \propto \exp(\log w_t)$
4: Output $\widehat{J}_{\mathrm{SN}}(\pi_e) = \frac{\sum_t w_t \ell_t}{\sum_t w_t}$

B.5 Doubly robust OPE

Algorithm 5 Doubly robust OPE

Require: Logged data, learned model $\hat{\mathcal{M}}$, policies π_b, π_e
Ensure: $\widehat{J}_{\mathrm{DR}}(\pi_e)$
1: Compute model-based baseline $\widehat{V}_{\hat{\mathcal{M}}}^{\pi_e}$
2: Compute weights w_t (as in Algorithm 4)
3: Compute correction $\Delta = \frac{1}{H} \sum_{t=0}^{H-1} w_t \left(\ell_t - \hat{\ell}_t^{\pi_e}\right)$
4: Output $\widehat{J}_{\mathrm{DR}}(\pi_e) = \widehat{V}_{\hat{\mathcal{M}}}^{\pi_e} + \Delta$

B.6 Effective sample size

Algorithm 6 Effective sample size (ESS)

Require: Normalized weights $\{w_t\}$
Ensure: ESS
1: $\mathrm{ESS} = \dfrac{(\sum_t w_t)^2}{\sum_t w_t^2}$