Variational Approach to Optimal IPS Estimator for Multi-logger Off-Policy Evaluation

Joon Suk Huh*
Department of Computer Science
University of Wisconsin–Madison, USA
joon@cs.wisc.edu

Junghoon Seo*
PIT IN Corp., Republic of Korea
sjh@pitin-ev.com

Abstract

We study off-policy evaluation (OPE) in contextual bandits with data collected from multiple logging policies. Inverse propensity scoring (IPS) is a standard approach to OPE and extends naturally to the multi-logger setting. However, as highlighted by Agarwal et al. [2017], neither balanced nor weighted IPS consistently outperforms the other in this setting. We resolve this dilemma by deriving an optimal IPS estimator with sample-dependent weights that minimize variance subject to unbiasedness. Using a variational calculus approach, we obtain closed-form optimal weights, yielding an estimator that is unbiased and achieves asymptotically optimal variance within a generalized weighted-IPS estimator class. Instead of a reward model, the estimator needs only a $K \times K$ matrix and a K-vector of moments, where K is the number of logging policies, and we quantify how much efficiency this costs relative to the semiparametric oracle. Experiments on benchmark datasets show that our estimator generally improves on existing multi-logger IPS methods. We also extend our estimator to a doubly robust form by incorporating a conditional reward estimator. We compare the resulting DR extension with the semiparametrically efficient DR estimator of Kallus et al. [2021], both theoretically and empirically. We show that the two estimators achieve the same asymptotic variance when the conditional reward estimator is consistent, while our estimator can retain semiparametric efficiency under certain forms of reward misspecification. Empirically, the proposed DR estimator often achieves lower estimation error than the existing DR baseline.

1 Introduction

Off-policy evaluation (OPE) is a core tool for data-driven decision making, with applications in recommender systems, healthcare, and advertising. The goal is to estimate the value of an evaluation policy using data collected by one or more different logging policies. While the single-logger setting is well understood, many practical datasets are collected from multiple logging policies, for example through heterogeneous systems, parallel deployments, or sequential changes to a production policy [Agarwal et al., 2017, Saito et al., 2021, Jung and Bellot, 2024]. This multi-logger structure creates a basic statistical challenge: different strata can have very different overlaps with the evaluation policy, so combining them poorly can lead to large variance.

Inverse propensity scoring (IPS) is a standard approach to OPE [Rosenbaum and Rubin, 1983, Precup et al., 2000, Bottou et al., 2013]. In the single-logger case, IPS reweights each observed reward by the propensity ratio between the evaluation and logging policies. In the multi-logger case, however, there is an additional design choice: one must decide how much each logger should contribute to the final estimate. Existing approaches answer this question in different ways. Naive IPS treats all logged

*These authors contributed equally to this work.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

samples with their own propensity ratios, balanced IPS uses the sample-size-weighted average logging policy, and weighted IPS chooses logger-level weights based on estimated variances [Agarwal et al., 2017]. As observed by Agarwal et al. [2017], the relative performance of balanced and weighted IPS is highly instance dependent: neither consistently dominates the other across problem instances, and naive IPS is dominated in their analysis. Kallus et al. [2021] later characterized the semiparametrically efficient estimator over a broader class, which serves as our benchmark.

We revisit this problem from a variational perspective. Instead of choosing among a fixed set of IPS estimators, we consider a generalized weighted-IPS class Γ_{wIPS} in which each sample may receive a weight depending on the logged tuple and its logger identity. That is, the weights are functions of the state, action, reward, and logger. We impose constraints that ensure unbiasedness and then directly minimize the variance functional over all feasible weight functions. This yields a convex variance minimization problem whose solution admits a closed-form characterization. The resulting estimator, denoted $\hat{J}_{\mathrm{oIPS}}$, can be viewed as a balanced-IPS estimator augmented with an optimal control-variate term built from the logging propensities. This construction does not require the conditional mean reward. Instead, it estimates a $K \times K$ matrix $\mathbf{T}$ and a vector $\mathbf{c} \in \mathbb{R}^K$ of moments (§3).

This perspective clarifies the apparent instance-dependence of existing IPS estimators. The optimal weights depend on the interaction among the evaluation policy, the logging policies, the reward distribution, and the stratum sizes. Thus, fixed IPS rules such as naive IPS, balanced IPS, and weighted IPS can each perform well on some instances and poorly on others. By optimizing over a class containing these IPS baselines, $\hat{J}_{\mathrm{oIPS}}$ adapts to the instance and achieves the smallest asymptotic variance within Γ_{wIPS}. We prove that the corresponding feasible cross-fitted implementation is unbiased when logging policies are known and is asymptotically normal with the optimal variance.

We also relate our estimator to the semiparametric efficiency theory of Kallus et al. [2021]. Their framework shows that the most efficient estimator uses the conditional mean reward as a control variate. Our IPS estimator can be viewed as a restricted, reward model-free version of this idea: instead of estimating the conditional mean reward directly, it constructs the best control variate available from the logging policies. When this logger-induced control variate is rich enough to represent the relevant reward structure, our estimator asymptotically recovers the efficient oracle. Otherwise, its excess variance is determined by the part of the reward structure that cannot be represented by the logging policies.

Finally, we extend our variational approach to a doubly robust estimator, denoted $\hat{J}_{\text{DR-oIPS}}$, by incorporating a reward model. This extension keeps the same idea as our IPS estimator: it uses the logging policies to choose an additional variance-reducing correction. This makes it directly comparable to the doubly robust balanced IPS estimator $\hat{J}_{\text{DR-bIPS}}$ of Kallus et al. [2021]. When the reward model is correctly specified, both estimators attain the same asymptotic variance. When the reward model is misspecified, however, $\hat{J}_{\text{DR-oIPS}}$ can remain efficient if the misspecification is representable by the logging policies, whereas $\hat{J}_{\text{DR-bIPS}}$ need not. This gives our DR extension an additional robustness advantage.

Summary of Contributions. We make the following contributions:

1. **A variationally optimal IPS estimator for multi-logger OPE.** We introduce a generalized weighted-IPS class Γ_{wIPS} and derive the variance-minimizing unbiased estimator within this class. The resulting estimator $\hat{J}_{\mathrm{oIPS}}$ has closed-form sample-dependent weights and estimates $\mathbf{T}$ and $\mathbf{c}$ in place of a reward model.
2. **Closed-form characterization, feasible estimation, and projection interpretation.** We show that the optimal weights are characterized by a linear system $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$, whose components can be estimated from data, leading to a practical cross-fitted implementation. We further show that this system has a geometric interpretation: it projects the reward-weighted evaluation policy onto the span of the logging policies, and the resulting projection residual exactly determines the variance gap to the semiparametrically efficient oracle.
3. **Doubly robust extension and comparison with efficient DR estimators.** We extend $\hat{J}_{\mathrm{oIPS}}$ to a doubly robust estimator $\hat{J}_{\text{DR-oIPS}}$ by applying the same variational construction to residuals based on $\hat{q}$. We show that $\hat{J}_{\text{DR-oIPS}}$ matches the asymptotic variance of $\hat{J}_{\text{DR-bIPS}}$ when $\hat{q}$ is consistent, and can retain the oracle variance under certain structured misspecifications of $\hat{q}$.

4. **Empirical validation.** Across benchmark datasets and a broad range of logger mixtures, our IPS estimator generally improves over existing multi-logger IPS baselines. The doubly robust extension also often achieves lower estimation error than the semiparametrically efficient baseline.

Related Work. Appendix G discusses related work on OPE with multiple loggers [He et al., 2019, Chen et al., 2020], connections to multiple importance sampling [Veach and Guibas, 1995, Elvira et al., 2019, Kondapaneni et al., 2019], and broader variance-reduction methods for IPS.

2 Preliminaries

This section introduces the problem setup and reviews baseline estimators used in our experiments.

Notation. For $n \in \mathbb{N}$, let $[n] \coloneqq \{1, \dots, n\}$. We use boldface for finite-dimensional vectors and matrices, e.g., $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{M} \in \mathbb{R}^{n \times n}$. Their entries are written in plain font (e.g., v_i and M_{ij}), and we likewise use plain symbols for scalar components such as α_i, c_i, and w_i. The symbol $\mathbb{1}[\,\cdot\,]$ denotes the indicator function.

Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space and let $f, g \in L^2(\mu)$. We write $\langle \cdot, \cdot \rangle_\mu$ for the $L^2(\mu)$ inner product:

$$\langle f, g \rangle_\mu \coloneqq \int_\Omega f(\omega)\, g(\omega)\, d\mu(\omega).$$

The associated $L^2(\mu)$-norm is $\|f\|_{L^2(\mu)} \coloneqq \langle f, f \rangle_\mu^{1/2}$. We use the same notation $\langle f, g \rangle_\mu$ whenever fg is μ-integrable, even if f or g is not in $L^2(\mu)$.

Throughout the paper, we fix a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$, where $\Omega \coloneqq \mathcal{S} \times \mathcal{A} \times \mathcal{R}$, with $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R}$ denoting the sets of states, actions, and rewards, respectively, as defined in §2.1. For $\omega = (s, a, r) \in \Omega$, let $\rho_r(\omega) \coloneqq r$ denote the projection onto the reward coordinate; R denotes the corresponding random variable. Here, $\mathcal{F}$ is a suitable σ-algebra and μ is a fixed reference measure. For brevity, we write $L^2 \coloneqq L^2(\mu, \Omega)$. For any measure $\nu \ll \mu$, we write $L^2(\nu)$ for the space of square-integrable functions with respect to ν. When Ω is a topological space, we denote by $\operatorname{supp}(\mu)$ the support of a measure μ, and by $\operatorname{supp}(f)$ the support of a measurable real-valued function f. For a policy π, we write $\operatorname{supp}(\pi(\cdot|s)) \coloneqq \operatorname{cl}(\{a \in \mathcal{A} : \pi(a \mid s) > 0\})$. For measurable functions f and g with $g \neq 0$ wherever $f \neq 0$, we define the ratio f/g pointwise, with the convention that $0/0 = 0$.

2.1 Problem Setting

We study OPE in contextual bandits, following the multi-logger setup in [Agarwal et al., 2017, He et al., 2019, Kallus et al., 2021]. A policy π maps a state (i.e., context) $s \in \mathcal{S}$ to a distribution over actions in $\mathcal{A}$, written as $\pi(a \mid s)$.

An environment is defined by the following *unknown* distributions: a context distribution $p_s(s)$ and reward distributions $p_r(r \mid s, a)$ over $\mathcal{R} \subseteq \mathbb{R}$ for each context-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. Interaction proceeds as follows: the environment samples $S \sim p_s$, the actor samples $A \sim \pi(\,\cdot \mid S)$, and the environment then samples the reward $R \sim p_r(\,\cdot \mid S, A)$. This induces a joint distribution μ_π on $\mathcal{S} \times \mathcal{A} \times \mathcal{R}$, whose density is $p_s(s)\, \pi(a|s)\, p_r(r|s, a)$ with respect to the dominating measure $\mu(ds, da, dr) = \mu_s(ds)\, \mu_a(da)\, \mu_r(dr \mid s, a)$, where $\mu_r(\cdot \mid s, a)$ is a reference kernel on $\mathcal{R}$. Equivalently, $\mu_\pi(ds, da, dr) = P_S(ds)\, \pi(da \mid s)\, P_R(dr \mid s, a)$, where $P_S(ds) = p_s(s)\, \mu_s(ds)$, $\pi(da \mid s) = \pi(a \mid s)\, \mu_a(da)$, and $P_R(dr \mid s, a) = p_r(r \mid s, a)\, \mu_r(dr \mid s, a)$. The expected reward of the policy π, denoted $J(\pi)$, is

$$J(\pi) \coloneqq \mathbb{E}_{\mu_\pi}[R],$$

where $\mathbb{E}_{\mu_\pi}$ denotes expectation with respect to μ_π.

The goal of OPE is to estimate the policy value $J(\pi_e)$ of an evaluation policy π_e from logged data. In the single-logger case, we observe i.i.d. samples $\mathcal{D} \coloneqq \{(S_j, A_j, R_j)\}_{j=1}^n$ drawn from μ_π, where π is the logging policy and typically $\pi \neq \pi_e$. In this work, we consider stratified OPE with multiple logging policies $(\pi_i)_{i \in [K]}$. The full dataset is $\mathcal{D} \coloneqq \{\mathcal{D}_i\}_{i \in [K]}$, where the strata are independent and each stratum $\mathcal{D}_i$ consists of n_i i.i.d. samples from μ_{π_i}, namely $\mathcal{D}_i \coloneqq \{(S_{ij}, A_{ij}, R_{ij})\}_{j=1}^{n_i}$, where $S_{ij} \sim p_s$, $A_{ij} \sim \pi_i(\cdot \mid S_{ij})$, and $R_{ij} \sim p_r(\cdot \mid S_{ij}, A_{ij})$.

2.2 Existing Inverse Propensity Scoring (IPS)-based Approaches

Naive IPS. Naive IPS reweights each observed reward by the propensity ratio between the evaluation and logging policies:

$$\hat{J}_{\mathrm{IPS}}(\pi_e; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R)\in\mathcal{D}_i} R \, \frac{\pi_e(A \mid S)}{\pi_i(A \mid S)}, \tag{2.1}$$

where $n \coloneqq \sum_{i=1}^{K} n_i$ is the total number of samples.

Balanced IPS. Agarwal et al. [2017] proposed a multiple importance sampling (MIS)-inspired estimator that uses the averaged logging policy as inverse propensity score:

$$\hat{J}_{\mathrm{bIPS}}(\pi_e; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R)\in\mathcal{D}_i} R \, \frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}, \tag{2.2}$$

where $\pi_{\mathrm{avg}} \coloneqq \frac{1}{n} \sum_{i=1}^{K} n_i \, \pi_i$ and $n_i \coloneqq |\mathcal{D}_i|$.

Weighted IPS. Agarwal et al. [2017] also introduced weighted IPS, which combines logger-specific IPS estimates using variance-based weights:

$$\hat{J}_{\mathrm{wIPS}}(\pi_e; \mathcal{D}) = \sum_{i=1}^{K} \hat{\gamma}_i \sum_{(S,A,R)\in\mathcal{D}_i} R \, \frac{\pi_e(A \mid S)}{\pi_i(A \mid S)}, \tag{2.3}$$

where $\hat{\gamma}_i \coloneqq \left(\widehat{\sigma}_r^2(\pi_e||\pi_i) \sum_{j=1}^{K} \frac{n_j}{\widehat{\sigma}_r^2(\pi_e||\pi_j)} \right)^{-1}$ and $\hat{\sigma}_r^2(\pi_e||\pi_i)$ is a positive consistent estimator of $\mathrm{Var}_{\mu_{\pi_i}}\left[R \, \frac{\pi_e(A|S)}{\pi_i(A|S)} \right]$, which quantifies the divergence between the target and logging policies.

Doubly Robust Balanced IPS. Kallus et al. [2021] proposed a doubly robust version of $\hat{J}_{\mathrm{bIPS}}$ using a conditional mean reward estimator $\hat{q}$:

$$\hat{J}_{\text{DR-bIPS}}(\pi_e; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R)\in\mathcal{D}_i} \left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} \big(R - \hat{q}(S,A)\big) + \hat{q}(S, \pi_e) \right], \tag{2.4}$$

where $\hat{q}(S, \pi_e) \coloneqq \int_{\mathcal{A}} \hat{q}(S,a) \pi_e(a \mid S) \, d\mu_a(a)$. When the logging policies are known and $\hat{q}$ consistently estimates the conditional mean reward $q(s,a) \coloneqq \mathbb{E}[R \mid S = s, A = a]$, $\hat{J}_{\text{DR-bIPS}}$ attains the semiparametric efficiency bound within the estimator class studied by Kallus et al. [2021], under their regularity conditions. In practice, $\hat{q}$ is estimated by cross-fitting: the reward model is trained on one part of the data and evaluated on held-out samples. The same cross-fitting strategy for nuisance quantities is also used in $\hat{J}_{\mathrm{wIPS}}$.

3 Variational Optimization Approach

In this section, we introduce our main estimator, which is derived by optimizing a convex variance functional and is asymptotically optimal within the generalized weighted-IPS class denoted by Γ_{wIPS}.

3.1 Generalized Weighted-IPS Estimator Class

Let $\mathbf{w} \coloneqq (w_i)_{i\in[K]}$, where each $w_i : \mathcal{S} \times \mathcal{A} \times \mathcal{R} \to \mathbb{R}$ is a sample-dependent weight function. The generalized weighted-IPS estimator class Γ_{wIPS} is

$$\boxed{\hat{J}(\pi_e, \mathbf{w}; \mathcal{D}) \coloneqq \sum_{i=1}^{K} \frac{1}{n_i} \sum_{(S,A,R)\in\mathcal{D}_i} w_i(S,A,R) \, \frac{\pi_e(A \mid S)}{\pi_i(A \mid S)}} \tag{3.1}$$

Estimators in Γ_{wIPS} are unbiased when the weights satisfy the following conditions for every $\omega = (s, a, r)$:

$$\textbf{Normalization: } \sum_{i=1}^{K} w_i(\omega) = \rho_r(\omega), \quad \textbf{Dominance: } \pi_i(a|s) = 0 \implies w_i(s, a, r) = 0.$$

Theorem 3.1. [Unbiasedness] *For any weights $\mathbf{w} \in (L^2)^K$ that satisfy the normalization and dominance conditions and for which each weighted score $w_i(S, A, R)\, \pi_e(A \mid S)/\pi_i(A \mid S)$ is μ_{π_i}-integrable, $\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})$ is unbiased, that is, $J(\pi_e) = \mathbb{E}[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})]$, where the expectation is over the sampling randomness of logged data $\mathcal{D} \coloneqq (\mathcal{D}_i)_{i\in[K]}$.*

In §4.3, we compare Γ_{wIPS} with the broader class of Kallus et al. [2021], whose efficient estimator requires an estimate of the conditional mean reward.

Variationally Optimized Weights. Since $\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})$ is a linear functional of $\mathbf{w}$, its variance is convex. We optimize this variance functional directly over feasible $\mathbf{w}$ satisfying the normalization and dominance constraints. The resulting optimal weight vector $\mathbf{w}^\star \coloneqq (w_i^\star)_{i\in[K]}$, whose optimality is established in Theorem 4.1, takes the form

$$w_i^\star(s, a, r) = \frac{\alpha_i^\star \pi_i(a|s)}{\pi_e(a|s)} + \frac{n_i\, \pi_i(a|s)}{n\, \pi_{\mathrm{avg}}(a|s)} \left(r - \frac{\sum_{j=1}^{K} \alpha_j^\star\, \pi_j(a|s)}{\pi_e(a|s)} \right) \tag{3.2}$$

where $\boldsymbol{\alpha}^\star \coloneqq (\alpha_i^\star)_{i\in[K]}$ is a solution of $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$, with $\mathbf{T} \in \mathbb{R}^{K\times K}$ and $\mathbf{c} \in \mathbb{R}^K$ whose components are given by

$$T_{ij} \coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[\frac{\pi_j(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} \right], \quad c_i \coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[R\, \frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} \right] \tag{3.3}$$

We call $\mathbf{T}$ the *technique matrix* and $\mathbf{c}$ the *contribution vector*. Estimating $\mathbf{T}$ and $\mathbf{c}$ thus replaces estimating the conditional mean reward function. We later show in this section that the equation $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ can be interpreted as the normal equation for an orthogonal projection.

Control-Variate Form. Plugging the expression for $\mathbf{w}^\star$ in (3.2) into $\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})$ in (3.1) yields the following form of the variationally optimal IPS estimator [2]:

$$\boxed{\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R)\in\mathcal{D}_i} \left[\frac{R\, \pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} + \sum_{k=1}^{K} \alpha_k^\star - \frac{g_{\boldsymbol{\alpha}^\star}(S, A)}{\pi_{\mathrm{avg}}(A \mid S)} \right]} \tag{3.4}$$

where $g_{\boldsymbol{\alpha}^\star}(s, a) \coloneqq \sum_{k\in[K]} \alpha_k^\star\, \pi_k(a|s)$. We refer to this as the control-variate form: the term involving $\sum_k \alpha_k^\star - g_{\boldsymbol{\alpha}^\star}/\pi_{\mathrm{avg}}$ is mean-zero and therefore acts as a control variate.

Geometric View. Let $q(s, a) \coloneqq \mathbb{E}[R \,|\, S = s, A = a]$ denote the conditional mean reward. Then, we can view $g_{\boldsymbol{\alpha}^\star}$ as the orthogonal projection of $q\, \pi_e$ onto the finite-dimensional space spanned by $\{\pi_i\}_{i\in[K]}$. Specifically, consider $L^2(\mu_{\log}, \mathcal{S} \times \mathcal{A})$, where $\mu_{\log}$ is given by

$$d\mu_{\log}(s, a) \coloneqq \frac{p_s(s)}{\pi_{\mathrm{avg}}(a|s)}\, d\mu_{s,a}(s, a),$$

with $\mu_{s,a} = \mu_s \otimes \mu_a$ denoting the base measure over $\mathcal{S} \times \mathcal{A}$. Equip $L^2(\mathcal{S} \times \mathcal{A}, \mu_{\log})$ with the inner product $\langle \cdot, \cdot \rangle_{\mu_{\log}}$. Then, one can rewrite (3.3) in terms of $\langle \cdot, \cdot \rangle_{\mu_{\log}}$ as

$$T_{ij} = \big\langle \pi_i, \pi_j \big\rangle_{\mu_{\log}}, \quad c_i = \big\langle q\, \pi_e, \pi_i \big\rangle_{\mu_{\log}}.$$

Thus, $\mathbf{T}\boldsymbol{\alpha}^\star = \mathbf{c}$ if and only if $\big\langle q\, \pi_e - \sum_j \alpha_j^\star \pi_j, \pi_i \big\rangle_{\mu_{\log}} = 0$ for all $i \in [K]$. This shows that $g_{\boldsymbol{\alpha}^\star}$ is the projection of $q\, \pi_e$ onto $\mathrm{span}\big(\{\pi_i\}_{i\in[K]}\big)$, with respect to the norm induced by $\langle \cdot, \cdot \rangle_{\mu_{\log}}$.

We formally state this observation in Lemma E.2. Later, in §4.3, we relate the squared $\| \cdot \|_{\mu_{\log}}$ norm of the residual component of $q\, \pi_e$ orthogonal to $\mathrm{span}\big(\{\pi_i\}_{i\in[K]}\big)$ to the excess variance of our estimator relative to the semiparametrically efficient oracle of Kallus et al. [2021].

[2] With a slight abuse of notation, we write $\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D}) = \hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})$.

3.2 Feasible Variationally Optimal Estimator $\hat{J}_{\mathrm{oIPS}}$

In this section, we derive a *feasible* variationally optimal estimator $\hat{J}_{\mathrm{oIPS}}$. Specifically, we employ the logger-wise cross-fitting strategy [Zheng and van der Laan, 2011, Chernozhukov et al., 2018, Kallus et al., 2021]: one part of the data is used to estimate the nuisance parameter $\boldsymbol{\alpha}$ by estimating $\mathbf{T}$ and $\mathbf{c}$, while the other part is used to compute the estimator (3.4).

Computing $\hat{J}_{\mathrm{oIPS}}$. For each $i \in [K]$, let $\{\mathcal{D}_i^{(z)}\}_{z\in[Z]}$ be a Z-fold partition ($Z \geq 2$) of $\mathcal{D}_i$ into nonempty folds chosen independently of the observed values, with $\mathcal{D}_i^{(-z)} \coloneqq \mathcal{D}_i \setminus \mathcal{D}_i^{(z)}$ and $n_i^{(-z)} \coloneqq |\mathcal{D}_i^{(-z)}|$. Then, for each fold $z \in [Z]$, estimate $\mathbf{T}, \mathbf{c}$ by

$$\hat{T}_{ij}^{(z)} \coloneqq \frac{1}{n_i^{(-z)}} \sum_{(S,A,R)\in\mathcal{D}_i^{(-z)}} \frac{\pi_j(A \mid S)}{\pi_{\mathrm{avg}}^{(-z)}(A \mid S)}, \quad \hat{c}_i^{(z)} \coloneqq \frac{1}{n_i^{(-z)}} \sum_{(S,A,R)\in\mathcal{D}_i^{(-z)}} \frac{R\,\pi_e(A \mid S)}{\pi_{\mathrm{avg}}^{(-z)}(A \mid S)} \tag{3.5}$$

where $\pi_{\mathrm{avg}}^{(-z)} \coloneqq \frac{1}{n^{(-z)}} \sum_{i\in[K]} n_i^{(-z)} \pi_i$ with $n^{(-z)} \coloneqq \sum_{i\in[K]} n_i^{(-z)}$. Then, we estimate $\boldsymbol{\alpha}^\star$ with $\hat{\boldsymbol{\alpha}}^{(z)} \coloneqq \big(\hat{\mathbf{T}}^{(z)}\big)^\dagger \hat{\mathbf{c}}^{(z)}$, where $\dagger$ denotes the Moore–Penrose pseudoinverse.

Let $n_i^{(z)} \coloneqq |\mathcal{D}_i^{(z)}|$ and $n^{(z)} \coloneqq \sum_{i\in[K]} n_i^{(z)}$. Then, the feasible variationally optimal estimator is

$$\boxed{\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D}) \coloneqq \frac{1}{n} \sum_{z=1}^{Z} n^{(z)} \hat{J}\big(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)}\big)} \tag{3.6}$$

where $\hat{J}\left(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)}\right)$ is (3.4) evaluated with $\hat{\boldsymbol{\alpha}}^{(z)}$ and $\mathcal{D}^{(z)} \coloneqq \{\mathcal{D}_i^{(z)}\}_{i\in[K]}$ in place of $\boldsymbol{\alpha}^\star$ and $\mathcal{D}$. Note that the quantity π_{avg} in (3.4) depends on $\mathcal{D}$ through n_i/n, so it is also replaced with $\pi_{\mathrm{avg}}^{(z)} \coloneqq \frac{1}{n^{(z)}} \sum_{i\in[K]} n_i^{(z)} \pi_i$.

When Logging Policies are Unknown. When the logging policies are unknown, we plug consistent estimators of $\pi_i(a|s)$ and $\pi_{\mathrm{avg}}(a|s)$ into the above procedure. The estimation procedures are given in Appendix A.

In Theorem 4.2, we show that $\hat{J}_{\mathrm{oIPS}}$ is unbiased and asymptotically normal with the optimal asymptotic variance $n\,\mathrm{Var}[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})]$ when the logging policies are known; Remark 2 discusses estimated logging policies.

4 Theoretical Analysis

In this section, we analyze $\hat{J}_{\mathrm{oIPS}}$ introduced in §3 and prove its asymptotic variance optimality within the generalized weighted-IPS class (3.1). We then compare $\hat{J}_{\mathrm{oIPS}}$ with other estimators, including the estimator $\hat{J}_{\text{DR-bIPS}}$ (2.4), which is asymptotically optimal within a class containing the three IPS baselines [Kallus et al., 2021] but requires an estimator of the conditional mean reward. We characterize the variance gap between $\hat{J}_{\mathrm{oIPS}}$ and $\hat{J}_{\text{DR-bIPS}}$ in terms of the residual from projecting the evaluation policy π_e, weighted by the conditional mean reward $q(s,a)$, namely $q\,\pi_e$, onto the logger basis space.

For simplicity of the analysis, we make the following assumptions.

Assumption 1 (Regularity)**.**

(i) All logging and evaluation policies have strictly positive densities over the action space at every state in the support of p_s.[3]

(ii) The weighted scores $w_i(S,A,R)\,\pi_e(A \mid S)/\pi_i(A \mid S)$ have finite second moments under μ_{π_i}, and the optimal weights in (3.2) *are square-integrable.*

[3] Consequently, we drop the dominance condition in the variance minimization problem, but our solution still satisfies it.

(iii) For the feasible estimator, K and Z are fixed and the folds of each stratum have equal sizes up to rounding. For every $i \in [K]$ and $z \in [Z]$, the summand of (3.4) *computed with $\hat{\boldsymbol{\alpha}}^{(z)}$ converges in probability to the summand computed with $\boldsymbol{\alpha}^\star$ in $L^2(\mu_{\pi_i})$, and both summands have finite second moments, including over the randomness of the training folds.*

Remark 1. *If π_e has zeros, as for the deterministic evaluation policy in §5, Assumption 1(i) fails: the weights* (3.2) *are undefined where $\pi_e(a \mid s) = 0$, and the control-variate form* (3.4) *is no longer a member of Γ_{wIPS}. The control-variate form remains unbiased for every $\boldsymbol{\alpha}$, and any solution of $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ still minimizes its variance over $\boldsymbol{\alpha} \in \mathbb{R}^K$. The comparisons in §4.3 extend to such π_e by continuity, by applying them to $\pi_e^\epsilon := (1-\epsilon)\pi_e + \epsilon\pi_u$, where π_u is the uniform policy, and letting $\epsilon \to 0$.*

4.1 Derivation of Optimal Weights

Let $\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})$ be a generalized weighted-IPS estimator (3.1). Then, $\mathrm{Var}\big[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})\big]$ is a convex functional of $\mathbf{w}$, so the variance can be optimized directly over the feasible weights. The main convex program of interest is

$$\begin{aligned} \underset{\mathbf{w} \in (L^2)^K}{\text{minimize}} \quad & \mathrm{Var}\big[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})\big] \\ \text{subject to} \quad & \sum_{i=1}^K w_i(\omega) = \rho_r(\omega), \quad \forall \omega \in \mathrm{supp}\,(\mu_{\pi_e}), \end{aligned} \tag{4.1}$$

Theorem 4.1. *The weight vector $\mathbf{w}^\star$ given in* (3.2) *solves the variance minimization problem* (4.1)*.*

The proof of this theorem is given in Appendix E.2.

By Theorem 4.1, computing $\hat{J}_{\mathrm{oIPS}}$ only requires estimating $\mathbf{T}$ and $\mathbf{c}$, whose i-th rows are averages over stratum i, and then solving the $K \times K$ system $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ for each fold. Assuming each propensity evaluation takes constant time, accumulating the moments takes $O(nK)$ time and each solve takes $O(K^3)$, so Z-fold cross-fitting costs $O(Z(nK + K^3))$, which is $O(nK + K^3)$ for fixed Z. If the per-fold moments are accumulated in one pass over the data, the extra memory is $O(ZK^2)$, and a second pass evaluates the estimator; storing every cross-logger propensity instead takes $O(nK)$. These counts exclude fitting nuisance models such as propensity models. Appendix F reports the corresponding runtimes.

4.2 Unbiasedness and Asymptotic Variance Optimality of $\hat{J}_{\mathrm{oIPS}}$ within Γ_{wIPS}

We next analyze the feasible cross-fitted estimator $\hat{J}_{\mathrm{oIPS}}$. Specifically, we prove its unbiasedness and asymptotic variance optimality within the generalized weighted-IPS estimator class Γ_{wIPS}. We follow the multi-logger asymptotic regime of Kallus et al. [2021]: for each logger, the number of observations is $n_i' = mn_i$, and we consider the limit $m \to \infty$.

Theorem 4.2. *Let $n' := \sum_{i=1}^K n_i'$ with $n_i' := mn_i$. Under Assumption 1, when the logging policies are known, $\hat{J}_{\mathrm{oIPS}}$ is unbiased, and*

$$\sqrt{n'}\left(\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D}) - J(\pi_e)\right) \xrightarrow{d} \mathcal{N}\left(0, n\,\mathrm{Var}[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})]\right)$$

as $m \to \infty$, where $n := \sum_{i=1}^K n_i$ and $\mathrm{Var}[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})]$ is computed at the base stratum sizes $(n_i)_{i=1}^K$.

That is, $\hat{J}_{\mathrm{oIPS}}$ is consistent and asymptotically normal, with the optimal asymptotic variance $n\,\mathrm{Var}[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})]$ achievable within the generalized weighted-IPS estimator class Γ_{wIPS}. The proof of Theorem 4.2 is given in Appendix E.3.

Remark 2 (Estimated logging policies)**.** *When the logging policies are unknown, we plug in estimates fitted on the training folds. Under mild regularity conditions, $\hat{J}_{\mathrm{oIPS}}$ remains consistent if these estimates are consistent, but its limiting distribution generally differs from that in Theorem 4.2: the score of $\hat{J}_{\mathrm{oIPS}}$ is not orthogonal to the propensities, so propensity estimation error enters at first order, as for IPS with estimated propensities [Hirano et al., 2003]. The limit in Theorem 4.2 still holds if the resulting fold-wise bias is $o_p(n^{-1/2})$, for example when $q\,\pi_e \in \mathrm{span}(\{\pi_i\}_{i \in [K]})$.*

4.3 Theoretical Comparison with Other Estimators

We provide a theoretical comparison of the estimators in terms of their asymptotic variance.

Comparison with $\hat{J}_{\mathrm{IPS}}$, $\hat{J}_{\mathrm{bIPS}}$ and $\hat{J}_{\mathrm{wIPS}}$. Γ_{wIPS} contains the naive IPS estimator ($\hat{J}_{\mathrm{IPS}}$), the balanced IPS estimator ($\hat{J}_{\mathrm{bIPS}}$), and the weighted IPS estimator ($\hat{J}_{\mathrm{wIPS}}$). Setting the weights in (3.1) to

$$w_i^{\mathrm{IPS}} = \frac{n_i}{n} r, \qquad w_i^{\mathrm{bIPS}} = \frac{n_i}{n} \frac{\pi_i}{\pi_{\mathrm{avg}}} r, \qquad w_i^{\mathrm{wIPS}} = n_i \hat{\gamma}_i r$$

recovers (2.1), (2.2), and (2.3), respectively. Each choice satisfies the normalization constraint because $\sum_i n_i/n = 1$, $\sum_i (n_i/n)\pi_i/\pi_{\mathrm{avg}} = 1$, and $\sum_i n_i \hat{\gamma}_i = 1$. The same identities hold for the cross-fitted version of $\hat{J}_{\mathrm{wIPS}}$, conditionally on the training folds, with the held-out stratum sizes in place of n_i. Hence, with known propensities, our asymptotic efficiency result in Theorem 4.2 implies that, for every problem instance,

$$\mathrm{AVar}\big[\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D})\big] \le \mathrm{AVar}\big[\hat{J}_{\mathrm{x}}(\pi_e; \mathcal{D})\big], \quad \forall \mathrm{x} \in \big\{\mathrm{IPS}, \mathrm{bIPS}, \mathrm{wIPS}\big\}$$

where $\mathrm{AVar}[\hat{J}]$ denotes the variance of the limiting distribution of $\sqrt{n'}\,(\hat{J} - J(\pi_e))$. This resolves the dilemma between balanced and weighted IPS noted by Agarwal et al. [2017]: asymptotically, $\hat{J}_{\mathrm{oIPS}}$ is at least as efficient as both.

Comparison with $\hat{J}_{\mathrm{DR\text{-}bIPS}}$. Rewriting our infeasible optimal estimator $\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})$ (3.4), we have

$$\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R) \in \mathcal{D}_i} \left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} \left(R - \frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e}(S, A) \right) + \frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e}(S, \pi_e) \right], \tag{4.2}$$

where $g_{\boldsymbol{\alpha}^\star} \coloneqq \sum_{i \in [K]} \alpha_i^\star \pi_i$ and $\frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e}(s, \pi_e) \coloneqq \int_{\mathcal{A}} \frac{g_{\boldsymbol{\alpha}^\star}(s,a)}{\pi_e(a|s)} \pi_e(a|s)\, d\mu_a(a) = \sum_{i \in [K]} \alpha_i^\star$.

Thus, our estimator belongs to the class of estimators studied by Kallus et al. [2021]:

$$\hat{J}_\Gamma(\pi_e, h, f; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R) \in \mathcal{D}_i} \big[h(i, S, A)\, \pi_e(A \mid S) \big(R - f(S, A)\big) + f(S, \pi_e) \big],$$

where h satisfies $\sum_i (n_i/n)\, \pi_i(a \mid s)\, h(i, s, a) = 1$, which makes $\hat{J}_\Gamma$ unbiased. The optimal variance is achieved by $h(i, s, a) = 1/\pi_{\mathrm{avg}}(a \mid s)$ for all $i \in [K]$ and $f(s, a) = q(s, a)$, with $q(s, a)$ denoting the conditional mean reward. This holds even though our weights w_i may depend on (s, a, r) while h depends only on (i, s, a): the representation (4.2) corresponds to $h = 1/\pi_{\mathrm{avg}}$ and $f = g_{\boldsymbol{\alpha}^\star}/\pi_e$.

The variance gap between our infeasible estimator $\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})$, which is optimal within Γ_{wIPS}, and the infeasible optimal estimator $\hat{J}_\Gamma^\star(\pi_e; \mathcal{D}) \coloneqq \hat{J}_\Gamma(\pi_e, 1/\pi_{\mathrm{avg}}, q; \mathcal{D})$ in this class is characterized in terms of the residual from the orthogonal projection of $q\,\pi_e$ onto the finite-dimensional space spanned by the logging policies, $\mathrm{span}\big(\{\pi_i\}_{i \in [K]}\big)$. The precise statement is given below, with the proof deferred to Appendix E.4.

Theorem 4.3. *The variance gap between $\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})$ and the semiparametrically efficient oracle estimator $\hat{J}_\Gamma^\star(\pi_e; \mathcal{D}) \coloneqq \hat{J}_\Gamma(\pi_e, 1/\pi_{\mathrm{avg}}, q; \mathcal{D})$ is given by*

$$\mathrm{Var}\big[\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})\big] - \mathrm{Var}\big[\hat{J}_\Gamma^\star(\pi_e; \mathcal{D})\big] = \frac{1}{n} \left(\inf_{\boldsymbol{\alpha} \in \mathbb{R}^K} \big\| q\,\pi_e - g_{\boldsymbol{\alpha}} \big\|_{\mu_{\log}}^2 - \mathrm{Var}_{p_s}\big[q(S, \pi_e)\big] \right) \ge 0,$$

where $\| \cdot \|_{\mu_{\log}}$ is the norm induced by the inner product $\langle \cdot, \cdot \rangle_{\mu_{\log}}$, with $d\mu_{\log} \coloneqq p_s/\pi_{\mathrm{avg}}\, d\mu_{s,a}$. Moreover, equality holds whenever $q\,\pi_e \in \mathrm{span}\big(\{\pi_i\}_{i \in [K]}\big)$.

Kallus et al. [2021] showed that, for any consistent estimator $\hat{q}$ of the conditional mean reward q, under their regularity conditions, $\hat{J}_{\mathrm{DR\text{-}bIPS}}$ (2.4) achieves the optimal asymptotic variance $n\mathrm{Var}\big[\hat{J}_\Gamma^\star(\pi_e; \mathcal{D})\big]$.

Combining this result with Theorem 4.2, we obtain

$$\mathrm{AVar}\big[\hat{J}_{\text{DR-bIPS}}(\pi_e;\mathcal{D})\big] \le \mathrm{AVar}\big[\hat{J}_{\text{oIPS}}(\pi_e;\mathcal{D})\big].$$

We close this gap in Appendix B with a doubly robust extension that uses the same reward model: it applies the control-variate form (3.4) to the residual $R - \hat{q}(S,A)$ and adds the direct-method term $\hat{q}(S,\pi_e)$. We show that, with known logging policies, $\hat{J}_{\text{DR-oIPS}}$ has asymptotic variance no larger than that of $\hat{J}_{\text{DR-bIPS}}$. Moreover, suppose that both estimators use the same misspecified reward model, $\hat{q} \to q' \neq q$. If $(q-q')\pi_e \in \mathrm{span}\{\pi_i\}_{i=1}^K$, then $\hat{J}_{\text{DR-oIPS}}$ still attains the oracle asymptotic variance $n\mathrm{Var}\big[\hat{J}_\Gamma^\star(\pi_e;\mathcal{D})\big]$, whereas $\hat{J}_{\text{DR-bIPS}}$ in general does not. The reason is that $\hat{J}_{\text{DR-oIPS}}$ projects $(q-q')\pi_e$ onto $\mathrm{span}\{\pi_i\}_{i=1}^K$ in $L^2(\mu_{\log})$ and effectively uses the corrected reward model $q' + g^\Delta/\pi_e$, where g^Δ denotes this projection; only the component of $(q-q')\pi_e$ orthogonal to the span contributes excess variance.

5 Numerical Experiments

5.1 Benchmark Datasets and Experiment Setup

Following prior multi-logger OPE protocols [Agarwal et al., 2017, Kallus et al., 2021], we evaluate on four UCI datasets (Table 1): *OptDigits* [Alpaydin and Kaynak, 1998], *SatImage* [Srinivasan, 1993], *PenDigits* [Alpaydin and Alimoglu, 1996], and *Letter* [Slate, 1991]. Each classification dataset is converted into a contextual bandit dataset by treating labels as actions and assigning reward 1 for a correct action and 0 otherwise.

Table 1: Dataset Statistics

Dataset Name	**OptDigits**	**SatImage**	**PenDigits**	**Letter**
#Classes (l)	10	6	10	26
#Samples	5620	6435	10992	20000

Mapping Classification Problems to Contextual Bandits. In deterministic-label multi-class classification, samples take the form $(S, y(S))$, where $S \sim p_s$ and $y(\cdot)$ is the ground-truth labeling function. We view this as a contextual bandit problem by treating the feature distribution p_s as the context distribution and defining the deterministic reward as $r(s,a) \coloneqq \mathbb{1}[y(s) = a]$.

To construct the evaluation policy, we train a multinomial logistic classifier on 30% of each dataset and define the deterministic policy $\pi_{\text{det}}(a \mid s) \coloneqq \mathbb{1}[a = \hat{y}(s)]$. We set the evaluation policy to $\pi_e = \pi_{\text{det}}$. Because π_e is deterministic, the weights (3.2) are undefined where $\pi_e(a \mid s) = 0$, so we compute $\hat{J}_{\text{oIPS}}$ through the control-variate form (3.4), which does not divide by π_e (Remark 1). The remaining 70% of the data is used as an evaluation pool. On this pool, the empirical accuracy of π_{det} is treated as the ground-truth policy value $J(\pi_e)$.

We generate logged bandit data from the evaluation pool using two logging policies,

$$\pi_1 = 0.95 \cdot \pi_{\text{det}} + 0.05 \cdot \pi_u, \qquad \pi_2 = 0.05 \cdot \pi_{\text{det}} + 0.95 \cdot \pi_u,$$

where π_u is the policy that chooses an action uniformly at random for any context. Thus, π_1 is well aligned with π_e, whereas π_2 is much less aligned with π_e. For each context in the evaluation pool, we assign it to either $\mathcal{D}_1$ or $\mathcal{D}_2$, sample an action from the corresponding logging policy, and observe the deterministic reward $r(s,a)$. We vary the stratum-size ratio over $n_1/n_2 \in \{0.1, 0.25, 0.5, 1, 2, 4, 10\}$. Smaller ratios make the logged dataset more dominated by π_2. Here the two loggers explore uniformly with probabilities 0.05 and 0.95 (the weights on π_u above). Appendix D uses $(0.20, 0.80)$ and $(0.35, 0.65)$, and Appendix C uses $(0.05, 0.50, 0.95)$ for $K = 3$ and $(0.05, 0.15, 0.30, 0.70, 0.95)$ for $K = 5$.

We benchmark each estimator $\hat{J}$ against the ground-truth value $J(\pi_e)$ using relative root mean squared error. For M independent trials with ground-truth values J_m and $\bar{J} = M^{-1}\sum_m J_m$, this metric is

$$\text{Relative-RMSE}(\hat{J}) \coloneqq \frac{1}{\bar{J}}\sqrt{\frac{1}{M}\sum_{m=1}^{M}\Big(J_m - \hat{J}_m\Big)^2}, \tag{5.1}$$

where $\hat{J}_m$ denotes the estimated policy value in the m-th trial. To separate bias from variance, we also report the squared bias $\big(M^{-1}\sum_{m=1}^M(\hat{J}_m - J_m)\big)^2$ and the empirical standard deviation

of $\hat{J}_1, \ldots, \hat{J}_M$ (Appendix A). We use $M = 200$ trials and report 95% paired percentile-bootstrap intervals. Each of the $B = 2{,}000$ bootstrap resamples draws M trial indices with replacement and recomputes Relative-RMSE for every estimator from the same indices and the matching ground-truth values J_m; the interval endpoints are the 2.5th and 97.5th percentiles. Because each trial redraws the data split, the logged actions, and the cross-fitting folds, the intervals reflect all three sources of variation. To compare two estimators, we apply the same procedure to the difference of their Relative-RMSEs computed on the same resampled trials.

Estimators Considered. We compare $\hat{J}_{\mathrm{IPS}}$ (Eq. 2.1), $\hat{J}_{\mathrm{bIPS}}$ (Eq. 2.2), $\hat{J}_{\mathrm{wIPS}}$ (Eq. 2.3), our $\hat{J}_{\mathrm{oIPS}}$ (§3), $\hat{J}_{\text{DR-bIPS}}$ [Kallus et al., 2021], and our $\hat{J}_{\text{DR-oIPS}}$ (Appendix B). The two DR estimators use the same reward model $\hat{q}$, a logistic regression fitted with two-fold cross-fitting, and the same folds: for both, every reward, propensity, and weight quantity is fitted on the training fold and evaluated on the same held-out observations. In all experiments, we compute $\hat{\boldsymbol{\alpha}}^{(z)}$ with NumPy's Moore–Penrose pseudoinverse (`numpy.linalg.pinv`, default relative cutoff 10^{-15}). In the designs with $K \geq 3$ loggers (Appendices C and F), every logger is a mixture of the same two policies, so $\mathbf{T}$ has rank two and the pseudoinverse returns the minimum-norm least-squares solution whenever $\hat{\mathbf{T}}^{(z)}$ is singular. Appendix A details estimation of unknown logging policies.

The experimental results are reported and discussed in Appendix A. In summary, $\hat{J}_{\mathrm{oIPS}}$ generally improves over existing multi-logger IPS baselines across benchmark datasets and logger mixtures, and the DR extension $\hat{J}_{\text{DR-oIPS}}$ achieves lower Relative-RMSE than $\hat{J}_{\text{DR-bIPS}}$ in many finite-sample settings. In a synthetic experiment with $n = 10^6$ samples (Appendix F.3), $\hat{J}_{\mathrm{oIPS}}$ attains a Relative-RMSE of 1.99×10^{-4}, compared with 5.44×10^{-4} for $\hat{J}_{\mathrm{wIPS}}$. Code to reproduce the experiments is included in the supplementary material and will be publicly released.

6 Conclusion

We studied off-policy evaluation with multiple logging policies, where existing IPS estimators can have highly instance-dependent performance. We addressed this by deriving $\hat{J}_{\mathrm{oIPS}}$, an asymptotically variance-minimizing estimator within the generalized weighted-IPS class Γ_{wIPS}, which contains previously considered IPS estimators. The optimal weights have a closed-form characterization through the linear system $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$, whose entries can be estimated from logged data, leading to a practical cross-fitted estimator. This gives a principled way to choose sample-dependent IPS weights rather than selecting among fixed rules such as naive IPS, balanced IPS, and weighted IPS.

We also related $\hat{J}_{\mathrm{oIPS}}$ to the semiparametric efficiency framework of Kallus et al. [2021]. The linear system $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ has a projection interpretation: it chooses the best control variate representable by the logging policies. When this space is rich enough, $\hat{J}_{\mathrm{oIPS}}$ asymptotically recovers the efficient oracle. Otherwise, the projection residual characterizes the variance gap. We further extended the estimator to a doubly robust version, $\hat{J}_{\text{DR-oIPS}}$, which matches the asymptotic variance of the efficient DR baseline when the reward model is correctly specified and can be more robust under certain structured misspecifications.

Empirically, $\hat{J}_{\mathrm{oIPS}}$ improves over existing multi-logger IPS baselines across benchmark datasets and logger mixtures, and $\hat{J}_{\text{DR-oIPS}}$ often improves over the DR balanced IPS baseline in finite samples. The main limitation is that our optimality result is restricted to Γ_{wIPS} and our asymptotic guarantees assume known logging policies; with estimated propensities, the limiting variance generally changes (Remark 2). We also do not analyze finite-sample behavior or the stability of the estimated weights, the method relies on sufficient overlap, and our experiments cover UCI benchmarks and one synthetic environment. Future work includes finite-sample analysis, richer basis classes beyond logging policies, and extensions to sequential settings.

Acknowledgments and Disclosure of Funding

This research was supported by the Deep-Tech TIPS Program (Grant No. RS-2024-0046898) funded by the Ministry of SMEs and Startups, Republic of Korea.

References

A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*, pages 687–696, 2017.

E. Alpaydin and F. Alimoglu. Pen-based recognition of handwritten digits. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5MG6K.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50P49.

I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Unified PAC-Bayesian study of pessimism for offline policy learning with regularized importance sampling. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

X. Chen, L. Wang, Y. Hang, H. Ge, and H. Zha. Infinite-horizon off-policy policy evaluation with multiple behavior policies. In *International Conference on Learning Representations*, 2020.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018.

J.-M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29:485—511, 2014.

V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

T. Hatt, D. Tschernutter, and S. Feuerriegel. Generalizing off-policy learning under sample selection bias. In *Uncertainty in Artificial Intelligence*, pages 769–779. PMLR, 2022.

L. He, L. Xia, W. Zeng, Z.-M. Ma, Y. Zhao, and D. Yin. Off-policy learning for multiple loggers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1184–1193, 2019.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

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

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Y. Jung and A. Bellot. Efficient policy evaluation across multiple different experimental datasets. *Advances in Neural Information Processing Systems*, 37:136361–136392, 2024.

N. Kallus, Y. Saito, and M. Uehara. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pages 5247–5256. PMLR, 2021.

I. Kondapaneni, P. Vévoda, P. Grittmann, T. Skřivan, P. Slusallek, and J. Křivánek. Optimal multiple importance sampling. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

I. Kuzborskij, C. Vernade, A. Gyorgy, and C. Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 640–648. PMLR, 2021.

S. D. Liu, C. Chen, and S. Zhang. Efficient multi-policy evaluation for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18951–18959, 2025.

A. M. Metelli, A. Russo, and M. Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in neural information processing systems*, 34: 8119–8132, 2021.

A. Owen and Y. Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

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

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Y. Saito, S. Aihara, M. Matsutani, and Y. Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

D. Slate. Letter recognition. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C5ZP40.

A. Srinivasan. Statlog (Landsat satellite). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C55887.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.

E. Veach and L. J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the annual conference on Computer graphics and interactive techniques*, pages 419–428, 1995.

J. M. Wooldridge. Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric theory*, 17(2):451–470, 2001.

W. Zheng and M. J. van der Laan. *Cross-Validated Targeted Minimum-Loss-Based Estimation*, pages 459–474. Springer New York, 2011.

A Main Experimental Results and Discussion

This appendix reports the full experimental results for the setup described in §5. Figures 1 and 2 show Relative-RMSE on *Letter*, *OptDigits*, *PenDigits*, and *SatImage* across the stratum-size ratios $n_1/n_2 \in \{0.1, 0.25, 0.5, 1, 2, 4, 10\}$. Each point is computed from $M = 200$ independent trials, and our estimator $\hat{J}_{\mathrm{oIPS}}$ uses $Z = 5$ folds.

When Logging Policies are Unknown. For non-DR estimators with unknown logging policies, we estimate each logging policy $\hat{\pi}_i$ separately using 2-fold cross-fitted multinomial logistic regression within each logger. The out-of-fold models are used to estimate the observed-action propensities $\hat{\pi}_i(A_{ij} \mid S_{ij})$ for samples in stratum i. We also fit a model on all samples from stratum i and use it to evaluate $\hat{\pi}_i(A_{kj} \mid S_{kj})$ for samples from all strata, which is needed when an estimator requires the propensity of a sample under a different logging policy.

For numerical stability, we blend each predicted action distribution with a 1% uniform distribution and clip all propensities below 10^{-3}. For degenerate strata, such as strata with too few samples or only one observed action class, we instead use the uniform policy. The resulting estimated logging policies are then plugged into IPS-style estimators.

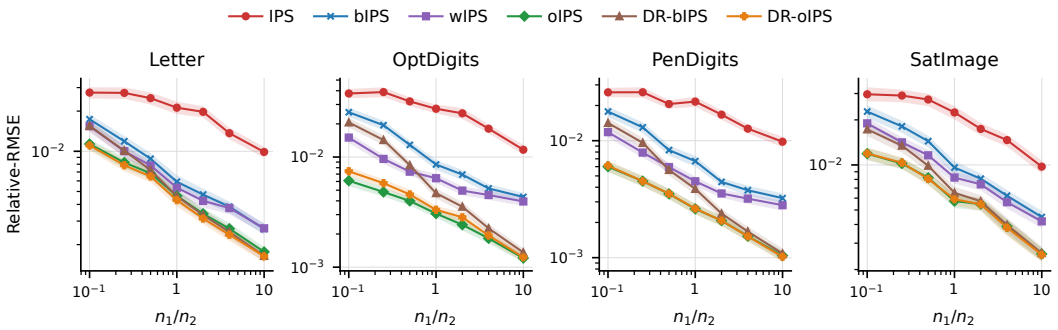


Figure 1: Relative-RMSEs (5.1), computed over $M = 200$ independent experiments, across datasets and estimators when the logging policies are **known**. Our proposed estimator $\hat{J}_{\mathrm{oIPS}}$ is implemented with $Z = 5$-fold cross-fitting. The x-axis denotes the stratum-size ratio (n_1/n_2) on a logarithmic scale, while the y-axis shows Relative-RMSE, also on a logarithmic scale. Lower values indicate better performance. Shaded bands show 95% paired percentile-bootstrap intervals (§5).

Known Logging Policies. When logging policies are known (Figure 1), $\hat{J}_{\mathrm{oIPS}}$ consistently improves over non-DR IPS baselines across datasets and stratum-size ratios. In several regimes, it is also competitive with or better than $\hat{J}_{\mathrm{DR\text{-}bIPS}}$. Adding a reward model ($\hat{J}_{\mathrm{DR\text{-}oIPS}}$) improves only slightly on $\hat{J}_{\mathrm{oIPS}}$, which suggests that little residual structure remains after the control-variate correction. These results support the practical value of the control-variate form (3.4), which uses the logger-basis structure to reduce variance without requiring a conditional reward estimator. By contrast, $\hat{J}_{\mathrm{IPS}}$ and $\hat{J}_{\mathrm{bIPS}}$ often show larger errors, especially at small n_1/n_2, where the logged data are dominated by the less aligned logger π_2.

Estimated Logging Policies. When logging policies are estimated (Figure 2), the same qualitative trend persists: $\hat{J}_{\mathrm{oIPS}}$ remains strong and generally outperforms non-DR baselines ($\hat{J}_{\mathrm{IPS}}$, $\hat{J}_{\mathrm{bIPS}}$, $\hat{J}_{\mathrm{wIPS}}$) across datasets and most stratum-size ratios. This is consistent with sample-dependent, cross-fitted weighting through $\hat{\mathbf{T}}$ and $\hat{\mathbf{c}}$, which stabilizes stratum balancing under propensity-estimation noise. In contrast, $\hat{J}_{\mathrm{wIPS}}$ can underperform $\hat{J}_{\mathrm{bIPS}}$ because its adaptive mixture weights depend on plug-in divergence estimates that are themselves sensitive to estimation noise.

Relative to the DR baseline $\hat{J}_{\mathrm{DR\text{-}bIPS}}$, $\hat{J}_{\mathrm{oIPS}}$ is not uniformly better when logging policies are estimated; this regime is not covered by Theorem 4.2 (Remark 2). However, the natural DR extension $\hat{J}_{\mathrm{DR\text{-}oIPS}}$ achieves lower Relative-RMSE than $\hat{J}_{\mathrm{DR\text{-}bIPS}}$ in many settings, particularly at low stratum-size ratios where the data are dominated by the less aligned logger. Subtracting the reward model

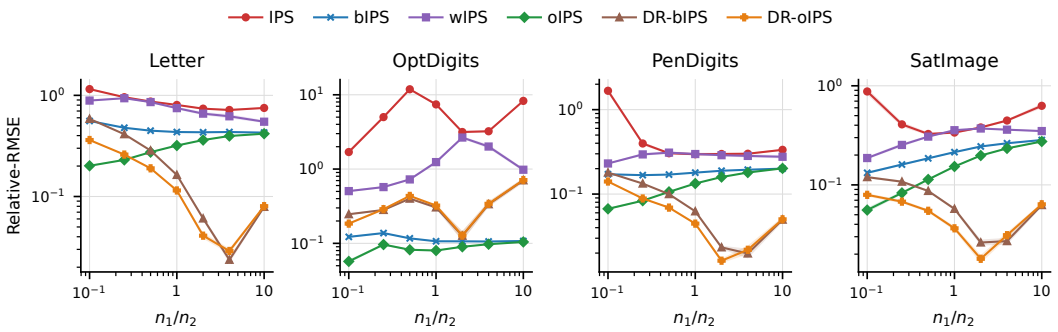


Figure 2: Relative-RMSEs (5.1), computed over $M = 200$ independent experiments, across datasets and estimators when the logging policies are **estimated**. Shaded bands show 95% paired percentile-bootstrap intervals (§5). Where bands are not visible, they are narrower than the markers.

first shrinks the inverse-propensity-weighted term $\pi_e(R - \hat{q})/\pi_{\text{avg}}$, and $\hat{J}_{\text{DR-oIPS}}$ then applies the control-variate correction to this smaller residual, which can improve accuracy in finite samples.

Summary. Overall, $\hat{J}_{\text{oIPS}}$ is a strong practical estimator for multi-logger OPE, especially against IPS baselines, and its DR extension $\hat{J}_{\text{DR-oIPS}}$ is often strongest when propensities are estimated. Appendices C and D show that these trends persist with more than two loggers and under alternative mixture settings. Across all settings, $\hat{J}_{\text{oIPS}}$ has a lower Relative-RMSE than each non-DR baseline in 276 of 276 configuration–baseline comparisons with known propensities and in 262 of 276 with estimated propensities. $\hat{J}_{\text{DR-oIPS}}$ has a lower Relative-RMSE than $\hat{J}_{\text{DR-bIPS}}$ in 92 of 92 and 66 of 92 configurations, respectively, and the paired 95% interval for the difference lies below zero in 89 and 64 of them.

Table 2 separates the error of each estimator into squared bias and empirical standard deviation at three representative stratum-size ratios. The squared bias is at most 1.5% of the variance in every cell, so the differences in Relative-RMSE come mainly from variance, and the smallest standard deviation is attained by $\hat{J}_{\text{oIPS}}$ or $\hat{J}_{\text{DR-oIPS}}$ in all twelve dataset–ratio pairs.

B Doubly Robust Extension of $\hat{J}_{\text{oIPS}}$

In this section, we present a doubly robust (DR) extension of our variationally optimal estimator $\hat{J}_{\text{oIPS}}$. Given any conditional reward estimator $\hat{q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we show how to incorporate $\hat{q}$ to obtain a DR version of $\hat{J}_{\text{oIPS}}$, denoted by $\hat{J}_{\text{DR-oIPS}}$. We then compare $\hat{J}_{\text{DR-oIPS}}$ both theoretically and empirically with the semiparametrically efficient estimator $\hat{J}_{\text{DR-bIPS}}$.

B.1 Generalized Weighted DR-IPS Estimator Class Γ_{DRwIPS}

We define the generalized weighted DR-IPS estimator class Γ_{DRwIPS} as follows. Let $\hat{q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be an estimator of the conditional mean reward $q(s, a) := \mathbb{E}[R \mid S = s, A = a]$, and define

$$\hat{J}_{\text{DM}}(\pi_e, \hat{q}; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R) \in \mathcal{D}_i} \int_{\mathcal{A}} \hat{q}(S, a)\, \pi_e(a \mid S)\, d\mu_a(a)$$

as the direct value estimator induced by $\hat{q}$.

Then, the estimator class Γ_{DRwIPS} consists of estimators of the following form:

$$\hat{J}(\pi_e, \mathbf{w}, \hat{q}; \mathcal{D}) := \hat{J}_{\text{DM}}(\pi_e, \hat{q}; \mathcal{D}) + \sum_{i=1}^{K} \frac{1}{n_i} \sum_{(S,A,R) \in \mathcal{D}_i} w_i(S, A, R) \cdot \frac{\pi_e(A \mid S)}{\pi_i(A \mid S)}. \tag{B.1}$$

where $\Delta(s, a, r) := r - \hat{q}(s, a)$ is the residual. The estimator is unbiased when $\hat{q}$ is a fixed function, or conditionally on the training data when $\hat{q}$ is fitted on data independent of $\mathcal{D}$, provided the weights satisfy the following constraints:

Table 2: Squared bias (top) and empirical standard deviation (bottom) of each estimator over $M = 200$ trials at three representative stratum-size ratios, with known logging policies ($\eta = 0.95$, $K = 2$).

n_1/n_2	Dataset	IPS	bIPS	wIPS	oIPS	DR-bIPS	DR-oIPS
10^{-1}	Letter	1.365e-6 0.021194	2.426e-7 0.013643	4.199e-9 0.012154	2.064e-7 0.009162	9.958e-8 0.012274	2.174e-7 0.008961
	OptDigits	9.845e-6 0.036553	6.086e-6 0.025019	1.955e-6 0.015421	6.030e-8 0.007236	3.771e-6 0.020381	1.661e-8 0.008195
	PenDigits	3.306e-6 0.024435	1.867e-6 0.016799	4.806e-7 0.011213	1.850e-7 0.006118	1.739e-6 0.013490	3.690e-7 0.006194
	SatImage	1.315e-7 0.023837	1.240e-7 0.018533	1.246e-7 0.015773	9.588e-8 0.010424	2.831e-7 0.014409	1.638e-7 0.010491
1	Letter	2.891e-9 0.016254	1.825e-8 0.005653	2.700e-8 0.005392	6.420e-9 0.004901	2.165e-8 0.004846	1.414e-8 0.004680
	OptDigits	3.120e-6 0.026398	5.313e-8 0.009119	7.139e-8 0.007522	5.952e-9 0.004654	1.070e-9 0.005835	2.516e-8 0.004749
	PenDigits	6.577e-7 0.020048	2.949e-8 0.006382	5.040e-10 0.004676	1.553e-8 0.003435	2.238e-10 0.004076	2.223e-8 0.003374
	SatImage	8.208e-8 0.018483	9.300e-8 0.009005	9.909e-8 0.008035	4.448e-10 0.006581	7.469e-9 0.007054	9.303e-10 0.006691
10	Letter	8.981e-7 0.007988	1.555e-10 0.004009	7.371e-9 0.004053	2.392e-10 0.003561	8.685e-10 0.003556	6.415e-10 0.003532
	OptDigits	1.723e-9 0.011590	8.947e-8 0.005328	1.261e-7 0.005066	1.026e-9 0.003627	3.153e-9 0.003636	7.132e-10 0.003605
	PenDigits	5.843e-7 0.009378	1.267e-8 0.003771	9.083e-13 0.003493	5.156e-10 0.002592	1.545e-9 0.002580	4.439e-10 0.002573
	SatImage	3.424e-7 0.008730	5.540e-8 0.005562	2.232e-8 0.005446	1.047e-8 0.005163	8.144e-9 0.005223	5.826e-9 0.005246

(i) **Normalization:** $\sum_{i=1}^{K} w_i(s,a,r) = \Delta(s,a,r)$,

(ii) **Dominance:** $\pi_i(a|s) = 0 \implies w_i(s,a,r) = 0$.

Intuitively, this form follows from the decomposition

$$\mathbb{E}_{\mu_{\pi_e}}[R] = \mathbb{E}_{\mu_{\pi_e}}[R - \hat{q}] + \mathbb{E}_{\mu_{\pi_e}}[\hat{q}].$$

That is, we apply weighted IPS to the residual term and use the direct method for the second term.

B.2 Feasible Extension: $\hat{J}_{\text{DR-oIPS}}$

We seek to minimize the variance of (B.1). The variational approach used for $\hat{J}_{\text{oIPS}}$ applies directly here, with the target variable r replaced by the residual $\Delta(s,a,r) \coloneqq r - \hat{q}(s,a)$.

Following the same cross-fitting procedure described in §3 for constructing the feasible estimator $\hat{J}_{\text{oIPS}}$, we obtain a feasible doubly robust extension, denoted $\hat{J}_{\text{DR-oIPS}}$. For each fold z, let $\hat{q}^{(z)}$ denote the conditional reward estimator trained on $\mathcal{D}^{(-z)}$. On the held-out fold $\mathcal{D}^{(z)}$, define the fold-specific direct-method term by

$$\hat{J}_{\text{DM}}\left(\pi_e, \hat{q}^{(z)}; \mathcal{D}^{(z)}\right) \coloneqq \frac{1}{n^{(z)}} \sum_{i=1}^{K} \sum_{(S,A,R)\in\mathcal{D}_i^{(z)}} \int_{\mathcal{A}} \hat{q}^{(z)}(S,a)\, \pi_e(a \mid S)\, d\mu_a(a).$$

The residual IPS component is obtained by applying the control-variate form (3.4) to the residuals $R - \hat{q}^{(z)}(S,A)$. Specifically, let

$$\hat{J}_{\Delta}\left(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)}\right)$$
$$\coloneqq \frac{1}{n^{(z)}} \sum_{i\in[K]} \sum_{(S,A,R)\in\mathcal{D}_i^{(z)}} \left[\frac{\big(R - \hat{q}^{(z)}(S,A)\big)\pi_e(A \mid S)}{\pi_{\text{avg}}^{(z)}(A \mid S)} - \frac{\sum_{k\in[K]} \hat{\alpha}_k^{(z)} \pi_k(A \mid S)}{\pi_{\text{avg}}^{(z)}(A \mid S)} + \sum_{k\in[K]} \hat{\alpha}_k^{(z)} \right],$$

where

$$\pi_{\text{avg}}^{(z)} \coloneqq \frac{1}{n^{(z)}} \sum_{i \in [K]} n_i^{(z)} \pi_i.$$

The feasible DR-oIPS estimator is then

$$\boxed{\hat{J}_{\text{DR-oIPS}}(\pi_e; \mathcal{D}) \coloneqq \frac{1}{n} \sum_{z=1}^{Z} n^{(z)} \left[\hat{J}_{\text{DM}} \left(\pi_e, \hat{q}^{(z)}; \mathcal{D}^{(z)} \right) + \hat{J}_{\Delta} \left(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)} \right) \right]} \tag{B.2}$$

Shared outer-fold procedure. Both DR estimators split each stratum into the same two folds ($Z = 2$). For each fold z:

1. On the training data $\mathcal{D}^{(-z)}$, fit the reward model $\hat{q}^{(z)}$, the logging-policy models when the propensities are unknown, and the moments $\hat{\mathbf{T}}^{(z)}$ and $\hat{\mathbf{c}}^{(z)}$ defined below, and set $\hat{\boldsymbol{\alpha}}^{(z)} = (\hat{\mathbf{T}}^{(z)})^{\dagger} \hat{\mathbf{c}}^{(z)}$.
2. Evaluate both estimators on the held-out data $\mathcal{D}^{(z)}$, using $\hat{\boldsymbol{\alpha}}^{(z)}$ for $\hat{J}_{\text{DR-oIPS}}$ and $\hat{\boldsymbol{\alpha}}^{(z)} = 0$ for $\hat{J}_{\text{DR-bIPS}}$.

The fold estimates are combined as in (B.2). With $\hat{\boldsymbol{\alpha}}^{(z)} = 0$, this yields the cross-fitted form of (2.4) with $\hat{q}^{(z)}$ and $\pi_{\text{avg}}^{(z)}$.

Estimation of Coefficients. As in $\hat{J}_{\text{oIPS}}$, the parameters $\hat{\boldsymbol{\alpha}}^{(z)}$ in (B.2) are obtained from $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ via cross-fitting. Specifically, $\hat{\boldsymbol{\alpha}}^{(z)} = (\hat{\mathbf{T}}^{(z)})^{\dagger} \hat{\mathbf{c}}^{(z)}$, where $\hat{\mathbf{T}}^{(z)}$ and $\hat{\mathbf{c}}^{(z)}$ are estimated on $\mathcal{D}^{(-z)}$ as follows.

(i) **Technique Matrix.** $\mathbf{T}$ depends only on propensities and sample sizes, so it is identical to the IPS case:

$$\hat{T}_{ij}^{(z)} \coloneqq \frac{1}{\left|\mathcal{D}_i^{(-z)}\right|} \sum_{(S,A,R) \in \mathcal{D}_i^{(-z)}} \frac{\pi_j(A \mid S)}{\pi_{\text{avg}}^{(-z)}(A \mid S)}.$$

(ii) **Contribution Vector.** The vector $\mathbf{c}$ depends on the target signal. For DR, we use residuals $R - \hat{q}^{(z)}(S, A)$ in place of R:

$$\hat{c}_i^{(z)} \coloneqq \frac{1}{\left|\mathcal{D}_i^{(-z)}\right|} \sum_{(S,A,R) \in \mathcal{D}_i^{(-z)}} \frac{\left(R - \hat{q}^{(z)}(S, A)\right) \pi_e(A \mid S)}{\pi_{\text{avg}}^{(-z)}(A \mid S)}.$$

Here, $\pi_{\text{avg}}^{(-z)} \coloneqq \frac{1}{n^{(-z)}} \sum_{i \in [K]} n_i^{(-z)} \pi_i$, with $n^{(-z)} \coloneqq \sum_{i \in [K]} n_i^{(-z)}$, as in the main text, §3.2. When the logging policies are unknown, we replace the propensities in the above expressions by their cross-fitted estimates.

B.3 Theoretical Comparison with the Optimal Estimator of Kallus et al. [2021]

In this section, we provide a theoretical comparison between $\hat{J}_{\text{DR-oIPS}}$ and $\hat{J}_{\text{DR-bIPS}}$. The purpose is to explain how the residual version of our variational construction relates to the semiparametrically efficient estimator studied in Kallus et al. [2021].

The comparison has two implications.

(i) When $\hat{q}$ is a consistent estimator of the conditional mean reward $q(s, a)$, both $\hat{J}_{\text{DR-oIPS}}$ and $\hat{J}_{\text{DR-bIPS}}$ attain the **same** semiparametrically efficient oracle variance asymptotically.

(ii) When $\hat{q}$ is **misspecified**, $\hat{J}_{\text{DR-oIPS}}$ can remain variance-optimal in cases where the misspecification lies in the logger-basis space, whereas $\hat{J}_{\text{DR-bIPS}}$ need not.

With known logging policies and under the conditions of Theorem 4.2, these conclusions follow from the following lemma, proved in Appendix E.

Lemma E.3 (Variance Gap)**.** *Let f be measurable and such that the summands of the estimator below have finite second moments under each μ_{π_i}, and let $\hat{J}_\Gamma(\pi_e, f; \mathcal{D})$ be the estimator defined by*

$$\hat{J}_\Gamma(\pi_e, f; \mathcal{D}) \coloneqq \frac{1}{n}\sum_{i=1}^{K} \sum_{(S,A,R)\in\mathcal{D}_i} \left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(R - f(S,A)\big) + f(S,\pi_e)\right], \tag{E.5}$$

where $f(s,\pi_e) \coloneqq \int_{\mathcal{A}} f(s,a)\pi_e(a|s)\, d\mu_a(a)$.

Let $q(s,a) \coloneqq \mathbb{E}[R \,|\, S = s, A = a]$ denote the conditional mean reward, and define $m_i(f) \coloneqq \mathbb{E}_{\mu_{\pi_i}}[(q-f)\pi_e/\pi_{\mathrm{avg}}] - \mathbb{E}_{p_s}[(q-f)(S,\pi_e)]$. Then, the variance gap between $\hat{J}_\Gamma(\pi_e, f; \mathcal{D})$ and $\hat{J}^\star_\Gamma(\pi_e; \mathcal{D}) \coloneqq \hat{J}_\Gamma(\pi_e, q; \mathcal{D})$ satisfies

$$\begin{aligned}
&\mathrm{Var}\big[\hat{J}_\Gamma(\pi_e, f; \mathcal{D})\big] - \mathrm{Var}\big[\hat{J}^\star_\Gamma(\pi_e; \mathcal{D})\big] \\
&\quad= \frac{1}{n}\left(\mathrm{Var}_{\mu_{\pi_{\mathrm{avg}}}}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big)\right] - \mathrm{Var}_{p_s}\big[q(S,\pi_e) - f(S,\pi_e)\big]\right) \\
&\qquad- \frac{1}{n}\sum_i \frac{n_i}{n} m_i(f)^2 \geq 0.
\end{aligned}$$

Below, the *asymptotic excess variance* of an estimator is its AVar minus $n\mathrm{Var}[\hat{J}^\star_\Gamma(\pi_e; \mathcal{D})]$, which equals n times the gap in Lemma E.3 evaluated at the limiting f.

Consistent reward estimation. We first explain how $\hat{J}_{\text{DR-oIPS}}$ fits into the $\hat{J}_\Gamma$ class. For fixed $\hat{q}$ and $\boldsymbol{\alpha}$, the infeasible version of $\hat{J}_{\text{DR-oIPS}}$ has summand

$$\hat{q}(S,\pi_e) + \frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(R - \hat{q}(S,A)\big) - \frac{g_{\boldsymbol{\alpha}}(S,A)}{\pi_{\mathrm{avg}}(A \mid S)} + \sum_{i\in[K]} \alpha_i.$$

Since

$$\sum_{i\in[K]} \alpha_i = \frac{g_{\boldsymbol{\alpha}}}{\pi_e}(S,\pi_e),$$

the last two terms simply replace $\hat{q}$ by $\hat{q} + g_{\boldsymbol{\alpha}}/\pi_e$ in the DR regression function. Indeed, the above summand is equal to

$$\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\left(R - \left(\hat{q}(S,A) + \frac{g_{\boldsymbol{\alpha}}(S,A)}{\pi_e(A \mid S)}\right)\right) + \left(\hat{q} + \frac{g_{\boldsymbol{\alpha}}}{\pi_e}\right)(S,\pi_e).$$

Thus, the infeasible DR-oIPS estimator corresponds to

$$\hat{J}_\Gamma\left(\pi_e, \hat{q} + \frac{g_{\boldsymbol{\alpha}}}{\pi_e}; \mathcal{D}\right).$$

The role of $\boldsymbol{\alpha}$ is to choose $g_{\boldsymbol{\alpha}}$ as the projection of $(q-\hat{q})\pi_e$ onto $\mathrm{span}\big(\{\pi_i\}_{i\in[K]}\big)$. In other words, $\hat{J}_{\text{DR-oIPS}}$ corrects $\hat{q}$ only in directions that can be represented by the logging policies.

If $\hat{q}$ is consistent for q, then $(q-\hat{q})\pi_e$ converges to zero, so its projection $g_{\boldsymbol{\alpha}}$ also converges to zero. Therefore,

$$\hat{q} + \frac{g_{\boldsymbol{\alpha}}}{\pi_e} \to q.$$

Applying Lemma E.3 to the limiting regression function $f = q$ shows that the asymptotic excess variance of $\hat{J}_{\text{DR-oIPS}}$ over the oracle estimator $\hat{J}^\star_\Gamma(\pi_e; \mathcal{D}) \coloneqq \hat{J}_\Gamma(\pi_e, q; \mathcal{D})$ vanishes. By Kallus et al. [2021], the same is true for $\hat{J}_{\text{DR-bIPS}}$ when $\hat{q}$ is consistent for q. Hence, in the well-specified case, $\hat{J}_{\text{DR-oIPS}}$ and $\hat{J}_{\text{DR-bIPS}}$ have the same asymptotic variance, namely $n\mathrm{Var}\big[\hat{J}^\star_\Gamma(\pi_e; \mathcal{D})\big]$.

Misspecified reward estimation. Now suppose that $\hat{q}$ converges to a misspecified limit q'. Then $\hat{J}_{\text{DR-bIPS}}$ corresponds asymptotically to

$$\hat{J}_\Gamma(\pi_e, q'; \mathcal{D}),$$

so Lemma E.3 characterizes its asymptotic excess variance over the oracle by taking $f = q'$.

By contrast, $\hat{J}_{\text{DR-oIPS}}$ corresponds asymptotically to

$$\hat{J}_\Gamma\left(\pi_e, q' + \frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e}; \mathcal{D}\right),$$

where $g_{\boldsymbol{\alpha}^\star}$ is the projection of $(q - q')\pi_e$ onto $\operatorname{span}\big(\{\pi_i\}_{i\in[K]}\big)$. Therefore, compared with $\hat{J}_{\text{DR-bIPS}}$, $\hat{J}_{\text{DR-oIPS}}$ replaces the misspecified regression function q' by the corrected regression function

$$q' + \frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e}.$$

This removes the component of the misspecification that lies in the span of the logging policies. Lemma E.3 then implies that the asymptotic excess variance of $\hat{J}_{\text{DR-oIPS}}$ is no larger than that of $\hat{J}_{\text{DR-bIPS}}$ when both use the same limiting reward model q'. Indeed, let $Q_f \coloneqq \pi_e(A \mid S)\big(R - f(S,A)\big)/\pi_{\text{avg}}(A \mid S) + f(S, \pi_e)$ denote the summand of (E.5) and $g^\star \coloneqq g_{\boldsymbol{\alpha}^\star}$. Then $Q_{q'} = Q_{q'+g^\star/\pi_e} + g^\star/\pi_{\text{avg}} - \sum_k \alpha_k^\star$, $m_i(q' + g^\star/\pi_e) = 0$ for every i, and $\operatorname{Cov}_{\mu_{\pi_{\text{avg}}}}\big[Q_{q'+g^\star/\pi_e},\, g^\star/\pi_{\text{avg}}\big] = 0$, so $n\big(\operatorname{Var}[\hat{J}_\Gamma(\pi_e, q'; \mathcal{D})] - \operatorname{Var}[\hat{J}_\Gamma(\pi_e, q' + g^\star/\pi_e; \mathcal{D})]\big) = \sum_i \frac{n_i}{n} \operatorname{Var}_{\mu_{\pi_i}}[g^\star/\pi_{\text{avg}}] \geq 0$.

In particular, if

$$(q - q')\pi_e \in \operatorname{span}\big(\{\pi_i\}_{i\in[K]}\big),$$

then $g_{\boldsymbol{\alpha}^\star} = (q - q')\pi_e$, and hence

$$q' + \frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e} = q.$$

Thus, $\hat{J}_{\text{DR-oIPS}}$ attains the oracle asymptotic variance even though $\hat{q}$ is misspecified. In contrast, $\hat{J}_{\text{DR-bIPS}}$ remains tied to q' and may have nonzero asymptotic excess variance. This is the sense in which $\hat{J}_{\text{DR-oIPS}}$ is more robust to misspecification in directions spanned by the logging policies.

B.4 Empirical Comparison with the Optimal Estimator of Kallus et al. [2021]

We provide an alternative, side-by-side presentation of the experimental results from §5, in Figure 3 to more directly compare $\hat{J}_{\text{DR-oIPS}}$ with $\hat{J}_{\text{DR-bIPS}}$, which is known to be asymptotically optimal within the class studied by Kallus et al. [2021]. We also summarize the nuisance estimation procedures used in these experiments.

Nuisance Estimation. We summarize how nuisance components (logging policies and conditional rewards) are estimated.

- **Logging policies ($\hat{\pi}_i$):** We estimate logging policies with multinomial logistic regression and 2-fold cross-fitting. To maintain numerical stability in both the technique-matrix estimate $\hat{\mathbf{T}}$ and the denominator of Eq. (B.2), we clip the estimated propensities to the range $[10^{-3}, 1]$ after 1% uniform smoothing.
- **Conditional rewards ($\hat{q}$):** We estimate $q(s,a) \coloneqq \mathbb{E}[r|s,a]$ via logistic regression with 2-fold cross-fitting, using the same protocol as for logging policies. Actions are represented with one-hot encodings.

Findings. In the setting of Figure 3, $\hat{J}_{\text{DR-oIPS}}$ has a lower Relative-RMSE than $\hat{J}_{\text{DR-bIPS}}$ in 16 of 28 configurations. Across all 184 configurations (92 with known and 92 with estimated propensities), it has a lower Relative-RMSE in 158, and the paired 95% interval for the difference lies below zero in 153.

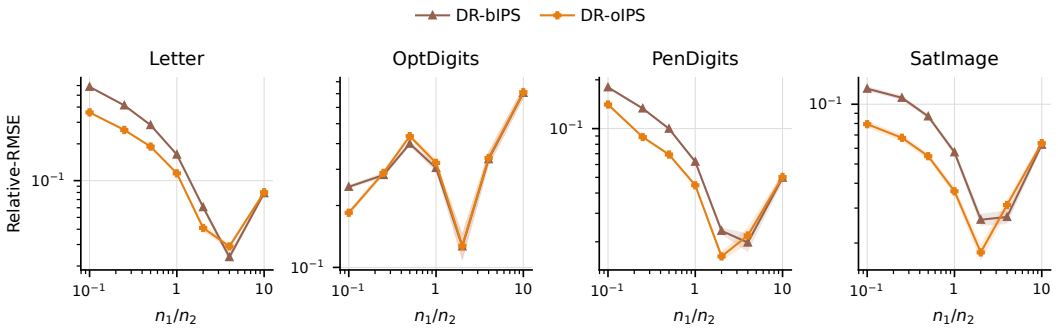


Figure 3: Relative-RMSEs, computed over $M = 200$ independent experiments, are reported across different datasets and estimators under the setting where the logging policies are **estimated**. Shaded bands show 95% paired percentile-bootstrap intervals (§5). Where bands are not visible, they are narrower than the markers.

C Empirical Results with $K > 2$ Logging Policies

We extend the two-logger experiment to heterogeneous mixtures with $K \in \{3, 5\}$ logging policies. As before, we train a multinomial logistic regression model on a training split to obtain a deterministic base policy π_{det} and set the evaluation policy to $\pi_e = \pi_{\text{det}}$. The reward is deterministic: $r(s, a) := \mathbb{1}[y(s) = a]$, where $y(s)$ is the label of context $s \in \mathbb{R}^d$. The ground-truth policy value is therefore the classification accuracy of π_e on the OPE sample.

$$J(\pi_e) = \mathbb{E}_{A\sim\pi_e(\cdot|S),S\sim p_s}\left[\mathbb{1}[y(S) = A]\right].$$

We construct $K \in \{3, 5\}$ logging policies by mixing π_{det} with the uniform policy π_u (which ignores context and selects actions uniformly). For $K = 5$, let $\beta_i \in \{0.05, 0.15, 0.30, 0.70, 0.95\}$ and define, for $i = 1, \ldots, K$,

$$\pi_i(a|s) := (1 - \beta_i) \cdot \pi_{\text{det}}(a|s) + \frac{\beta_i}{|\mathcal{A}|}.$$

The OPE sample is generated from an equal-weight logger mixture, with stratum weights $\tau_i = 0.2$ for $K = 5$ and $\tau_i = \frac{1}{3}$ for $K = 3$ (so $n_i/n \approx \tau_i$, up to integer rounding). We partition the OPE sample into these strata, sample $a \sim \pi_i(\cdot \mid s)$ in stratum i, and assign reward $r(s, a) := \mathbb{1}[y(s) = a]$. For $K = 3$, we use $\beta_i \in \{0.05, 0.5, 0.95\}$.

Results. Across settings with up to five loggers (Figure 4 and Figure 5), our logger-basis IPS estimator $\hat{J}_{\text{oIPS}}$ generally achieves strong accuracy when the logging policies are known, while its doubly robust extension $\hat{J}_{\text{DR-oIPS}}$ tends to perform best when the logging policies are estimated.

When logging policies are estimated, $\hat{J}_{\text{DR-oIPS}}$ outperforms $\hat{J}_{\text{DR-bIPS}}$ on all four datasets for $K = 5$ and three of four for $K = 3$. Among the non-DR estimators, $\hat{J}_{\text{IPS}}$ and $\hat{J}_{\text{wIPS}}$ can become unstable (particularly on OptDigits). This instability appears when estimated logging propensities are near zero or noisy in specific strata, yielding extreme $1/\hat{\pi}_i(a|s)$ weights and heavy-tailed variance. By comparison, the sample-dependent weighting in $\hat{J}_{\text{oIPS}}$ and $\hat{J}_{\text{DR-oIPS}}$ better controls variance and produces substantially more stable estimates. The same qualitative pattern holds for three loggers (Figures 6 and 7).

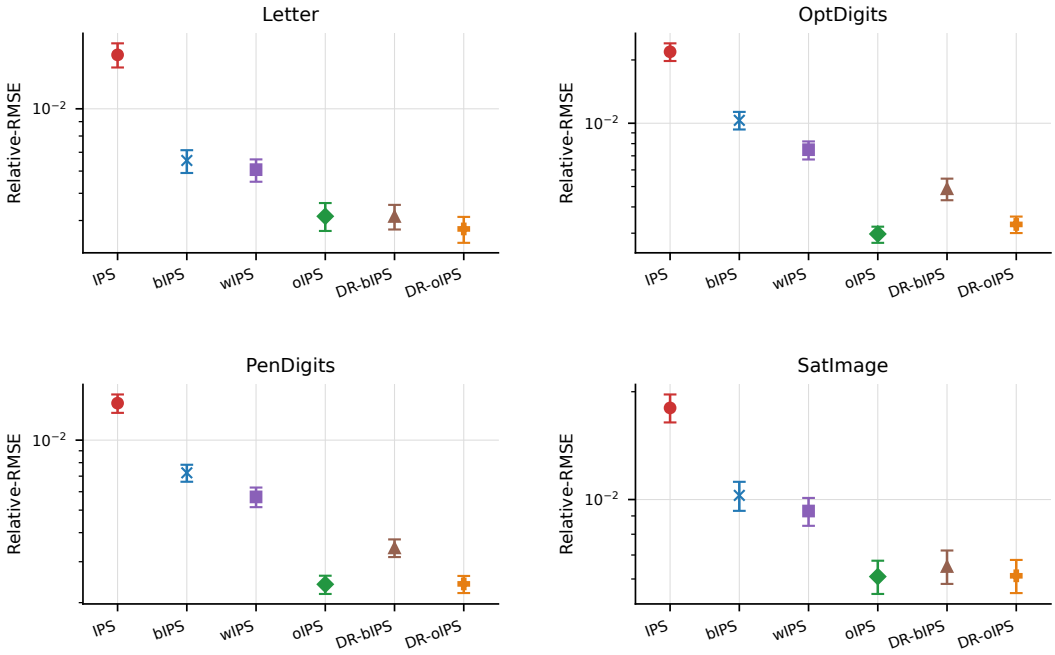


Figure 4: Rel-RMSE across five heterogeneous logging policies where they are known. Lower is better. Error bars show 95% paired percentile-bootstrap intervals (§5).

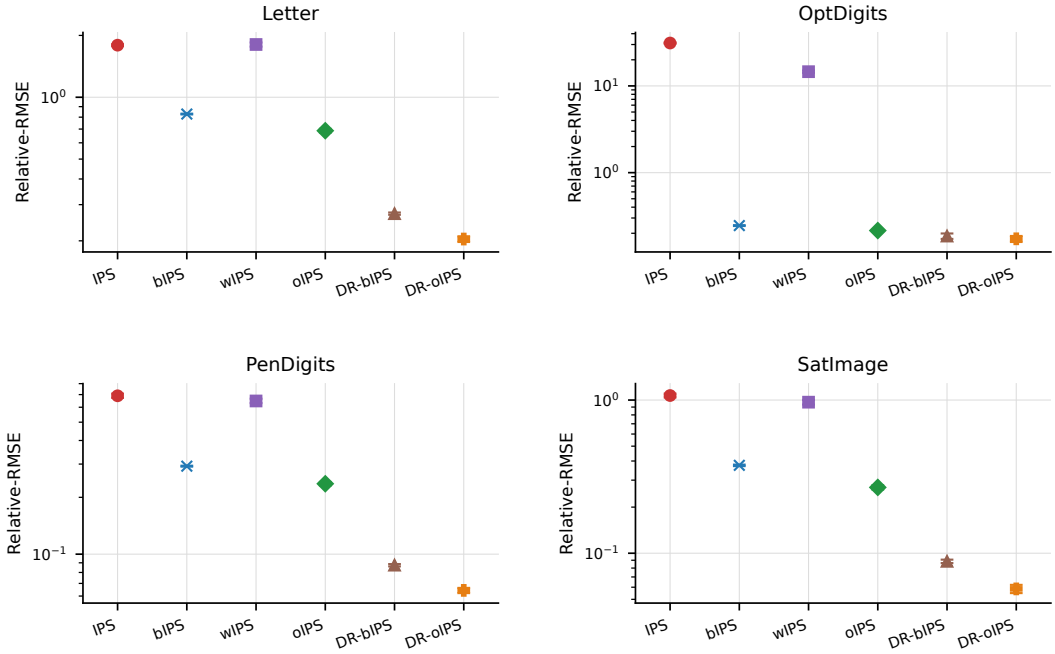


Figure 5: Rel-RMSE across five heterogeneous logging policies where they are estimated. Lower is better. Error bars show 95% paired percentile-bootstrap intervals (§5). Where error bars are not visible, they are shorter than the markers.

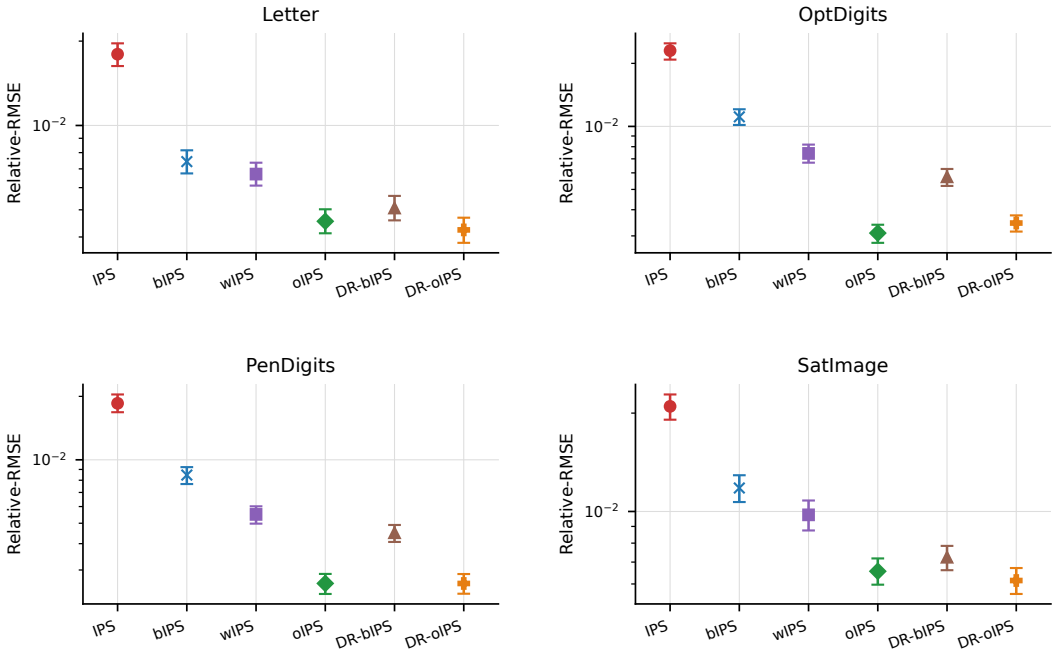


Figure 6: Rel-RMSE across three heterogeneous logging policies where they are known. Lower is better. Error bars show 95% paired percentile-bootstrap intervals (§5).

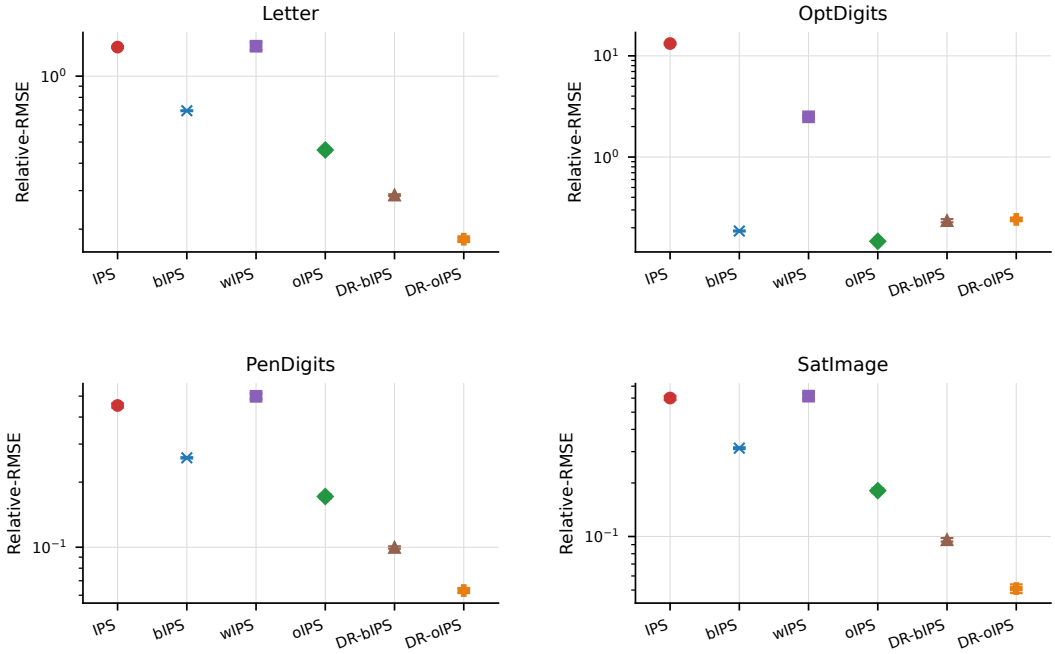


Figure 7: Rel-RMSE across three heterogeneous logging policies where they are estimated. Lower is better. Error bars show 95% paired percentile-bootstrap intervals (§5). Where error bars are not visible, they are shorter than the markers.

D Sensitivity to the Additive Mixture of Logging Policies

We repeat the main experiments with alternative mixture ratios between the deterministic policy π_{det} and the uniform policy π_u. For $\eta \in \{0.80, 0.65\}$, define

$$\pi_1^{(\eta)} \;=\; \eta\,\pi_{\mathrm{det}} + (1-\eta)\,\pi_u, \quad \pi_2^{(\eta)} \;=\; (1-\eta)\,\pi_{\mathrm{det}} + \eta\,\pi_u.$$

All other settings, including the evaluation policy $\pi_e = \pi_{\mathrm{det}}$, datasets, train/evaluation split ratio, stratum-size ratios n_1/n_2, number of repeats $M = 200$, and non-DR cross-fitting folds $Z = 5$, match the main-text protocol. Both DR estimators use the shared two-fold procedure of Appendix B.

Results. Across both additive mixtures ($\eta = 0.80$ and $\eta = 0.65$), the qualitative trends in Figures 8, 9, 10, and 11 remain consistent with the main-text results. Our logger-basis IPS estimator and its DR extension ($\hat{J}_{\mathrm{oIPS}}$ and $\hat{J}_{\mathrm{DR\text{-}oIPS}}$) remain among the strongest methods in known- and estimated-logging settings, respectively. Improvements are most pronounced in low-overlap regimes (small n_1/n_2), while gaps narrow as overlap increases. Estimated propensities increase absolute errors and can change which estimator ranks first.

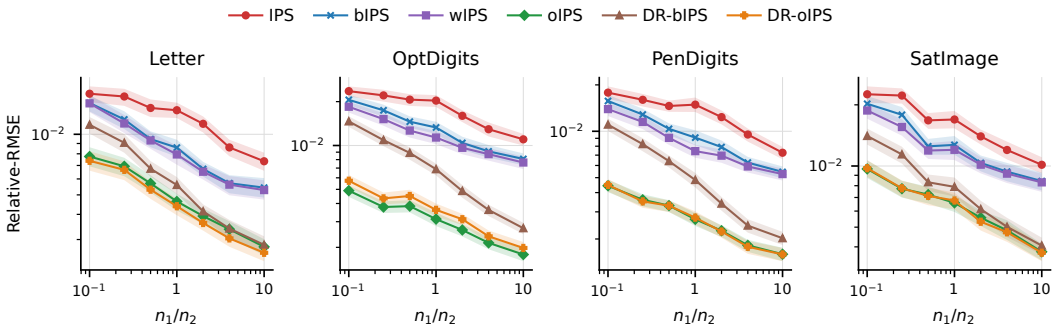


Figure 8: **Mixture** $\eta = 0.80$**.** Relative-RMSEs, computed over $M = 200$ independent experiments, are reported across different datasets and estimators under the setting where the logging policies are **known**. Shaded bands show 95% paired percentile-bootstrap intervals (§5).

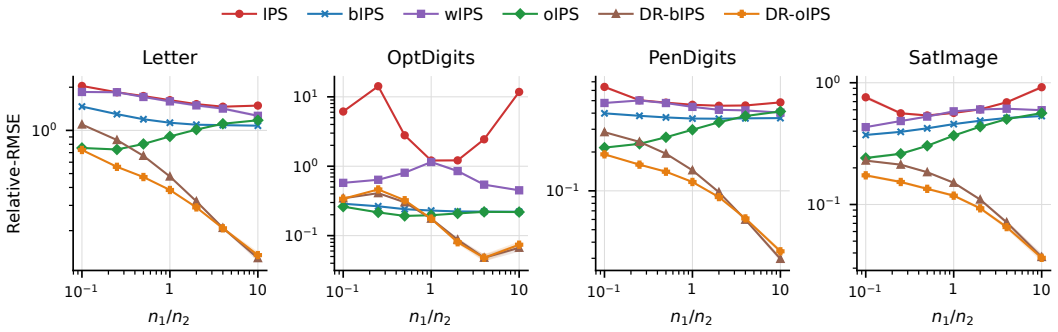


Figure 9: **Mixture** $\eta = 0.80$**.** Relative-RMSEs, computed over $M = 200$ independent experiments, are reported across different datasets and estimators under the setting where the logging policies are **estimated**. Shaded bands show 95% paired percentile-bootstrap intervals (§5). Where bands are not visible, they are narrower than the markers.

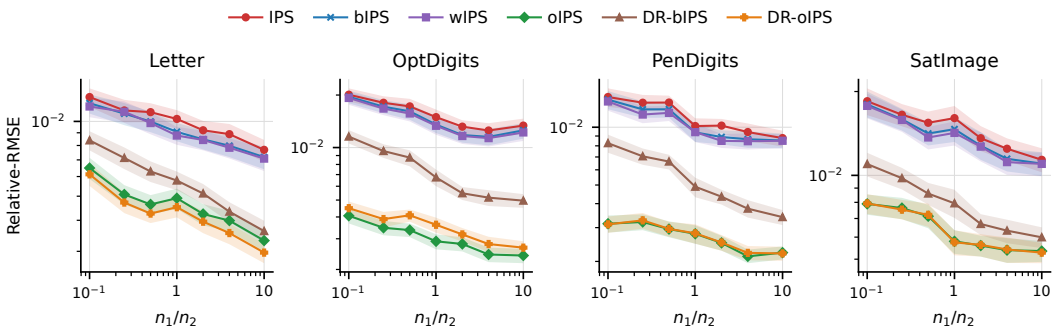


Figure 10: **Mixture** $\eta = 0.65$**.** Relative-RMSEs, computed over $M = 200$ independent experiments, are reported across different datasets and estimators under the setting where the logging policies are **known**. Shaded bands show 95% paired percentile-bootstrap intervals (§5).

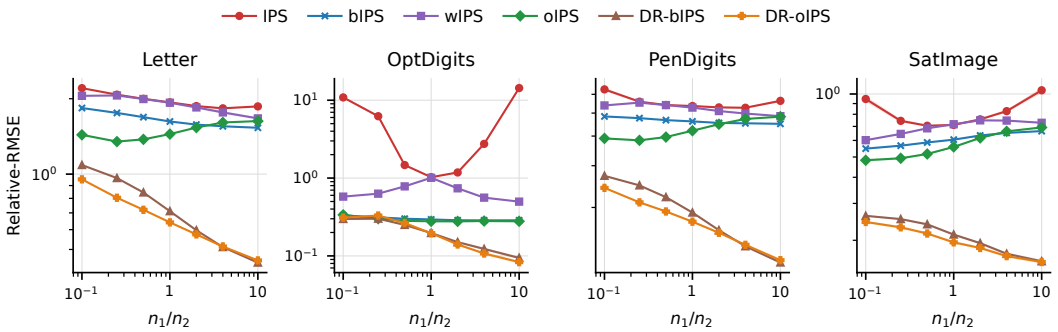


Figure 11: **Mixture** $\eta = 0.65$**.** Relative-RMSEs, computed over $M = 200$ independent experiments, are reported across different datasets and estimators under the setting where the logging policies are **estimated**. Shaded bands show 95% paired percentile-bootstrap intervals (§5). Where bands are not visible, they are narrower than the markers.

E Deferred Proofs

In this section, we present detailed proofs of the main theorems/lemmas.

E.1 Proof of Unbiasedness of Generalized Weighted-IPS Estimators

Theorem 3.1. [Unbiasedness] *For any weights* $\mathbf{w} \in (L^2)^K$ *that satisfy the normalization and dominance conditions and for which each weighted score* $w_i(S,A,R)\,\pi_e(A \mid S)/\pi_i(A \mid S)$ *is* μ_{π_i}*-integrable,* $\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})$ *is unbiased, that is,* $J(\pi_e) = \mathbb{E}[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})]$*, where the expectation is over the sampling randomness of logged data* $\mathcal{D} \coloneqq (\mathcal{D}_i)_{i\in[K]}$.

Proof of Theorem 3.1. Let $\omega \coloneqq (s,a,r) \in \Omega$ and μ_{π_i} be the probability distribution with density $p_r(r|s,a)\,\pi_i(a|s)\,p_s(s)$ with respect to μ. Recall the general weighted IPS estimator,

$$\hat{J}(\pi_e, \mathbf{w}; \mathcal{D}) \coloneqq \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{w_i(S_{ij}, A_{ij}, R_{ij})\,\pi_e(A_{ij} \mid S_{ij})}{\pi_i(A_{ij} \mid S_{ij})} \coloneqq \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{J}_{ij}(w_i; \mathcal{D}_i).$$

Taking the expectation over data,

$$\begin{aligned}
\mathbb{E}\big[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})\big] &= \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E}\big[\hat{J}_{ij}(w_i; \mathcal{D}_i)\big] \\
&= \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j=1}^{n_i} \int_{\mathrm{supp}(w_i)} w_i(\omega)\,\frac{\pi_e(a|s)}{\pi_i(a|s)}\,d\mu_{\pi_i}(\omega) \\
&\qquad\qquad \text{(as } \pi_i(a|s) = 0 \implies w_i(s,a,r) = 0\text{)} \\
&= \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j=1}^{n_i} \int_{\mathrm{supp}(w_i)} w_i(\omega)\,\frac{\pi_e(a|s)}{\pi_i(a|s)}\,p_r(r|s,a)\,\pi_i(a|s)\,p_s(s)\,d\mu(\omega) \\
&= \sum_{i=1}^{K} \int_{\mathrm{supp}(w_i)} w_i(\omega)\,p_r(r|s,a)\,\pi_e(a|s)\,p_s(s)\,d\mu(\omega) \\
&= \sum_{i=1}^{K} \int_{\Omega} w_i(\omega)\,d\mu_{\pi_e}(\omega) \\
&= \int_{\Omega} \left(\sum_{i=1}^{K} w_i(\omega)\right) d\mu_{\pi_e}(\omega) \\
&= \int_{\Omega} r\; d\mu_{\pi_e}(\omega) \qquad\qquad \text{(as } \textstyle\sum_{i=1}^{K} w_i(\omega) = r \text{ over } \omega \in \mathrm{supp}(\mu_{\pi_e})\text{)} \\
&= \mathbb{E}_{\mu_{\pi_e}}[R].
\end{aligned}$$

□

E.2 Proof of Theorem 4.1.

In this section, we prove Theorem 4.1. Let $\mathbf{w}^\star \coloneqq (w_i^\star)_{i\in[K]}$ be defined by

$$w_i^\star(s,a,r) = \frac{\alpha_i^\star \pi_i(a|s)}{\pi_e(a|s)} + \frac{n_i\,\pi_i(a|s)}{n\,\pi_{\mathrm{avg}}(a|s)} \left(r - \frac{\sum_{j=1}^{K} \alpha_j^\star\,\pi_j(a|s)}{\pi_e(a|s)}\right)$$

where $\boldsymbol{\alpha}^\star \coloneqq (\alpha_i^\star)_{i\in[K]}$ is a solution of $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$. Here, $\mathbf{T} \in \mathbb{R}^{K\times K}$ and $\mathbf{c} \in \mathbb{R}^K$ have components

$$T_{ij} \coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[\frac{\pi_j(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right], \quad c_i \coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[R\,\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right].$$

The theorem states that $\mathbf{w}^\star$ is an optimal solution to the variance minimization problem (4.1):

$$\begin{aligned} &\underset{\mathbf{w}\in(L^2)^K}{\text{minimize}} && \mathrm{Var}\big[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})\big] \\ &\text{subject to} && \sum_{i=1}^K w_i(\omega) = r, \quad \forall \omega \in \mathrm{supp}\,(\mu_{\pi_e}) \end{aligned}$$

Variance Characterization. We begin with a variance characterization of the generalized weighted-IPS estimator $\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})$ in (3.1). This characterization enables functional differentiation with respect to $\mathbf{w}$.

Lemma E.1. *For any weights $\mathbf{w} \in (L^2)^K$ whose weighted scores $w_i(S,A,R)\,\pi_e(A \mid S)/\pi_i(A \mid S)$ have finite second moments under μ_{π_i},*

$$\mathrm{Var}\big[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})\big] = \sum_{i=1}^K \frac{1}{n_i}\left(\left\langle w_i^2,\, \frac{\pi_e}{\pi_i}\right\rangle_{\mu_{\pi_e}} - \langle w_i,\, 1\rangle^2_{\mu_{\pi_e}}\right).$$

Proof of Lemma E.1. As $\hat{J}_{ij}(w_i; \mathcal{D}_i)$ are independent to each other,

$$\begin{aligned} \mathrm{Var}[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})] &= \mathrm{Var}\left[\sum_{i=1}^K \frac{1}{n_i}\sum_{j=1}^{n_i} \hat{J}_{ij}(w_i; \mathcal{D}_i)\right] = \sum_{i=1}^K \frac{1}{n_i^2}\sum_{j=1}^{n_i} \mathrm{Var}\left[\hat{J}_{ij}(w_i; \mathcal{D}_i)\right] \\ &= \sum_{i=1}^K \frac{1}{n_i^2}\sum_{j=1}^{n_i} \mathbb{E}\left[\hat{J}_{ij}^2(w_i; \mathcal{D}_i)\right] - \sum_{i=1}^K \frac{1}{n_i^2}\sum_{j=1}^{n_i} \mathbb{E}\left[\hat{J}_{ij}(w_i; \mathcal{D}_i)\right]^2. \end{aligned}$$

The first term is

$$\begin{aligned} \sum_{i=1}^K \frac{1}{n_i^2}\sum_{j=1}^{n_i} \mathbb{E}\left[\hat{J}_{ij}^2(w_i; \mathcal{D}_i)\right] &= \sum_{i=1}^K \frac{1}{n_i^2}\sum_{j=1}^{n_i} \int_\Omega \frac{w_i^2(\omega)\pi_e^2(a|s)}{\pi_i^2(a|s)}\, p_r(r|s,a)\,\pi_i(a|s)\,p_s(s)\,d\mu(\omega) \\ &= \sum_{i=1}^K \frac{1}{n_i}\int_\Omega \frac{w_i^2(\omega)\pi_e(a|s)}{\pi_i(a|s)}\, p_r(r|s,a)\,\pi_e(a|s)\,p_s(s)\,d\mu(\omega) \\ &= \sum_{i=1}^K \frac{1}{n_i}\int_\Omega w_i^2(\omega)\,\frac{\pi_e(a|s)}{\pi_i(a|s)}\,d\mu_{\pi_e}(\omega) \\ &= \sum_{i=1}^K \frac{1}{n_i}\left\langle w_i^2,\, \frac{\pi_e}{\pi_i}\right\rangle_{\mu_{\pi_e}}. \end{aligned}$$

The second term is

$$\begin{aligned} \sum_{i=1}^K \frac{1}{n_i^2}\sum_{j=1}^{n_i} \mathbb{E}\big[\hat{J}_{ij}(w_i; \mathcal{D}_i)\big]^2 &= \sum_{i=1}^K \frac{1}{n_i^2}\sum_{j=1}^{n_i} \left(\int_\Omega \frac{w_i(\omega)\pi_e(a|s)}{\pi_i(a|s)}\, p_r(r|s,a)\,\pi_i(a|s)\,p_s(s)\,d\mu(\omega)\right)^2 \\ &= \sum_{i=1}^K \frac{1}{n_i}\left(\int_\Omega w_i(\omega)\,p_r(r|s,a)\,\pi_e(a|s)\,p_s(s)\,d\mu(\omega)\right)^2 \\ &= \sum_{i=1}^K \frac{1}{n_i}\left(\int_\Omega w_i(\omega)\,d\mu_{\pi_e}(\omega)\right)^2 \\ &= \sum_{i=1}^K \frac{1}{n_i}\langle w_i, 1\rangle^2_{\mu_{\pi_e}}. \end{aligned}$$

□

Proof of Optimality of $\mathbf{w}^\star$. We now prove Theorem 4.1, which is restated below.

Theorem 4.1. *The weight vector $\mathbf{w}^\star$ given in* (3.2) *solves the variance minimization problem* (4.1).

Proof of Theorem 4.1. Since the objective in (4.1) is the variance of a linear functional of $\mathbf{w}$, the variance minimization problem subject to the normalization constraint is convex. We therefore derive the first-order stationarity condition for the Lagrange-augmented functional $I[\mathbf{w}, \lambda]$ defined below.

Recall $\rho_r(\omega) = r$ for $\omega = (s, a, r)$. Define the functional $I[\mathbf{w}, \lambda]$ as

$$\begin{aligned} I[\mathbf{w}, \lambda] &\coloneqq \mathrm{Var}\big[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})\big] - \left\langle \lambda, \sum_{i=1}^{K} w_i - \rho_r \right\rangle_{\mu_{\pi_e}} \\ &= \mathrm{Var}\big[\hat{J}(\pi_e, \mathbf{w}; \mathcal{D})\big] - \sum_{i=1}^{K} \langle \lambda, w_i \rangle_{\mu_{\pi_e}} + \langle \lambda, \rho_r \rangle_{\mu_{\pi_e}}, \end{aligned}$$

where $\lambda : \Omega \to \mathbb{R}$ is the Lagrange multiplier enforcing the normalization constraint.

Next, for a perturbation direction $\boldsymbol{\delta}(\omega) \coloneqq (\delta_i(\omega))_{i \in [K]}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^K$, define $I(\boldsymbol{\epsilon}; \boldsymbol{\delta}, \mathbf{w}, \lambda)$ as

$$I(\boldsymbol{\epsilon}; \boldsymbol{\delta}, \mathbf{w}, \lambda) \coloneqq I[w_1 + \epsilon_1 \delta_1, \ldots, w_K + \epsilon_K \delta_K, \lambda],$$

which can be interpreted as the variation of $I[\mathbf{w}, \lambda]$ under perturbations along the direction $\boldsymbol{\delta}$.

To verify optimality, it suffices to exhibit feasible $\mathbf{w}^\star$ and $\lambda^\star$ such that, for every perturbation $\boldsymbol{\delta}$ with finite weighted second moments and each $i \in [K]$,

$$\left. \frac{\partial I(\boldsymbol{\epsilon}; \boldsymbol{\delta}, \mathbf{w}^\star, \lambda^\star)}{\partial \epsilon_i} \right|_{\boldsymbol{\epsilon}=0} = 0.$$

By Lemma E.1, this is equivalent to

$$\left\langle \delta_i, \frac{2 w_i^\star \pi_e}{n_i \pi_i} - \frac{2 \langle w_i^\star, 1 \rangle_{\mu_{\pi_e}}}{n_i} - \lambda^\star \right\rangle_{\mu_{\pi_e}} = 0. \tag{E.1}$$

Next, we deduce $\mathbf{w}^\star$ and $\lambda^\star$ from (E.1) and verify that the deduced $\mathbf{w}^\star$ is indeed a minimizer.

To proceed, let us define $\alpha_i \coloneqq \langle w_i^\star, 1 \rangle_{\mu_{\pi_e}}$. Then, (E.1) implies that, for almost every $\omega \coloneqq (s, a, r) \in \mathrm{supp}(\mu_{\pi_e})$,

$$\frac{2 w_i^\star(\omega) \pi_e(a|s)}{n_i \pi_i(a|s)} - \frac{2 \alpha_i}{n_i} - \lambda^\star(\omega) = 0.$$

Rearranging the above and summing over $i \in [K]$, we obtain

$$\begin{aligned} w_i^\star(\omega) &= \frac{\alpha_i \pi_i(a|s)}{\pi_e(a|s)} + \frac{n_i \pi_i(a|s)}{2 \pi_e(a|s)} \lambda^\star(\omega) \qquad \text{(E.2)} \\ \implies r &= \sum_{i=1}^{K} w_i^\star(\omega) = \frac{\sum_{i=1}^{K} \alpha_i \pi_i(a|s)}{\pi_e(a|s)} + \frac{\sum_{i=1}^{K} n_i \pi_i(a|s)}{2 \pi_e(a|s)} \lambda^\star(\omega), \end{aligned}$$

where we used the fact that $\sum_{i=1}^{K} w_i^\star(\omega) = r$ for all $\omega \in \mathrm{supp}\,(\mu_{\pi_e})$.

Solving the above for $\lambda^\star(\omega)$,

$$\lambda^\star(\omega) = \frac{2r\, \pi_e(a|s) - 2 \sum_{i=1}^{K} \alpha_i \pi_i(a|s)}{\sum_{i=1}^{K} n_i \pi_i(a|s)}.$$

Plugging $\lambda^\star(\omega)$ back into (E.2) results in

$$\begin{aligned} w_i^\star(\omega) &= \frac{\alpha_i \pi_i(a|s)}{\pi_e(a|s)} + \frac{n_i\, \pi_i(a|s)}{\sum_{j=1}^{K} n_j \pi_j(a|s)} \left(r - \frac{\sum_{j=1}^{K} \alpha_j \pi_j(a|s)}{\pi_e(a|s)} \right), \\ &= \frac{\alpha_i \pi_i(a|s)}{\pi_e(a|s)} + \frac{n_i}{n} \frac{\pi_i(a|s)}{\pi_{\mathrm{avg}}(a|s)} \left(r - \frac{\sum_{j=1}^{K} \alpha_j \pi_j(a|s)}{\pi_e(a|s)} \right) \qquad \text{(E.3)} \end{aligned}$$

where one can easily verify that the omitted constraint "$\pi_i(a|s) = 0 \implies w_i(\omega) = 0$ for all $i \in [K]$" is satisfied for any $\boldsymbol{\alpha}$.

Next, we derive a self-consistency equation for $\boldsymbol{\alpha}$ by plugging the above expression for $w_i^\star(\omega)$ into $\alpha_i \coloneqq \langle w_i^\star, 1\rangle_{\mu_{\pi_e}}$,

$$\begin{aligned}\alpha_i = \alpha_i \, \mathbb{E}_{\mu_{\pi_e}}\left[\frac{\pi_i(A \mid S)}{\pi_e(A \mid S)}\right] + \frac{n_i}{n}\, \mathbb{E}_{\mu_{\pi_e}}\left[\frac{R\,\pi_i(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right] \\ - \frac{n_i}{n} \sum_{j=1}^{K} \alpha_j \, \mathbb{E}_{\mu_{\pi_e}}\left[\frac{\pi_i(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} \, \frac{\pi_j(A \mid S)}{\pi_e(A \mid S)}\right]. \end{aligned} \tag{E.4}$$

Since

$$\begin{aligned}\mathbb{E}_{\mu_{\pi_e}}\left[\frac{\pi_i(A \mid S)}{\pi_e(A \mid S)}\right] &= \int_\Omega \frac{\pi_i(a|s)}{\pi_e(a|s)}\, p_r(r|s,a)\, \pi_e(a|s)\, p_s(s)\, d\mu(\omega) \\ &= \int_\Omega p_r(r|s,a)\, \pi_i(a|s)\, p_s(s)\, d\mu(\omega) \\ &= \int_\Omega \mu_{\pi_i}(\omega) = 1,\end{aligned}$$

and

$$\begin{aligned}T_{ij} &\coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[\frac{\pi_j(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right] = \mathbb{E}_{\mu_{\pi_e}}\left[\frac{\pi_i(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\, \frac{\pi_j(A \mid S)}{\pi_e(A \mid S)}\right] \\ c_i &\coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[\frac{R\,\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right] = \mathbb{E}_{\mu_{\pi_e}}\left[\frac{R\,\pi_i(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right]\end{aligned}$$

The self-consistency equation (E.4) can be written as $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ with $\mathbf{T}$ and $\mathbf{c}$ given in (3.3).

Finally, by Lemma E.2, $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ admits a solution. Hence, the corresponding weights $\mathbf{w}$ satisfy the first-order optimality condition (E.1) and therefore solve the variance minimization convex program (4.1). □

E.3 Proof of Unbiasedness and Asymptotic Variance Optimality of $\hat{J}_{\mathrm{oIPS}}$

In this section, we prove that $\hat{J}_{\mathrm{oIPS}}$ is unbiased and asymptotically normal, with asymptotic variance $n\mathrm{Var}\big[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})\big]$, which is optimal within Γ_{wIPS}.

Theorem 4.2. *Let $n' \coloneqq \sum_{i=1}^K n_i'$ with $n_i' \coloneqq mn_i$. Under Assumption 1, when the logging policies are known, $\hat{J}_{\mathrm{oIPS}}$ is unbiased, and*

$$\sqrt{n'}\left(\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D}) - J(\pi_e)\right) \xrightarrow{d} \mathcal{N}\left(0, n\,\mathrm{Var}[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})]\right)$$

as $m \to \infty$, where $n \coloneqq \sum_{i=1}^K n_i$ and $\mathrm{Var}[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})]$ is computed at the base stratum sizes $(n_i)_{i=1}^K$.

Proof of Theorem 4.2. We first prove the unbiasedness of $\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D})$ when the logging policies are known.

Unbiasedness. We begin by rewriting the expectation of $\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D})$ as an iterated expectation, conditioning on the data $\mathcal{D}^{(-z)}$ used to estimate the nuisance parameter $\boldsymbol{\alpha}$:

$$\begin{aligned}\mathbb{E}\left[\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D})\right] &= \frac{1}{n}\sum_{z=1}^{Z} n^{(z)}\, \mathbb{E}\left[\hat{J}(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)})\right] \\ &= \frac{1}{n}\sum_{z=1}^{Z} n^{(z)}\, \mathbb{E}\left[\mathbb{E}\big[\hat{J}(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)}) \,\big|\, \mathcal{D}^{(-z)}\big]\right]\end{aligned}$$

It therefore suffices to show that

$$\mathbb{E}\big[\hat{J}(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)}) \big| \mathcal{D}^{(-z)}\big] = J(\pi_e),$$

for each fold z. To see this, let $\hat{\mathbf{w}}^{(z)}$ be defined by

$$\hat{w}_i^{(z)}(s,a,r) \coloneqq \frac{\hat{\alpha}_i^{(z)} \pi_i(a|s)}{\pi_e(a|s)} + \frac{n_i^{(z)} \pi_i(a|s)}{\sum_{j=1}^K n_j^{(z)} \pi_j(a|s)} \left(r - \frac{\sum_{k=1}^K \hat{\alpha}_k^{(z)} \pi_k(a|s)}{\pi_e(a|s)} \right)$$

Then, conditional on $\hat{\boldsymbol{\alpha}}^{(z)}$, these weights satisfy the normalization condition $\sum_{i=1}^K \hat{w}_i^{(z)}(s,a,r) = r$ and the dominance condition $\pi_i(a|s) = 0 \implies \hat{w}_i^{(z)}(s,a,r) = 0$.

Moreover, $\hat{J}(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)})$ in (3.4) can be rewritten in the canonical generalized weighted-IPS form (3.1), with $\hat{\mathbf{w}}^{(z)}$ as its weights:

$$\hat{J}(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)}) = \sum_{i=1}^K \frac{1}{n_i^{(z)}} \sum_{(S,A,R) \in \mathcal{D}_i^{(z)}} \hat{w}_i^{(z)}(S,A,R) \, \frac{\pi_e(A \mid S)}{\pi_i(A \mid S)}.$$

Since $\hat{\mathbf{w}}^{(z)}$ satisfies the normalization and dominance conditions conditional on $\hat{\boldsymbol{\alpha}}^{(z)}$, Theorem 3.1 implies that $\hat{J}(\pi_e, \hat{\boldsymbol{\alpha}}^{(z)}; \mathcal{D}^{(z)})$ is unbiased conditional on $\hat{\boldsymbol{\alpha}}^{(z)}$, equivalently conditional on $\mathcal{D}^{(-z)}$.

Asymptotic Normality. Next, we show that

$$\sqrt{n'} \left(\hat{J}_{\mathrm{oIPS}}(\pi_e; \mathcal{D}) - J(\pi_e) \right) \xrightarrow{d} \mathcal{N}\left(0, n \operatorname{Var}[\hat{J}(\pi_e, \mathbf{w}^\star; \mathcal{D})]\right).$$

We work with known logging policies under Assumption 1: K and Z are fixed, the logger-wise folds are balanced up to rounding, the fitted scores converge in probability to the oracle score in each $L^2(\mu_{\pi_i})$, and all fitted and oracle scores have finite second moments, including over training randomness.

We rewrite our infeasible optimal estimator $\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})$ in (3.4) in the form considered by Kallus et al. [2021]:

$$\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^K \sum_{(S,A,R) \in \mathcal{D}_i} \left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} \left(R - \frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e}(S,A) \right) + \frac{g_{\boldsymbol{\alpha}^\star}}{\pi_e}(S, \pi_e) \right].$$

The normal equation gives the oracle summand mean $J(\pi_e)$ under every logger. Grouping the oracle observations into m independent blocks, each containing n_i observations from logger i, the i.i.d. central limit theorem gives the displayed limit.

Conditional independence and the assumed L^2 score convergence make each fold's contribution to the centered error relative to the oracle, scaled by $\sqrt{n'}$, $o_p(1)$ by conditional Chebyshev. With known policies, the conditional bias is zero. Since Z is fixed and fold weights are sample proportions, summing the contributions and applying Slutsky's theorem gives the desired asymptotic normality result. If the logging policies are instead estimated by normalized policies fitted on the training folds, with positive mixtures, the same argument applies whenever each fold mean has conditional bias $o_p(n'^{-1/2})$; this is the condition in Remark 2, and it is not implied by $\sqrt{n}$-consistency of the estimated propensities. The same argument applies to DR scores under the corresponding moment, score-convergence, and bias conditions. □

E.4 Additional Proofs

Lemma E.2 (Normal-Equation Interpretation of $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$)**.** *For* $\mathbf{T}$ *and* $\mathbf{c}$ *defined in* (3.3)*, their entries can be written as*

$$T_{ij} = \big\langle \pi_i, \pi_j \big\rangle_{\mu_{\log}}, \quad c_i = \big\langle q\, \pi_e, \pi_i \big\rangle_{\mu_{\log}},$$

where $\mu_{\log}$ *is given by*

$$\frac{d\mu_{\log}}{d\mu_{s,a}}(s,a) \coloneqq \frac{p_s(s)}{\pi_{\mathrm{avg}}(s,a)},$$

with $\mu_{s,a}$ denoting the base measure over $\mathcal{S} \times \mathcal{A}$.

Consequently, $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ is the normal equation for the orthogonal projection of $q\,\pi_e$ onto $\mathcal{V} \coloneqq \operatorname{span}\big(\{\pi_i\}_{i\in[K]}\big)$. In particular, $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ admits a solution, and any solution $\boldsymbol{\alpha}^\star$ defines the orthogonal projection of $q\,\pi_e$ onto $\mathcal{V}$ through $\sum_{i\in[K]} \alpha_i^\star \pi_i$.

Proof of Lemma E.2. Recall the definitions of $\mathbf{T}$ and $\mathbf{c}$ in (3.3):

$$T_{ij} \coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[\frac{\pi_j(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right], \;\; c_i \coloneqq \mathbb{E}_{\mu_{\pi_i}}\left[R\,\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right].$$

The first identity follows by direct computation:

$$\begin{aligned}
T_{ij} &= \mathbb{E}_{\mu_{\pi_i}}\left[\frac{\pi_j(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right] \\
&= \int_\Omega \frac{\pi_i(a|s)\pi_j(a|s)}{\pi_{\mathrm{avg}}(a|s)}\, p_r(r|s,a)\, p_s(s)\, d\mu(\omega) \\
&= \int_{\mathcal{S}\times\mathcal{A}} \pi_i(a|s)\, \pi_j(a|s)\, \frac{p_s(s)}{\pi_{\mathrm{avg}}(a|s)}\, d\mu_{s,a}(s,a) \\
&= \big\langle \pi_i, \pi_j \big\rangle_{\mu_{\log}}.
\end{aligned}$$

Similarly, for c_i, we have

$$\begin{aligned}
c_i &= \mathbb{E}_{\mu_{\pi_i}}\left[R\,\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right] \\
&= \mathbb{E}_{\mu_{\pi_i}}\left[\mathbb{E}_{\mu_{\pi_i}}\left[R\,\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)} \,\middle|\, S, A\right]\right] \\
&= \mathbb{E}_{\mu_{\pi_i}}\left[\mathbb{E}_{\mu_{\pi_i}}[R \mid S, A]\;\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right] \\
&= \mathbb{E}_{\mu_{\pi_i}}\left[q(S,A)\,\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\right] \\
&= \int_\Omega q(s,a)\,\frac{\pi_e(a|s)\,\pi_i(a|s)}{\pi_{\mathrm{avg}}(a|s)}\, p_r(r|s,a)\, p_s(s)\, d\mu(\omega) \\
&= \int_{\mathcal{S}\times\mathcal{A}} q(s,a)\,\pi_e(a|s)\,\pi_i(a|s)\frac{p_s(s)}{\pi_{\mathrm{avg}}(a|s)}\, d\mu_{s,a}(s,a) \\
&= \big\langle q\,\pi_e, \pi_i \big\rangle_{\mu_{\log}}.
\end{aligned}$$

It remains to show that $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ is the normal equation for the orthogonal projection. Let

$$\mathcal{V} \coloneqq \operatorname{span}\big(\{\pi_i\}_{i\in[K]}\big) \subset L^2(\mu_{\log}).$$

Since $\mathcal{V}$ is finite-dimensional, it is closed. Hence, the orthogonal projection of $q\,\pi_e$ onto $\mathcal{V}$ exists. Write this projection as

$$g_{\boldsymbol{\alpha}^\star} \coloneqq \sum_{k\in[K]} \alpha_k^\star \pi_k.$$

By the characterization of orthogonal projections, $g_{\boldsymbol{\alpha}^\star}$ satisfies

$$\big\langle q\,\pi_e - g_{\boldsymbol{\alpha}^\star}, \pi_i \big\rangle_{\mu_{\log}} = 0, \quad \forall i \in [K].$$

Expanding this condition gives

$$\big\langle q\,\pi_e, \pi_i \big\rangle_{\mu_{\log}} = \sum_{k\in[K]} \alpha_k^\star \big\langle \pi_k, \pi_i \big\rangle_{\mu_{\log}}, \quad \forall i \in [K].$$

By the identities proved above, this system is exactly $\mathbf{T}\boldsymbol{\alpha}^\star = \mathbf{c}$. Therefore, $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ admits a solution, and any solution $\boldsymbol{\alpha}^\star$ defines the orthogonal projection of $q\,\pi_e$ onto $\mathcal{V}$ through $g_{\boldsymbol{\alpha}^\star} = \sum_{i\in[K]} \alpha_i^\star \pi_i$. □

Lemma E.3 (Variance Gap)**.** *Let f be measurable and such that the summands of the estimator below have finite second moments under each μ_{π_i}, and let $\hat{J}_\Gamma(\pi_e, f; \mathcal{D})$ be the estimator defined by*

$$\hat{J}_\Gamma(\pi_e, f; \mathcal{D}) \coloneqq \frac{1}{n}\sum_{i=1}^{K}\sum_{(S,A,R)\in\mathcal{D}_i}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(R - f(S,A)\big) + f(S,\pi_e)\right], \tag{E.5}$$

where $f(s,\pi_e) \coloneqq \int_{\mathcal{A}} f(s,a)\pi_e(a|s)\, d\mu_a(a)$.

Let $q(s,a) \coloneqq \mathbb{E}[R \,|\, S = s, A = a]$ denote the conditional mean reward, and define $m_i(f) \coloneqq \mathbb{E}_{\mu_{\pi_i}}[(q-f)\pi_e/\pi_{\mathrm{avg}}] - \mathbb{E}_{p_s}[(q-f)(S,\pi_e)]$. Then, the variance gap between $\hat{J}_\Gamma(\pi_e, f; \mathcal{D})$ and $\hat{J}^\star_\Gamma(\pi_e; \mathcal{D}) \coloneqq \hat{J}_\Gamma(\pi_e, q; \mathcal{D})$ satisfies

$$\begin{aligned}
&\mathrm{Var}\big[\hat{J}_\Gamma(\pi_e, f; \mathcal{D})\big] - \mathrm{Var}\big[\hat{J}^\star_\Gamma(\pi_e; \mathcal{D})\big] \\
&\quad= \frac{1}{n}\left(\mathrm{Var}_{\mu_{\pi_{\mathrm{avg}}}}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big)\right] - \mathrm{Var}_{p_s}\big[q(S,\pi_e) - f(S,\pi_e)\big]\right) \\
&\qquad- \frac{1}{n}\sum_i \frac{n_i}{n} m_i(f)^2 \geq 0.
\end{aligned}$$

Proof of Lemma E.3. Let $Q_f = \pi_e(R-f)/\pi_{\mathrm{avg}} + f(S,\pi_e)$ and $V(f) = \mathrm{Var}_{\mu_{\pi_{\mathrm{avg}}}}[Q_f]$. Independence across samples and the variance decomposition over loggers give

$$\mathrm{Var}[\hat{J}_\Gamma(\pi_e, f; \mathcal{D})] = \frac{1}{n^2}\sum_i n_i \mathrm{Var}_{\mu_{\pi_i}}[Q_f] = \frac{1}{n}\left(V(f) - \sum_i \frac{n_i}{n} m_i(f)^2\right).$$

Here $m_i(f) = \mathbb{E}_{\mu_{\pi_i}}[Q_f] - J(\pi_e)$, which agrees with the definition in the lemma, the second equality uses $\mathbb{E}_{\mu_{\pi_{\mathrm{avg}}}}[Q_f] = J(\pi_e)$, and $m_i(q) = 0$. For the remainder of the proof, all variances, covariances, and expectations are taken with respect to $\mu_{\pi_{\mathrm{avg}}}$, so we omit the corresponding subscript.

Using $\mathrm{Var}[X] - \mathrm{Var}[Y] = \mathrm{Var}[X-Y] + 2\mathrm{Cov}[X-Y, Y]$, we obtain

$$\begin{aligned}
&\frac{V(f) - V(q)}{n} \\
&= \frac{1}{n}\mathrm{Var}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big) - \big(q(S,\pi_e) - f(S,\pi_e)\big)\right] \\
&\quad+ \frac{2}{n}\mathrm{Cov}\left[\begin{array}{l}\dfrac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big) - \big(q(S,\pi_e) - f(S,\pi_e)\big), \\ \dfrac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(R - q(S,A)\big) + q(S,\pi_e)\end{array}\right].
\end{aligned}$$

We now show that the covariance term is zero. It decomposes into

$$\begin{aligned}
&\mathrm{Cov}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big) - \big(q(S,\pi_e) - f(S,\pi_e)\big), \frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(R - q(S,A)\big)\right] \\
&\qquad+ \mathrm{Cov}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big) - \big(q(S,\pi_e) - f(S,\pi_e)\big), q(S,\pi_e)\right].
\end{aligned}$$

The first covariance vanishes because the first argument is measurable with respect to (S,A), while

$$\mathbb{E}[R - q(S,A) \mid S, A] = 0.$$

Thus, by iterated expectation, its covariance with the centered residual term is zero.

For the second covariance, note that $q(S,\pi_e)$ depends only on S. Moreover, conditional on S,

$$\mathbb{E}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big)\,\middle|\, S\right] = q(S,\pi_e) - f(S,\pi_e).$$

Hence,

$$\mathbb{E}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big) - \big(q(S,\pi_e) - f(S,\pi_e)\big) \,\middle|\, S\right] = 0.$$

Since the second argument $q(S,\pi_e)$ is a function only of S, the second covariance also vanishes. Therefore, the covariance term is zero, and the difference of mixture variances reduces to

$$\begin{aligned}&\frac{V(f) - V(q)}{n}\\ &= \frac{1}{n}\mathrm{Var}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big) - \big(q(S,\pi_e) - f(S,\pi_e)\big)\right] \geq 0.\end{aligned}$$

We proceed further to put the above expression into the claimed form. Define

$$X \coloneqq \frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big), \qquad Y \coloneqq q(S,\pi_e) - f(S,\pi_e).$$

Under $\mu_{\pi_{\mathrm{avg}}}$, the conditional distribution of A given S is $\pi_{\mathrm{avg}}(\cdot \mid S)$. Hence,

$$\begin{aligned}\mathbb{E}\left[X \mid S\right] &= \int_{\mathcal{A}} \frac{\pi_e(a \mid S)}{\pi_{\mathrm{avg}}(a \mid S)}\big(q(S,a) - f(S,a)\big)\pi_{\mathrm{avg}}(a \mid S)\, d\mu_a(a)\\ &= \int_{\mathcal{A}} \pi_e(a \mid S)\big(q(S,a) - f(S,a)\big)\, d\mu_a(a)\\ &= q(S,\pi_e) - f(S,\pi_e) = Y.\end{aligned}$$

Therefore, the last display can be written as

$$\frac{V(f) - V(q)}{n} = \frac{1}{n}\mathrm{Var}[X - \mathbb{E}[X \mid S]]\,.$$

By the law of total variance,

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[\mathrm{Var}\left[X \mid S\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[X \mid S\right]\right].$$

Since $\mathbb{E}[X \mid S] = Y$, this gives

$$\mathrm{Var}\left[X - \mathbb{E}[X \mid S]\right] = \mathrm{Var}\left[X\right] - \mathrm{Var}\left[Y\right] \geq 0.$$

Substituting the definitions of X and Y, and noting that the marginal distribution of S under $\mu_{\pi_{\mathrm{avg}}}$ is p_s, we obtain

$$\begin{aligned}&\frac{V(f) - V(q)}{n}\\ &= \frac{1}{n}\left(\mathrm{Var}\left[\frac{\pi_e(A \mid S)}{\pi_{\mathrm{avg}}(A \mid S)}\big(q(S,A) - f(S,A)\big)\right] - \mathrm{Var}_{p_s}\big[q(S,\pi_e) - f(S,\pi_e)\big]\right) \geq 0.\end{aligned}$$

Subtracting $n^{-1}\sum_i (n_i/n) m_i(f)^2$ proves the claimed identity. Nonnegativity follows since the result equals $n^{-1}\sum_i (n_i/n)\mathrm{Var}_{\mu_{\pi_i}}[X - Y]$. Here we use $\sum_i (n_i/n)\mathrm{Cov}_{\mu_{\pi_i}}[X - Y, Q_q] = \mathrm{Cov}_{\mu_{\pi_{\mathrm{avg}}}}[X - Y, Q_q] = 0$, where the first equality holds because $\mathbb{E}_{\mu_{\pi_i}}[Q_q] = J(\pi_e)$ for every i. □

Theorem 4.3. *The variance gap between $\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})$ and the semiparametrically efficient oracle estimator $\hat{J}^\star_\Gamma(\pi_e; \mathcal{D}) \coloneqq \hat{J}_\Gamma(\pi_e, 1/\pi_{\mathrm{avg}}, q; \mathcal{D})$ is given by*

$$\mathrm{Var}\big[\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})\big] - \mathrm{Var}\big[\hat{J}^\star_\Gamma(\pi_e; \mathcal{D})\big] = \frac{1}{n}\left(\inf_{\boldsymbol{\alpha} \in \mathbb{R}^K} \big\|q\,\pi_e - g_{\boldsymbol{\alpha}}\big\|^2_{\mu_{\log}} - \mathrm{Var}_{p_s}\big[q(S,\pi_e)\big]\right) \geq 0,$$

where $\|\cdot\|_{\mu_{\log}}$ is the norm induced by the inner product $\langle\cdot,\cdot\rangle_{\mu_{\log}}$, with $d\mu_{\log} \coloneqq p_s/\pi_{\mathrm{avg}}\, d\mu_{s,a}$. Moreover, equality holds whenever $q\,\pi_e \in \mathrm{span}\big(\{\pi_i\}_{i\in[K]}\big)$.

Proof of Theorem 4.3. First, rewrite $\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})$ in (3.4) as

$$\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{(S,A,R)\in\mathcal{D}_i} \left[\frac{\pi_e(A \mid S)}{\pi_{\text{avg}}(A \mid S)} \left(R - f_{\boldsymbol{\alpha}^\star}(S,A)\right) + f_{\boldsymbol{\alpha}^\star}(S, \pi_e) \right],$$

where $f_{\boldsymbol{\alpha}^\star} \coloneqq g_{\boldsymbol{\alpha}^\star}/\pi_e = \sum_{i=1}^{K} \alpha_i^\star \pi_i / \pi_e$. Hence, $\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D}) = \hat{J}_\Gamma(\pi_e, f_{\boldsymbol{\alpha}^\star}; \mathcal{D})$.

The normal equation gives $\mathbb{E}_{\mu_{\pi_i}}[\pi_e(q - f_{\boldsymbol{\alpha}^\star})/\pi_{\text{avg}}] = 0$ for every i. Also, weighting the i-th normal equation by n_i/n and using $\sum_i \frac{n_i}{n} T_{ij} = 1$ and $\sum_i \frac{n_i}{n} c_i = J(\pi_e)$ gives $\sum_j \alpha_j^\star = J(\pi_e)$, so $m_i(f_{\boldsymbol{\alpha}^\star}) = 0$. Applying Lemma E.3 therefore gives

$$\begin{aligned}
&\operatorname{Var}\big[\hat{J}(\pi_e, \boldsymbol{\alpha}^\star; \mathcal{D})\big] - \operatorname{Var}\big[\hat{J}_\Gamma^\star(\pi_e; \mathcal{D})\big] \\
&\quad = \frac{1}{n} \left(\operatorname{Var}_{\mu_{\pi_{\text{avg}}}} \left[\frac{\pi_e(A \mid S)}{\pi_{\text{avg}}(A \mid S)} \big(q(S,A) - f_{\boldsymbol{\alpha}^\star}(S,A)\big) \right] - \operatorname{Var}_{p_s}\big[q(S,\pi_e) - f_{\boldsymbol{\alpha}^\star}(S,\pi_e)\big] \right) \\
&\quad = \frac{1}{n} \left(\operatorname{Var}_{\mu_{\pi_{\text{avg}}}} \left[\frac{\pi_e(A \mid S)}{\pi_{\text{avg}}(A \mid S)} \big(q(S,A) - f_{\boldsymbol{\alpha}^\star}(S,A)\big) \right] - \operatorname{Var}_{p_s}\big[q(S,\pi_e)\big] \right).
\end{aligned}$$

The last equality follows because

$$f_{\boldsymbol{\alpha}^\star}(S, \pi_e) = \int_{\mathcal{A}} \sum_{i=1}^{K} \alpha_i^\star \pi_i(a|s)\, d\mu_a(a) = \sum_{i=1}^{K} \alpha_i^\star,$$

which is deterministic and therefore does not contribute to the variance.

It remains to rewrite the first variance term in terms of the norm $\|\cdot\|_{\mu_{\log}}$. First, observe that the term inside $\operatorname{Var}_{\mu_{\pi_{\text{avg}}}}$ has mean zero:

$$\begin{aligned}
\mathbb{E}_{\mu_{\pi_{\text{avg}}}} \left[\frac{\pi_e(A \mid S)}{\pi_{\text{avg}}(A \mid S)} \big(q(S,A) - f_{\boldsymbol{\alpha}^\star}(S,A)\big) \right] &= \mathbb{E}_{\mu_{\pi_e}}\big[q(S,A) - f_{\boldsymbol{\alpha}^\star}(S,A)\big] \\
&= \mathbb{E}_{\mu_{\pi_e}}\big[q(S,A)\big] - \mathbb{E}_{\mu_{\pi_e}} \left[\sum_{i=1}^{K} \alpha_i^\star \frac{\pi_i(A|S)}{\pi_e(A|S)} \right] \\
&= J(\pi_e) - \sum_{i=1}^{K} \alpha_i^\star = 0.
\end{aligned}$$

Here, the last equality follows from the definition $\alpha_i^\star = \langle w_i, 1 \rangle_{\mu_{\pi_e}}$ in the proof of Theorem 4.1 and the normalization condition $\sum_{i\in[K]} w_i(s,a,r) = r$, which imply

$$\sum_{i\in[K]} \alpha_i^\star = \Big\langle \sum_{i\in[K]} w_i, 1 \Big\rangle_{\mu_{\pi_e}} = \mathbb{E}_{\mu_{\pi_e}}[R] = J(\pi_e).$$

Therefore,

$$\begin{aligned}
\operatorname{Var}_{\mu_{\pi_{\text{avg}}}} \left[\frac{\pi_e(A \mid S)}{\pi_{\text{avg}}(A \mid S)} \big(q(S,A) - f_{\boldsymbol{\alpha}^\star}(S,A)\big) \right] &= \mathbb{E}_{\mu_{\pi_{\text{avg}}}} \left[\frac{\pi_e^2(A \mid S)}{\pi_{\text{avg}}^2(A \mid S)} \big(q(S,A) - f_{\boldsymbol{\alpha}^\star}(S,A)\big)^2 \right] \\
&= \mathbb{E}_{\mu_{\pi_{\text{avg}}}} \left[\frac{\big(q(S,A)\pi_e(A \mid S) - g_{\boldsymbol{\alpha}^\star}(S,A)\big)^2}{\pi_{\text{avg}}^2(A \mid S)} \right] \\
&= \big\| q\,\pi_e - g_{\boldsymbol{\alpha}^\star} \big\|_{\mu_{\log}}^2 \\
&= \inf_{\boldsymbol{\alpha}\in\mathbb{R}^K} \big\| q\,\pi_e - g_{\boldsymbol{\alpha}} \big\|_{\mu_{\log}}^2,
\end{aligned}$$

where the last equality follows from the fact that $g_{\boldsymbol{\alpha}^\star}$ is the orthogonal projection of $q\,\pi_e$ with respect to the norm $\|\cdot\|_{\mu_{\log}}$. This proves the theorem. □

F Computational Complexity and Scalability

In this section, we analyze the computational complexity of $\hat{J}_{\mathrm{oIPS}}$, especially when the number of loggers is large ($K \approx 100$). Although restricted-optimal coefficients require solving a linear system with technique matrix $\mathbf{T} \in \mathbb{R}^{K \times K}$, we show both theoretically and empirically that the added cost is modest for practical K. In this appendix, N denotes the total sample size, written n in the main text.

F.1 Theoretical Complexity Analysis

Computing $\hat{J}_{\mathrm{oIPS}}$ within cross-fitting has two main steps: (i) estimating the technique matrix $\mathbf{T}$ and contribution vector $\mathbf{c}$, and (ii) solving for $\boldsymbol{\alpha} = \mathbf{T}^{-1}\mathbf{c}$ (or its pseudoinverse analogue).

(i) **Matrix construction:** Constructing $\mathbf{T}$ aggregates propensity terms over N samples and K loggers, yielding $O(NK)$ complexity. This matches the dominant complexity class of balanced IPS ($\hat{J}_{\mathrm{bIPS}}$), which also computes aggregate propensity statistics (e.g., $\sum_k n_k \pi_k$) per sample.

(ii) **Linear solve / pseudoinversion:** Solving $\mathbf{T}\boldsymbol{\alpha} = \mathbf{c}$ has worst-case complexity $O(K^3)$ with standard dense linear algebra (e.g., LU or SVD).

The total time is $O(Z(NK + K^3))$, which is $O(NK + K^3)$ for fixed Z. Accumulating the per-fold moments in one pass uses $O(ZK^2)$ additional memory, while storing all cross-logger propensities uses $O(NK)$. These counts exclude fitting nuisance models such as propensity models. We compute $\boldsymbol{\alpha}$ with the SVD-based Moore–Penrose pseudoinverse, which also handles a rank-deficient $\mathbf{T}$; this occurs in the runtime study below, where all loggers are mixtures of the deterministic and uniform policies.

F.2 Empirical Runtime Analysis

To validate practicality, we measured runtime on the *Letter* dataset ($N = 20{,}000$, 26 actions) with $K = 10$ and $K = 100$ loggers. Experiments ran on an Apple MacBook Air M1 (base model, 8GB RAM). Logging policies were generated by mixing the deterministic policy with a uniform policy. We compared Naive IPS, Balanced IPS, Weighted IPS [Agarwal et al., 2017], and our Optimal IPS. Methods requiring sample-dependent weights (Weighted IPS and Optimal IPS) used $Z = 5$ cross-fitting folds.

Table 3 reports mean inference time per trial (milliseconds) over 200 independent runs. Each $\pm$ value is the half-width of a 95% confidence interval based on the normal approximation.

Table 3: Comparison of Relative RMSE and wall-clock Time (in milliseconds) on the Letter dataset with $N = 20{,}000$. We compare settings with $K = 10$ and $K = 100$ logging policies. Our method achieves the lowest error while maintaining a runtime comparable to the existing Weighted IPS baseline.

Estimator	$K = 10$ **Loggers**		$K = 100$ **Loggers**	
	Rel-RMSE	**Time (ms)**	**Rel-RMSE**	**Time (ms)**
Naive IPS ($\hat{J}_{\mathrm{IPS}}$)	0.0113	0.16 ± 0.02	0.0104	0.18 ± 0.02
Balanced IPS ($\hat{J}_{\mathrm{bIPS}}$)	0.0133	0.13 ± 0.02	0.0087	0.51 ± 0.01
Weighted IPS ($\hat{J}_{\mathrm{wIPS}}$)	0.0122	2.16 ± 0.03	0.0075	15.49 ± 0.43
Optimal IPS ($\hat{J}_{\mathrm{oIPS}}$)	**0.0045**	2.52 ± 0.05	**0.0045**	20.06 ± 0.13

Observations.

- **Scalability:** Even at $K = 100$, Optimal IPS takes about 20ms per trial. This is slower than heuristic balanced IPS but still inexpensive for offline evaluation.
- **Comparison with Weighted IPS:** Runtime is comparable to Weighted IPS. This is expected because both methods use cross-fitting and dataset passes to estimate nuisance quantities (variance terms for Weighted IPS, technique matrix for Optimal IPS). The linear solve adds limited overhead relative to those passes.

- **Efficiency–Compute Trade-off:** For a small runtime increase (milliseconds), Optimal IPS reduces Relative-RMSE by about 40% versus Weighted IPS and about 48% versus Balanced IPS in the $K = 100$ setting.

Practical Takeaway. These numbers should be interpreted in the context of the full OPE pipeline. In our runs, end-to-end processing, including data loading, preprocessing, and fitting logger models, dominates total cost. Against that baseline, the additional time for $\hat{J}_{\mathrm{oIPS}}$ remains a moderate fraction of wall-clock time rather than a bottleneck, while providing substantial accuracy gains.

F.3 Large-N Synthetic Experiment

We use the synthetic data generator of the Open Bandit Pipeline (OBP, version 0.5.7) [Saito et al., 2021] with 10 actions, five-dimensional Gaussian contexts, and binary rewards from a logistic model with fixed coefficients. The evaluation policy chooses the action with the highest conditional mean reward. Each of the five known logging policies mixes the evaluation policy with the uniform policy, with uniform-exploration probabilities 0.05, 0.25, 0.50, 0.75, and 0.95, and all strata have the same size. For each $N \in \{10^4, 10^5, 10^6\}$, we draw contexts, actions, and rewards with 20 independent random seeds. The ground-truth value is the evaluation policy's conditional mean reward averaged over the sampled contexts. Weighted IPS and $\hat{J}_{\mathrm{oIPS}}$ use five-fold cross-fitting. Confidence intervals are computed as in §5, with the 20 seeds as the resampled units.

Table 4: Relative-RMSE ($\times 10^{-3}$) in the large-N synthetic experiment. Parentheses show 95% bootstrap CIs.

Estimator	$N = 10^4$	$N = 10^5$	$N = 10^6$
IPS	13.0 (9.71, 16.2)	3.92 (2.88, 4.84)	1.40 (0.941, 1.79)
bIPS	6.14 (4.32, 7.94)	2.15 (1.53, 2.72)	0.729 (0.555, 0.904)
wIPS	3.83 (2.77, 4.82)	1.85 (1.23, 2.43)	0.544 (0.370, 0.718)
oIPS	**2.70** (1.89, 3.43)	**0.775** (0.570, 0.948)	**0.199** (0.144, 0.251)

Table 5: Mean runtime of each estimator in milliseconds, and the largest peak resident memory over 20 seeds. Memory is measured for a single process that generates the data and runs all four estimators.

N	IPS	bIPS	wIPS	oIPS	Peak memory (MiB)
10^4	0.178	0.161	4.626	6.000	124.652
10^5	0.994	1.220	15.372	25.435	164.668
10^6	18.053	23.462	162.890	346.201	584.516

We ran these measurements on an x86-64 Linux machine with eight allocated CPU cores and 12 GiB of RAM. Runtimes exclude data generation and the computation of propensities; peak memory includes the stored cross-logger propensities.

G Related Work

OPE evaluates counterfactual policies from logged data [Precup et al., 2000, Bottou et al., 2013]. Although classical treatments focus on a single logger, many real applications aggregate data from multiple behavior policies, such as parallel A/B systems, leading to a stratified design in which logger identity defines strata [Agarwal et al., 2017, He et al., 2019, Kallus et al., 2021, Chen et al., 2020, Wooldridge, 2001].

OPE with Multiple Loggers. Agarwal et al. [2017] initiated the multi-logger OPE literature with MIS-inspired IPS estimators [Veach and Guibas, 1995, Cornuet et al., 2012, Elvira et al., 2019] and

showed that no single estimator uniformly dominates across instances. Kallus et al. [2021] formalized this setting, derived the semiparametric efficiency bound, and proposed doubly robust estimators that asymptotically attain it. Related directions include off-policy learning with multiple loggers [He et al., 2019], infinite-horizon analysis [Chen et al., 2020], and generalization under selection bias [Hatt et al., 2022]. Liu et al. [2025] studied multi-policy evaluation in RL via behavior-policy design.

Multiple Importance Sampling. MIS was originally developed to combine multiple proposal distributions in Monte Carlo rendering [Veach and Guibas, 1995, Owen and Zhou, 2000, Elvira et al., 2019, Kondapaneni et al., 2019]. The balance and power heuristics [Veach and Guibas, 1995] provide robust performance under nonnegative weights, and later work studied broader weighting families [Elvira et al., 2019]. Kondapaneni et al. [2019] derived variance-optimal MIS weights that may be negative, yielding further variance reduction in some regimes. Our method adapts this variance-minimization perspective to multi-logger OPE.

Broader Context. Beyond the multi-logger setting, there is extensive work on variance control for IPS in standard OPE. Classical IPS traces back to Horvitz–Thompson [Horvitz and Thompson, 1952]. Subsequent work studied weight truncation [Ionides, 2008] and doubly robust estimation in contextual bandits [Dudík et al., 2014]. More recent lines include self-normalization and implicit-exploration objectives [Swaminathan and Joachims, 2015, Kuzborskij et al., 2021], transformations yielding sub-Gaussian concentration [Metelli et al., 2021], and pessimistic regularization through PAC-Bayesian analysis [Aouali et al., 2024]. These methods are complementary to, rather than replacements for, our multi-logger focus.

H Broader Impacts

This work is methodological research on off-policy evaluation with multiple logging policies. Its main positive impact is to make retrospective policy evaluation more reliable in domains where direct online testing may be costly or risky, such as recommender systems, healthcare decision support, and advertising. The main potential negative impacts arise from downstream use: more accurate OPE could still be used to justify the deployment of policies trained on biased, privacy-sensitive, or otherwise unrepresentative logs, and an incorrectly specified evaluation pipeline could give false confidence in a harmful policy. These risks are not unique to our estimator, but they motivate using it together with domain review, data-governance checks, privacy controls, fairness and subgroup analyses, and prospective validation before deployment in high-stakes settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the estimator class, closed-form coefficient system, cross-fitted implementation, restricted optimality claims, and empirical scope. These claims are matched by the method in Section 3, the analysis in Section 4, and the experiments in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 states the overlap, support, moment, and logger-independence assumptions (Assumption 1), explains the scope of the restricted-optimality result, and discusses estimated logging policies (Remark 2). Section 6 identifies the remaining limitations, including the known-propensity assumption of our asymptotic guarantees, the lack of finite-sample analysis, and the scope of the benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions for the main theoretical statements are collected in Section 4, and the main theorems and corollaries are stated with their conditions. Full deferred proofs are provided in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and Appendix A specify the datasets, train/evaluation split, logging-policy construction, stratum ratios, estimators, number of trials, and nuisance estimation procedures. The supplementary materials include trial-level results and reproduction instructions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary materials provide the public benchmark data, reproduction code, saved trial results, and reproduction instructions. The code will be publicly released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 specifies the four UCI datasets, the 30%/70% split, deterministic evaluation policy, synthetic logging policies, stratum ratios, Relative-RMSE metric, $M = 200$ trials, and $Z = 5$ cross-fitting. Appendix A describes 2-fold cross-fitted multinomial logistic regression, smoothing, and propensity clipping for estimated logging policies.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Section 5 defines the paired percentile-bootstrap 95% intervals shown in the experimental figures, using $2{,}000$ joint resamples of 200 UCI trials or 20 synthetic trials. Appendix A also reports squared bias and empirical standard deviation at representative settings (Table 2).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F provides asymptotic computational complexity, empirical runtime measurements, and the hardware used for runtime experiments: an Apple MacBook Air M1 with 8GB RAM. Appendix F.3 reports the large-N resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is methodological research on off-policy evaluation and uses public benchmark datasets, synthetic logged bandit feedback, and reproducible code. It does not involve human-subject experiments, private data, or deployment of a high-risk model.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H discusses positive impacts from more reliable retrospective policy evaluation and negative risks from using OPE on biased, privacy-sensitive, or unrepresentative logs. It also notes mitigations through domain review, governance, privacy controls, fairness analysis, and prospective validation for high-stakes deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained generative models, scraped datasets, or other assets with high misuse or dual-use risk. The released materials are experiment code, benchmark classification data, generated contextual-bandit logs, and aggregate outputs.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section 5 cites the four UCI datasets and the UCI repository lists them under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The supplemental code archive releases the authors' code under the MIT License and preserves CC BY 4.0 attribution for the included UCI-derived data. The large-N experiment uses the Open Bandit Pipeline (version 0.5.7) [Saito et al., 2021], released under the Apache 2.0 License.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The supplementary materials document the experimental procedures, generated bandit observations, and saved trial results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects. All experiments use public benchmark datasets converted into synthetic contextual-bandit feedback.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve crowdsourcing or human-subject experiments, so IRB or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology, theory, and experiments do not use LLMs as a component of the proposed estimator or evaluation pipeline. Any writing, editing, or formatting assistance would not affect the scientific method or results and therefore does not require declaration under this question.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.