An Empirical Evaluation of Doubly Robust Learning for Recommendation

OLIVIER JEUNEN, Adrem Data Lab, University of Antwerp
BART GOETHALS, Adrem Data Lab, University of Antwerp; Faculty of Information Technology, Monash University

Methods for bandit learning from user interactions can broadly be divided into two camps: value-based methods that model the likelihood of an action leading to a positive reward, and policy-based methods that model a counterfactual estimate of the reward a given policy will accumulate. A unifying family of "Doubly Robust" approaches brings an explicit reward model into the counterfactual estimator in order to reduce its variance, and has been shown to consistently attain competitive results for various machine learning applications. Theory suggests that the reward estimator should be independent of the logged bandit feedback that is used to train the policy, which can impede its adoption in real-world environments where carefully logged samples are expensive to collect. Recent work provides an empirical analysis of the policy- and value-based families of approaches, but it remains unclear under which circumstances doubly robust learning can lead to a superior recommendation policy.

This work aims to fill that gap. We briefly present the doubly robust estimation framework and its extensions in a recommendation context, and present a wide range of empirical results using the RecoGym simulation framework, focusing on the use-case where logging propensities are known and the number of training samples is limited. In line with previous work, our results highlight that the stochasticity of the logging policy is the main factor deciding between the superiority of value- or policy-based methods. In contrast with previous work, our results indicate that recommendation policies learned via standard doubly robust estimation can often be outperformed by either their standalone value- or policy-based component. We discuss the implications of our results for the application of doubly robust learning methods in practice, and propose a scope for future research to further validate our findings.

ACM Reference Format:
Olivier Jeunen and Bart Goethals. 2020. An Empirical Evaluation of Doubly Robust Learning for Recommendation. In *Proceedings of the REVEAL Workshop on Bandit and Reinforcement Learning from User Interactions (REVEAL'20), September 26, 2020, Virtual Event, Brazil.* ACM, New York, NY, USA, 3 pages.

EXTENDED ABSTRACT

This work focuses on counterfactual learning for recommendation [5, 6, 9, 19]. We assume logged bandit feedback samples $\mathcal{D}$ consisting of N tuples $\{\boldsymbol{x}_i, a_i, \pi_0(a_i|\boldsymbol{x}_i), c_i\}$, respectively denoting contexts, actions (drawn from an action space $\mathcal{A}$), logging propensities and observed binary rewards. We represent the context or user state as a vector of length n containing counts of historical *organic* interactions with items. The goal at hand is to learn a fixed recommendation policy from this offline dataset that is able to obtain the largest cumulative reward when deployed. To this end, we simulate A/B-tests using the RecoGym environment [12] and evaluate policies on their best guesses: $a^* = \arg\max_{a \in \mathcal{A}} \pi_{\boldsymbol{\theta}}(a|\boldsymbol{x})$.

A common approach is to use the logged bandit feedback samples for Maximum Likelihood Estimation (MLE) to infer a logistic regression model that models the probability of a click for every given context-action pair: $P(C|\boldsymbol{x}, a) \approx \hat{c}(a|\boldsymbol{x})$. This reward estimator can either be used directly to obtain a^* (as done in [5, 6, 9]), or used to estimate the *value* of a policy per the Direct Method (DM; as done in [2] and shown in Eq. 1). For more details on the parameterisation of these methods, we refer to the work of Jeunen et al. [6]; additionally noting that this is in line with various "traditional" approaches to recommendation such as SLIM [10], EASE$^{\text{R}}$ [15] and extensions [7]. Moving away from the Direct Method, another widespread approach to counterfactual learning is to directly optimise a counterfactual estimate of the expected reward for $\pi_{\boldsymbol{\theta}}$ using importance sampling or Inverse Propensity Scoring (IPS) techniques, also shown in Eq. 1 [1, 11, 18].

2020. Manuscript submitted to ACM

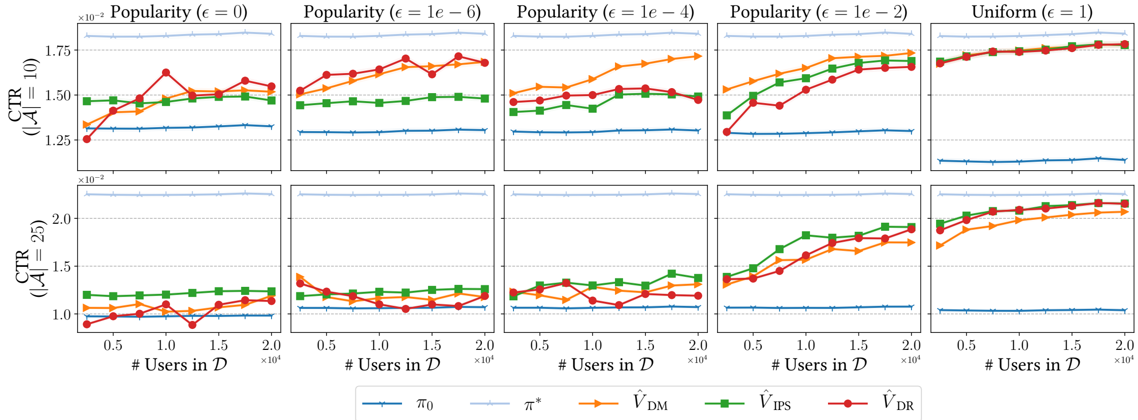

Fig. 1. Results from a range of A/B-tests for different settings in the RecoGym environment. Every column corresponds to a different level of stochasticity in an epsilon-greedy logging policy; rows correspond to $|\mathcal{A}| \in \{10, 25\}$. The x-axis shows the number of users in the training sample, the y-axis shows the measured CTR over 5 runs with 10 000 users, and the 95% confidence interval.

$$\hat{V}_{\text{DM}} = \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} \pi_{\boldsymbol{\theta}}(a|\boldsymbol{x}_i)\hat{c}(a|\boldsymbol{x}_i) \qquad\qquad \hat{V}_{\text{IPS}} = \sum_{i=1}^{N} \frac{\pi_{\boldsymbol{\theta}}(a_i|\boldsymbol{x}_i)}{\pi_0(a_i|\boldsymbol{x}_i)} c_i \tag{1}$$

The choice of π_0 has a direct impact on the quality of the fit for the reward estimator $\hat{c}(\cdot)$, as well as on the variance of the IPS weights $\frac{\pi_{\boldsymbol{\theta}}(\cdot)}{\pi_0(\cdot)}$. Doubly Robust (DR) estimators (Eq. 2) have been proposed to combine the DM and IPS methods, decreasing the variance in the value estimates and leading to better policies in mulit-class classification settings with simulated partial feedback [2]. Such settings are very different from the recommendation use-case, in terms of stochasticity of the rewards and the effect sizes between different actions. When and whether policies optimised for DR estimators actually lead to superior recommendation policies is the main research question we aim to answer.

$$\hat{V}_{\text{DR}} = \sum_{i=1}^{N} \left(\frac{\pi_{\boldsymbol{\theta}}(a|\boldsymbol{x}_i)}{\pi_0(a|\boldsymbol{x}_i)} \left(c_i - \hat{c}(a|\boldsymbol{x}_i)\right) + \sum_{a \in \mathcal{A}} \pi_{\boldsymbol{\theta}}(a|\boldsymbol{x}_i)\hat{c}(a|\boldsymbol{x}_i) \right) \tag{2}$$

Theory suggests that the logged samples used to optimise the estimator $\hat{c}$ should be independent from those used to optimise the policy $\pi_{\boldsymbol{\theta}}$. As carefully logged samples are often expensive to collect in real-world environments, we report results for the setting where the full training sample $\mathcal{D}$ is reused for both. Although this violates the theoretical assumptions made by Dudík et al. [2] in terms of quantifying variance reduction, we obtain superior results compared to randomly splitting the data in two parts to be used for optimising $\hat{c}$ and $\pi_{\boldsymbol{\theta}}$. Recent work shows that jointly optimising a value- and policy-based approach with a shared parameterisation on the same sample can attain state-of-the-art performance [6]. In future work, we wish to explore the benefits of a shared parameterisation for DR learning as well.

Figure 1 shows results for a range of epsilon-greedy based logging policies, highlighting how the superiority of either the DM or IPS estimator depends on the stochasticity of π_0, whilst demonstrating that the doubly robust approach is not guaranteed to beat both at the same time. In small action spaces with a decent amount of randomisation, for example, the variance reduction from DR *hurts* performance instead of improving it.

On top of theoretical justifications for our observations, we wish to extend our analysis to include: (1) wider ranges of settings in RecoGym such as larger action spaces; (2) recent extensions to the Doubly Robust paradigm, including MRDR [3], CAB [17] and DRs [16]; and (3) a comparison with competing state-of-the-art approaches such as POEM [18], BanditNet [8], BLOB [13], Dual Bandit [6] and a promising recent family of distributionally robust approaches [4, 14].

REFERENCES

[1] L. Bottou, J. Peters, J. Quiñonero-Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14, 1 (2013), 3207–3260.

[2] M. Dudík, J. Langford, and L. Li. 2011. Doubly Robust Policy Evaluation and Learning. In *Proc. of the 28th International Conference on International Conference on Machine Learning (ICML'11)*. 1097–1104.

[3] M. Farajtabar, Y. Chow, and M. Ghavamzadeh. 2018. More Robust Doubly Robust Off-policy Evaluation. In *Proc. of the 35th International Conference on Machine Learning (ICML'18, Vol. 80)*. PMLR, 1447–1456.

[4] L. Faury, U. Tanielian, F. Vasile, E. Smirnova, and E. Dohmatob. 2020. Distributionally Robust Counterfactual Risk Minimization. In *Proc. of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*. AAAI Press.

[5] O. Jeunen, D. Mykhaylov, D. Rohde, F. Vasile, A. Gilotte, and M. Bompaire. 2019. Learning from Bandit Feedback: An Overview of the State-of-the-art. arXiv:1909.08471 [cs.IR]

[6] O. Jeunen, D. Rohde, F. Vasile, and M. Bompaire. 2020. Joint Policy-Value Learning for Recommendation. In *Proc. of the 26th ACM Conference on Knowledge Discovery & Data Mining (KDD '20)*. ACM.

[7] O. Jeunen, J. Van Balen, and B. Goethals. 2020. Closed-Form Models for Collaborative Filtering with Side-Information. In *Proc. of the 14th ACM Conference onRecommender Systems (RecSys'20)*. ACM.

[8] T. Joachims, A. Swaminathan, and M. de Rijke. 2018. Deep Learning with Logged Bandit Feedback. In *Proc. of the 6th International Conference on Learning Representations (ICLR '18)*.

[9] D. Mykhaylov, D. Rohde, F. Vasile, M. Bompaire, and O. Jeunen. 2019. Three Methods for Training on Bandit Feedback. arXiv:1904.10799 [cs.IR]

[10] X. Ning and G. Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *Proc. of the 2011 IEEE 11th International Conference on Data Mining (ICDM '11)*. IEEE Computer Society, 497–506.

[11] A. B. Owen. 2013. *Monte Carlo theory, methods and examples*.

[12] D. Rohde, S. Bonner, T. Dunlop, F. Vasile, and A. Karatzoglou. 2018. RecoGym: A Reinforcement Learning Environment for the problem of Product Recommendation in Online Advertising. *ArXiv e-prints* (Aug. 2018). arXiv:1808.00720 [cs.IR]

[13] O. Sakhi, S. Bonner, D. Rohde, and F. Vasile. 2020. BLOB : A Probabilistic Model for Recommendation that Combines Organic and Bandit Signals. In *Proc. of the 26th ACM Conference on Knowledge Discovery & Data Mining (KDD '20)*. ACM.

[14] N. Si, F. Zhang, Z. Zhou, and J. Blanchet. 2020. Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits. In *International Conference on Machine Learning (ICML'20)*.

[15] H. Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference (WWW '19)*. ACM, 3251–3257.

[16] Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. 2019. Doubly robust off-policy evaluation with shrinkage. arXiv:1907.09623 [cs.LG]

[17] Y. Su, L. Wang, M. Santacatterina, and T. Joachims. 2019. CAB: Continuous Adaptive Blending for Policy Evaluation and Learning. In *International Conference on Machine Learning (ICML'19)*. 6005–6014.

[18] A. Swaminathan and T. Joachims. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research* 16, 1 (2015), 1731–1755.

[19] F. Vasile, D. Rohde, O. Jeunen, and A. Benhalloum. 2020. A Gentle Introduction to Recommendation as Counterfactual Policy Learning. In *Proc. of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. 392–393.