CAB: Continuous Adaptive Blending for Policy Evaluation and Learning

Yi Su [*1] **Lequn Wang** [*1] **Michele Santacatterina** [2] **Thorsten Joachims** [1]

Abstract

The ability to perform offline A/B-testing and off-policy learning using logged contextual bandit feedback is highly desirable in a broad range of applications, including recommender systems, search engines, ad placement, and personalized health care. Both offline A/B-testing and off-policy learning require a counterfactual estimator that evaluates how some new policy would have performed, if it had been used instead of the logging policy. In this paper, we present and analyze a family of counterfactual estimators which subsumes most estimators proposed to date. Most importantly, this analysis identifies a new estimator – called Continuous Adaptive Blending (CAB) – which enjoys many advantageous theoretical and practical properties. In particular, it can be substantially less biased than clipped Inverse Propensity Score (IPS) weighting and the Direct Method, and it can have less variance than Doubly Robust and IPS estimators. In addition, it is subdifferentiable such that it can be used for learning, unlike the SWITCH estimator. Experimental results show that CAB provides excellent evaluation accuracy and outperforms other counterfactual estimators in terms of learning performance.

1. Introduction

Contextual bandit feedback is ubiquitous in a wide range of intelligent systems that interact with their users through the following process. The system observes a context, takes an action, and then observes feedback under the chosen action. The logs of search engines, recommender systems, ad-placement systems, and many other systems contain terabytes of this partial-information feedback data, and it is highly desirable to use this historic data for offline evaluation and learning. What makes evaluation and learning challenging, however, is that we only get to see the feedback for the chosen action, but we do not observe how the user would have responded if the system had taken a different action. This means the feedback is both partial (e.g. only observed for the selected action) and biased (e.g. by the choices of the policy that logged the data), which makes batch learning from contextual bandit feedback substantially different from typical supervised learning, where the correct label and a loss function provide full-information feedback.

Both learning and evaluation can be viewed as examples of *counterfactual reasoning*, where we need to estimate from the historic log data how well some other policy would have performed, if we had used it instead of the policy that logged the data. Three main approaches have been proposed for this counterfactual or off-policy evaluation problem. First, the Direct Method (DM) (Schafer, 1997; Rubin, 2004; Dudík et al., 2011; Little & Rubin, 2019) uses regression to learn a model of the reward and imputes the missing feedback. Second, inverse propensity score (IPS) weighting (Horvitz & Thompson, 1952; Strehl et al., 2011) models the selection bias in the assignment mechanism and directly provides an unbiased estimate of the quality of a policy under suitable common support conditions. Both approaches are complementary and have different strengths and drawbacks. On the one hand, DM typically has low variance but can lead to highly biased results due to model misspecification. On the other hand, since we are typically controlling the logging policy and can log propensities, IPS-based methods can be provably unbiased but often suffer from large variance when the IPS weights are large. The third class is a hybrid of them. The most prominent one is the Doubly Robust (DR) estimator (Robins & Rotnitzky, 1995; Kang et al., 2007; Dudík et al., 2011), which is based on DM but also uses IPS weighting and an additive control variate to reduce variance.

Generalizing these existing counterfactual estimators, we present a parametric family of estimators for off-policy evaluation that covers most of the off-policy estimators proposed to date — including IPS (Horvitz & Thompson, 1952), clipped IPS (Strehl et al., 2011), DM (Dudík et al., 2011), DR (Dudík et al., 2011) and its MRDR variant (Farajtabar et al., 2018), SWITCH (Wang et al., 2017), and Static Blend-

[*]Equal contribution [1]Cornell University, Ithaca, USA [2]Cornell TRIPODS Center for Data Science, Ithaca, USA. Correspondence to: Yi Su <ys756@cornell.edu>, Lequn Wang <lw633@cornell.edu>, Michele Santacatterina <santacatterina@cornell.edu>, Thorsten Joachims <tj@cs.cornell.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

ing (SB) (Thomas & Brunskill, 2016). Providing a general bias-variance analysis for this family of estimators, we find that there is a particular new estimator in this family that has many desirable properties. We call this new estimator Continuous Adaptive Blending (CAB) and show how it blends the complementary strengths of DM and IPS, thus providing an effective tool for optimizing the bias of DM against the variance of IPS. Compared to existing estimators, we find that CAB can be substantially less biased than clipped IPS and DM while having lower variance compared to IPS and DR estimators. Furthermore, compared to estimators that perform static blending (SB) (Thomas & Brunskill, 2016), CAB is adaptive to the IPS weights and handles violations of the support condition gracefully. Unlike SWITCH (Wang et al., 2017), CAB is sub-differentiable which allows its use as the training objective in off-policy learning algorithms like POEM (Swaminathan & Joachims, 2015a) and BanditNet (Joachims et al., 2018). Finally, unlike the DR estimator, CAB can be used in off-policy Learning to Rank (LTR) algorithms like (Joachims et al., 2017), and CAB is specifically designed to control the bias/variance trade-off. We evaluate CAB both theoretically and empirically. In particular, we present theoretical results that characterize the bias and variance of CAB. Furthermore, we provide an extensive empirical evaluation of CAB on both contextual-bandit problems and partial-information ranking problems, including real-world data from Amazon Music.

2. Off-policy Evaluation in Contextual Bandits

Before presenting our general family of counterfactual estimators, we begin with a formal definition of off-policy evaluation and learning in the contextual-bandit setting. To keep notation and exposure simple, we focus on the contextual bandit setting and do not explicitly consider other partial-information settings like counterfactual LTR (Joachims et al., 2017). However, most estimators can be translated into that setting as well, and we discuss further details in the context of the ranking experiments in Section 4 and in Appendix A.

2.1. Contextual-Bandit Setting and Learning

In the contextual-bandit setting, a context $x \in \mathcal{X}$ (e.g., user profile, query) is drawn i.i.d. from some unknown $P(\mathcal{X})$, the deployed (stochastic) policy $\pi_0(y|x)$ then selects an action $y \in \mathcal{Y}$, and the system receives feedback $r \sim D(r|x,y)$ for this particular (context, action) pair. However, we do not observe feedback for any of the other actions. The logged contextual bandit data we get from logging policy π_0 is of the form

$$\mathcal{S} = \{(x_i, y_i, r_i, \pi_0(\cdot|x_i))\}_{i=1}^n, \tag{1}$$

where $r_i := r(x_i, y_i)$ is the observed reward. If the logging policy π_0 is unknown, an estimate $\hat{\pi}_0$ of the logging policy is used. Off-policy evaluation refers to the problem of using $\mathcal{S}$ for estimating the expected reward (or loss) R of a new policy π

$$R(\pi) = \mathbb{E}_{x\sim P(x)}\, \mathbb{E}_{\bar{y}\sim\pi(\bar{y}|x)}\, \mathbb{E}_{r\sim D(r|x,\bar{y})}[r]. \tag{2}$$

Off-policy learning uses $\mathcal{S}$ for finding an optimal policy $\pi^* \in \Pi$ that maximizes the expected reward (or minimizes the expected loss)

$$\pi^* = \operatorname*{argmax}_{\pi\in\Pi} \left[R(\pi)\right]. \tag{3}$$

The expected reward $R(\pi)$ cannot be computed directly, and it is typically replaced with a counterfactual estimate $\hat{R}(\pi)$ that can be computed from the logged bandit feedback $\mathcal{S}$ (Bottou et al., 2013). This enables Empirical Risk Minimization (ERM) for batch learning from bandit feedback (BLBF) (Zadrozny et al., 2003; Beygelzimer & Langford, 2009; Strehl et al., 2011; Swaminathan & Joachims, 2015a;b), where the algorithm finds the policy in Π that maximizes the estimated expected reward

$$\hat{\pi}^* = \operatorname*{argmax}_{\pi\in\Pi} \left[\hat{R}(\pi)\right], \tag{4}$$

possibly subject to capacity and variance regularization (Swaminathan & Joachims, 2015a). Since the counterfactual estimator $\hat{R}(\pi)$ is at the core of ERM learning, it is expected that an improved estimator will also lead to improved learning performance (Strehl et al., 2011). In particular, unlike in supervised learning, the counterfactual estimator can have vastly different bias and variance for different policies in Π, such that trading off bias and variance of the estimator becomes important for effective learning (Swaminathan & Joachims, 2015b).

2.2. Interpolated Counterfactual Estimator Family (ICE Family)

In this section, we present a general family of estimators for off-policy evaluation in the contextual bandit setting. An analogous family of estimators for the setting of partial-information learning-to-rank (Joachims et al., 2017) is described in Appendix A.

Let $\hat{\delta}(x, \bar{y})$ be the estimated reward for action $\bar{y}$ given context x, and let $\hat{\pi}_0$ be the estimated (or known) logging policy. Then our family of off-policy estimators is defined as follows, where $\mathbf{w} = (\mathrm{w}^\alpha, \mathrm{w}^\beta, \mathrm{w}^\gamma)$ is a triplet of weighting functions that parameterizes the family.

Definition (Interpolated Counterfactual Estimator Family)**.**

Table 1. Overview of how the family of estimators $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ subsumes existing off-policy estimators.

Estimator	$w_i^{\alpha}(\bar{y})$	w_i^{β}	w_i^{γ}
DM	1	0	0
IPS	0	1	0
cIPS	0	$\min\left\{\frac{M\hat{\pi}_0(y_i\|x_i)}{\pi(y_i\|x_i)}, 1\right\}$	0
DR	1	1	-1
SB	$1-\tau$	τ	0
SWITCH	$\mathbb{1}\left\{\frac{\pi(\bar{y}\|x_i)}{\hat{\pi}_0(\bar{y}\|x_i)} > M\right\}$	$\mathbb{1}\left\{\frac{\pi(y_i\|x_i)}{\hat{\pi}_0(y_i\|x_i)} \leq M\right\}$	0

Given a triplet $\mathbf{w} = (\mathrm{w}^{\alpha}, \mathrm{w}^{\beta}, \mathrm{w}^{\gamma})$ *of weighting functions,*

$$\hat{\mathrm{R}}^{\mathbf{w}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x_i)\ \mathrm{w}^{\alpha}_{i\bar{y}}\ \alpha_{i\bar{y}} + \frac{1}{n}\sum_{i=1}^{n}\pi(y_i|x_i)\ \mathrm{w}^{\beta}_{\mathrm{i}}\ \beta_i + \frac{1}{n}\sum_{i=1}^{n}\pi(y_i|x_i)\ \mathrm{w}^{\gamma}_{\mathrm{i}}\ \gamma_i$$

$$\text{with} \quad \alpha_{i\bar{y}} := \alpha(x_i,\bar{y}) := \hat{\delta}(x_i,\bar{y}),$$
$$\beta_i := \beta(x_i,y_i) := \frac{r(x_i,y_i)}{\hat{\pi}_0(y_i|x_i)},$$
$$\gamma_i := \gamma(x_i,y_i) := \frac{\hat{\delta}(x_i,y_i)}{\hat{\pi}_0(y_i|x_i)}.$$

We will see in the following that different choices of the weighting functions $\mathbf{w} = (\mathrm{w}^{\alpha}, \mathrm{w}^{\beta}, \mathrm{w}^{\gamma})$ recover a wide variety of existing estimators (see Table 1), and that our bias-variance analysis of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ suggests new estimators with desirable properties. This class of counterfactual estimators builds upon three basic estimators for off-policy evaluation, and we will now introduce the motivation for the construction of the (context, action) level functions $\alpha(x_i,\bar{y}), \beta(x_i,y_i)$ and $\gamma(x_i,y_i)$.

The first component $\alpha(x_i,\bar{y})$ follows a "Model the World" approach (Schafer, 1997; Rubin, 2004; Dudík et al., 2011), which exploits a model $\hat{\delta}(x,\bar{y})$ of the reward. The estimator that purely relies on this component is the DM estimator that is a special case of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ with static weights $\mathbf{w} = (1,0,0)$ for all $(x_i,y_i,\bar{y})$ (see Table 1). The reward model $\hat{\delta}(x,\bar{y})$ is typically learned via regression, and it serves as an estimate of $\mathbb{E}_r[r|x,\bar{y}]$ in (2) to be imputed in place of the unobserved rewards. The reward model typically has low variability, so the variance of DM is typically small. However, due to often unavoidable misspecification of the reward model, DM is often statistically inconsistent and can have a large bias.

The second component $\beta(x_i,y_i)$ follows a "Model the Bias" approach for the assignment mechanism, which is particularly attractive in many applications where we control the assignment mechanism by design (Horvitz & Thompson, 1952; Swaminathan & Joachims, 2015a; Nikos Vlassis, 2018). The estimator fully based on this term is the widely used inverse propensity score (IPS) weighting estimator (Horvitz & Thompson, 1952; Strehl et al., 2011; Bottou et al., 2013; Swaminathan & Joachims, 2015a), which is the special case of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ with static weights $\mathbf{w} = (0,1,0)$ for all $(x_i,y_i,\bar{y})$ (see Table 1). If $\hat{\pi}_0(y_i|x_i)$ is known and the logging policy has sufficient support w.r.t. the new policy π, then the IPS estimator is unbiased (see Section 2.3). However, it can have large variance if the IPS weights $\pi(y_i|x_i)/\hat{\pi}_0(y_i|x_i)$ are large.

The third component $\gamma(x_i,y_i)$ combines the "Model the World" approach and "Model the Bias" approach by debiasing the estimated reward term. This component is not necessarily an attractive estimator itself, but can be used as part of a control variate for variance reduction as in the DR estimator (Robins & Rotnitzky, 1995; Bang & Robins, 2005; Kang et al., 2007; Dudík et al., 2011). DR treats DM as a baseline while correcting the baseline when data is available, and it is a special case of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ with static weights $\mathbf{w} = (1,1,-1)$ for all $(x_i,y_i,\bar{y})$. It is unbiased when either the reward model or the propensity model is correct. However, by maintaining unbiasedness, we will discuss in Section 3.1 that it can still suffer from excessive variance. Furthermore, due to the non-zero weight put on the control variate term $\gamma(x_i,y_i)$, DR cannot be used for LTR from implicit feedback, since the analog of y_i is only partially observable in that setting (see Appendix A).

Another hybrid estimator was proposed by (Thomas & Brunskill, 2016) for off-policy evaluation in the more general setting of reinforcement learning. When we translate the key idea behind their MAGIC estimator to the contextual bandit setting, we see that it is a special case of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ with tunable weights $\mathbf{w} = (1-\tau,\tau,0)$ for all $(x_i,y_i,\bar{y})$, where $\tau \in [0,1]$ is a parameter. We call this estimator Static Blending (SB), since τ is static and does not depend on the importance weights.

The SWITCH Estimator (Wang et al., 2017), in contrast, is more adaptive. As the name implies, it switches between $\alpha(x_i,\bar{y})$ and $\beta(x_i,y_i)$ depending on a hard threshold M on the IPS weights. It is a special case of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ using the weights given in Table 1. A drawback of SWITCH is that its hard switching makes it discontinuous with respect to any parameters of the target policy π. This not only creates more erratic behavior when the threshold M is changed, but it also means that SWITCH is not differentiable and thus cannot be used in gradient-based learning algorithms like POEM (Swaminathan & Joachims, 2015a) or BanditNet (Joachims et al., 2018) for BLBF.

2.3. Theoretical Analysis

We will now provide a general characterization of the bias and variance for the Interpolated Counterfactual Estimator

Family, where both the propensity model and the reward model may be misspecified. Following Dudík et al. (2011), let $\zeta(x,y)$ denote the multiplicative deviation of the propensity estimates from the true propensity model, and $\Delta(x,y)$ be the additive deviation of the reward estimates from the true expected reward.

$$\zeta := \zeta(x,y) = 1 - \frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)} \tag{5}$$

$$\Delta := \Delta(x,y) = \hat{\delta}(x,y) - \delta(x,y). \tag{6}$$

Moreover, let $\sigma_r^2(x,y)$ denote the randomness of the reward.

$$\sigma_r^2 := \sigma_r^2(x,y) = \mathbb{V}_r(r(x,y)|x,y) \tag{7}$$

Note that $\zeta(x,y)$ is zero when the logging policy π_0 is known. For brevity, we denote the true IPS weight as $c(x,y) := \frac{\pi(y|x)}{\pi_0(y|x)}$ and its estimated version as $\hat{c}(x,y) := \frac{\pi(y|x)}{\hat{\pi}_0(y|x)}$. For known propensities, $\hat{c}(x,y) = c(x,y)$. As usual, we posit that the following support/positivity condition holds.

Condition 1 (Common Support)**.** *The logging policy π_0 has full support for the target policy π, which means $\pi(y|x) > 0 \rightarrow \pi_0(y|x) > 0$ for all x and y.*

Theorem 1 (Bias of the Interpolated Counterfactual Estimator Family)**.** *For contexts $x_1, x_2, \cdots, x_n$ drawn i.i.d from some distribution $P(\mathcal{X})$ and for actions $y_i \sim \pi_0(\mathcal{Y}|x_i)$, under Condition 1 the bias of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ with weighting functions $\mathbf{w} = (\mathrm{w}^\alpha, \mathrm{w}^\gamma, \mathrm{w}^\gamma)$ is*

$$\mathbb{E}_x \mathbb{E}_{y\sim\pi}\Big[\mathrm{w}^\alpha\,\Delta - \mathrm{w}^\beta\,\zeta\delta + \mathrm{w}^\gamma(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^\alpha + \mathrm{w}^\beta + \mathrm{w}^\gamma)\delta - \delta\Big] \tag{8}$$

Proof: See Appendix B.1.

Theorem 2 (Variance of the Interpolated Counterfactual Estimator Family)**.** *Under the same conditions as in Theorem 1, the variance of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ with weighting functions $\mathbf{w} = (\mathrm{w}^\alpha, \mathrm{w}^\beta, \mathrm{w}^\gamma)$ is*

$$\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\mathrm{w}^\alpha\,\Delta - \mathrm{w}^\beta\,\zeta\delta + \mathrm{w}^\gamma(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^\alpha + \mathrm{w}^\beta + \mathrm{w}^\gamma)\delta]\Big) + \mathbb{E}_x\mathbb{E}_\pi\Big[(\mathrm{w}^\beta)^2 c(1-\zeta)^2\sigma_r^2\Big] + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(\mathrm{w}^\beta\,c(1-\zeta)\delta + \mathrm{w}^\gamma\,c(1-\zeta)(\delta+\Delta))\Big]\Big\} \tag{9}$$

Proof: See Appendix B.2.

Throughout the rest of the paper, these general results will guide the design of new estimators, and they will allow us to compare different estimators within the family with respect to their bias/variance trade-offs.

3. Continuous Adaptive Blending Estimator (CAB)

In this section, we identify a new estimator within this family that has many desirable properties. In particular, the estimator arises naturally by filtering our family of estimators according to these properties. First, we would like the estimator to be unbiased if the reward model and the propensity model are correct, which can be achieved through constraining the weights $(\mathrm{w}^\alpha_{\mathrm{i}\bar{\mathrm{y}}}, \mathrm{w}^\beta_{\mathrm{i}}, \mathrm{w}^\gamma_{\mathrm{i}})$ to sum to 1 for each context-action pair. Second, the estimator should be applicable to a wide range of partial information settings, including learning to rank, which requires $\mathrm{w}^\gamma_{\mathrm{i}} = 0$. Third, we would like to achieve low MSE, which argues for data-dependent weights that allow an instance dependent trade-off between bias and variance. And, fourth, we would like to use the estimator for gradient-based learning, which implies that the weighting functions need to be (sub-)differentiable.

These desiderata and constraints lead us to the following new Continuous Adaptive Blending estimator (CAB).

$$\hat{R}_{CAB}(\pi) = \hat{\mathrm{R}}^{\mathbf{w}}(\pi) \text{ with } \begin{cases} \mathrm{w}^\alpha_{\mathrm{i}\bar{\mathrm{y}}} = 1 - \min\Big\{M\frac{\pi_0(\bar{y}|x_i)}{\pi(\bar{y}|x_i)}, 1\Big\} \\ \mathrm{w}^\beta_{\mathrm{i}} = \min\Big\{M\frac{\pi_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\Big\} \\ \mathrm{w}^\gamma_{\mathrm{i}} = 0 \end{cases}$$

It is easy to see that CAB interpolates between DM and IPS in an example-dependent way, which allows trade-off between bias and variance by controlling M. In particular, CAB inherits the idea that clipping is a data-dependent way to achieve smaller MSE. However, CAB imputes a regression estimate $\hat{\delta}(x_i, \bar{y})$ proportional to the clipped-off portion of the IPS weight – unlike clipped IPS (see Table 1) that implicitly imputes zero. Note that the particular choice of weights $(\mathrm{w}^\alpha_{\mathrm{i}\bar{\mathrm{y}}}, \mathrm{w}^\beta_{\mathrm{i}}, \mathrm{w}^\gamma_{\mathrm{i}})$ makes $\hat{R}_{CAB}(\pi)$ continuous and subdifferentiable with respect to the parameters of the policy π.

3.1. Bias and Variance Analysis

We now analyze the bias and variance of CAB as an instance of the counterfactual estimator family, and we will compare them to those of IPS, cIPS, DR, and DM.

Theorem 3 (Bias of CAB)**.** *For contexts $x_1, x_2, \cdots, x_n$ drawn i.i.d from some distribution $P(\mathcal{X})$ and for actions $y_i \sim \pi_0(\mathcal{Y}\,|x_i)$, under Condition 1 the bias of $\hat{R}_{CAB}(\pi)$ is*

$$\mathbb{E}_x\mathbb{E}_\pi\Big[-\delta\zeta\mathbb{1}\{\hat{c}\le M\} + \{\Delta(1-\frac{M}{c(1-\zeta)}) - \frac{M}{c(1-\zeta)}\delta\zeta\}\mathbb{1}\{\hat{c}>M\}\Big] \tag{10}$$

Proof: See Appendix B.3

Note that the first part of the bias results from the use of IPS when the IPS weight is small, while the second part results

from the convex combination of IPS and DM when the IPS weight is large.

Theorem 4 (Variance of CAB)**.** *Under the same conditions as in Theorem 3, the variance of $\hat{R}_{CAB}(\pi)$ is*

$$\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\delta - \delta\zeta\mathbb{1}\{\hat{c}\le M\} + (\Delta(1-\frac{M}{c(1-\zeta)}) - \frac{M}{c(1-\zeta)}\delta\zeta)\mathbb{1}\{\hat{c}>M\}]\Big) + \mathbb{E}_x\mathbb{E}_\pi\Big[c(1-\zeta)^2\sigma_r^2\mathbb{1}\{\hat{c}\le M\} + \frac{M^2}{c}\sigma_r^2\mathbb{1}\{\hat{c}>M\}\Big] + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(c(1-\zeta)\delta\mathbb{1}\{\hat{c}\le M\} + M\delta\mathbb{1}\{\hat{c}>M\})\Big]\Big\} \tag{11}$$

Proof: See Appendix B.4

The first term of the variance is due to the randomness in context x, the second term results from the randomness in the rewards compounded with *bounded* IPS weights. The third term is, in expectation, the variability in the expected reward compounded with *bounded* IPS weights.

Bias improvements over cIPS and DM. We can now compare the bias of CAB to that of cIPS, which we can derive as a special case of Theorem 3 with $\hat{\delta}(x_i,\bar{y}) = 0$ for all $(x_i,\bar{y})$ pair. Focusing on the case of logged propensities for conciseness and its real-world prevalence in online systems, this reduces to $Bias(\hat{R}_{cIPS}(\pi)) = \mathbb{E}_x\mathbb{E}_\pi[-\delta(1-\frac{M}{c})\mathbb{1}\{c>M\}]$ for cIPS and $Bias(\hat{R}_{CAB}(\pi)) = \mathbb{E}_x\mathbb{E}_\pi[\Delta(1-\frac{M}{c})\mathbb{1}\{c>M\}]$ for CAB. It can be seen that if we have a moderately good predictor of the expected reward $\delta(x,\bar{y})$, CAB will have an advantage as long as the predictor is better than imputing the constant 0 everywhere. In practice, it is sensible to assume that the reward estimation error $\Delta(x,\bar{y})$ is substantially smaller than $\delta(x,\bar{y})$, such that CAB enjoys a substantial amount of bias reduction.

In comparison to DM, CAB can also enjoy smaller bias when the propensity is known, since $Bias(\hat{R}_{DM}(\pi)) = \mathbb{E}_x\mathbb{E}_\pi[\Delta]$ while $Bias(\hat{R}_{CAB}(\pi)) = \mathbb{E}_x\mathbb{E}_\pi[\Delta(1-\frac{M}{c})\mathbb{1}\{c>M\}]$ which reflects that CAB incurs bias only on the clipped portion of the importance sample weights.

Variance improvements over IPS and DR. Comparing the variance of CAB to that of IPS and DR, we again focus on the case of logged propensities. From Theorem 2 we can deduce that the variance of IPS is

$$\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\delta]\Big) + \mathbb{E}_x\mathbb{E}_\pi\Big[c\sigma_r^2\Big] + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(c\delta)\Big]\Big\} \tag{12}$$

while the variance for CAB is

$$\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\delta + \Delta(1-\frac{M}{c})\mathbb{1}\{c>M\}]\Big) + \mathbb{E}_x\mathbb{E}_\pi\Big[c\sigma_r^2\mathbb{1}\{c\le M\} + \frac{M^2}{c}\sigma_r^2\mathbb{1}\{c>M\}\Big] + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}[(c\delta)\mathbb{1}\{c\le M\} + M\delta\mathbb{1}\{c>M\}]\Big]\Big\}. \tag{13}$$

The first term is similar for both estimators since δ can be expected to dominate Δ. The second and the third terms, which are the variance of the reward $r(x,\bar{y})$ and the expected reward $\delta(x_i,\bar{y})$ compounded with the IPS weights c, can be very large for IPS when the logging policy π_0 and the target policy π are very different. In contrast, for CAB these two terms are bounded by $M\mathbb{E}_\pi[\sigma_r^2] + M^2\mathbb{E}_x[\mathbb{V}_{\pi_0}(\delta)]$, and thus will be smaller than those for IPS.

Comparing CAB to DR, note that DR intends to reduce the variance of IPS by putting weight 1 on the observed loss term $\beta(x_i,y_i)$ and -1 on the estimated loss term $\gamma(x_i,y_i)$. However, this "residual term" is still compounded with the IPS weights c and can blow up the variance either when we have a poor estimate Δ or when the target policy is very different from the logging policy. This is apparent in the second and third terms of the variance of DR as derived from Theorem 2.

$$\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\delta]\Big) + \mathbb{E}_x\mathbb{E}_\pi\Big[c\sigma_r^2\Big] + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(c\Delta)\Big]\Big\} \tag{14}$$

This is again different from CAB, where the IPS weights are all bounded by M.

3.2. CAB-DR

Finally, we present a variant of CAB that incorporates the DR estimator, called CAB-DR. While this leads to an estimator that cannot be used for ranking, we investigate whether the control variate of DR leads to even better estimates than CAB. The key idea is to substitute the IPS part of CAB with the DR estimator, leading to

$$\hat{R}_{CABDR}(\pi) = \hat{\mathrm{R}}^{\mathbf{w}}(\pi) \text{ with } \begin{cases} \mathrm{w}_{\mathrm{i}\bar{\mathrm{y}}}^{\alpha} = 1 \\ \mathrm{w}_{\mathrm{i}}^{\beta} = \min\Big\{M\frac{\pi_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\Big\} \\ \mathrm{w}_{\mathrm{i}}^{\gamma} = -\min\Big\{M\frac{\pi_0(\bar{y}|x_i)}{\pi(\bar{y}|x_i)}, 1\Big\} \end{cases}$$

as a clipped version of DR. Using Theorem 1, it is easy to derive the bias of CAB-DR to be

$$\mathbb{E}_x\,\mathbb{E}_\pi\Big[\zeta\Delta\mathbb{1}\{\hat{c}\le M\} + \Delta(1-\frac{M}{c})\mathbb{1}\{\hat{c}>M\}\Big].$$

If the propensities are logged, then the bias terms for CAB and CAB-DR are identical. However, if the propensities are only approximates, CAB will suffer from more bias from the term $\mathbb{E}_x\,\mathbb{E}_\pi[\zeta\delta\mathbb{1}\{\hat{c}\le M\}]$ compared to $\mathbb{E}_x\,\mathbb{E}_\pi[\zeta\Delta\mathbb{1}\{\hat{c}\le M\}]$ for CAB-DR.

The variance of CAB-DR is given in Appendix B.5. Given logged propensities, the variance of CAB-DR differs from that of CAB in only a single term. For CAB-DR we have

$$\mathbb{E}_x\left[\mathbb{V}_{\pi_0}\left(c(-\Delta)\mathbb{1}\{c \le M\} + M(-\Delta)\mathbb{1}\{c > M\}\right)\right],$$

while for CAB we have

$$\mathbb{E}_x\left[\mathbb{V}_{\pi_0}\left(c(\delta)\mathbb{1}\{c \le M\} + M(\delta)\mathbb{1}\{c > M\}\right)\right].$$

In general cases, by adopting the idea of choosing opposite weights for $\beta(x_i, y_i)$ and $\gamma(x_i, y_i)$ from DR, the third variance term for CAB-DR becomes $\mathbb{E}_x[\mathbb{V}_\pi(-\mathrm{w}^\beta\, c(1-\zeta)\Delta)]$, which is no longer on the order of $\mathbb{E}_x[\mathbb{V}_\pi(\mathrm{w}^\beta\, c(1-\zeta)\delta)]$.

4. Experiments

We empirically examine the evaluation accuracy and learning performance of CAB in two different partial-information settings. In the BLBF setting, we conduct the experiments on bandit feedback data for multi-class classification. This setting is extensively used in the off-policy evaluation literature (Dudík et al., 2011; Wang et al., 2017). In the LTR setting, the experiments are based on user feedback with position bias for ranking (Joachims et al., 2017). In both cases, we use real datasets from which we sample synthetic bandit or click data. This increases the external validity of the experiments, while at the same time providing ground truth for a bias/variance analysis. Furthermore, it allows us to vary the properties of both data and logging policy π_0 to explore the robustness of the estimators.

4.1. Experiment Setup

For the BLBF setting, our experiment setup follows Dudík et al. (2011) and Wang et al. (2017) using the standard supervised → bandit conversion (Agarwal et al., 2014) for several multiclass classification datasets from the UCI repository (Asuncion & Newman, 2007). In the LTR setting, we follow the experiment setup of Joachims et al. (2017) and conduct experiments on the YAHOO! LTR Challenge corpus (set 1), which comes with a train/validation/test split. More details are given in Appendix A.

In both settings, we use a small amount of the full-information training data to train a logger π_0 and a regression model $\hat{\delta}$. The policy π to be evaluated is trained on the whole training set. The partial feedback data is generated from the full-information test set. We evaluate the policy π with different estimators on the partial feedback data of different sizes and treat the performance of the full-information test set as the ground truth. The performance is measured by the expected test error and the average rank of positive results for BLBF and LTR respectively. We repeat the experiments 500 times for BLBF and 100 times for LTR to calculate bias, variance and MSE.

For the BLBF learning experiments, we use POEM (Swaminathan & Joachims, 2015a) to learn stochastic linear policies. For LTR, Appendix C.2 derives a generalized version of propensity SVM-Rank (Joachims et al., 2017) that enables the use of CAB and other estimators with $\mathrm{w}_\mathrm{i}^\gamma = 0$ from our family. As input to the learning algorithms, different amounts of partial feedback data are simulated from the full-information training data. To avoid biases from the regression model, we adopt 90 percentile cIPS to conduct hyperparameter selection for M(or τ for SB) and regularization parameter on the partial feedback data simulated from the validation set. The experiments are run for 10 and 5 times on BLBF and LTR respectively and the average is reported. Details are shown in Appendix C.

4.2. Experiment Results

Can CAB achieve improved estimation accuracy by trading bias for variance through M? We first verify that CAB can indeed achieve improved MSE by adjusting the bias-variance trade-off. Figure 1 shows how the choice of M affects bias and variance of CAB on the SATIMAGE and YAHOO! LTR datasets using different amounts of data (qualitatively similar results are obtained for the other datasets). For each dataset, the bias decreases as we increase M as expected, since CAB moves towards the unbiased IPS estimator. However, the variance increases as more data points rely on IPS weighting. For all data set sizes, the best MSE always falls in the middle of the range of M which confirms that CAB can effectively trade-off between bias and variance through controlling M for a range of data-set sizes. Moreover, the MSE curve suggests that the performance of CAB is pretty robust to the choice of M.

How does CAB compare to other off-policy estimators? Figure 2 compares different off-policy estimators on the 4 UCI datasets (selecting those with dataset size larger than 5000) and the YAHOO! LTR dataset. For each UCI dataset, we keep the test data size as 2,000. For the LTR dataset, we present the results with 0.5 sweeps of the test set. Notice that CAB , CAB-DR, cIPS and SWITCH all have a clipping parameter $M \in [0, \infty)$, while for SB the blending is achieved through the static weight parameter $\tau \in [0, 1]$. To be able to plot SB together with the other estimators, we rescale τ in the plot for comparison. DM and DR do not have any hyperparameter, so we use two horizontal lines to represent them.

For both SB and CAB (CAB-DR), and across all datasets, we observe a U-shape curve for MSE with the optimum value in the middle. However, on most datasets CAB (and CAB-DR) substantially outperforms SB. Furthermore, CAB outperforms cIPS in the full range of M on all datasets, indicating that imputing a reasonably good regression estimate is indeed consistently better than naively imputing

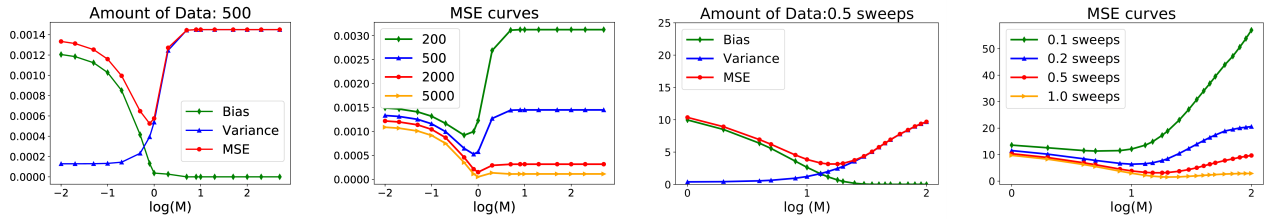

Figure 1. The Bias, Variance and MSE graph for CAB on the SATIMAGE and YAHOO! LTR dataset. From left to right: (a) The bias, variance and MSE curves for the SATIMAGE dataset. (b) MSE curves for SATIMAGE when we vary the amount of log data. (c) The bias, variance and MSE curves for the YAHOO! LTR dataset. (d) MSE curves for YAHOO! LTR when we vary the amount of log data.

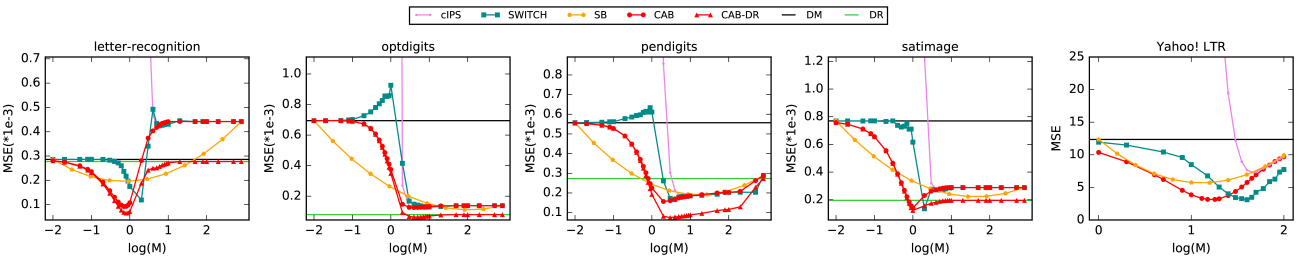

Figure 2. MSE comparison of various off-policy estimators for different datasets

zero. For the SWITCH estimator, the MSE curve is somewhat more erratic than that of CAB especially on the UCI datasets, which we conjecture is due to the hard switch it makes and the discontinuities this implies. While SWITCH can be used in LTR algorithms like Propensity SVM-Rank (Joachims et al., 2017) (details are shown in Appendix A), we show in the next section that this behavior may make model selection during learning more stable for CAB than for SWITCH. Furthermore, CAB performs at least comparable to SWITCH across all datasets, and on some it can be substantially more accurate than SWITCH.

For all datasets, DR outperforms IPS and DM as expected. However, CAB (also CAB-DR) still outperforms DR on most datasets, which validates the idea that estimators outside the class of unbiased estimators can have advantages on this problem. Furthermore, when used in learning, one already faces a bias/variance trade-off due to the capacity of the policy space, such that it seems unjustified to insist on the unbiasedness of the empirical risk to begin with.

How robust is CAB on real-world data? We evaluated CAB on data from a contextual bandit problem at Amazon Music. Both the logging policy and the target policy are a Thompson sampling contextual bandit algorithm for which we estimated the respective policy $\pi_0^t(y|x)$ and $\pi_{target}^t(y|x)$ at each time step t through Monte Carlo sampling. When analyzing the logging distribution, we found that the logging policy does not provide full support as the Thompson sampler converges, and on average only 33% of the available actions have non-zero support. We used the model

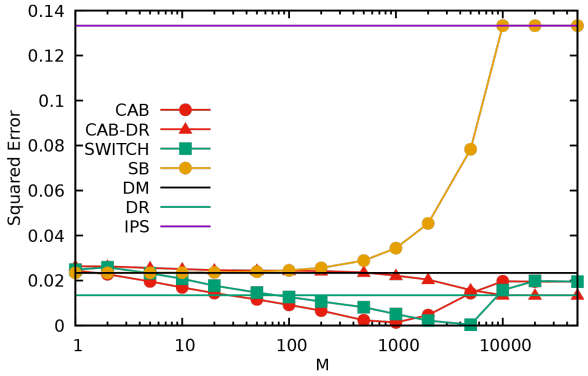

Figure 3. Error of the estimates on the Amazon Music contextual bandit problem.

learned by the Thompson sampler of the logging policy as the regression imputation model, $\hat{\delta}(x, y)$.

Figure 3 shows the error of the estimates depending on the clipping constant M, where we use the average reward of $\pi_{target}^t(y|x)$ measured during online A/B testing as the gold standard. For the majority of the values of the clipping constant M, CAB and SWITCH show a higher level of accuracy than DM, IPS and SB. Both IPS and SB perform poorly. This is due to the fact that IPS has no mechanism for detecting and correcting for the missing support of the logging policy, while SB has a static blending constant. For a large range of M, CAB and SWITCH also outperform DR and CAB-DR. Overall, we find that both CAB and SWITCH outperform the other methods and can provide

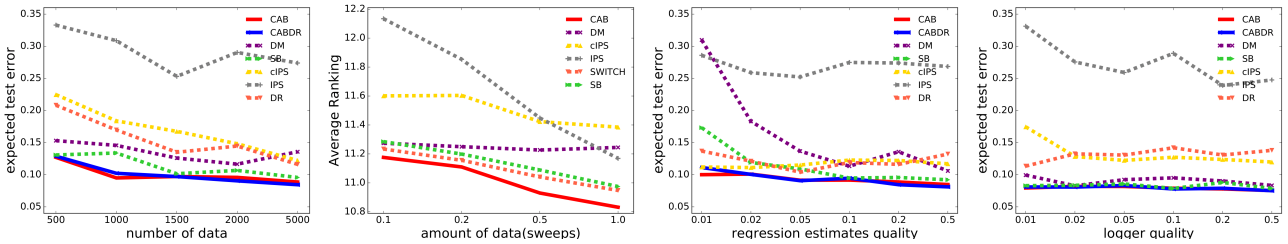

Figure 4. Test set learning performance under various scenarios. (1)-(2) Performance vs. amount of data for PENDIGITS and YAHOO! LTR. (3) Performance vs. regression model quality for PENDIGITS. (4) Performance vs. logger quality for PENDIGITS.

robust solutions to the contextual bandit problem at Amazon Music.

How effective is learning with CAB as empirical risk across datasets? Table 2 shows the learning performance when the various estimators are used as empirical risk in POEM and Propensity SVM-Rank. We use the same datasets as in the evaluation experiments. For the 4 UCI datasets (with 5000 training data for each) we report the average test set error, while for YAHOO! LTR (with 1 sweep of data) we report the average rank of the relevant results for the test queries. For both lower is better. Across all datasets, learning with CAB performs very well, providing one of the best prediction performances on all datasets which validates that learning with an improved estimator will also lead to improved learning performance. Comparing CAB with CAB-DR, we don't see a real benefit in using the DR model in CAB-DR instead of using IPS in CAB. We conjecture that it is due to the over-reliance on the regression model, as DR relies more on it than IPS.

Table 2. Test set learning performance on various datasets.

DATA	LETTER	OPTDIGITS	SATIMAGE	PENDIGITS	YAHOO! LTR
DM	0.6372	0.0649	0.3083	0.1133	11.25
DR	0.6852	0.0471	0.2762	0.1191	-
IPS	0.8969	0.0695	0.3266	0.2748	11.17
CIPS	0.8504	0.0447	**0.2415**	0.1228	11.39
SB	0.6091	0.0460	0.2481	0.0949	10.98
SWITCH	-	-	-	-	10.95
CAB	**0.5740**	**0.0445**	0.2442	**0.0917**	**10.83**
CAB-DR	0.5877	0.0461	0.2762	0.0946	-

How does learning performance change when we increase the amount of training data? Plots (1) and (2) in Figure 4 show test set performance when we increase the amount of training data for PENDIGITS and YAHOO! LTR. We find that performance improves for all estimators as the amount of data increases. However, for all training data sizes, we observe that CAB and CAB-DR perform very well compared to the other estimators, with a substantial improvement over IPS, cIPS and DR especially in the small sample setting. Furthermore, CAB performs better than SWITCH in the ranking setting, which we attribute to the more erratic behavior of SWITCH during model selection.

How does learning performance change when we vary the quality of the estimated regression model? In order to vary the quality of the regression model used by DM, DR, CAB, CAB-DR, SWITCH and SB, we vary the fraction of full-information data that is used to learn the regression model. The results are shown in plot (3) of Figure 4. IPS and cIPS result in two flat lines (up to variance), as they do not rely on the regression model. The other methods improve with the quality of the regression model, and both CAB and CAB-DR do well over the whole range.

How does learning performance change when we vary the quality of the logging policy? Plot (4) of Figure 4 shows the results when changing the logger quality. To do so, we used different fractions of the full-information data to train the logging policy, from 0.01 to 0.5. Since the DM estimator is independent of the logging policy, it results in a flat line (up to variance). IPS and cIPS are heavily affected by the logger quality, due to their dependency on the IPS weights, while CAB, CAB-DR and SB are only moderately affected. Overall, CAB and CAB-DR perform well across the whole range.

5. Conclusion

This paper proposed a parametric family of estimators for off-policy evaluation, which unifies and characterizes a number of popular off-policy estimators. The theoretical analysis also motivates the CAB estimator, which not only provides a controllable bias/variance trade-off for off-policy evaluation, but is also continuous with respect to the target policy to enable gradient-based learning. We argue theoretically that CAB can be less biased than cIPS and DM and often enjoys smaller variance than IPS and DR. Experiment results on two different partial-information settings – contextual bandit and partial-information LTR – confirm that CAB can consistently achieve improved evaluation performance over other counterfactual estimators, and that it also leads to excellent learning performance.

Acknowledgements

This research was supported in part by NSF Awards IIS-1615706 and IIS-1513692, as well as a gift from Amazon. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2014.

Agarwal, A., Zaitsev, I., Wang, X., Li, C., Najork, M., and Joachims, T. Estimating position bias without intrusive interventions. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 474–482. ACM, 2019.

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Beygelzimer, A. and Langford, J. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 129–138. ACM, 2009.

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

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2011.

Fang, Z., Agarwal, A., and Joachims, T. Intervention harvesting for context-dependent examination-bias estimation. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning (ICML)*, 2018.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Joachims, T. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

Joachims, T. Training linear SVMs in linear time. In *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD)*, 2006.

Joachims, T., Swaminathan, A., and Schnabel, T. Unbiased learning-to-rank with biased feedback. In *ACM Conference on Web Search and Data Mining (WSDM)*, 2017.

Joachims, T., Swaminathan, A., and de Rijke, M. Deep learning with logged bandit feedback. In *International Conference on Learning Representations (ICLR)*, 2018.

Kang, J. D., Schafer, J. L., et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. Wiley, 2019.

Nikos Vlassis, Aurelien Bibaut, T. J. On the design of estimators for off-policy evaluation. In *REVEAL Workshop at RecSys*, 2018.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Rubin, D. B. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley and Sons, 2004.

Schafer, J. L. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

Strehl, A., Langford, J., Li, L., and Kakade, S. M. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research (JMLR)*, 16:1731–1755, Sep 2015a. Special Issue in Memory of Alexey Chervonenkis.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Neural Information Processing Systems (NIPS)*, 2015b.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., and Najork, M. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 610–618. ACM, 2018.

Wang, Y.-X., Agarwal, A., and Dudik, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning (ICML)*, 2017.

Zadrozny, B., Langford, J., and Abe, N. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 435–442. IEEE, 2003.