Reward-Shaping Control Variates for Off-Policy Evaluation Under Sparse Rewards

Ritam Majumdar*[1] Finale Doshi-Velez[2] Sonali Parbhoo*[1]

[1]Imperial College London [2]Harvard University *Equal contribution

Corresponding Author: r.majumdar24@imperial.ac.uk

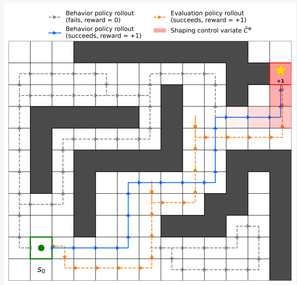
ICML International Conference On Machine Learning

Aim of the Paper

Develop **unbiased** OPE estimators for **sparse-reward** settings using **potential-based reward shaping** as zero-mean control variates.

- Reduce variance **without adding bias**.
- Extract signal from **every** trajectory not just rare successes.
- Practical Applications: Healthcare, Education, Finance, where sparse and delayed rewards are prevalent.

The Sparse Reward OPE Problem



- Most trajectories give **zero reward**.
- OPE dominated by rare successes ⇒ **high variance**.
- **Shaped** estimators use **dense state** signal as control variates.

Why Existing Methods Fail

- **IS:** Exploding IS-weights on rare successes
- **DR:** Must learn Q from ~ 0 TD targets.
- **MIS:** Density-ratio objective ill-conditioned when reward support is sparse

Reward Shaping Control Variates: Intuition

- Rewards are sparse, but **state visitations are dense**.
- Learn $\Phi(s)$: a **progress score** over states.
- Use Φ to build a **zero-mean control variate** for the IS return.
- Trained *only* from how well it correlates with returns, no reward labels needed.
- Signal from **all** trajectories, not just successes.

Proposed Shaping Estimator

$$\widehat{V}^{\pi_e}_{\text{Shaped-PDIS}}(\lambda) = \frac{1}{N}\sum_{n=1}^{N}\Big[\underbrace{Y^{(n)}}_{\text{IS return}} + \lambda^\top \underbrace{\widetilde{C}^{\Phi,(n)}}_{\text{shaping control variate}}\Big]$$

Building blocks

- **PDIS return:** $Y^{(n)} = \sum_{t=0}^{T-1}\gamma^t W_t r_t$
  Standard PDIS estimator. High-variance under sparse rewards.
- **Shaping feature:** $C_t^\Phi = \gamma^t W_t\big(\gamma\Phi(s_{t+1}) - \Phi(s_t)\big)$
  Potential-based shaping, responsible for variance reduction.
- **Boundary:** $\Phi(s_T) = 0$
  Anchors the potential and forces telescoping.

Why it works

- **Centered CV:** $\widetilde{C}^\Phi = C^\Phi - \mathbb{E}_{\pi_b}[C^\Phi]$
  Zero-mean *by construction* ⇒ adding $\lambda^\top\widetilde{C}^\Phi$ never changes the expectation.
- **Optimal weight:** $\lambda^\star = -\Sigma_{CC}^{-1}\Sigma_{CY}$
  Closed-form: projection of Y onto the span of shaping columns.
- **Variance reduction:** $\mathrm{Var}(Y + \lambda^{\star\top}\widetilde{C}^\Phi) = \mathrm{Var}(Y) - \Sigma_{CY}^\top\Sigma_{CC}^{-1}\Sigma_{CY}$
  Strict reduction whenever $\widetilde{C}^\Phi$ correlates with Y. Never larger than $\mathrm{Var}(Y)$.

Training Objective: Learning Φ_β for Variance Reduction: $\max_\beta J(\beta) = \Sigma_{CY}(\beta)^\top\,\Sigma_{CC}(\beta)^{-1}\,\Sigma_{CY}(\beta)$

No dense reward labels required: Φ_β is trained *only* to make its shaping columns correlate with the noisy IS return Y.

Same construction extends to **Shaped-DR** and **Shaped-MRDR** by replacing Y with Y_{DR} or Y_{MRDR}.

Advantages over DR / MRDR

- **Easier to learn.**
  - DR needs $Q(s,a)$ satisfying *Bellman consistency* across long horizons, under sparse TD targets ~ 0.
  - $\Phi(s)$ only needs to *correlate* with Y with sparse reward labels.
- **Interpretable.**
  - $\Phi(s)$ is a scalar over states
  - Learned λ_t reveal *which timesteps* drive variance reduction.

Algorithm

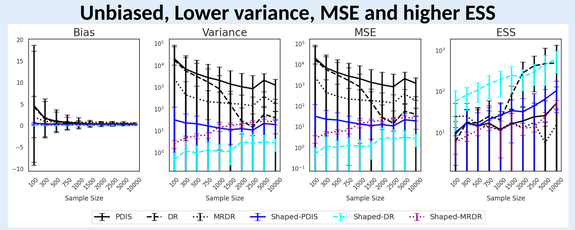
Input: trajectories from π_b, target π_e, γ, ridge α
For $k = 1, \ldots, K$:

1. Split into fit set ξ_{fit} and eval set ξ_k.
2. On ξ_{fit}: build IS returns $Y^{(n)}$ and shaping columns $C^{\Phi_\beta,(n)}$.
3. Estimate covariances $\widehat{\Sigma}_{CC}, \widehat{\Sigma}_{CY}$.
4. Update β by gradient ascent on $J(\beta) = \Sigma_{CY}^\top\Sigma_{CC}^{-1}\Sigma_{CY}$.
5. Compute $\lambda_k = -(\widehat{\Sigma}_{CC} + \alpha I)^{-1}\widehat{\Sigma}_{CY}$.
6. Evaluate on ξ_k: $\widehat{V}_k^{\pi_e} = \frac{1}{|\xi_k|}\sum_{n\in\xi_k}\big(Y^{(n)} + \lambda_k^\top C^{\Phi_\beta,(n)}\big)$.

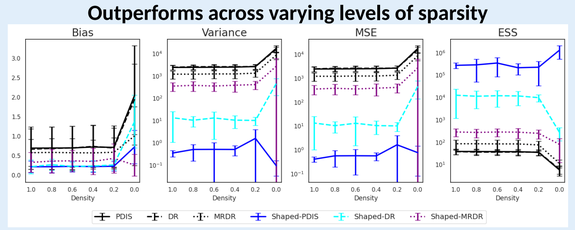
Output: $\widehat{V}^{\pi_e}_{\text{Shaped-PDIS}} = \frac{1}{K}\sum_k \widehat{V}_k^{\pi_e}$.

Quantitative Improvements

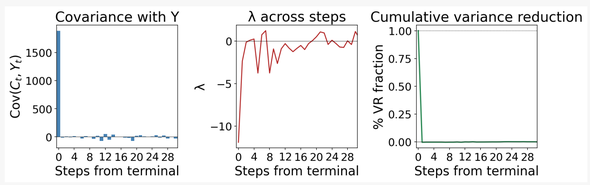
Unbiased, Lower variance, MSE and higher ESS



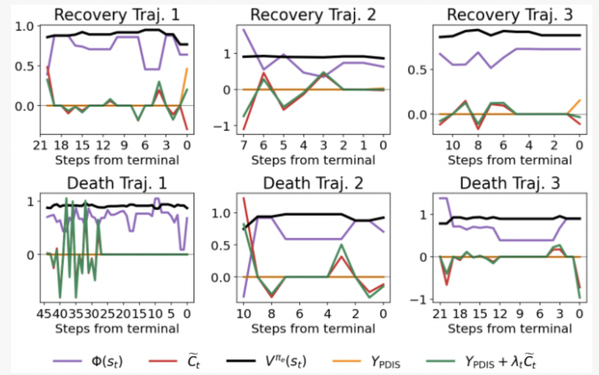
Outperforms across varying levels of sparsity



Variance Reduction Concentrates Near Terminal Reward

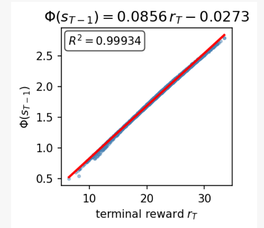


Learnt Control Variate $\widetilde{C}^\Phi$ is not the same as value function V^{π_e}



Φ satisfies **no Bellman constraint**, it is optimised purely for variance reduction.

Shaping potential Φ can reveal meaningful structure on state quality



- Slope (~0.08) scales returns down and drives variance reduction.
- Sign encodes **state quality**: $\Phi > 0$ high-portfolio, $\Phi < 0$ low.
- Holds when IS-weight dispersion doesn't dominate.

Limitations & Future Work

- Limited shaping of earlier timesteps.
- Long-horizon IS-weights **limit** Φ interpretability.
- State-density **not** explicitly modelled in variance reduction.

Conclusions

- Reward-shaping control variates are **unbiased** by construction.
- Shaped estimators achieve **strict variance reduction** over traditional counterparts.
- **State-of-the-art** on sparse, noisy, and misspecified OPE benchmarks across healthcare, finance, and tabular domains.
- Open a path to **interpretable** OPE: Φ and λ expose *where* and *when* variance reduction happens.