Off-Policy Evaluation and Learning with Data from Bandits

Susan Athey, Ruohan Zhan

Introduction

It has been increasingly common for data to be collected using adaptive experimentation, for example, using bandit exploration. Historical data of this type can be used to evaluate alternative treatment assignment policies and learn optimal policies to guide future innovation or experiments. However, offline evaluation and learning in these settings can be challenging. Popular approaches can be plagued by bias, excessive variance, or both.

We introduce novel estimators that are consistent and asymptotically normal for policy evaluation, based on which we further develop algorithms for offline optimal policy learning with a guaranteed regret bound. Using synthetic data and public benchmarks, we provide empirical evidence for the effectiveness of our estimators in evaluation and learning relative to existing alternatives.

This is based on joint work with Vitor Hadad, David A. Hirshberg, Stefan Wager, Zhimei Ren, and Zhengyuan Zhou.

Set Up

Consider a contextual bandit setting with K arms . At each time, the experimenter collects a sample $(X_t, W_t, Y_t, e_t(X_t, W_t))$, where X_t is covariate, W_t is pulled arm, Y_t is the observed outcome, and $e_t(X_t, W_t)$ is the probability of assigning arm W_t to covariate X_t. Besides, we use $\mathrm{H}_t = \{(X_s, W_s, Y_s)\}_{s=1}^{t}$ to represent observations up to time t. We make the following assumptions throughout.

- $(X_t, Y_t(1), \dots, Y_t(K))$ are i.i.d.
- $(Y_t(1), \dots, Y_t(K)) \perp W_t \mid X_t, H_{t-1}$.
- $e_t(x, w) > 0$ for all (x, w).

The provided observational data for policy evaluation and learning is $\{(X_t, W_t, Y_t, e_t(X_t, W_t))\}_{t=1}^{T}$.

Task I – Off-Policy Evaluation
Given a policy π, evaluate the policy value $Q(\pi) = E[\sum_w Y_t(w)\pi(X_t, w)]$ and construct confidence intervals around the estimate.

Task II – Off-Policy Learning
Given a policy class Π, learn a policy that achieves as large policy value as possible and characterize the regret bound, where regret $R(\pi) = Q(\pi^*) - Q(\pi)$, and π^* is the optimal policy in Π that achieves the largest policy value.

Generalized Augmented Inverse Propensity Weighted Estimator

Given a policy π, for each sample $(X_t, W_t, Y_t, e_t(X_t, W_t))$. The augmented inverse propensity weighted (AIPW) score is

$$\widehat{\Gamma}_t = \sum_w \pi(X_t, w)\left(\hat{\mu}_t(X_t, w) + \frac{1[W_t = w]}{e_t(X_t, W_t)}\left(Y_t - \hat{\mu}_t(X_t, w)\right)\right),$$

where $\hat{\mu}_t(X_t, w)$ is the nuisance estimator of $\mu(X_t, w) = E[Y(w)|X_t]$ fitted on historical data $\mathrm{H}_{t-1} = \{(X_s, W_s, Y_s)\}_{s=1}^{t-1}$.

- **Conditional unbiasedness**: $E\left[\widehat{\Gamma}_t \middle| H_{t-1}\right] = Q(\pi)$.
- **Conditional variance**: $Var\left(\widehat{\Gamma}_t \middle| H_{t-1}\right) = \Omega\left(E\left[\sum_w \frac{\pi^2(X_t, w)}{e_t(X_t, w)} \middle| H_{t-1}\right]\right)$.

The generalized AIPW estimator for policy value $Q(\pi)$ is $\hat{Q}_T(\pi) = \frac{\sum_{t=1}^{T} h_t \widehat{\Gamma}_t}{\sum_{t=1}^{T} h_t}$, where h_t are weights to stabilize the estimate.

Off-Policy Evaluation

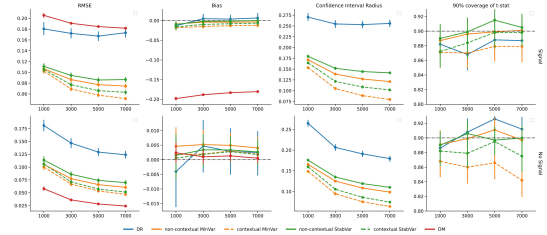
To evaluate a policy and construct valid confidence intervals, weights h_t are determined adaptively by policy π and historical samples H_{t-1}.

$$h_t = \phi\left(E\left[\sum_w \frac{\pi^2(X_t, w)}{e_t(X_t, w)} \middle| H_{t-1}\right]\right), \text{ where}$$

- StableVar: $\phi(v) = \sqrt{1/v}$, weights h_t approximately standardize $\widehat{\Gamma}_t$;
- MinVar: $\phi(v) = 1/v$, weights h_t approximately minimize the variance of $\hat{Q}_T$.

[Central Limit Theorem] Consider a fixed-arm policy w. Suppose that $\mathrm{Var}(Y_t(w)|X_t) \in [L, U]$ for some positive L, U, and $||\hat{\mu}_t||_\infty$ bounded, then $\hat{Q}_T(w)$ with StableVar weights is consistent for $Q(w)$. Suppose that in addition $\mathrm{E}[Y_t^4(w)|X_t] < \infty$, $e_t(\,, w) \geq Ct^{-\alpha}$ for some constants C and $\alpha \in [0, 0.5)$, $\hat{\mu}_t$ converges to μ_∞ a.s., and that either (i) e_t converges or (ii) $\mu_\infty = Q(w)$ and $E[(Y_t(w) - Q(w))^2|X_t]$ is constant a.e. Stablevar weighting yields an asymptotically normal studentized statistic

$$\frac{\hat{Q}_T(w) - Q(w)}{\sqrt{\hat{V}_T(w)}} \Rightarrow \mathcal{N}(0,1), \quad \text{where } \widehat{V_T}(w) = \frac{\sum_{t=1}^{T} h_t^2 (\widehat{\Gamma}_t - \hat{Q}_T(w))^2}{(\sum_{t=1}^{T} h_t)^2}.$$



Off-Policy Learning

To learn the optimal policy out of a class Π, we use pre-specified (*not* adaptive) weights h_t such as

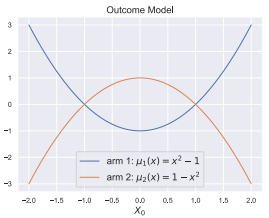
- uniform weighting: $h_t = 1$;
- optimal weighting (minimizes the regret bound): $h_t = g_t$ where g_t is the assignment probability floor $e_t \geq g_t$.

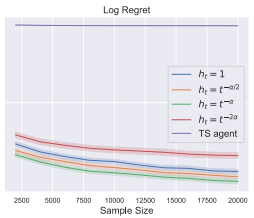
We output the policy $\hat{\pi}$ that maximizes $\hat{Q}_T(\pi)$ over the class Π.

[Finite-Sample Regret Bound] Suppose that Y_t is bounded uniformly by a positive constant M. Suppose that $e_t \geq Ct^{-\alpha}$ for some constants C. For any $\delta > 0$, with probability at least $1 - \delta$, the regret incurred by policy $\hat{\pi}$ with uniform weighting satisfies

$$R(\hat{\pi}) \leq \frac{M}{C} T^{\alpha - 0.5}\left(475\kappa(\Pi) + 1180 + 160\sqrt{\log\left(\frac{1}{\delta}\right)} + 160T^{-0.5}\right),$$

where $\kappa(\Pi)$ is the entropy integral defined under Hamming distance of policy class Π.





Acknowledgements

We are grateful for the generous financial support provided by the Sloan Foundation, Office of Naval Research grant N00014-17-1-2131, Office of Naval Research grant N0014-19-1-2468, National Science Foundation grant DMS-1916163, Schmidt Futures, Golub Capital Social Impact Lab, and the Stanford Institute for Human-Centered Artificial Intelligence. R. Z. also acknowledges generous support from the Total Fellowship and PayPal Fellowship. In addition, we thank Steve Howard, Sylvia Klosin, Sanath Kumar Krishnamurthy, Ruoxuan Xiong, and Aaditya Ramdas for helpful conversations.

References

1. Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, *118*(15).
2. Zhan, R., Hadad, V., Hirshberg, D. A., & Athey, S. (2021). Off-Policy Evaluation via Adaptive Weighting with Data from Contextual Bandits.
3. Zhan, R., Ren, Z., Athey, S., & Zhou, Z. (2021). Policy Learning with Adaptively Collected Data. *arXiv preprint arXiv:2105.02344*.