Optimal Off-Policy Evaluation from Multiple Logging Policies

Nathan Kallus [*1] **Yuta Saito** [*1] **Masatoshi Uehara** [*1]

Abstract

We study off-policy evaluation (OPE) from multiple logging policies, each generating a dataset of fixed size, *i.e.*, stratified sampling. Previous work noted that in this setting the ordering of the variances of different importance sampling estimators is instance-dependent, which brings up a dilemma as to which importance sampling weights to use. In this paper, we resolve this dilemma by finding the OPE estimator for multiple loggers with *minimum* variance for any instance, *i.e.*, the *efficient* one. In particular, we establish the efficiency bound under stratified sampling and propose an estimator achieving this bound when given consistent q-estimates. To guard against misspecification of q-functions, we also provide a way to choose the control variate in a hypothesis class to minimize variance. Extensive experiments demonstrate the benefits of our methods' efficiently leveraging of the stratified sampling of off-policy data from multiple loggers.

1. Introduction

In many applications where personalized and dynamic decision making is of interest, exploration is costly, risky, unethical, or otherwise infeasible ruling out the use of online algorithms for contextual bandits (CB) and reinforcement learning (RL) that need to explore in order to learn. This includes both healthcare, where we fear bad patient outcomes, and e-commerce, where we fear alienating users. This motivates the study of off-policy evaluation (OPE), which is the task of estimating the value of a given policy using only historical data, which is generated by current decision policies. Given how invaluable this is, OPE has been studied extensively both in CB (Kallus, 2018; Wang et al., 2017; Dudík et al., 2014; Swaminathan et al., 2017) and in RL (Farajtabar et al., 2018; Liu et al., 2018; Kallus and Uehara, 2019a; Jiang and Li, 2016; Yin et al., 2020).

[*] Alphabetical Order [1]Cornell University, NY, USA . Correspondence to: Masatoshi Uehara <mu223@cornell.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

In most of the above studies, the observations used to evaluate a new policy are assumed generated by a *single* logging policy. Often, however, we have the opportunity to leverage multiple datasets, each potentially generated by a different logging policy (Agarwal et al., 2017; He et al., 2019; Strehl et al., 2010; Bareinboim and Pearl, 2016). For example, in hospitals, we might have the two-types datasets (clinical trials data) by different design (logging policies) to evaluate the efficacy of some drugs. The goal here is to combine these two datasets efficiently. In these cases, the size of each dataset is generally fixed by design, which distinguishes this setting from a single logging policy given by the mixture of logging policies. Such fixed dataset sizes is an example of *stratified sampling* (Wooldridge, 2001), where the identity of the logging policies constitute the stratum.

The distinction of these two settings is crucial since the same estimator may have varying precision in each setting (a fact well-known in Monte Carlo integration, Geyer, 1994; Kong et al., 2003, noise contrastive estimation, Gutmann and Hyvärinen, 2010; Uehara et al., 2018, and survey sampling (Fuller, 2009)). Thus, many results in the standard *un*stratified OPE setting cannot be directly translated to a multiple logger setting, most crucially the efficiency lower bound on mean-squared error (MSE) and estimators that achieve this lower bound (Kallus and Uehara, 2020a; Narita et al., 2019). In the multiple logger setting, we may additionally consider a much greater variety of estimators that can utilize the logger identity as data. In this paper, we study a wide range of such estimators, establish the efficiency lower bound, and propose estimators that achieve it.

Previous work on OPE with multiple loggers proposed various importance sampling (IS) estimators that use the logger identity (Agarwal et al., 2017). However, they arrived at a *dilemma*: there is no strict ordering between the IS estimate with marginalized logging probabilities and a precision-weighted combination of the IS estimates in each dataset. That is, which estimate has lower MSE depends on the problem instance and is not known a priori, and therefore it is not clear which should be preferred. Our analysis resolves this dilemma by developing an *efficient* estimator, which has MSE better (or not worse) than both of the above.

Our contributions are as follows.

- (I) When the logging policies are known, we study the

variances of a new class of unbiased estimators that includes and is much bigger than the class considered in Agarwal et al. (2017). This new class incorporates both control variates and flexible weights that may depend on logger identity. In this way the class is special to the multiple-logger setting. We show that a single estimator has minimum (non-asymptotic) MSE in this class (Sections 3.1 and 3.2).

- (II) Considering the case where the logging policies are possibly unknown, we derive the lower bound (i.e. efficiency bound) of the asymptotic MSEs among the class of *all* regular estimators, which is a larger class than the above new class (Section 3.3). This derivation is fundamentally different than the single-logger setting because the data in the multiple-logger setting is not independent and identically distributed (iid). We show how to construct an efficient estimator. We also extend this result to the RL case (Section 7).

- (III) We investigate the differences between in the stratified and unstratified cases by showing that the variances of the estimators are generally different under the two settings (Section 5). We use this insight to choose optimal control variates to directly minimize variance, extending the More Robust Doubly Robust (MRDR) estimator of Rubin and der Laan (2008); Farajtabar et al. (2018) to the stratified setting (Section 6).

2. Background

We start by setting up the problem and summarizing the relevant literature.

2.1. Problem Setup

We focus on the CB setting as was the topic of previous work (Agarwal et al., 2017).

We are concerned with the average reward of taking an action $a \in \mathcal{A}$ in context (state) $s \in \mathcal{S}$ when following the policy $\pi^e(a \mid s)$, known as the evaluation policy. Both $\mathcal{A}$ and $\mathcal{S}$ may be discrete or continuous. Rewards $r \in [0, R_{\max}]$ are described by the (unknown) reward emission probability density $p_{R|S,A}(r \mid s, a)$, and contexts are drawn from the (unknown) density $p_S(s)$. Thus, the average reward under π^e, which is our target estimand is

$$J := \mathbb{E}_{\pi_e}[r],$$

where the subscript π_e refers to the joint density $p_S(s)\pi^e(a \mid s)p_{R|S,A}(r \mid s, a)$ over (s, a, r).

To help estimate J, we consider observing K datasets, $\mathcal{D} = \{\mathcal{D}_1, \cdots, \mathcal{D}_K\}$, each of (fixed) size n_k and associated with the logging policy $\pi_k(a \mid s)$, for $k \in [K] = \{1, \ldots, K\}$. (We consider both the cases where π_k are known and unknown.) Each dataset consists of observations of state-action-reward triplets, $\mathcal{D}_k = \{(S_{kj}, A_{kj}, R_{kj})\}_{j=1}^{n_k}$, drawn independently according to the product density

$$(S_{kj}, A_{kj}, R_{kj}) \sim p_S(s)\pi_k(a \mid s)p_{R|S,A}(r \mid s, a).$$

Notice that the distribution above differs from the distribution in the definition of J in the policy used to generate actions. We let $n = n_1 + \cdots + n_K$ be the total dataset size. We often reindex the whole data as $\mathcal{D} = \bigcup_{k=1}^K \{(k, s, a, r) : (s, a, r) \in \mathcal{D}_k\} = \{(k_i, S_i, A_i, R_i) : i = 1, \ldots, n\}$, treating the logger identity k_i as an additional component of an observation in one big pooled dataset. For a function $f(s, a, r)$ we let $\mathbb{E}_{n_k}[f] = \frac{1}{n_k}\sum_{(s,a,r)\in\mathcal{D}_k} f(s, a, r)$ and for a function $f(k, s, a, r)$ we let $\mathbb{E}_n[f] = \frac{1}{n}\sum_{(k,s,a,r)\in\mathcal{D}} f(k, s, a, r)$. As mentioned above, we let $\mathbb{E}_\pi$ refer to expectations with respect to the distribution on (s, a, r) induced by playing π (similarly, var_π). Unsubscripted expectations and variances are with respect to the data generation (such as the variance of an estimator).

We let $\rho_k = n_k/n$ be the dataset proportions and $\pi_*(a \mid s) = \sum_{k=1}^K \rho_k \pi_k(a \mid s)$ be the marginal logging policy (as a policy, it corresponds to randomizing the choice of logger with weights ρ_k and then playing the chosen logger, but note this is *not* how the data is generated, as n_k are fixed). For any function $f(s, a)$, let $f(s, \pi) = \mathbb{E}_\pi[f(s, a) \mid s] = \int f(s, a)\pi(a \mid s)\mathrm{d}(a)$. We let $q(s, a) = \mathbb{E}_{p_{R|S,A}}[r \mid s, a]$, $v(s) = q(s, \pi^{\mathrm{e}})$, $\sigma_r^2(s, a) = \mathrm{var}_{p_{R|S,A}}[r \mid s, a]$. We define the L_2 norm by $\|f\|_2 = \{\mathbb{E}_{\pi_*}[f^2(s, a, r)]\}^{1/2}$. We denote the normal distribution with mean μ and variance σ^2 by $\mathcal{N}(\mu, \sigma^2)$.

We always let $n, n_1, \ldots, n_K$ be fixed and finite. When we discuss asymptotic behavior we consider sample sizes $n' = mn, n'_k = mn_k$ and $m \to \infty$ such that sample proportions $\rho_k = n_k/n = n'_k/n'$ remain fixed. This setting is generally called a stratified sampling (Wooldridge, 2001; Imbens and Lancaster, 1996). When combining datasets, it is most natural to assume n_k is fixed rather than random just as when dealing with one dataset we treat n as fixed rather than random. The same perspective is generally taken in causal inference (Yang and Ding, 2020).

2.2. Previous Work and the Multiple Logger Dilemma

In the *un*stratified setting, wherein the logging policy first chooses k at random from $[K]$ with weights ρ_k and then plays the logging policy π_k, the standard IS estimator is

$$\hat{J}_{\mathrm{IS}} := \mathbb{E}_n\left[\frac{\pi^{\mathrm{e}}(a|s)r}{\pi_*(a|s)}\right].$$

This estimator can still be applied in the stratified setting in the sense that is unbiased under weak overlap.

Assumption 1 (Weak Overlap)**.** For almost all $s \in \mathcal{S}$, $\{a : \pi^{\mathrm{e}}(a \mid s) > 0\} \subset \bigcup_{k\in[K]}\{a : \pi_k(a \mid s) > 0\}$.

Agarwal et al. (2017) study the multiple logger setting and propose estimators that combine the IS estimators in each of the K datasets: given simplex weights $\lambda \in \Delta^K = \{\lambda \in \mathbb{R}^K : \lambda_k \geq 0, \sum_{k=1}^K \lambda_k = 1\}$, they let

$$\Upsilon(\mathcal{D};\lambda) = \sum_{k=1}^K \lambda_k \mathbb{E}_{n_k}\left[\frac{\pi^{\mathrm{e}}(a|s)r}{\pi_k(a|s)}\right]. \tag{1}$$

For any $\lambda \in \Delta^K$, $\Upsilon(\mathcal{D};\lambda)$ is unbiased under whole weak overlap.

Assumption 2 (Whole Weak Overlap)**.** For almost all $s \in \mathcal{S}$, $\{a : \pi^{\mathrm{e}}(a \mid s) > 0\} \subset \bigcap_{k\in[K]}\{a : \pi_k(a \mid s) > 0\}$.

Clearly Assumption 2 implies Assumption 1.

Then, they consider two important special cases: the naïve average of the K IS estimates,

$$\hat{J}_{\text{IS-Avg}} := \Upsilon(\mathcal{D}; (n_1/n, \dots, n_K/n)),$$

and a precision-weighted average,

$$\hat{J}_{\text{IS-PW}} := \Upsilon(\mathcal{D};\lambda^*),\ \lambda^*_k = \frac{n_k/\mathrm{var}_{\pi_k}[\{\pi^{\mathrm{e}}(a|s)r\}/\{\pi_k(a|s)\}]}{\sum_{k'} n_{k'}/\mathrm{var}_{\pi_{k'}}[\{\pi^{\mathrm{e}}(a|s)r\}/\{\pi_{k'}(a|s)\}]}.$$

Notice that $\lambda^* = \arg\min_{\lambda\in\Delta^K} \mathrm{var}[\Upsilon(\mathcal{D};\lambda)]$. Unlike $\hat{J}_{\text{IS}}$ and $\hat{J}_{\text{IS-Avg}}$, the estimator $\hat{J}_{\text{IS-PW}}$ is not feasible in practice since λ^* needs to be estimated from data first (we discuss this in more detail in Section 3.2 and show that asymptotically there is no inflation in variance).

Agarwal et al. (2017) established two relationships about the above:

$$\mathrm{var}[\hat{J}_{\text{IS-Avg}}] \geq \mathrm{var}[\hat{J}_{\text{IS}}], \quad \mathrm{var}[\hat{J}_{\text{IS-Avg}}] \geq \mathrm{var}[\hat{J}_{\text{IS-PW}}].$$

However, they noted that they cannot find a theoretical relationship between $\mathrm{var}[\hat{J}_{\text{IS}}]$ and $\mathrm{var}[\hat{J}_{\text{IS-PW}}]$. In fact, unlike the above two relationships, which of these two estimators has smaller variance *depends* on the problem instance. This brings up an apparent *dilemma*: which one should we use? We resolve this dilemma by showing another estimator dominates both. In fact, it dominates a much bigger class of estimators, that includes $\hat{J}_{\text{IS}}, \Upsilon(\mathcal{D};\lambda), \hat{J}_{\text{IS-Avg}}, \hat{J}_{\text{IS-PW}}$.

3. Optimality

We next tackle the question of what would be the *optimal* estimator. We tackle this from three perspectives. First, we study a class of estimators like $\Upsilon(\mathcal{D};\lambda)$ but larger, allowing for control variates, and determine the single estimator with minimal (non-asymptotic) MSE among these. Second, since not all estimators (including this optimum) are feasible in practice as they may involve unknown nuisances (just like $\hat{J}_{\text{IS-PW}}$ depends on the unknown λ^*), we then consider a class of feasible estimators given by plugging in these nuisances and we show that asymptotically the minimum MSE is the same and achievable. Third, we show that this minimum is in fact the efficiency lower bound, that is, the minimum asymptotic MSE among all regular estimators. Fig. 1 illustrates the relationship between these different classes of estimators.

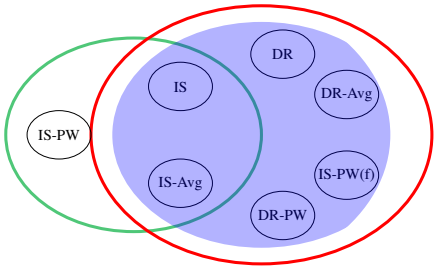

Figure 1. Relationship between the classes of estimators considered in Section 3. The green circle represents the class $\{\Gamma(\mathcal{D};h,g)\}$. The red circle is regular estimators. The blue shaded region is the estimators $\hat{J}_{\text{BI}}(\hat{h},\hat{g})$ with feasible and consistent estimators $\hat{h},\hat{g}$ (see Theorem 2). The minimal asymptotic MSE in *any one* of these sets is the *same* and achievable by a feasible estimator.

3.1. A Class of (Possibly Infeasible) Unbiased Estimators

Consider the class of estimators given by

$$\Gamma(\mathcal{D};h,g) = \mathbb{E}_n[h(k,s,a)\pi^{\mathrm{e}}(a \mid s)(r - g(s,a)) + g(s,\pi^{\mathrm{e}})],$$

for any choice of functions $h(k,s,a)$, $g(s,a)$, where we restrict to functions h that satisfy

$$\sum_{k=1}^K n_k\pi_k(a \mid s)h(k,s,a) = n \quad \forall s,a : \pi^{\mathrm{e}}(a \mid s) > 0. \tag{2}$$

This is designed to satisfy the following property:

Lemma 1. $\mathbb{E}[\Gamma(\mathcal{D};h,g)] = J$.

This is a fairly large class in the sense that it allows both for flexible weights that depend on logger identity and for control variates. In fact, it includes the class $\Upsilon(\mathcal{D};\lambda)$ as a subclass (including $\hat{J}_{\text{IS-Avg}}, \hat{J}_{\text{IS-PW}}$) by letting $h(k,s,a) = 1/\pi_k$ or $h(k,s,a) = n\lambda^*_k/(n_k\pi_k(a \mid s))$, and $g = 0$. It also includes $\hat{J}_{\text{IS}}$ by letting $h_k(k,s,a) = 1/\pi_*(a \mid s)$ and $g = 0$. This class of estimators is unbiased, *i.e.*, $\mathbb{E}[\Gamma(\mathcal{D};h,g)] = J$. But notice that the restriction on h (Eq. (2)) implicitly requires a form of h-specific overlap. *E.g.*, for $h(k,s,a) = 1/\pi_*(a \mid s)$, it corresponds to Assumption 1, and for $h(k,s,a) = n\lambda^*_k/(n_k\pi_k(a \mid s))$, it is implied by Assumption 2.

Here h, g may depend on unknown aspects of the data generating distribution (*e.g.*, $g = q$). Thus, certain choices may

be infeasible in practice. Feasible analogs may be derived by estimating h, g and plugging the estimates in as we will do in the next section. We refer to the class of estimator as we range over h, g satisfying Eq. (2) as $\{\Gamma(\mathcal{D}; h, g)\}$, and we refer to h as "weights" and g as "control variates."

We have the following optimality result.

Theorem 1. *Suppose Assumption 1 holds. The minimum of the variances (MSEs) among estimators in the class $\{\Gamma(\mathcal{D}; h, g)\}$ is V^*/n where*

$$V^* := \mathbb{E}_{\pi_*(a|s)p_S(s)}\left[\left\{\frac{\pi^{\mathrm{e}}(a|s)}{\pi_*(a|s)}\right\}^2 \sigma_r^2(s,a)\right] + \mathrm{var}_{p_S}[v(s)].$$

This minimum is achieved by $\Gamma(\mathcal{D}; 1/\pi_*, q)$.

The result is remarkable in two ways. First, it gives an answer to the dilemma outlined in Section 2. In the end, none of the three estimators $\hat{J}_{\text{IS-PW}}, \hat{J}_{\text{IS}}, \hat{J}_{\text{IS-Avg}}$ studied by (Agarwal et al., 2017) are optimal. Second, it states the surprising fact that logger identity information does *not* contribute to the lower bound. In other words, whether we allow different weights in different strata (allow h to depend on k), the minimum variance is unchanged since it is achieved by a stratum-*independent* weight function.

3.2. A Class of Feasible Unbiased Estimators

When h, g depend on unknowns, such as $g = q$ as in the optimal estimator in Theorem 1, the estimator $\Gamma(\mathcal{D}; h, g)$ is actually infeasible in practice. We therefore next study what happens when we estimate g, h and plug them in. Generally, when we plug nuisance estimates in, the variance may inflate due to the additional uncertainty associated with these estimates, both in finite samples *and* asymptotically: for example, when we consider a direct method estimator $\mathbb{E}_n[\hat{q}(S, \pi^{\mathrm{e}})]$, the asymptotic variance is much larger than $\mathbb{E}_n[q(S, \pi^{\mathrm{e}})]$. Interestingly, for the current case, this inflation does not occur asymptotically.

Specifically, we propose the feasible estimators $\hat{J}_{\text{BI}}(\hat{h}, \hat{g})$ given by the meta-algorithm in Algorithm 1, which uses a cross-fitting technique (Zheng and van Der Laan, 2011; Chernozhukov et al., 2018). The idea is to split the sample into a part where we estimate g, h and a part where we plug them in and then averaging over different roles of the splits. If each $\hat{h}^{(z)}$ satisfies Eq. (2), then this *feasible* estimator is still unbiased since

$$\mathbb{E}[\Gamma(\mathcal{L}_z; \hat{h}^{(z)}, \hat{g}^{(z)})] = \mathbb{E}[\mathbb{E}[\Gamma(\mathcal{L}_z; \hat{h}^{(z)}, \hat{g}^{(z)}) \mid \mathcal{U}_z]] = J.$$

If we do not use cross-fitting, this unbiasedness cannot be ensured.

In addition, in the asymptotic regime (recall that in the asymptotic regime we consider $n' = mn$, $n'_k = mn_k$ observations and $m \to \infty$) we can show that whenever $\hat{h}, \hat{g}$ are consistent, the feasible estimator $\hat{J}_{\text{BI}}(\hat{h}, \hat{g})$ is also asymptotically normal with the *same* variance as the possibly infeasible $\Gamma(\mathcal{D}; h, g)$.

Algorithm 1 Feasible Cross-Fold Version of $\Gamma(\mathcal{D}; h, g)$

Input: Estimators $\hat{h}(k, s, a), \hat{g}(s, a)$
Fix a positive integer Z. For each $k \in [K]$, take a Z-fold random even partition $(I_{kz})_{z=1}^Z$ of the observation indices $\{1, \dots, n_k\}$ such that the size of each fold, $|I_{kz}|$, is within 1 of n_k/Z
Let $\mathcal{L}_z = \{(S_{ki}, A_{ki}, R_{ki}) : k = 1, \dots, K, i \in I_{kz}\}$, $\mathcal{U}_z = \{(S_{ki}, A_{ki}, R_{ki}) : k = 1, \dots, K, i \notin I_{kz}\}$
for $z = 1, \cdots, Z$ **do**
    Construct estimators $\hat{h}^{(z)} = \hat{h}(k, s, a; \mathcal{U}_z)$, $\hat{g}^{(z)} = \hat{g}(s, a; \mathcal{U}_z)$ of h, g using only $\mathcal{U}_z$ as data
    Set $\hat{J}_z = \Gamma(\mathcal{L}_z; \hat{h}^{(z)}, \hat{g}^{(z)})$
end for
Return: $\hat{J}_{\text{BI}}(\hat{h}, \hat{g}) = \frac{1}{n}\sum_{z=1}^Z |\mathcal{L}_z|\, \hat{J}_z$.

Theorem 2. *Suppose $\|\pi^{\mathrm{e}}/\hat{h}^{(z)} - \pi^{\mathrm{e}}/h\|_2 = \mathrm{o}_p(1)$, $\|\hat{g}^{(z)} - g\|_2 = \mathrm{o}_p(1)$, $\pi^{\mathrm{e}}/\hat{h}^{(z)}, \hat{g}^{(z)}, \pi^{\mathrm{e}}/h, g$ are uniformly bounded by some constants, and $h, \hat{h}^{(z)}$ satisfy Eq. (2). Then, $\hat{J}_{BI}(\hat{h}, \hat{g})$ is unbiased and*

$$\sqrt{n'}(\hat{J}_{BI}(\hat{h}, \hat{g}) - J) \xrightarrow{d} \mathcal{N}(0, n\mathrm{var}[\Gamma(\mathcal{D}; h, g)]).$$

Note the restriction on $\hat{h}^{(z)}$ implicitly assumes we know logging policies. Theorems 1 and 2 together immediately lead to two important corollaries:

Corollary 1. *Under the assumptions of Theorem 2, $\hat{J}_{BI}(\hat{h}, \hat{g})$ is asymptotically normal and has asymptotic MSE lower bounded by V^*.*

Corollary 1 shows that among the class $\{\hat{J}_{\text{BI}}(\hat{h}, \hat{g})\}$, V^* is also an MSE lower bound. This class is *larger* than $\{\Gamma(\mathcal{D}; h, g)\}$ since we can always take $\hat{h} = h, \hat{g} = g$ although it may be infeasible in practice.

Corollary 2. *Suppose $g = q$ and $\|\hat{q}^{(z)} - q\|_2 = \mathrm{o}_p(1)$, Assumption 1 holds, and $\hat{q}^{(z)}, \pi^{\mathrm{e}}/\pi_*, q$ are uniformly bounded by some constants. Then, the cross-fitting doubly robust estimator*

$$\hat{J}_{\mathrm{DR}} := \hat{J}_{BI}(1/\pi_*, \hat{q})$$

is asymptotically normal and achieves the asymptotic variance lower bound V^*.

Corollary 2 shows that, when the logging policies are known, the minimum MSE is achievable by the cross-fitting doubly robust estimator $\hat{J}_{\mathrm{DR}}$. In Section 6, we discuss how to estimate $\hat{q}$, which is a necessary ingredient in constructing $\hat{J}_{\mathrm{DR}}$.

Theorem 2 can also be used to establish new theoretical results about other (suboptimal) estimators.

For example, we can consider a feasible version of $\hat{J}_{\text{IS-PW}}$, which we call $\hat{J}_{\text{IS-PW}(f)}$, where we use $\hat{\lambda}_k^* = \frac{n_k/\text{var}_{n_k}[\pi^e(a|s)r/\pi_k(a|s)]}{\sum_{k'} n_{k'}/\text{var}_{n_{k'}}[\pi^e(a|s)r/\pi_{k'}(a|s)]}$. Theorem 2 shows it has the *same* asymptotic variance as $\hat{J}_{\text{IS-PW}}$, which was not established in Agarwal et al. (2017). Additionally, we can consider the naively weighted and precision-weighted average of the doubly robust estimators in each dataset, respectively:

$$\hat{J}_{\text{DR-Avg}} := \hat{J}_{\text{BI}}(1/\pi_k(a \mid s), \hat{q}),$$
$$\hat{J}_{\text{DR-PW}} := \hat{J}_{\text{BI}}(n_k\hat{\lambda}_k^\dagger/(n\pi_k(a \mid s)), \hat{q}),$$
$$\hat{\lambda}_k^\dagger := \frac{n_k \text{var}_{n_k}[\pi^e r/\pi_k\{r-\hat{q}(s,a)\}+\hat{q}(s,\pi^e)]}{\sum_{k'} n_{k'}\text{var}_{n_{k'}}[\pi^e r/\pi_{k'}\{r-\hat{q}(s,a)\}+\hat{q}(s,\pi^e)]}.$$

These have the same asymptotic variance as $\Gamma(\mathcal{D}; 1/\pi_k(a \mid s), q), \Gamma(\mathcal{D}; n_k\lambda_k^\dagger/(n\pi_k(a \mid s)), q)$, respectively, where $\lambda_k^\dagger$ is the same as $\hat{\lambda}_k^\dagger$ with var_{n_k} replaced with var_{π_k}. Neither, however, is optimal and $\hat{J}_{\text{BI}}(1/\pi_*, \hat{q})$ outperforms these both.

Even if the estimators $\hat{h}^{(z)}$ does not satisfy Eq. (2), as long as the convergence point h satisfies Eq. (2), the final estimator is consistent, but it may not be asymptotically normal. In this case, we need additional conditions on the convergence rates to ensure $\sqrt{n'}$-consistency. This is relevant when the logging policies are not known. We explore this in Section 4.

3.3. The Class of Regular Estimators

The previous sections considered the minimal MSE in a class of estimators given explicitly by a certain structure or by a meta-algorithm. We now show that the same minimum in fact reigns among the asymptotic MSE of (almost) *all* estimators that are feasible in that they "work" for all data-generating processes (DGPs).

Recall our data is drawn from

$$\mathcal{D} \sim \prod_{k=1,i=1}^{K,n_k} p_S(s_{ki})\pi_k(a_{ki} \mid s_{ki})p_{R|S,A}(r_{ki} \mid s_{ki}, a_{ki}),$$

and that in the asymptotic regime we consider observing m independent copies of $\mathcal{D}$ (for total data size $n' = mn$). Consider first the case where π_k are known. Then, p_S and $p_{R|S,A}$ are the only unknowns in the above DGP. That is, different instances of the problem are given by setting these two to different densities. Thus, in the known-logger case, we consider the model (*i.e.*, class of instances) given by all DGPs where p_S and $p_{R|S,A}$ vary arbitrarily and π_k are fixed. (This is a *nonparametric* model in that these distributions are unrestricted.) Regular estimators are those that are $\sqrt{n'}$-consistent and remain so under perturbations to of size $1/\sqrt{n'}$ to the DGP that remain inside the model (for exact definition see van der Vaart, 1998). When $\hat{h}, \hat{g}$ satisfy the conditions of Theorem 2, $\hat{J}_{\text{BI}}(\hat{h}, \hat{g})$ is a regular estimator, as a consequence of Theorem 2 and van der Vaart (1998, Lemma 8.14) as in Fig. 1.

We next establish the efficiency bound in this model, meaning the minimal possible asymptotic variance among regular estimators. We paraphrase the key characteristic of the efficiency bound and provide additional detail in Appendix C.

Theorem 3. *(van der Vaart, 1998, Theorem 25.20 and Lemma 25.23) Given a model and a DGP in that model, if the efficiency bound $V_{\text{eff}} < \infty$ exists for the estimand J, then: (I) for any estimator $\hat{J}$ that is regular with respect to the model, the variance of the limiting distribution of $\sqrt{n'}(\hat{J} - J)$ is at least V_{eff}. (II) There exists an (unknown) estimator that is regular at the DGP with respect to the model achieving this bound.*

We next derive the efficiency bound for our problem. That is, for our average-reward estimand J in the model given by varying p_S, $p_{R|S,A}$ arbitrarily.

Theorem 4. *In the model with $p_S, p_{R|S,A}$ varying and π_k fixed, the efficiency bound is V^* as in Theorem 1, if finite.*

This shows that, remarkably, V^* is the lower bound in this large class of estimators, and that $\hat{J}_{\text{DR}}$ is in fact also optimal in the much broader sense of semiparametric efficiency. Moreover, while Theorem 3 only ensures some unknown (hence infeasible) efficient estimator exists for each DGP, we have shown that a single feasible estimator, $\hat{J}_{\text{DR}}$, is efficient in any DGP satisfying the conditions of Corollary 2.

Our derivation of the efficiency bounds is quite different from the one under the standard *un*stratified case. In efficiency theory for OPE in the standard *un*stratified case (Kallus and Uehara, 2020a) and in other standard semiparametric theory (Bickel et al., 1998; Tsiatis, 2006), we must consider iid sampling of observations. However, in the stratified case the data are *not* iid, since n_k are fixed. To be able to tackle the stratified case meaningfully we consider a dataset of size $n' \to \infty$ where the proportions of data from each logger, ρ_k, are always *fixed*. We achieve this in a new way via the equivalent construction of observing m independent copies of $\mathcal{D}$ with $m \to \infty$. Finally, note we will also discuss the difference between these two settings in more detail in Section 5.

4. Unknown Logging Policies Case

We now consider the case where the logging policies π_k are *not* known. Namely, we consider the model where we allow *all* of $p_S, p_{R|S,A}, \pi_1, \dots, \pi_K$ can vary arbitrarily. In this larger model, the efficiency bound is again the same.

Theorem 5. *In the model where $p_S, p_{R|S,A}, \pi_1, \dots, \pi_K$ all vary, the efficiency bound is V^* as in Theorem 1, if finite.*

Next, we prove double robustness of $\hat{J}_{\text{DR-}\hat{\pi}_*}$. This suggests when we posit parametric models for $\hat{q}, \hat{\pi}$, as long as either model is well-specified, the final estimator $\hat{J}_{\text{DR-}\hat{\pi}_*}$ is

$\sqrt{n'}$-consistent though might not be efficient. This is formalized as follows noting that well-specified parametric models converge at rate $n'^{-1/2}$.

Theorem 6 (Double Robustness). *Suppose Assumption 1 holds. Assume $\forall z \in [Z]$, for some $q^\dagger, \pi_*^\dagger$, $\|\hat{q}^{(z)} - q^\dagger\|_2 = \mathcal{O}_p(n'^{-1/2})$ and $\|\pi^{\mathrm{e}}/\hat{\pi}_*^{(z)} - \pi^{\mathrm{e}}/\pi_*^\dagger\|_2 = \mathcal{O}_p(n'^{-1/2})$, and $1/\pi_*^\dagger, q^\dagger, \hat{q}^{(z)}, 1/\hat{\pi}_*^{(z)}$ are uniformly bounded by some constants. Then, as long as either $q^\dagger = q$ or $\pi_*^\dagger = \pi_*$, $\hat{J}_{\mathrm{DR}}$-$\hat{\pi}_*$ is $\sqrt{n'}$-consistent.*

Finally, we consider how to achieve efficient estimation. The efficient estimator proposed in Section 3, $\hat{J}_{\mathrm{DR}} = \hat{J}_{\mathrm{BI}}(1/\pi_*, \hat{q})$, only works when logging policies, hence π_*, are known. A natural estimation approach when we do not know logging policies is to estimate π_* and plug it in:

$$\hat{J}_{\mathrm{DR}}\text{-}\hat{\pi}_* := \hat{J}_{\mathrm{BI}}(1/\hat{\pi}_*, \hat{q}).$$

We next prove the efficiency of $\hat{J}_{\mathrm{DR}\text{-}\hat{\pi}_*}$ under lax nonparametric rate conditions for the nuisance estimators.

Theorem 7 (Efficiency). *Suppose $\pi^{\mathrm{e}}/\pi_*, q, \hat{q}^{(z)}, \pi^{\mathrm{e}}/\hat{\pi}_*^{(z)}$ are uniformly bounded by some constants and that Assumption 1 holds. Assume $\forall z \in [Z]$, $\|\hat{q}^{(z)} - q\|_2 = \mathrm{o}_p(1)$, $\|\pi^{\mathrm{e}}/\hat{\pi}_*^{(z)} - \pi^{\mathrm{e}}/\pi_*\|_2 = \mathrm{o}_p(1)$, and $\|\hat{q}^{(z)} - q\|_2 \|\pi^{\mathrm{e}}/\hat{\pi}_*^{(z)} - \pi^{\mathrm{e}}/\pi_*\|_2 = \mathrm{o}_p(n'^{-1/2})$. Then, $\hat{J}_{\mathrm{DR}}$-$\hat{\pi}_*$ is efficient: $\sqrt{n'}(\hat{J}_{\mathrm{DR}}\text{-}\hat{\pi}_* - J) \xrightarrow{d} \mathcal{N}(0, V^*)$.*

First, notice that Corollary 2 can also be seen as corollary of Theorem 7 by noting that if we set $\hat{\pi}_* = \pi_*$ then $\|\pi^{\mathrm{e}}/\hat{\pi}_*^{(z)} - \pi^{\mathrm{e}}/\pi_*\|_2 = 0$. Second, notice that unlike Theorem 2, we do *not* restrict $\hat{h} = 1/\hat{\pi}_*$ to satisfy Eq. (2), as indeed satisfying it would be impossible when π_k are unknown. At the same time, $\hat{J}_{\mathrm{DR}}$-$\hat{\pi}_*$ is not unbiased (only asymptotically). Finally, notice that again $\hat{J}_{\mathrm{DR}}$-$\hat{\pi}_*$, an efficient estimator, does not appear to use logger identity data. We will, however, use it in Section 6 to improve q-estimation.

5. Stratified vs iid Sampling

We next discuss in more detail the differences and similarities between stratified and iid sampling. To make comparisons, consider the alternative iid DGP: $\mathcal{D}' = \{(S_i, A_i, R_i) : i = 1, \dots, n\}$, where $(S_i, A_i, R_i) \sim p_S(s)\pi_*(a \mid s)p_{R|S,A}(r \mid s, a)$ independently [1] for $i = 1, \dots, n$. That is, we observe n iid samples from the logging policy π_*. This is equivalent to randomizing the dataset sizes as $(n_1, \dots, n_K) \sim \mathrm{Multinomial}(n, \rho_1, \dots, \rho_K)$. In this iid setting, the results of Kallus and Uehara (2020a) show that the efficiency bound is the *same* V^* as in Theorems 1, 2 and 4.

[1] This assumption is primarily assumed to simplify the discussion. We can relax it by assuming some mixing conditions.

This is surprising since usually an estimator has different variances in different DGPs. For example, the variance of $\hat{J}_{\mathrm{IS}}$ under the two different sampling settings are *different*, *i.e.*:

$$\begin{aligned}\mathrm{var}_{\mathcal{D}}[\hat{J}_{\mathrm{IS}}] &= \tfrac{1}{n}\textstyle\sum_{k=1}^K \rho_k \mathrm{var}_{\pi_k}\left[\frac{\pi^{\mathrm{e}}(a|s)r}{\pi_*(a|s)}\right] \\ &\le \tfrac{1}{n}\mathrm{var}_{\pi_*}\left[\frac{\pi^{\mathrm{e}}(a|s)r}{\pi_*(a|s)}\right] = \mathrm{var}_{\mathcal{D}'}[\hat{J}_{\mathrm{IS}}].\end{aligned}$$

This inequality is easily proved by law of total variance and shows that the variance under stratified sampling is *lower*. The inequality is generally strict when π_k are distinct. This observation generalizes.

Theorem 8. *Suppose Assumption 1 holds. Consider the class of estimators $\{\mathbb{E}_n[\phi(s, a, r; g)]\}$, where ϕ is given by*

$$\phi(s, a, r; g) = \tfrac{\pi^{\mathrm{e}}(a|s)}{\pi_*(a|s)}(r - g(s, a)) + g(s, \pi^{\mathrm{e}})$$

and g is any function. Then: (I) Estimators in this class are unbiased. (II) We have

$$\mathrm{var}_{\mathcal{D}}[\mathbb{E}_n[\phi(s, a, r; g)]] \le \mathrm{var}_{\mathcal{D}'}[\mathbb{E}_n[\phi(s, a, r; g)]]. \quad (3)$$

(III) The above holds with equality if $g(s, a) = q(s, a)$ (IV) Conversely, if equality holds then $\mathbb{E}_{\pi^{\mathrm{e}}}[g(s, a) - q(s, a)] = 0$.

We have already seen the statement (III). The intuition for the statement (IV) is that the difference in Eq. (3), $\mathrm{var}[\mathbb{E}[\mathbb{E}_n[\phi(s, a, r; g)] \mid \{n_k\}_{k=1}^K]]$, is zero exactly when $\mathbb{E}_{\pi_k}[\phi(s, a, r; g)] = J \ \forall k \in [K]$, which leads to $\mathbb{E}_{\pi^{\mathrm{e}}}[g(s, a) - q(s, a)] = 0$. This conveys two things: stratification is still beneficial in reducing variance in finite samples since we never know the true q exactly, while at the same time the efficiency bound is the same in the two settings so this reduction washes out asymptotically when we use an efficient estimator.

This is related to but different from the benefit of stratification in survey sampling (Athey and Imbens, 2017; Fuller, 2009). In survey sampling, one considers stratification on the *covariates* s. Instead, we consider stratification on the treatment-assignment policies. Our result is distinct from the former and unique to our setting.

6. Stratified More Robust Doubly Robust Estimation

We have so far considered a meta-algorithm for efficient estimation given a q-estimator, which can be constructed by applying any type of off-the-shelf nonparametric or machine learning regression method to the whole dataset $\mathcal{D}$. However, if $\hat{q}$ is misspecified and inconsistent, the theoretical guarantees such as efficiency fail to hold. This a serious concern in practice as we always risk some level of model misspecification. We therefore next consider a more tailored

loss function for q-estimation that can still provide intrinsic efficiency guarantees regardless of specification.

Specifically, following Rubin and der Laan (2008); Cao et al. (2009); Farajtabar et al. (2018), we consider choosing the control variate g in a hypothesis class $\mathcal{Q}$ to minimize the variance of $\Gamma(\mathcal{D}; 1/\pi_*, g) = \mathbb{E}_n[\phi(s,a,r;g)]$. Specifically, we are interested in $\tilde{q}$:

$$\tilde{q} \in \arg\min_{g\in\mathcal{Q}} V(g),\ V(g) = n\mathrm{var}[\mathbb{E}_n[\phi(s,a,r;g)]] = \sum_{k=1}^{K} \rho_k \mathrm{var}_{\pi_k}[\phi(s,a,r;g)].$$

Of course, per Theorem 1, if $q \in \mathcal{Q}$ then $\tilde{q} = q$, but the concern is that $q \notin \mathcal{Q}$. In this case, $\tilde{q}$ will ensure best-in-class variance and will generally perform better than the best-in-class regression function $\bar{q} = \arg\min_{g\in\mathcal{Q}} \mathbb{E}_{\pi_*}[(r - g(s,a))^2]$, which empirical risk minimization would estimate.

In practice, we need to estimate $\mathrm{var}_{\pi_k}[\phi(s,a,r;g)]$. A feasible estimator is

$$\hat{q}_{\mathrm{SMRDR}} := \arg\min_{g\in\mathcal{Q}} \textstyle\sum_{k=1}^{K} \rho_k \mathrm{var}_{n_k}[\phi(s,a,r;g)].$$

Then, we define the *Stratified More Robust Doubly Robust* estimator as $\hat{J}_{\mathrm{SMRDR}} := \hat{J}_{\mathrm{BI}}(1/\pi_*, \hat{q}_{\mathrm{SMRDR}})$.

Theorem 9. *Suppose $\tilde{q}$ satisfies $\|\hat{q}_{SMRDR} - \tilde{q}\|_2 = \mathrm{o}_p(1)$. Also, suppose $\pi^{\mathrm{e}}/\pi_*, \sup_{g\in\mathcal{Q}} \|g\|_\infty$ are uniformly bounded by some constants. Then, $\sqrt{n'}(\hat{J}_{SMRDR} - J) \xrightarrow{d} \mathcal{N}(0, \min_{g\in\mathcal{Q}} V(g))$.*

Here, the assumption $\|\hat{q}_{\mathrm{SMRDR}} - \tilde{q}\|_2 = \mathrm{o}_p(1)$ is essentially satisfied by identifiability and assuming that ϕ belongs to a Glivenko–Cantelli class, i.e., the function class where uniform law of large number is satisfied (van der Vaart, 1998, Chapter 19).

Notice that if we had ignored the stratification and used the standard MRDR estimator (Cao et al., 2009), we would end up minimizing the *wrong* objective:

$$\hat{q}_{\mathrm{MRDR}} := \arg\min_{g\in\mathcal{Q}} \mathrm{var}_n[\phi(s,a,r;g)],$$

which targets the variance under iid sampling. In particular, we will *not* obtain the best-in-class variance. This is again a consequence of Theorem 8: when the control variates is not *exactly* q, the variances under stratified and iid setting are generally *different*.

Remark 1. When π_* is unknown, though we can just plug $\hat{\pi}_*$ in, the variance estimator has some bias. When π^* is parametrically estimated, we might be able to correct it following Cao et al. (2009). We leave it to future work regarding its formalization. We will show that the simple plug-in method empirically works in Section 8.

7. Extension to Reinforcement Learning

OPE with a single logging policy has been extensively studied in the RL setting (Precup et al., 2000; Liu et al., 2018; Jiang and Li, 2016; Kallus and Uehara, 2019a). We discuss the RL case with multiple loggers (Chen et al., 2020).

We consider observing $\mathcal{D} = \{D_1, \cdots, D_K\}$, where

$$\mathcal{D}_k = \{S_{kj}, A_{kj}, R_{kj}, S'_{kj}\}_{j=1}^{n_k} \overset{\text{i.i.d}}{\sim} p_k(s)\pi_k(a \mid s) p_{R|S,A}(r \mid s,a) p_{S'|S,A}(s' \mid s,a),$$

When $K = 1$, this is the standard DGP assumed in RL. Here, we consider K different loggers. The state densities $p_k(s)$ for each logger are also possibly different for each dataset. Our target is the policy value $J(\gamma)$ defined by the same MDP, an initial known density $p_e(s)$ and an evaluation policy π^{e} with a discount factor γ:

$$J(\gamma) = (1-\gamma)\lim_{T\to\infty} \mathbb{E}_{\pi^{\mathrm{e}}}[\textstyle\sum_{t=1}^{T} \gamma^{t-1} r_t],$$

where the expectation is taken w.r.t $s_1 \sim p_e(s), a_1 \sim \pi^e(a \mid s_1), r_1 \sim p_{R|S,A}(r \mid s_1, a_1), s_2 \sim p_{S'|S,A}(s \mid s_1, a_1), a_2 \sim \pi^e(a \mid s_2), \ldots$. We define the q-function as $q(s,a) := \mathbb{E}_{\pi^{\mathrm{e}}}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, a_1 = a]$.

We next derive efficiency bound in this model.

Theorem 10. *Let $\pi^*(s,a) := \sum_{k=1}^{K} (n_k/n)\pi_k(a|s)p_k(s)$. Whether the logging policies are known or unknown, the efficiency bounds is*

$$\mathbb{E}_{\pi_*(s,a)}\left[\left\{\frac{p^{(\infty)}_{\pi^{\mathrm{e}},\gamma}(s)\pi^{\mathrm{e}}(a|s)}{\pi_*(s,a)}\right\}^2 \mathrm{var}[r + \gamma q(s', \pi^{\mathrm{e}}) \mid s, a]\right],$$

where $p^{(\infty)}_{\pi^{\mathrm{e}},\gamma}(s)$ is the average visitation density with discount factor γ and an initial density $p_e(s)$.

If $\gamma = 0$, RL reduces to the bandit setting, and indeed the above would match the first term in V^* in Theorem 1. The second term in V^* does not appear here since we assume that the initial density $p_e(s)$ is known.

Next, we propose an efficient estimator. We define

$$w(s,a) := \frac{p^{(\infty)}_{\pi^{\mathrm{e}},\gamma}(s)\pi^{\mathrm{e}}(a|s)}{\pi_*(s,a)}.$$

We can estimate q using fitted q-iteration (Antos et al., 2008) and q or w using minimax methods (Liu et al., 2018; Zhang et al., 2020; Uehara et al., 2020). Given some estimates $\hat{w}(s,a), \hat{q}(s,a)$ for $w(s,a), q(s,a)$, we set

$$\hat{J}(\gamma) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \hat{w}(S_i, A_i)\{R_i + \gamma \hat{q}(S'_i, \pi^{\mathrm{e}}) - \hat{q}(S_i, A_i)\} + \mathbb{E}_{p_e(s)}[\hat{q}(s, \pi^{\mathrm{e}})]$$

The efficiency is proved as in Theorem 7. For details, refer to Appendix B.

Table 1. The evaluation and logging policies used in the experiments.

evaluation policy (π_e)	$1.00\pi_{\text{det}} + 0.00\pi_u$
logging policy 1 (π_1)	$0.95\pi_{\text{det}} + 0.05\pi_u$
logging policy 2 (π_2)	$0.05\pi_{\text{det}} + 0.95\pi_u$

We next compare to Chen et al. (2020). Their estimator is given by setting $\hat{q} = 0$, *i.e.*, the RL extension of the IS estimator. In contrast, we incorporate control variates and can therefore obtain efficiency, similarly to our key point in the bandit setting. Additionally, unlike Chen et al. (2020), we do not assume that $p_k(s)$ is a stationary distribution.

8. Experimental Results

We next empirically compare our methods with the existing estimators for OPE with multiple loggers. To focus on main takeaways, we restrict our attention to the bandit setting.

Setup. Following previous work on OPE (Farajtabar et al., 2018; Wang et al., 2017; Kallus and Uehara, 2019b) we evaluate our estimators using multiclass classification datasets from the UCI repository. Here we consider the optdigits and pendigits datasets (see Table 3 in Appendix E.). We transform each classification dataset into a contextual bandit dataset by treating the labels as actions and recording reward of 1 if the correct label is chosen by a classifier, and 0 otherwise. This lets us evaluate and compare several different estimators with the ground-truth policy value of an evaluation policy.

We split the original data into training (30%) and evaluation (70%) sets. We first obtain a deterministic policy π_{det} by training a logistic regression model on the training set. Then, following Table 1, we construct evaluation and logging policies as mixtures of one of the deterministic policy and the uniform random policy π_u. We vary $\rho_1/(1-\rho_1) = n_1/n_2$ in $\{0.1, 0.25, 0.5, 1, 2, 4, 10\}$. Since π_1 is closer to π^{e} than π_2, larger ρ_1/ρ_2 corresponds to an easier problem. We then split the evaluation dataset into two according to proportions ρ_1, ρ_2 and in each dataset we use the corresponding policy to make decisions and generate reward observations (the true label is then omitted). Using the resulting dataset we consider various estimators $\hat{J}$ for J. We describe additional details of the experimental setup in Appendix E.

We repeat the process $M = 200$ times with different random seeds and report the *relative root MSE*:

$$\text{Relative-RMSE}\ (\hat{J}) = \frac{1}{J\sqrt{M}} \sqrt{\sum_{m=1}^{M} \left(J - \hat{J}_m\right)^2}$$

where $\hat{J}_m$ is an estimated policy value with m-th data.

Estimators considered. We consider the following estimators:

- Our proposed estimators, J_{DR}-$\hat{\pi}_*$, $\hat{J}_{\text{SMRDR}}$.
- Standard estimators in the iid setting, $\hat{J}_{\text{IS}}$, $\hat{J}_{\text{MRDR}}$.
- (Feasible versions of) the two estimators proposed by (Agarwal et al., 2017), $\hat{J}_{\text{IS-Avg}}$, $\hat{J}_{\text{IS-PW}}$.
- The natural doubly robust extension of these as discussed in Section 3.2, $\hat{J}_{\text{DR-Avg}}$, $\hat{J}_{\text{DR-PW}}$.

We suppose we do not know logging policies. For all estimators, we estimate the logging policies using logistic regression on the evaluation set with 2-fold cross-fitting as in Algorithm 1. Most of the estimators above are introduced with known logging densities in the previous sections. Here, we just replace each π_k with their estimates. For DR, DR-Avg, and DR-PW, we construct q-estimates using logistic regression again using 2-fold cross-fitting as in Algorithm 1. For SMRDR and MRDR, we optimize their respective estimated variance objectives over the class of logistic regression $\mathcal{Q}$. We use *tensorflow* and the same hyperparameter setting for DR, DR-Avg, DR-PW, SMRDR, and MRDR to ensure a fair comparison.

Results. The resulting Relative-RMSEs on optdigits and pendigits datasets with varying values of n_1/n_2 are given in Figs. 2 and 3. Several findings emerge from the results. First, we see the dilemma pointed out by Agarwal et al. (2017): Specifically, the ordering of the variances of IS-Avg and IS-PW depend on the instance. More generally, there is no clear ordering between IS, IS-Avg, IS-PW, DR-Avg, and DR-PW. For example, on the optdigits data, DR-PW performs best among baselines with small values of n_1/n_2, while IS performs better with large values of n_1/n_2. This behavior is predicted by our analysis showing none of these estimators are optimal.

Second, our proposed estimators successfully resolve the dilemma and are superior to the above suboptimal estimators. Moreover, we see SMRDR generally performs better than DR, especially when the overlap is weak (n_1/n_2 is small), which exacerbates issues of misspecification. Recall DR is generally *not* efficient when the q-hypothesis class is misspecified; on the other hand, SMRDR is the best among the hypothesis class as in Theorem 9. It does appear that DR outperforms SMRDR in the specific example of optdigits when overlap is strong (n_1/n_2 is large), which might be attributed to bad optimization of the non-convex objective compared to a reasonably good-enough plug-in q-estimate.

Finally, we directly compare the performances of SMRDR and MRDR in Figure 3. We observe that SMRDR significantly outperforms MRDR in the stratified setting, leading to up to 45% reduction in error. This strongly highlights that even though the asymptotic efficiency bounds are the same

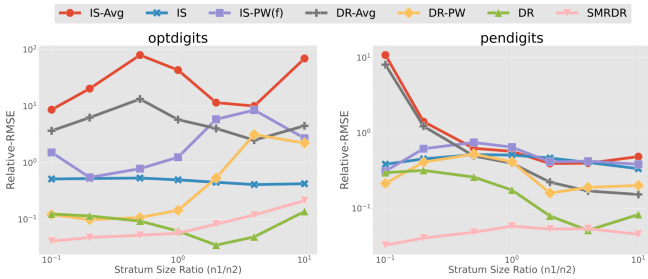

Figure 2. Comparing proposed estimators to some variants of IS type estimators.

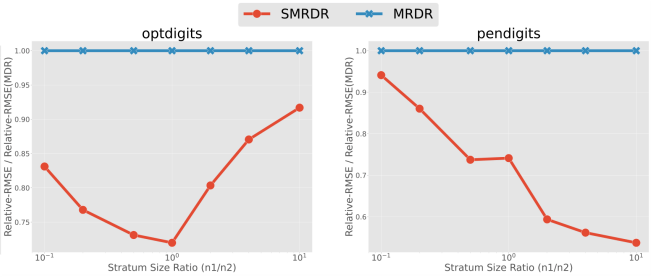

Figure 3. Comparing SMRDR (leveraging the stratification) and MRDR (ignoring the stratification).

in the stratified and iid settings, leveraging the stratification structure can still offer significant gains in the multiple logger setting.

9. Conclusions

We studied OPE in the multiple logger setting, framing it as a form of stratified sampling. We then studied optimality in several classes of estimators and showed that, at least asymptotically, the minimum MSE is the same among all of them. We proposed feasible estimators that can achieve this minimum, whether logging policies are known or not. This gives a concrete and positive resolution to the multiple logger dilemma posed in Agarwal et al. (2017). We further discuss how to take stratification into account when choosing best-in-class control variates.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1846210.

Masatoshi Uehara is partially supported by Masason foundation

References

Agarwal, A., S. Basu, T. Schnabel, and T. Joachims (2017). Effective evaluation using logged bandit feedback from multiple loggers. KDD '17, pp. 687–696.

Antos, A., C. Szepesvári, and R. Munos (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning 71*, 89–129.

Athey, S. and G. Imbens (2017). *The Econometrics of Randomized Experiments*. Elsevier.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer.

Cao, W., A. A. Tsiatis, and M. Davidian (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*, 723–734.

Chen, X., L. Wang, Y. Hang, H. Ge, and H. Zha (2020). Infinite-horizon off-policy policy evaluation with multiple behavior policies. *In Proceedings of the 8th International Conference on Learning Representations (ICLR), 2020*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–C68.

Dudík, M., D. Erhan, J. Langford, L. Li, et al. (2014). Doubly robust policy evaluation and optimization. *Statistical Science 29*(4), 485–511.

Farajtabar, M., Y. Chow, and M. Ghavamzadeh (2018). More robust doubly robust off-policy evaluation. *In Proceedings of the 35th International Conference on Machine Learning*, 1447–1456.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in markov chain monte carlo. *Technical Report 568. School of Statistics, University of Minnesota, Minneapolis*.

Gutmann, M. and A. Hyvärinen (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304.

He, L., L. Xia, W. Zeng, Z.-M. Ma, Y. Zhao, and D. Yin (2019). Off-policy learning for multiple loggers. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Imbens, G. W. and T. Lancaster (1996). Efficient estimation and stratified sampling. *Journal of Econometrics 74*(2), 289–318.

Jiang, N. and L. Li (2016). Doubly robust off-policy value evaluation for reinforcement learning. *In Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume*, 652–661.

Kallus, N. (2018). Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pp. 8895–8906.

Kallus, N. and M. Uehara (2019a). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*.

Kallus, N. and M. Uehara (2019b). Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, pp. 3320–3329.

Kallus, N. and M. Uehara (2020a). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research 21*, 1–63.

Kallus, N. and M. Uehara (2020b). Efficient evaluation of natural stochastic policies in offline reinforcement learning.

Kong, A., P. McCullagh, X. L. Meng, D. Nicolae, and Z. Tan (2003). A theory of statistical models for monte carlo integration. *Journal of the Royal Statistical Society. Series B, Statistical methodology 65*(3), 585–618.

Liu, Q., L. Li, Z. Tang, and D. Zhou (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31*, pp. 5356–5366.

Narita, Y., S. Yasui, and K. Yata (2019). Efficient counterfactual learning from bandit feedback. *AAAI*.

Precup, D., R. Sutton, and S. Singh (2000). Eligibility traces for off-policy policy evaluation. *In Proceedings of the 17th International Conference on Machine Learning*, 759–766.

Rubin, D. B. and M. J. V. der Laan (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randmized experiments and survival analysis. *International Journal of Biostatistics 4*, Article 5.

Strehl, A. L., J. Langford, L. Li, and S. M. Kakade (2010). Learning from Logged Implicit Exploration Data. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pp. 2217–2225.

Swaminathan, A., A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni (2017). Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems 30*, pp. 3632–3642.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. New York, NY: Springer New York.

Uehara, M., J. Huang, and N. Jiang (2020). Minimax weight and q-function learning for off-policy evaluation. *ICML 2020 (To appear)*.

Uehara, M., T. Matsuda, and H. Komaki (2018). Analysis of noise contrastive estimation from the perspective of asymptotic variance. *arXiv preprint arXiv:1808.07983*.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.

Wang, Y.-X., A. Agarwal, and M. Dudik (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3589–3597.

Wooldridge, J. M. (2001). Asymptotic properties of weighted m -estimators for standard stratified samples. *Econometric Theory 17*, 451–470.

Yang, S. and P. Ding (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association 115*(531), 1540 – 1554.

Yin, M., Y. Bai, and Y.-X. Wang (2020). Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning.

Zhang, R., B. Dai, L. Li, and D. Schuurmans (2020). Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*.

Zheng, W. and M. J. van Der Laan (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, pp. 459–474. New York, NY: Springer New York.