Efficient Counterfactual Learning from Bandit Feedback

Yusuke Narita
Yale University
yusuke.narita@yale.edu

Shota Yasui
CyberAgent Inc.
yasui_shota@cyberagent.co.jp

Kohei Yata
Yale University
kohei.yata@yale.edu

Abstract

What is the most statistically efficient way to do off-policy optimization with batch data from bandit feedback? For log data generated by contextual bandit algorithms, we consider offline estimators for the expected reward from a counterfactual policy. Our estimators are shown to have lowest variance in a wide class of estimators, achieving variance reduction relative to standard estimators. We then apply our estimators to improve advertisement design by a major advertisement company. Consistent with the theoretical result, our estimators allow us to improve on the existing bandit algorithm with more statistical confidence compared to a state-of-the-art benchmark.

1 Introduction

Interactive bandit systems (e.g. personalized education and medicine, ad/news/recommendation/search platforms) produce log data valuable for evaluating and redesigning the systems. For example, the logs of a news recommendation system record which news article was presented and whether the user read it, giving the system designer a chance to make its recommendation more relevant. Exploiting log data is, however, more difficult than conventional supervised machine learning: the result of each log is only observed for the action chosen by the system (e.g. the presented news) but not for all the other actions the system could have taken. Moreover, the log entries are biased in that the logs over-represent actions favored by the system.

A potential solution to this problem is an A/B test that compares the performance of counterfactual systems. However, A/B testing counterfactual systems is often technically or managerially infeasible, since deploying a new policy is time- and money-consuming, and entails a risk of failure.

This leads us to the problem of *counterfactual (off-policy) evaluation and learning*, where one aims to use batch data collected by a logging policy to estimate the value of a counterfactual policy or algorithm without employing it (Li et al. 2010; Strehl et al. 2010; Li et al. 2011; 2012; Bottou et al. 2013; Swaminathan and Joachims 2015a; 2015b; Wang, Agarwal, and Dudik 2017; Swaminathan et al. 2017). Such evaluation allows us to compare the performance of counterfactual policies to decide which policy should be deployed in the field. This alternative approach thus solves the above problem with the naive A/B test approach.

Method. For off-policy evaluation with log data of bandit feedback, this paper develops and empirically implements a variance minimization technique. Variance reduction and statistical efficiency are important for minimizing the uncertainty we face in decision making. Indeed, an important open question raised by Li (2015) is how to achieve "statistically more efficient (even optimal) offline estimation" from batch bandit data. This question motivates a set of studies that bound and characterize the variances of particular estimators (Dudík et al. 2014; Li, Munos, and Szepesvári 2015; Thomas, Theocharous, and Ghavamzadeh 2015; Munos et al. 2016; Thomas and Brunskill 2016; Agarwal et al. 2017).

We study this statistical efficiency question in the context of offline policy value estimation with log data from a class of contextual bandit algorithms. This class includes most of the widely-used algorithms such as contextual ϵ-Greedy and Thompson Sampling, as well as their non-contextual analogs and random A/B testing. We allow the logging policy to be unknown, degenerate (non-stochastic), and time-varying, all of which are salient in real-world bandit applications. We also allow the evaluation target policy to be degenerate, again a key feature of real-life situations.

We consider offline estimators for the expected reward from a counterfactual policy. Our estimators can also be used for estimating the average treatment effect. Our estimators are variations of well-known inverse probability weighting estimators (Horvitz and Thompson (1952), Rosenbaum and Rubin (1983), and modern studies cited above) except that we use an *estimated* propensity score (logging policy) even if we know the true propensity score. We show the following result, building upon Bickel et al. (1993), Hirano, Imbens, and Ridder (2003), and Ackerberg et al. (2014) among others:

Theoretical Result 1. Our estimators minimize the variance among all reasonable estimators. More precisely, our estimators minimize the asymptotic variance among all "asymptotically normal" estimators (in the standard statistical sense defined in Section 3).

We also provide estimators for the asymptotic variances of our estimators, thus allowing analysts to calculate the variance in practice. In contrast to Result 1, we also find:

> **Theoretical Result 2.** Standard estimators using the true propensity score (logging policy) have larger asymptotic variances than our estimators.

Perhaps counterintuitively, therefore, the policy-maker should use an estimated propensity score even when she knows the true one.

Application. We empirically apply our estimators to evaluate and optimize the design of online advertisement formats. Our application is based on proprietary data provided by CyberAgent Inc., the second largest Japanese advertisement company with about 6 billion USD market capitalization (as of November 2018). This company uses a contextual bandit algorithm to determine the visual design of advertisements assigned to users. Their algorithm produces logged bandit data.

We use this data and our estimators to optimize the advertisement design for maximizing the click through rates (CTR). In particular, we estimate how much the CTR would be improved by a counterfactual policy of choosing the best action (advertisement) for each context (user characteristics). We first obtain the following result:

> **Empirical Result A.** Consistent with Theoretical Results 1-2, our estimators produce narrower confidence intervals about the counterfactual policy's CTR than a benchmark using the known propensity score (Swaminathan and Joachims 2015b).

This result is reported in Figure 1, where the confidence intervals using "True Propensity Score (Benchmark)" are wider than other confidence intervals using propensity scores estimated either by the Gradient Boosting, Random Forest, or Ridge Logistic Regression.

Thanks to this variance reduction, we conclude that the logging policy's CTR is below the confidence interval of the hypothetical policy of choosing the best advertisement for each context. This leads us to obtain the following bottom-line:

> **Empirical Result B.** Unlike the benchmark estimator, our estimator predicts the hypothetical policy to statistically significantly improve the CTR by 10-15% (compared to the logging policy).

Empirical Results A and B therefore show that our estimator can substantially reduce uncertainty we face in real-world policy-making.

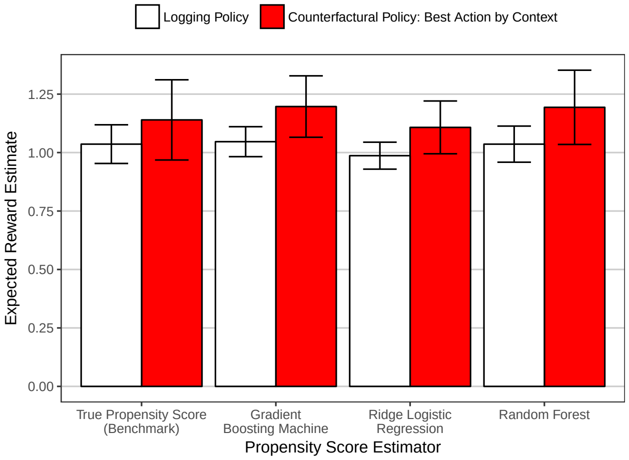

Figure 1: Improving Ad Design with Lower Uncertainty

Notes: This figure shows estimates of the expected CTRs of the logging policy and a counterfactual policy of choosing the best action for each context. CTRs are multiplied by a constant for confidentiality reasons. We obtain these estimates by the "self-normalized inverse probability weighting estimator" using the true propensity score (benchmark thanks to Swaminathan and Joachims (2015b)) or estimated propensity scores (our proposal), both of which are defined and analyzed in Section 3. Bars indicate 95% confidence intervals based on our asymptotic variance estimators developed in Section 4.

2 Setup

2.1 Data Generating Process

We consider a general multi-armed contextual bandit setting. There is a set of $m+1$ *actions* (equivalently, *arms* or *treatments*), $\mathcal{A} = \{0, ..., m\}$, that the decision maker can choose from. Let $Y(\cdot) : \mathcal{A} \to \mathbb{R}$ denote a potential reward function that maps actions into rewards or outcomes, where $Y(a)$ is the reward when action a is chosen (e.g., whether an advertisement as an action results in a click). Let X denote *context* or *covariates* (e.g., the user's demographic profile and browsing history) that the decision maker observes when picking an action. We denote the set of contexts by $\mathcal{X}$. We think of $(Y(\cdot), X)$ as a random vector with unknown distribution G.

We consider log data coming from the following data generating process (DGP), which is similar to those used in the literature on the offline evaluation of contextual bandit algorithms (Li et al. 2010; Strehl et al. 2010; Li et al. 2011; 2012; Swaminathan and Joachims 2015a; 2015b; Swaminathan et al. 2017). We observe data $\{(Y_t, X_t, D_t)\}_{t=1}^{T}$ with T observations. $D_t \equiv (D_{t0}, ..., D_{tm})'$ where D_{ta} is a binary variable indicating whether action a is chosen in round t. Y_t denotes the reward observed in round t, i.e., $Y_t \equiv \sum_{a=0}^{m} D_{ta} Y_t(a)$. X_t denotes the context observed in round t.

A key feature of our DGP is that the data $\{(Y_t, X_t, D_t)\}_{t=1}^{T}$ are divided into B batches, where different batches may use different choice probabilities (propensity scores). Let $X_t^b \in \{1, 2, ..., B\}$ denote a random variable indicating the batch to which round t belongs. We treat this batch number as one of context variables and write $X_t = (\tilde{X}_t, X_t^b)$, where $\tilde{X}_t$ is the vector of context variables other than X_t^b.

Let $p_t = (p_{t0}, ..., p_{tm})' \in \Delta(\mathcal{A})$ denote the potentially unknown probability vector indicating the probability that each action is chosen in round t. Here $\Delta(\mathcal{A}) \equiv \{(p_a) \in$

$\mathbb{R}_+^{m+1} | \sum_a p_a = 1\}$ with p_a being the probability that action a is chosen. A *contextual bandit algorithm* is a sequence $\{F_b\}_{b=1}^B$ of distribution functions of choice probabilities p_t conditional on $\tilde{X}_t$, where $F_b : \tilde{\mathcal{X}} \rightarrow \Delta(\Delta(\mathcal{A}))$ for $b \in \{1, 2, ..., B\}$ and $\tilde{\mathcal{X}}$ is the support of $\tilde{X}$, where $\Delta(\Delta(\mathcal{A}))$ is the set of distributions over $\Delta(\mathcal{A})$. F_b takes context $\tilde{X}_t$ as input and returns a distribution of probability vector p_t in rounds of batch b. F_b can vary across batches but does not change across rounds within batch b. We assume that the log data are generated by a contextual bandit algorithm $\{F_b\}_{b=1}^B$ as follows:

- In each round $t = 1, ..., T$, $(Y_t(\cdot), X_t)$ is i.i.d. drawn from distribution G. Re-order round numbers so that they are monotonically increasing in their batch numbers X_t^b.
- In each round t within batch $b \in \{1, 2, ..., B\}$ and given $\tilde{X}_t$, probability vector $p_t = (p_{t0}, ..., p_{tm})'$ is drawn from $F_b(\cdot|\tilde{X}_t)$. Action is randomly chosen based on probability vector p_t, creating the action choice D_t and the associated reward Y_t.

Here, the contextual bandit algorithm $\{F_b\}_{b=1}^B$ and the realized probability vectors $\{p_t\}_{t=1}^T$ may or may not be known to the analyst. We also allow for the realization of p_t to be degenerate, i.e., a certain action may be chosen with probability 1 at a point in time.

Examples. This DGP allows for many popular bandit algorithms, as the following examples illustrate. In each of the examples below, the contextual bandit algorithm F_b is degenerate and produces a particular probability vector p_t for sure.

Example 1 (Random A/B testing)**.** *We always choose each action uniformly at random:* $p_{ta} = \frac{1}{m+1}$ *always holds for any* $a \in \mathcal{A}$ *and any* $t = 1, ..., T$.

In the remaining examples, at every batch b, the algorithm uses the history of observations from the previous $b - 1$ batches to estimate the mean and the variance of the potential reward under each action conditional on each context: $\mu(a|x) \equiv \mathbb{E}[Y(a)|\tilde{X} = x]$ and $\sigma^2(a|x) \equiv \mathbb{V}[Y(a)|\tilde{X} = x]$. We denote the estimates using the history up to batch $b - 1$ by $\hat{\mu}_{b-1}(a|x)$ and $\hat{\sigma}^2_{b-1}(a|x)$. See Li et al. (2012) and Dimakopoulou, Athey, and Imbens (2017) for possible estimators based on generalized linear models and generalized random forest, respectively. The initial estimates, $\hat{\mu}_0(a|x)$ and $\hat{\sigma}^2_0(a|x)$, are set to any values.

Example 2 (ϵ-Greedy)**.** *In each round within batch b, we choose the best action based on* $\hat{\mu}_{b-1}(a|\tilde{X}_t)$ *with probability* $1-\epsilon_b$ *and choose the other actions uniformly at random with probability* ϵ_b*:*

$$p_{ta} = \begin{cases} 1 - \epsilon_b & \textit{if } a = \operatorname*{argmax}_{a' \in \mathcal{A}} \hat{\mu}_{b-1}(a'|\tilde{X}_t) \\ \dfrac{\epsilon_b}{m} & \textit{otherwise}. \end{cases}$$

Example 3 (Thompson Sampling using Gaussian priors)**.** *In each round within batch b, we sample the potential reward* $y_t(a)$ *from distribution* $\mathcal{N}(\hat{\mu}_{b-1}(a|\tilde{X}_t), \hat{\sigma}^2_{b-1}(a|\tilde{X}_t))$ *for each action, and choose the action with the highest sampled potential reward,* $\operatorname*{argmax}_{a' \in \mathcal{A}} y_t(a')$*. As a result, this algorithm chooses actions with the following probabilities:*

$$p_{ta} = \Pr\{a = \operatorname*{argmax}_{a' \in \mathcal{A}} y_t(a')\},$$

where $(y_t(0), ..., y_t(m))' \sim \mathcal{N}(\hat{\mu}_{b-1}(\tilde{X}_t), \hat{\Sigma}_{b-1}(\tilde{X}_t))$, $\hat{\mu}_{b-1}(x) = (\hat{\mu}_{b-1}(0|x), ..., \hat{\mu}_{b-1}(m|x))'$*, and*

$$\hat{\Sigma}_{b-1}(x) = \begin{pmatrix} \hat{\sigma}^2_{b-1}(0|x) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \hat{\sigma}^2_{b-1}(m|x) \end{pmatrix}.$$

In Examples 2 and 3, p_t depends on the random realization of the estimates $\hat{\mu}_{b-1}(a|x)$ and $\hat{\sigma}^2_{b-1}(a|x)$, and so does the associated F_b. If the data are sufficiently large, the uncertainty in the estimates vanishes: $\hat{\mu}_{b-1}(a|x)$ and $\hat{\sigma}^2_{b-1}(a|x)$ converge to $\mu_{b-1}(a|x) \equiv \mathbb{E}[Y(a)|\tilde{X} = x, X^b \leq b-1]$ and $\sigma^2_{b-1}(a|x) \equiv \mathbb{V}[Y(a)|\tilde{X} = x, X^b \leq b-1]$, respectively. In this case, F_b becomes nonrandom since it depends on the fixed realizations $\mu_{b-1}(a|x)$ and $\sigma^2_{b-1}(a|x)$. In the following analysis, we consider this large-sample scenario and assume that F_b is nonrandom.

To make the notation simpler, we put $\{F_b\}_{b=1}^B$ together into a single distribution $F : \mathcal{X} \rightarrow \Delta(\Delta(\mathcal{A}))$ obtained by $F(\cdot|\tilde{X}, X^b = b) = F_b(\cdot|\tilde{X})$ for each $b \in \{1, 2, ..., B\}$. We use this to rewrite our DGP as follows:

- In each round $t = 1, ..., T$, $(Y_t(\cdot), X_t)$ is i.i.d. drawn from distribution G. Given X_t, probability vector $p_t = (p_{t0}, ..., p_{tm})'$ is drawn from $F(\cdot|X_t)$. Action is randomly chosen based on probability vector p_t, creating the action choice D_t and the associated reward Y_t.

Define

$$p_{0a}(x) \equiv \Pr_{D \sim p,\ p \sim F}(D_a = 1|X = x)$$

for each a, and let $p_0(x) = (p_{00}(x), ..., p_{0m}(x))'$. This is the choice probability vector conditional on each context. We call $p_0(\cdot)$ the *logging policy* or the *propensity score*.

F is common for all rounds regardless of the batch to which they belong. Thus p_t and D_t are i.i.d. across rounds. Because $(Y_t(\cdot), X_t)$ is i.i.d. and $Y_t = \sum_{a=0}^m D_{ta} Y_t(a)$, each observation (Y_t, X_t, D_t) is i.i.d.. Note also that D_t is independent of $Y_t(\cdot)$ conditional on X_t.

2.2 Parameters of Interest

We are interested in using the log data to estimate the expected reward from any given *counterfactual policy* $\pi : \mathcal{X} \rightarrow \bar{\Delta}(\mathcal{A})$, which chooses a distribution of actions given each context:

$$\begin{aligned} V^\pi &\equiv \mathbb{E}_{(Y(\cdot), X) \sim G}[\sum_{a=0}^m Y(a) \pi(a|X)] \\ &= \mathbb{E}_{(Y(\cdot), X) \sim G,\ D \sim p_0(X)}[\sum_{a=0}^m Y(a) D_a \frac{\pi(a|X)}{p_{0a}(X)}], \quad (1) \end{aligned}$$

where the last equality uses the independence of D and $Y(\cdot)$ conditional on X and the definition of $p_0(\cdot)$. Here, $\bar{\Delta}(\mathcal{A}) \equiv \{(p_a) \in \mathbb{R}^{m+1} | \sum_a p_a \leq 1\}$. We allow the counterfactual policy π to be degenerate, i.e., π may choose a particular action with probability 1.

Depending on the choice of π, V^π represents a variety of parameters of interest. When we set $\pi(a|x) = 1$ for a particular action a and $\pi(a'|x) = 0$ for all $a' \in \mathcal{A} \backslash \{a\}$ for all $x \in \mathcal{X}$, V^π equals $\mathbb{E}_{(Y(\cdot),X)\sim G}[Y(a)]$, the expected reward from action a. When we set $\pi(a|x) = 1$, $\pi(0|x) = -1$ and $\pi(a'|x) = 0$ for all $a' \in \mathcal{A} \backslash \{0, a\}$ for all $x \in \mathcal{X}$, V^π equals $\mathbb{E}_{(Y(\cdot),X)\sim G}[Y(a) - Y(0)]$, the average treatment effect of action a over action 0. Such treatment effects are of scientific and policy interests in medical and social sciences. Business and managerial interests also motivate treatment effect estimation. For example, when a company implements a bandit algorithm using a particular reward measure like an immediate purchase, the company is often interested in treatment effects on other outcomes like long-term user retention.

3 Efficient Value Estimation

We consider the efficient estimation of the expected reward from a counterfactual policy, V^π. We consider an estimator consisting of two steps. In the first step, we nonparametrically estimate the propensity score vector $p_0(\cdot)$ by a consistent estimator. Possible estimators include machine learning algorithms such as gradient boosting, as well as nonparametric sieve estimators and kernel regression estimators, as detailed in Section 3.2. In the second step, we plug the estimated propensity $\hat{p}(\cdot)$ into the sample analogue of expression (1) to estimate V^π (in practice, some trimming or thresholding may be desirable for numerical stability):

$$\hat{V}^\pi = \frac{1}{T} \sum_{t=1}^{T} \sum_{a=0}^{m} Y_t D_{ta} \frac{\pi(a|X_t)}{\hat{p}_a(X_t)}.$$

Alternatively, one can use a "self-normalized" estimator inspired by Swaminathan and Joachims (2015b) when $\sum_{a=0}^{m} \pi(a|x) = 1$ for all $x \in \mathcal{X}$:

$$\hat{V}^\pi_{SN} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m} Y_t D_{ta} \frac{\pi(a|X_t)}{\hat{p}_a(X_t)}}{\frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m} D_{ta} \frac{\pi(a|X_t)}{\hat{p}_a(X_t)}}.$$

Swaminathan and Joachims (2015b) suggest that $\hat{V}^\pi_{SN}$ tends to be less biased than $\hat{V}^\pi$ in small sample. Unlike Swaminathan and Joachims (2015b), however, we use the estimated propensity score rather than the true one.

The above estimators estimate a scalar parameter V^π defined as a function of the distribution of $(Y(\cdot), X)$, on which we impose no parametric assumption. Our estimators therefore attempt to solve a semiparametric estimation problem, i.e., a partly-parametric and partly-nonparametric estimation problem. For this semiparametric estimation problem, we first derive the *semiparametric efficiency bound* on how efficient and precise the estimation of the parameter can be, which is a semiparametric analog of the Cramer-Rao bound (Bickel et al. 1993). The asymptotic variance of any asymptotically normal estimator is no smaller than the semiparametric efficiency bound. Following the standard statistics terminology, we say that estimator $\hat{\theta}$ for parameter θ is *asymptotically normal* if $\sqrt{T}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, \Sigma)$ as $T \to \infty$, where $\rightsquigarrow$ denotes convergence in distribution, and $\mathcal{N}(0, \Sigma)$ denotes a normally distributed random variable with mean 0 and variance Σ. We call Σ the *asymptotic variance* of $\hat{\theta}$. The semiparametric efficiency bound for θ is a lower bound on the asymptotic variance of asymptotically normal estimators; the Supplementary Material provides a formal definition of the semiparametric efficiency bound.

We show the above estimators achieve the semiparametric efficiency bound, i.e., they minimize the asymptotic variance among all asymptotically normal estimators. Our analysis uses a couple of regularity conditions. We first assume that the logging policy $p_0(\cdot)$ ex ante chooses every action with a positive probability for every context.

Assumption 1. *There exists some* $\underline{p}$ *such that* $0 < \underline{p} \leq \Pr_{D\sim p,\, p\sim F}(D_a = 1|X = x) \equiv p_{0a}(x)$ *for any* $x \in \mathcal{X}$ *and for* $a = 0, ..., m$.

Note that Assumption 1 is consistent with the possibility that the realization of p_{ta} takes on value 0 or 1 (as long as it takes on positive values with a positive probability).

We also assume the existence of finite second moments of potential rewards.

Assumption 2. $\mathbb{E}[Y(a)^2] < \infty$ *for* $a = 0, ..., m$.

The following proposition provides the semiparametric efficiency bound for V^π. All the proofs are in the Supplementary Material.

Lemma 1 (Semiparametric Efficiency Bound)**.** *Under Assumptions 1 and 2, the semiparametric efficiency bound for* V^π*, the expected reward from counterfactual policy* π*, is*

$$\mathbb{E}[\sum_{a=0}^{m} \mathbb{V}[Y(a)|X] \frac{\pi(a|X)^2}{p_{0a}(X)} + (\theta(X) - V^\pi)^2],$$

where $\theta(X) = \sum_{a=0}^{m} \mathbb{E}[Y(a)|X]\pi(a|X)$ *is the expected reward from policy* π *conditional on* X.

Lemma 1 implies the semiparametric efficiency bounds for the expected reward from each action and for the average treatment effect, since they are special cases of V^π.

Corollary 1. *Suppose that Assumptions 1 and 2 hold. Then, the semiparametric efficiency bound for the expected reward from each action,* $\mathbb{E}[Y(a)]$*, is*

$$\mathbb{E}\big[\frac{\mathbb{V}[Y(a)|X]}{p_{0a}(X)} + (\mathbb{E}[Y(a)|X] - \mathbb{E}[Y(a)])^2\big].$$

The semiparametric efficiency bound for the average treatment effect, $\mathbb{E}[Y(a) - Y(0)]$*, is*

$$\mathbb{E}\big[\frac{\mathbb{V}[Y(0)|X]}{p_{00}(X)} + \frac{\mathbb{V}[Y(a)|X]}{p_{0a}(X)} + (\mathbb{E}[Y(a) - Y(0)|X] - \mathbb{E}[Y(a) - Y(0)])^2\big].$$

Our proposed estimators are two-step generalized-method-of-moment estimators and are asymptotically normal under some regularity conditions, one of which requires that the convergence rate of $\hat{p}(\cdot)$ be faster than $n^{1/4}$ (Newey 1994; Chen 2007). Given the asympotic normality of the estimators, we find that they achieve the semiparametric efficiency bound, building upon Ackerberg et al. (2014) among others.

Theorem 1 (Efficient Estimators)**.** *Suppose that Assumptions 1 and 2 hold and that $\hat{p}(\cdot)$ is a consistent estimator for $p_0(\cdot)$. Then, the variance of $\hat{V}^\pi$ and $\hat{V}^\pi_{SN}$ achieves the semiparametric efficiency bound for V^π (provided in Lemma 1).*

3.1 Inefficient Value Estimation

In some environments, we know the true $p_0(\cdot)$ or observe the realization of the probability vectors $\{p_t\}_{t=1}^T$. In this case, an alternative way to estimate V^π is to use the sample analogue of the expression (1) without estimating the propensity score. If we know $p_0(\cdot)$, a possible estimator is

$$\tilde{V}^\pi = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m} Y_t D_{ta}\frac{\pi(a|X_t)}{p_{0a}(X_t)}.$$

If we observe the realization of $\{p_t\}_{t=1}^T$, we may use

$$\ddot{V}^\pi = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m} Y_t D_{ta}\frac{\pi(a|X_t)}{p_{ta}}.$$

When $\sum_{a=0}^m \pi(a|x) = 1$ for all $x \in \mathcal{X}$, it is again possible to use their self-normalized versions:

$$\tilde{V}^\pi_{SN} = \frac{\frac{1}{T}\sum_{t=1}^T\sum_{a=0}^m Y_t D_{ta}\frac{\pi(a|X_t)}{p_{0a}(X_t)}}{\frac{1}{T}\sum_{t=1}^T\sum_{a=0}^m D_{ta}\frac{\pi(a|X_t)}{p_{0a}(X_t)}}.$$

$$\ddot{V}^\pi_{SN} = \frac{\frac{1}{T}\sum_{t=1}^T\sum_{a=0}^m Y_t D_{ta}\frac{\pi(a|X_t)}{p_{ta}}}{\frac{1}{T}\sum_{t=1}^T\sum_{a=0}^m D_{ta}\frac{\pi(a|X_t)}{p_{ta}}}.$$

These intuitive estimators turn out to be less efficient than the estimators with the estimated propensity score, as the following result shows.

Theorem 2 (Inefficient Estimators)**.** *Suppose that the propensity score $p_0(\cdot)$ is known and we observe the realization of $\{p_t\}_{t=1}^T$. Suppose also that Assumptions 1 and 2 hold and that $\hat{p}(\cdot)$ is a consistent estimator for $p_0(\cdot)$. Then, the asymptotic variances of $\tilde{V}^\pi, \ddot{V}^\pi, \tilde{V}^\pi_{SN}$, and $\ddot{V}^\pi_{SN}$ are no smaller than that of $\hat{V}^\pi$ and $\hat{V}^\pi_{SN}$. Generically, $\tilde{V}^\pi, \ddot{V}^\pi, \tilde{V}^\pi_{SN}$, and $\ddot{V}^\pi_{SN}$ are strictly less efficient than $\hat{V}^\pi$ and $\hat{V}^\pi_{SN}$ in the following sense.*

1. *If $\Pr(\mathbb{E}[Y(a)|X]\frac{\pi(a|X)}{p_{0a}(X)} \neq \theta(X)$ for some $a) > 0$, then the asymptotic variances of $\tilde{V}^\pi$, $\ddot{V}^\pi$, $\tilde{V}^\pi_{SN}$ and $\ddot{V}^\pi_{SN}$ are strictly larger than that of $\hat{V}^\pi$ and $\hat{V}^\pi_{SN}$.*
2. *If $\Pr(\mathbb{E}[Y(a)^2|X]\pi(a|X)^2(\mathbb{E}[\frac{1}{p_a}|X] - \frac{1}{p_{0a}(X)}) \neq 0$ for some $a) > 0$, then the asymptotic variance of $\ddot{V}^\pi$ and $\ddot{V}^\pi_{SN}$ is strictly larger than that of $\hat{V}^\pi$ and $\hat{V}^\pi_{SN}$.*

The condition in Part 1 of Theorem 2 is about the dominating term in the difference between $\hat{V}^\pi$ and $\tilde{V}^\pi$. The proofs of Theorems 1 and 2 show that the asymptotic variance of $\hat{V}^\pi$ is the asymptotic variance of $\tilde{V}^\pi - \frac{1}{T}\sum_{t=1}^T\left[\sum_{a=0}^m \mathbb{E}[Y(a)|X_t]\frac{\pi(a|X_t)}{p_{0a}(X_t)}D_{ta} - \theta(X_t)\right]$. Part 1 of Theorem 2 requires that the second term be not always zero so that the asymptotic variance of $\hat{V}^\pi$ is different from that of $\tilde{V}^\pi$. As long as the two variances are not the same, $\hat{V}^\pi$ achieves variance reduction.

Part 2 of Theorem 2 requires that $\mathbb{E}[\frac{1}{p_{ta}}|X_t] - \frac{1}{\mathbb{E}[p_{ta}|X_t]}(= \mathbb{E}[\frac{1}{p_{ta}}|X_t] - \frac{1}{p_{0a}(X_t)}) \neq 0$ with a positive probability. This means that p_t is not always the same as the true propensity score $p_0(X_t)$, i.e., $F(\cdot|X_t)$ is not degenerate (recall that p_t is drawn from $F(\cdot|X_t)$ whose expected value is $p_0(X_t)$). Under this condition, $\ddot{V}^\pi$ has a strictly larger asymptotic variance than $\tilde{V}^\pi$ and $\hat{V}^\pi$.

Theorems 1 and 2 suggest that we should use an estimated score regardless of whether the propensity score is known. To develop some intuition for this result, consider a simple situation where the context X_t always takes some constant value x. Suppose that we are interested in estimation of the expected reward from action a, $\mathbb{E}[Y(a)]$. Since X is constant across rounds, a natural nonparametric estimator for $p_{0a}(x)$ is the proportion of rounds in which action a was chosen: $\hat{p}_a(x) = \frac{\sum_{t=1}^T D_{ta}}{T}$. The estimator using the estimated propensity score is

$$\hat{V}^\pi = \frac{1}{T}\sum_{t=1}^T Y_t\frac{D_{ta}}{\hat{p}_a(x)} = \frac{1}{\sum_{t=1}^T D_{ta}}\sum_{t=1}^T Y_t D_{ta}.$$

The estimator using the true propensity score is

$$\tilde{V}^\pi = \frac{1}{T}\sum_{t=1}^T Y_t\frac{D_{ta}}{p_{0a}(x)} = \frac{1}{Tp_{0a}(x)}\sum_{t=1}^T Y_t D_{ta}.$$

When action a happens to be chosen frequently in a sample so that $\sum_{t=1}^T D_{ta}$ is larger, the absolute value of $\sum_{t=1}^T Y_t D_{ta}$ tends to be larger in the sample. Because of this positive correlation between $\sum_{t=1}^T D_{ta}$ and the absolute value of $\sum_{t=1}^T Y_t D_{ta}$, $\hat{V}^\pi$ has a smaller variance than $\tilde{V}^\pi$, which produces no correlation between the numerator and the denominator. Similar intuition applies to the comparison between $\ddot{V}^\pi$ and $\hat{V}^\pi$.

3.2 How to Estimate Propensity Scores?

There are several options for the first step estimation of the propensity score.

1. A sieve Least Squares (LS) estimator:

$$\hat{p}_a(\cdot) = \underset{p_a(\cdot)\in\mathcal{H}_{aT}}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^T (D_{ta} - p_a(X_t))^2,$$

where $\mathcal{H}_{aT} = \{p_a(x) = \sum_{j=1}^{k_{aT}} q_{aj}(x)\lambda_{aj} = q^{k_{aT}}(x)'\lambda_a\}$ and $k_{aT} \to \infty$ as $T \to \infty$. Here $\{q_{aj}\}_{j=1}^\infty$ is some known basis functions defined on $\mathcal{X}$ and $q^{k_{aT}}(\cdot) = (q_{a1}(\cdot), ..., q_{ak_{aT}}(\cdot))'$.

2. A sieve Logit Maximum Likelihood estimator:

$$\hat{p}(\cdot) = \underset{p(\cdot)\in\mathcal{H}_T}{\operatorname{argmax}} \frac{1}{T}\sum_{t=1}^{T}\sum_{a=0}^{m} D_{at}\log p_a(X_t),$$

   where $\mathcal{H}_T = \{p : \mathcal{X} \to (0,1)^{m+1} : p_a(x) = \frac{\exp(R^{k_T}(x)'\lambda_a)}{1+\sum_{a=1}^{m}\exp(R^{k_T}(x)'\lambda_a)}$ for $a = 1, ..., m, p_0(x) = 1 - \sum_{a=1}^{m} p_a(x)\}$. Here $R^{k_T}(\cdot) = (R_1(\cdot), ..., R_{k_T}(\cdot))'$ and $\{R_j\}_{j=1}^{\infty}$ is the set of some basis functions.

3. Prediction of D_{ta} by X_t using a modern machine learning algorithm like random forest, ridge logistic regression, and gradient boosting.

The above estimators are known to satisfy consistency with a convergence rate faster than $n^{1/4}$ under regularity conditions (Newey 1997; Cattaneo 2010; Knight and Fu 2000; Blanchard, Lugosi, and Vayatis 2003; Bühlmann and Van De Geer 2011; Wager and Athey 2018).

How should one choose a propensity score estimator? We prefer an estimated score to the true one because it corrects the discrepancy between the realized action assignment in the data and the assignment predicted by the true score. To achieve this goal, a good propensity score estimator should fit the data better than the true one, which means that the estimator should overfit to some extent. As a concrete example, in our empirical analysis, random forest produces a larger (worse) variance than gradient boosting and ridge logistic regression (see Figure 1 and Table 1). This is because random forest fits the data worse, which is due to its bagging aspect preventing random forest from overfitting. In general, however, we do not know which propensity score estimator achieves the best degree of overfitting. We would therefore suggest that the analyst try different estimators to determine which one is most efficient.

4 Estimating Asymptotic Variance

We often need to estimate the asymptotic variance of the above estimators. For example, variance estimation is crucial for determining whether a counterfactual policy is statistically significantly better than the logging policy. We propose an estimator that uses the sample analogue of an expression of the asymptotic variance. As shown in the proof of Theorem 1, the asymptotic variance of $\hat{V}^{\pi}$ and $\hat{V}^{\pi}_{SN}$ is $\mathbb{E}[(g(Y, X, D, V^{\pi}, p_0) + \alpha(X, D, p_0, \mu_0))^2]$, where $\mu_0 : \mathcal{X} \to \mathbb{R}^{m+1}$ such that $\mu_0(a|x) = \mathbb{E}[Y(a)|X = x]$ for each a and x,

$$g(Y, X, D, \theta, p) = \sum_{a=0}^{m} Y D_a \frac{\pi(a|X)}{p_a(X)} - \theta,$$

and

$$\alpha(X, D, p, \mu) = -\sum_{a=0}^{m} \mu(a|X)\frac{\pi(a|X)}{p_a(X)}(D_a - p_a(X)).$$

We estimate this asymptotic variance in two steps. In the first step, we obtain estimates of V^{π} and p_0 using the method in Section 3. In addition, we estimate $\mu_0(a|x)$ by nonparametric regression of Y_t on X_t using the subsample with $D_{ta} = 1$ for each a. Denote the estimate by $\hat{\mu}$. For this regression, one may use a sieve Least Squares estimator and machine learning algorithms. In the empirical application below, we use ridge logistic regression.

In the second step, we plug the estimates of V^{π}, p_0 and μ_0 into the sample analogue of $\mathbb{E}[(g(Y, X, D, V^{\pi}, p_0) + \alpha(X, D, p_0, \mu_0))^2]$ to estimate the asymptotic variance: when we use $\hat{V}^{\pi}$:

$$\begin{aligned}&\widehat{AVar}(\hat{V}^{\pi})\\ =&\frac{1}{T}\sum_{t=1}^{T}(g(Y_t, X_t, D_t, \hat{V}^{\pi}, \hat{p}) + \alpha(X_t, D_t, \hat{p}, \hat{\mu}))^2.\end{aligned}$$

When we use $\hat{V}^{\pi}_{SN}$, its asymptotic variance estimator is obtained by replacing $\hat{V}^{\pi}$ with $\hat{V}^{\pi}_{SN}$ in the above expression.

This asymptotic variance estimator is a two-step generalized-method-of-moment estimator, and is shown to be a consistent estimator under the condition that the first step estimator of (V^{π}, p_0, μ_0) is consistent and some regularity conditions (Newey 1994).

It is easier to estimate the asymptotic variance of $\tilde{V}^{\pi}$ and $\tilde{V}^{\pi}_{SN}$ with the true propensity score. Their asymptotic variance is $\mathbb{E}[g(Y, X, D, V^{\pi}, p_0)^2]$ by the standard central limit theorem. When we use $\tilde{V}^{\pi}$, we estimate this asymptotic variance by

$$\widehat{AVar}(\tilde{V}^{\pi}) = \frac{1}{T}\sum_{t=1}^{T} g(Y_t, X_t, D_t, \tilde{V}^{\pi}, p_0)^2$$

When we use $\tilde{V}^{\pi}_{SN}$, its asymptotic variance estimator is obtained by replacing $\tilde{V}^{\pi}$ with $\tilde{V}^{\pi}_{SN}$ in the above expression.

5 Real-World Application

We apply our estimators described in Sections 3 and 4 to empirically evaluate and optimize the design of online advertisements. This application uses proprietary data provided by CyberAgent Inc., which we described in the introduction. This company uses a contextual bandit algorithm to determine the visual design of advertisements assigned to user impressions (there are four design choices). This algorithm produces logged bandit data. We use this logged bandit data and our estimators to improve their advertisement design for maximizing the click through rates (CTR). In the notation of our theoretical framework, reward Y is a click, action a is one of the four possible individual advertisement designs, and context X is user and ad characteristics used by the company's logging policy.

The logging policy (the company's existing contextual bandit algorithm) works as follows. For each round, the logging policy first randomly samples each action's predicted reward from a beta distribution. This beta distribution is parametrized by the predicted CTR for each context, where the CTR prediction is based on a Factorization Machine (Rendle 2010). The logging policy then chooses the action (advertisement) with the largest sampled reward prediction. The logging policy and the underlying CTR prediction stay

	Existing Logging Policy		Policy of Choosing Best Action by Context	
Propensity Score Estimator	CI for Expected Reward	Shrinkage in CI	CI	Shrinkage in CI
True Score (Benchmark)	1.036 ± 0.083	0	1.140 ± 0.171	0
Gradient Boosting Machine	1.047 ± 0.064	-22.5%	1.197 ± 0.131	-23.4%
Ridge Logistic Regression	0.987 ± 0.058	-30.2%	1.108 ± 0.113	-34.1%
Random Forest	1.036 ± 0.077	-6.62%	1.194 ± 0.159	-7.41%
Sample Size	$57,619$			

Table 1: Improving Ad Design with Lower Uncertainty

Notes: The first and third columns of this table show 95% confidence intervals of the expected CTRs V^π of the logging policy and a hypothetical policy of choosing the best action (ad) for each context. CTRs are multiplied by a constant for confidentiality reasons. We obtain the CTR estimates by the self-normalized inverse probability weighting estimator $\tilde{V}^\pi_{SN}$ using the true propensity score (Swaminathan and Joachims 2015b) or the estimated propensity score ($\hat{V}^\pi_{SN}$ in Section 3). We estimate standard errors and confidence intervals based on the method described in Section 4. The second and fourth columns show the size of reductions in confidence interval length, i.e., value α such that the length of the confidence interval is equal to $100-\alpha\%$ of the length of the confidence interval using the true propensity score.

the same for all rounds in each day. Each day therefore performs the role of a batch in the model in Section 2. This somewhat nonstandard logging policy and the resulting log data are an example of our DGP in Section 2.

This logging policy may have room for improvement for several reasons. First, the logging policy randomly samples advertisements and does not necessarily choose the advertisement with the best predicted CTR. Also, the logging policy uses a predictive Factorization Machine for its CTR prediction, which may be different from the causal CTR (the causal effect of each advertisement on the probability of a click).

To improve on the logging policy, we first estimate the propensity score by random forest, ridge logistic regression, or gradient boosting (implemented by XGBoost). These estimators are known to satisfy the regularity conditions (e.g. consistency) required for our theoretical results, as explained in Section 3.2.

With the estimated propensity score, we then use our estimator $\hat{V}^\pi_{SN}$ to estimate the expected reward from two possible policies: (1) the logging policy and (2) a counterfactual policy that chooses the best action (advertisement) that is predicted to maximize the CTR conditional on each context. To implement this counterfactual policy, we estimate $\mathbb{E}[Y(a)|X]$ by ridge logistic regression for each action a and context X used by the logging policy (we apply one-hot encoding to categorical variables in X). Given each context X, the counterfactual policy then chooses the action with the highest estimated value of $\mathbb{E}[Y(a)|X]$.

Importantly, we use separate data sets for the two estimation tasks (one for the best actions and the other for the expected reward from the hypothetical policy). Specifically, we use data logged during April 20-26, 2018 for estimating the best actions and data during April 27-29 for estimating the expected reward. This data separation allows us to avoid overfitting and overestimation of the CTR gains from the counterfactual policy.

As a benchmark, we also estimate the same expected rewards based on Swaminathan and Joachims (2015b)'s self-normalized estimator $\tilde{V}^\pi_{SN}$, which uses the true propensity score. The resulting estimates show the following result:

Empirical Result A. Consistent with Theorems 1-2, our estimator $\hat{V}^\pi_{SN}$ with the estimated score is statistically more efficient than the benchmark $\tilde{V}^\pi_{SN}$ with the true score.

This result is reported in Figure 1 and Table 1, where the confidence intervals about the predicted CTR using "True Propensity Score (Benchmark)" are less precise (wider) than those using estimated propensity scores (regardless of which one of the three score estimators to use). The magnitude of this shrinkage in the confidence intervals and standard errors is 6-34%, depending on how to estimate the propensity score.

This variance reduction allows us to conclude that the logging policy is below the lower bound of the confidence interval of the hypothetical policy, giving us confidence in the following implication:

Empirical Result B. Compared to the logging policy, the hypothetical policy (choosing the best advertisement given each context) improves the CTR by 10-15% statistically significantly at the 5% significance level.

6 Conclusion

We have investigated the most statistically efficient use of batch bandit data for estimating the expected reward from a counterfactual policy. Our estimators minimize the asymptotic variance among all asymptotically normal estimators (Theorem 1). By contrast, standard estimators have larger asymptotic variances (Theorem 2).

We have also applied our estimators to improve online advertisement design. Compared to the frontier benchmark $\tilde{V}^\pi_{SN}$, our reward estimator $\hat{V}^\pi_{SN}$ provides the company with more statistical confidence in how to improve on its existing bandit algorithm (Empirical Results A and B). The hypothetical policy of choosing the best advertisement given user characteristics would improve the click through rate by 10-15% at the 5% significance level. These empirical results thus highlight the practical values of Theorems 1-2.

Acknowledgments. We are grateful to seminar participants at ICML/IJCAI/AAMAS Workshop “Machine Learning for Causal Inference, Counterfactual Prediction, and Autonomous Action (CausalML)” and RIKEN Center for Advanced Intelligence Project, especially Junya Honda, Masaaki Imaizumi, Atsushi Iwasaki, Kohei Kawaguchi, and Junpei Komiyama.

References

Ackerberg, D.; Chen, X.; Hahn, J.; and Liao, Z. 2014. Asymptotic Efficiency of Semiparametric Two-step GMM. *Review of Economic Studies* 81(3):919–943.

Agarwal, A.; Basu, S.; Schnabel, T.; and Joachims, T. 2017. Effective Evaluation Using Logged Bandit Feedback from Multiple Loggers. *KDD* 687–696.

Bickel, P. J.; Klaassen, C. A. J.; Ritov, Y.; and Wellner, J. A. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

Blanchard, G.; Lugosi, G.; and Vayatis, N. 2003. On the Rate of Convergence of Regularized Boosting Classifiers. *Journal of Machine Learning Research* 4(Oct):861–894.

Bottou, L.; Peters, J.; Quiñonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14(1):3207–3260.

Bühlmann, P., and Van De Geer, S. 2011. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Cattaneo, M. D. 2010. Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability. *Journal of Econometrics* 155(2):138–154.

Chen, X. 2007. Large Sample Sieve Estimation of Semi-nonparametric Models. *Handbook of Econometrics* 6:5549–5632.

Dimakopoulou, M.; Athey, S.; and Imbens, G. 2017. Estimation Considerations in Contextual Bandits. *ArXiv*.

Dudík, M.; Erhan, D.; Langford, J.; and Li, L. 2014. Doubly Robust Policy Evaluation and Optimization. *Statistical Science* 29:485–511.

Hirano, K.; Imbens, G. W.; and Ridder, G. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71(4):1161–1189.

Horvitz, D. G., and Thompson, D. J. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association* 47(260):663–685.

Knight, K., and Fu, W. 2000. Asymptotics for Lasso-type Estimators. *Annals of Statistics* 1356–1378.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A Contextual-bandit Approach to Personalized News Article Recommendation. *WWW* 661–670.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms. *WSDM* 297–306.

Li, L.; Chu, W.; Langford, J.; Moon, T.; and Wang, X. 2012. An Unbiased Offline Evaluation of Contextual Bandit Algorithms with Generalized Linear Models. *Journal of Machine Learning Research: Workshop and Conference Proceedings* 26:19–36.

Li, L.; Munos, R.; and Szepesvári, C. 2015. Toward Minimax Off-policy Value Estimation. *AISTATS* 608–616.

Li, L. 2015. Offline Evaluation and Optimization for Interactive Systems. *WSDM*.

Munos, R.; Stepleton, T.; Harutyunyan, A.; and Bellemare, M. 2016. Safe and Efficient Off-policy Reinforcement Learning. *NIPS* 1054–1062.

Newey, W. K. 1994. The Asymptotic Variance of Semiparametric Estimators. *Econometrica* 62(6):1349–1382.

Newey, W. K. 1997. Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics* 79(1):147–168.

Rendle, S. 2010. Factorization Machines. *ICDM* 995–1000.

Rosenbaum, P. R., and Rubin, D. B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1):41–55.

Strehl, A.; Langford, J.; Li, L.; and Kakade, S. M. 2010. Learning from Logged Implicit Exploration Data. *NIPS* 2217–2225.

Swaminathan, A., and Joachims, T. 2015a. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research* 16:1731–1755.

Swaminathan, A., and Joachims, T. 2015b. The Self-normalized Estimator for Counterfactual Learning. *NIPS* 3231–3239.

Swaminathan, A.; Krishnamurthy, A.; Agarwal, A.; Dudik, M.; Langford, J.; Jose, D.; and Zitouni, I. 2017. Off-policy Evaluation for Slate Recommendation. *NIPS* 3635–3645.

Thomas, P., and Brunskill, E. 2016. Data-efficient Off-policy Policy Evaluation for Reinforcement Learning. *ICML* 2139–2148.

Thomas, P. S.; Theocharous, G.; and Ghavamzadeh, M. 2015. High-Confidence Off-Policy Evaluation. *AAAI* 3000–3006.

Wager, S., and Athey, S. 2018. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wang, Y.-X.; Agarwal, A.; and Dudik, M. 2017. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. *ICML* 3589–3597.