Robust Offline Policy Learning with Observational Data from Multiple Sources

Aldo Gael Carranza
Stanford University

Susan Athey
Stanford University

Abstract

We consider the problem of using observational bandit feedback data from multiple heterogeneous data sources to learn a personalized decision policy that robustly generalizes across diverse target settings. To achieve this, we propose a minimax regret optimization objective to ensure uniformly low regret under general mixtures of the source distributions. We develop a policy learning algorithm tailored to this objective, combining doubly robust offline policy evaluation techniques and no-regret learning algorithms for minimax optimization. Our regret analysis shows that this approach achieves the minimal worst-case mixture regret up to a moderated vanishing rate of the total data across all sources. Our analysis, extensions, and experimental results demonstrate the benefits of this approach for learning robust decision policies from multiple data sources.

1 INTRODUCTION

Offline policy learning from observational bandit feedback data is an effective approach for learning personalized decision policies in applications where obtaining real-time data is impractical (Swaminathan and Joachims, 2015; Kitagawa and Tetenov, 2018; Athey and Wager, 2021). Typically, the observational data used in offline policy learning is assumed to originate from a single source. However, in practice, multiple datasets collected from various experiments under different populations, environments, or logging policies are often available (Kallus et al., 2021). For example, a healthcare policymaker aiming to design a targeted medical intervention policy might have access to data from various hospitals, each having conducted different clinical trials on distinct patient populations. Leveraging heterogeneous observational datasets effectively, with their broader and more diverse coverage of the decision space, can lead to more generalizable policies across a wider range of settings.

However, simply training a policy on aggregated data does not guarantee robust performance. While such a policy may perform well on the uniform source mixture distribution, it may fail entirely in the original individual source settings or similar environments, particularly when significant distribution shifts exist across sources. This undermines the goal of using multiple observational datasets to generalize effectively across settings captured by any of the source distributions. Developing a policy that performs well across diverse target distributions requires robust domain adaptation. The key question is: given a family of target distributions that can be modeled by the source distributions or combinations thereof, can we train a policy that reliably generalizes across this entire family?

In this work, we address this challenge by framing the problem as one of minimax regret optimization over a family of target distributions represented as mixtures of the source distributions. We propose a novel offline policy learning algorithm that integrates techniques from doubly robust policy evaluation, offline policy optimization oracles, and no-regret learning algorithms for minimax optimization. Our analysis establishes finite-sample regret bounds, showing that the robustly trained policy achieves the minimal worst-case regret across the family of target distributions up to a vanishing rate of the total data across all sources. Additionally, we extend these bounds to target distributions that are not fully captured by any mixture of the source distributions. We characterize these regret bounds in terms of source heterogeneity and source distribution shift. Our theoretical analysis and experimental validation demonstrate the benefits of multi-source adaptation for robust offline policy learning across diverse environments.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

2 RELATED WORK

Offline Policy Learning. There have been many recent advancements in offline policy learning from observational bandit feedback data. Swaminathan and Joachims (2015); Kitagawa and Tetenov (2018) introduced foundational frameworks for learning structured decision policies via offline policy evaluation strategies. Athey and Wager (2021) achieved optimal regret rates under unknown propensities through doubly robust estimators, while Zhou et al. (2023) extended these results to the multi-action setting. Kallus (2018) directly derived optimal weights for target policies from the data, and Zhan et al. (2021) ensured optimal regret guarantees under adaptively collected data with diminishing propensities. Jin et al. (2022) relaxed the uniform overlap assumption, allowing for partial overlap under the optimal policy. We also mention that many contextual bandit methods often utilize offline policy learning oracles when designing adaptive action-assignment rules (Bietti et al., 2021; Simchi-Levi and Xu, 2022; Carranza et al., 2022).

Multiple Task Learning & Source Adaptation. Our problem is closely related to multi-task learning, which has been extensively studied in the supervised learning setting (Zhang and Yang, 2018) and, more pertinently, in offline policy learning (Hong et al., 2023). Additionally, in the context of offline policy learning with heterogeneous data sources, prior work Agarwal et al. (2017); He et al. (2019); Kallus et al. (2021) has leveraged data from multiple historical logging policies but assumes a shared underlying population and environment. Another closely related area is multiple source adaptation. In particular, Mohri et al. (2019) presented a framework for robust supervised learning across multiple sources, introducing the concepts of weighted Rademacher complexity and mixture skewness measures—which are essential to our approach.

Minimax Excess Risk Optimization. Agarwal and Zhang (2022) proposed minimax excess risk minimization in the context of distributionally robust supervised learning, aiming to achieve uniformly low regret across a family of test distributions. We adapt analytical and algorithmic techniques from this framework for minimax regret optimization in robust multi-source offline policy learning for a family of target distributions.

3 PRELIMINARIES

3.1 Setting

We introduce the setting of offline policy learning from observational bandit feedback data across multiple data sources. Let $\mathcal{X} \subset \mathbb{R}^p$ be the context space, $\mathcal{A} = \{a_1, \ldots, a_d\}$ be the finite action space with d actions, and $\mathcal{Y} \subset \mathbb{R}$ be the reward space. A *decision policy* $\pi : \mathcal{X} \to \mathcal{A}$ is a deterministic mapping from the context space $\mathcal{X}$ to actions $\mathcal{A}$. We assume there is a finite set of data sources $\mathcal{S}$, with each source $s \in \mathcal{S}$ possessing a *data-generating distribution* $\mathcal{D}_s$ defined over $\mathcal{X} \times \mathcal{Y}^d$ which governs how the source contexts X^s and potential reward outcomes $Y^s(a_1), \ldots, Y^s(a_d)$ are generated. These source distributions may all be different from each other.

We assume access to observational bandit feedback data from each of the data sources. The aim is to use this data learn a policy that performs well uniformly across target distributions that can be captured by the source distributions. We consider the natural scenario where the target distribution can be modeled as a *mixture* of the source distributions, i.e., $\mathcal{D}_\lambda := \sum_{s \in \mathcal{S}} \lambda_s \mathcal{D}_s$ for some unknown mixture weights λ. As an extension, we will account for the case where there may be a *discrepancy* between the target distribution and the mixture distributions. Since the the right choice of mixture weights is unknown, we must come up with a solution that is favorable for any weights in a specified set of *valid mixture weights* in the simplex over the source set, i.e., $\lambda \in \Lambda \subset \Delta(\mathcal{S})$. In the following section, we introduce policy performance measures that capture these objectives.

3.2 Objective

First, we define the expected reward of a decision policy gained under a given source mixture distribution.

Definition 1 (Mixture Policy Value)**.** For any policy π, the *mixture policy value* under the mixture weights $\lambda \in \Lambda$ is

$$Q_\lambda(\pi) := \mathop{\mathbb{E}}_{Z \sim \mathcal{D}_\lambda} [Y(\pi(X))],$$

where the expectation is taken with respect to the mixture distribution $Z = (X, Y(a_1), \ldots, Y(a_d)) \sim \mathcal{D}_\lambda$.

The performance of a policy is typically characterized by the notion of regret against an optimal policy in a specified *policy class* $\Pi \subset \{\pi : \mathcal{X} \to \mathcal{A}\}$, which we assume to be fixed throughout the paper.

Definition 2 (Mixture Regret)**.** For any policy π, the *mixture regret* under the mixture weights $\lambda \in \Lambda$ relative to the given policy class Π is

$$R_\lambda(\pi) := \max_{\pi' \in \Pi} Q_\lambda(\pi') - Q_\lambda(\pi).$$

Thus, the objective is to determine a policy in the specified policy class that minimizes the *worst-case mixture regret* under the specified set of valid mixture weights. In this paper, we propose a policy learning procedure that approximately achieves the minimal worst-case mixture regret for any choice of source mixture $\lambda' \in \Lambda$ up to a rate that vanishes with the total data. In particular, we describe a procedure that learns a policy $\hat{\pi} \in \Pi$ that achieves a bound of the form

$$R_{\lambda'}(\hat{\pi}) \lesssim \min_{\pi \in \Pi} \max_{\lambda \in \Lambda} R_{\lambda}(\pi) + \xi(n)$$

for some root-n vanishing rate $\xi(n)$ of the total data size n across all sources. We will then discuss extensions to regret bounds under general target distributions not captured by the class of mixture distributions.

3.3 Data-Generating Processes

We assume each source $s \in \mathcal{S}$ has a *local observational data set* $\{(X_i^s, A_i^s, Y_i^s)\}_{i=1}^{n_s} \subset \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ consisting of $n_s \in \mathbb{N}$ triples of contexts, actions, and rewards collected using a *local experimental stochastic policy* $e_s : \mathcal{X} \to \Delta(\mathcal{A})$ in the following manner. For the i-th data point of source $s \in \mathcal{S}$,

1. nature samples $(X_i^s, Y_i^s(a_1), \dots, Y_i^s(a_d)) \sim \mathcal{D}_s$;
2. source s is assigned action $A_i^s \sim e_s(\cdot|X_i^s)$;
3. source s observes $Y_i^s = Y_i^s(A_i^s)$;
4. source s logs the data tuple (X_i^s, A_i^s, Y_i^s).[1]

Remark. We will let $n := \sum_{s \in \mathcal{S}} n_s$ denote the *total sample size* across sources, and we will derive regret bounds that scale with the total sample size.

Note that although the counterfactual reward outcomes $Y_i^s(a)$ for all $a \in \mathcal{A} \backslash \{A_i^s\}$ exist in the source data-generating process, they are not observed in the realized data. All sources only observe the outcomes associated to their assigned treatments. For this reason, such observational data is also referred to as *bandit feedback data* (Swaminathan and Joachims, 2015).

Given these data-generating processes, it will also be useful to introduce the induced data-generating distributions that incorporate how actions are sampled. For each source $s \in \mathcal{S}$, the local historical policy e_s induces a *complete data-generating distribution* $\bar{\mathcal{D}}_s$ defined over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}^d$ that dictates how the entire local contexts, actions, and potential outcomes were sampled in the local data-generating process, i.e., $(X_i^s, A_i^s, Y_i^s(a_1), \dots, Y_i^s(a_d)) \sim \bar{\mathcal{D}}_s$.

3.4 Data Assumptions

We make the following standard assumptions on the data-generating process of any given source.

Assumption 1 (Local Ignorability)**.** For any source $s \in \mathcal{S}$, the local complete data-generating distribution $(X^s, A^s, Y^s(a_1), \dots, Y^s(a_d)) \sim \bar{\mathcal{D}}_s$ satisfies:

(a) *Boundedness*: The marginal distribution of $\bar{\mathcal{D}}_s$ on the set of potential outcomes $\mathcal{Y}^d$ has a bounded support, i.e., there exists some $B_s > 0$ such that $|Y^s(a)| \leqslant B_s$ for all $a \in \mathcal{A}$.

(b) *Unconfoundedness*: Potential outcomes are independent of the observed action given the observed context, i.e., $(Y^s(a_1), \dots, Y^s(a_d)) \perp\!\!\!\perp A^s \mid X^s$.

(c) *Overlap*: For any given context, every action has a non-zero probability of being sampled, i.e., there exists some $\eta_s > 0$ such that $\mathbb{P}(A^s = a|X^s = x) \geqslant \eta_s$ for any $a \in \mathcal{A}$ and $x \in \mathcal{X}$.

Note that the *boundedness* assumption is not essential and we only impose it for simplicity in our analysis. With additional effort, we can instead rely on light-tail distributional assumptions such as sub-Gaussian potential outcomes as in (Athey and Wager, 2021). *Unconfoundedness* ensures that action assignment is as good as random after accounting for measured covariates, and it is necessary to ensure consistent policy value estimation using inverse propensity-weighted strategies. The *uniform overlap* condition ensures that the decision space is sufficiently explored to guarantee accurate evaluation of any policy. However, we note that this assumption may not be entirely necessary as recent work (Jin et al., 2022) has introduced a pessimism-based approach that does away with the uniform overlap assumption for all actions and only relies on overlap for the optimal policy, assuming the behavior policy from the data collection process is known. Nevertheless, we decided to impose this uniform overlap assumption since pessimistic offline policy learning still requires known propensities, whereas we consider a more general setting where propensities can be estimated. Moreover, we made the above assumptions to simplify our analysis and maintain the focus of our contributions on policy learning under multiple sources. In any case, these stated assumptions are standard and they are satisfied in many practical settings such as randomized controlled trials or A/B tests.

Next, we also impose the following local data scaling assumption on each source.

Assumption 2 (Local Data Scaling)**.** All local sample sizes asymptotically increase with the total sample size, i.e., for each $s \in \mathcal{S}$, $n_s = \Omega(\nu_s(n))$ where ν_s is an increasing function.

This assumption states that, asymptotically, the total sample size cannot increase without increasing across all data sources. We emphasize that this assumption is quite benign since ν_s could be any slowly increasing

[1] If the propensity $e_s(A_i^s|X_i^s) = \mathbb{P}_{\bar{\mathcal{D}}_s}(A_i^s|X_i^s)$ is known, it also locally logged as it can facilitate subsequent policy value estimation.

function (e.g., an iterated logarithm) and the asymptotic lower bound condition even allows step-wise increments. We only impose this assumption to ensure that the regret bounds in our analysis scale with respect to the total sample size with sensible constants. However, it does come at the cost of excluding scenarios in which a source always contributes $O(1)$ amount of data relative to the total data, no matter how much more total data is made available in aggregate, in which case it may be better to exclude any such source, assuming no additional information.

4 APPROACH

Our general approach is to use the available observational data across sources to construct an appropriate estimator of the mixture regret and use this estimator to inform an appropriate optimization objective for determining a robust policy under any mixture weights.

4.1 Nuisance Parameters

We define the following functions which are referred to as *nuisance parameters* since they are required to be separately known or estimated for the policy value estimates.

Definition 3 (Nuisance Parameters)**.** For any source $s \in \mathcal{S}$, their *conditional response* function μ_s and *inverse conditional propensity* function ω_s are defined, respectively, for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$, as

$$\mu_s(x; a) := \mathbb{E}_{\bar{\mathcal{D}}_s}[Y^s(a) | X^s = x],$$
$$\omega_s(x; a) := 1/\mathbb{P}_{\bar{\mathcal{D}}_s}(A^s = a | X^s = x).$$

For convenience, we denote $\mu_s(x) = (\mu_s(x;a))_{a \in \mathcal{A}}$ and $\omega_s(x) = (\omega_s(x;a))_{a \in \mathcal{A}}$.

In our estimation strategy, we must separately estimate the source conditional response and inverse conditional propensity functions when they are unknown. However, we can also pool data across sources known to be equivalent for improved nuisance parameter estimation. Following the literature on double machine learning Chernozhukov et al. (2018), we make the following high-level assumption on the estimators of these source nuisance parameters.

Assumption 3. For any source $s \in \mathcal{S}$, the source estimates $\hat{\mu}_s$ and $\hat{\omega}_s$ of the nuisance parameters μ_s and ω_s, respectively, trained on n_s source data points satisfy the following squared error bounds:

$$\mathbb{E}_{\mathcal{D}_s}\big[\, \|\hat{\mu}_s(X^s) - \mu_s(X^s)\|_2^2 \,\big] \leqslant \frac{o(1)}{\sqrt{n_s}},$$
$$\mathbb{E}_{\mathcal{D}_s}\big[\, \|\hat{\omega}_s(X^s) - \omega_s(X^s)\|_2^2 \,\big] \leqslant \frac{o(1)}{\sqrt{n_s}}.$$

We emphasize this is a standard assumption in the double machine learning literature, and we can easily construct estimators that satisfy these rate conditions, given sufficient regularity on the nuisance parameters (Zhou et al., 2023). See Appendix E.2 for more details. They can be estimated with widely available out-of-the-box regression and classification implementations.

4.2 Policy Value and Regret Estimators

Next, we define our policy value estimators. For any $s \in \mathcal{S}$, consider the *source augmented inverse propensity weighted* (AIPW) score for each $a \in \mathcal{A}$ to be

$$\Gamma^s(a) := \mu_s(X^s; a) + \bar{Y}^s \cdot \omega_s(X^s; a) \cdot \mathbf{1}\{A^s = a\},$$

where $\bar{Y}^s = Y^s(A^s) - \mu_s(X^s; a)$ are centered outcomes and $(X^s, A^s, Y^s(a_1), \ldots, Y^s(a_d)) \sim \bar{\mathcal{D}}_s$. One can readily show that this is an unbiased estimate of the mixture policy value, i.e., $Q_\lambda(\pi) = \mathbb{E}_\lambda \mathbb{E}_{\bar{\mathcal{D}}_s}[\Gamma^s(\pi(X^s))]$ (see the proof in Lemma 7). Accordingly, our procedure is to estimate the source AIPW scores and appropriately aggregate them to form the mixture policy value estimator.

To this end, we assume we have constructed nuisance parameter estimates $\hat{\mu}_s$ and $\hat{\omega}_s$ that satisfy Assumption 3. Then, for each data point (X_i^s, A_i^s, Y_i^s) in the observational data set of source $s \in \mathcal{S}$, we define the *approximate source AIPW* score for each $a \in \mathcal{A}$ to be

$$\hat{\Gamma}_i^s(a) := \hat{\mu}_s(X_i^s; a) + \hat{\bar{Y}}_i^s \cdot \hat{\omega}_s(X_i^s; a) \cdot \mathbf{1}\{A_i^s = a\}.$$

where $\hat{\bar{Y}}_i^s = Y_i^s - \hat{\mu}_s(X_i^s; a)$ are the approximate centered outcomes. Using these estimated scores, we introduce the following *mixture policy value estimate* and the corresponding *mixture regret estimate*:

$$\hat{Q}_\lambda(\pi) := \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \hat{\Gamma}_i^s(\pi(X_i^s))$$
$$\hat{R}_\lambda(\pi) := \max_{\pi' \in \Pi} \hat{Q}_\lambda(\pi') - \hat{Q}_\lambda(\pi).$$

Our proposed estimator is a generalized aggregate version of the doubly robust estimator introduced in the standard offline policy learning setting (Zhou et al., 2023). It is doubly robust in the sense that it is accurate as long as one of the nuisance parameter estimates is accurate for each source. To ensure we can use the same data to estimate the nuisance parameters and to construct the policy value estimates, we utilize a *cross-fitting* strategy for each source. See Appendix E.3 for more details on the cross-fitting strategy to estimate AIPW scores.

4.3 Optimization Objective

Using our mixture regret estimator, we propose an optimization objective to determine a decision policy that robustly performs well under any valid mixture. The optimization objective is to find a policy $\hat{\pi} \in \Pi$ that minimizes the worst-case mixture regret estimate over valid mixture weights $\lambda \in \Lambda$:

$$\hat{\pi} = \arg\min_{\pi \in \Pi} \max_{\lambda \in \Lambda} \hat{R}_\lambda(\pi)$$

The primary difficulty with this optimization problem is that it is, in general, non-convex-concave so the order of optimization cannot generally be interchanged, and it is non-differentiable so standard gradient-based minimax optimization procedures are not viable. Thus, in the following section, we introduce an algorithm based on no-regret dynamics to solve the stochastic reformulation of this minimax optimization problem.

5 ALGORITHM

5.1 OPO Oracle

First, we assume that we have access to a standard offline policy optimization (OPO) oracle.

Definition 4 (OPO Oracle)**.** Given a dataset of the form $\{(x_i, \gamma_i(a_1), \ldots, \gamma_i(a_d))\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}^d$, an *OPO oracle* for a policy class Π solves the following optimization problem:

$$\arg\max_{\pi \in \Pi} \sum_{i=1}^{N} \gamma_i(\pi(x_i)).$$

Any procedure that solves the standard offline policy learning problem is a candidate oracle. For example, we can use the PolicyTree method for finite-depth tree policy classes (Sverdrup et al., 2020). For linear classes and other parametric policy classes, we can use cost-sensitive classification methods (Beygelzimer et al., 2008, 2009; Krishnamurthy et al., 2017).

5.2 Exponentiated Gradient OPO

Assume for now that Λ is a finite set. Let $\Delta(\Lambda)$ denote the set of distributions over Λ, and let $\Delta(\Pi)$ denote the set of distributions over Π. We can readily rewrite the minimax optimization problem presented in Section 4.3 as a convex-concave optimization over these probability spaces:

$$\min_{\pi \in \Pi} \max_{\lambda \in \Lambda} \hat{R}_\lambda(\pi) = \min_{P \in \Delta(\Pi)} \max_{\rho \in \Delta(\Lambda)} \mathop{\mathbb{E}}_{\pi \sim P,\, \lambda \sim \rho} [\hat{R}_\lambda(\pi)].$$

We can approximately solve this zero-sum game through a no-regret dynamics approach (Freund and Schapire, 1996). In particular, note that for a given choice of strategy $\rho \in \Delta(\Lambda)$ by the maximizing player, their corresponding adversarial reward $r(\rho)$ defined by the minimizing player has the following equivalence:

$$r(\rho) := \min_{P \in \Delta(\Pi)} \mathop{\mathbb{E}}_{\pi \sim P,\, \lambda \sim \rho} [\hat{R}_\lambda(\pi)] = \min_{\pi \in \Pi} \mathop{\mathbb{E}}_{\lambda \sim \rho} [\hat{R}_\lambda(\pi)].$$

Therefore, we can use the following best response strategy for the minimizing player:

$$\begin{aligned} &\arg\min_{\pi \in \Pi} \mathop{\mathbb{E}}_{\lambda \sim \rho} [\hat{R}_\lambda(\pi)] \\ &= \arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{\lambda \sim \rho} [\hat{Q}_\lambda(\pi)] \\ &= \arg\max_{\pi \in \Pi} \sum_{\lambda \in \Lambda_\epsilon} \rho_\lambda \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \hat{\Gamma}_i^s(\pi(X_i^s)). \end{aligned}$$

The last equality shows that this best response for the minimizing player can be computed through a single call to the OPO oracle on the following dataset: $\{(X_i^s, \alpha_\lambda^s \Gamma_i^s(a_1), \ldots, \alpha_\lambda^s \Gamma_i^s(a_d)) \mid \lambda \in \Lambda, s \in \mathcal{S}, i \in [n_s]\}$ where $\alpha_\lambda^s = \rho_\lambda \cdot \lambda_s / n_s$. Therefore, given we can compute the corresponding adversarial reward for the maximizing player, we can use the exponentiated gradient (EG) algorithm (Kivinen and Warmuth, 1997) to sequentially learn a maximizing distribution over a finite collection of experts with each $\lambda \in \Lambda$ being an expert. More formally, we initialize the expert distribution to $\rho^{(1)} \propto 1$ and sequentially update:

$$\pi^{(t)} = \arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{\lambda \sim \rho^{(t)}} [\hat{Q}_\lambda(\pi)]$$
$$\rho_\lambda^{(t+1)} \propto \rho_\lambda^{(t)} \cdot \exp\left(\eta \cdot g_\lambda^{(t)}\right),$$

where $g_\lambda^{(t)}$ is the λ-th element of the gradient of the observed adversarial reward $r^{(t)}(\rho^{(t)})$, i.e.,

$$g_\lambda^{(t)} := \frac{\partial}{\partial \rho_\lambda^{(t)}} r^{(t)}(\rho^{(t)}) = \hat{R}_\lambda(\pi^{(t)}) = \max_{\pi' \in \Pi} \hat{Q}_\lambda(\pi') - \hat{Q}_\lambda(\pi^{(t)})$$

and $\eta = \sqrt{\frac{\log |\Lambda|}{\bar{B}^2 T}}$ with a constant $\bar{B}$ that uniformly bounds the empirical mixture regret (see Corollary 1 in Appendix C.6 for a suitable high probability bound based on B). We run this repeated procedure for T steps and return the policy $\hat{\pi} = \pi^{(T)}$. Note that the gradient $g_\lambda^{(t)}$ can also be computed with a call to the OPO oracle for each $\lambda \in \Lambda$ to determine $\max_{\pi' \in \Pi} \hat{Q}_\lambda(\pi')$ on the following dataset $\{(X_i^s, \alpha_\lambda^s \Gamma_i^s(a_1), \ldots, \alpha_\lambda^s \Gamma_i^s(a_d)) \mid s \in \mathcal{S}, i \in [n_s]\}$ where $\alpha_\lambda^s = \lambda_s / n_s$.

In the case where Λ is not finite, we can simply perform the same procedure over a *minimal covering set* of Λ. In our regret bound, we account for any incurred approximation error from this discretization. Additionally, we will account for the approximation error for terminating the algorithm after finitely many steps. See Algorithm 1 for a pseudocode outline of the algorithm, referred to as EG-OPO, and Appendix E for more details on the algorithm complexity, the choice of OPO oracles, and nuisance parameter estimation.

Algorithm 1 EG-OPO

```
Require: OPO oracle, time horizon $T$, approximation error threshold $\epsilon$, uniform regret bound $\bar{B}$
Compute minimal $\epsilon$-covering $\Lambda_\epsilon$ of $\Lambda$ under the $\ell_1$ distance
Set $\eta = \sqrt{\frac{\log|\Lambda_\epsilon|}{\bar{B}^2 T}}$
Set $\rho_\lambda^{(1)} \propto 1$ for all $\lambda \in \Lambda_\epsilon$
for each round $t = 1, 2, \dots, T$ do
    Utilize the OPO oracle to solve for the minimizing player's best response to $\rho^{(t)}$:
        $$\pi^{(t)} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\lambda \sim \rho^{(t)}} [\hat{Q}_\lambda(\pi)]$$
    Utilize OPO oracle to compute gradient of the adversarial reward for the maximizing player:
        $$g_\lambda^{(t)} = \hat{R}_\lambda(\pi^{(t)}), \ \forall \lambda \in \Lambda_\epsilon$$
    Update the maximizing player's distribution with the exponentiated gradient ascent update:
        $$\rho_\lambda^{(t+1)} \propto \rho_\lambda^{(t)} \cdot \exp(\eta \cdot g_\lambda^{(t)}), \ \forall \lambda \in \Lambda_\epsilon$$
end for
return $\hat{\pi} = \pi^{(T)}$
```

6 REGRET BOUNDS

In this section, we establish regret bounds for the policy solution to our optimization objective in Section 4.3. Refer to Appendices A, B, C for detailed discussions and proofs of the results presented in this section.

6.1 Complexity and Skewness

First, we introduce important quantities that appear in our regret bounds.

Policy Class Complexity The following quantity provides a measure of policy class complexity based on a variation of the classical entropy integral introduced by Dudley (1967), and it is useful in establishing a class-dependent regret bound. See Appendix B.2 for more details on its definition.

Definition 5 (Entropy integral). Let $\mathrm{H}(\pi_1, \pi_2; x) := \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{\pi_1(x_i) \neq \pi_2(x_i)\}$ be the Hamming distance between any two policies $\pi_1, \pi_2 \in \Pi$ given a covariate set $x \subset \mathcal{X}$ of size $m \in \mathbb{N}$. The *entropy integral* of a policy class Π is

$$\kappa(\Pi) := \int_0^1 \sqrt{\log N_{\mathrm{H}}(\epsilon^2, \Pi)} d\epsilon,$$

where $N_{\mathrm{H}}(\epsilon^2, \Pi)$ is the maximal ϵ^2-covering number of Π under the Hamming distance over covariate sets of arbitrary size.

The entropy integral is constant for a fixed policy class, and rather weak assumptions on the class are sufficient to ensure it is finite such as sub-exponential growth on its Hamming covering number, which is satisfied by many policy classes including parametric and finite-depth tree policy classes (Zhou et al., 2023). In the binary action setting, the entropy integral of a policy class relates to its VC-dimension with $\kappa(\Pi) = \sqrt{\mathrm{VC}(\Pi)}$, and for D-dimensional linear classes $\kappa(\Pi) = \mathcal{O}(\sqrt{D})$.

Source Skewness The following quantity measures the imbalance of the source sampling distribution λ relative to the *empirical distribution of samples across sources.* This quantity naturally arises in the generalization bounds of weighted mixture distributions (Mohri et al., 2019; Mansour et al., 2021).

Definition 6 (Skewness). The *skewness* of a given set of mixture weights λ relative to the empirical distribution of samples $\bar{n} := (n_s/n)_{s \in \mathcal{S}}$ is

$$\mathfrak{s}(\lambda \| \bar{n}) := 1 + \chi^2(\lambda || \bar{n}),$$

where $\chi^2(\lambda||\bar{n})$ is the chi-squared divergence of λ from $\bar{n}$. Additionally, the *mixture-agnostic skewness* is

$$\mathfrak{s}(\Lambda \| \bar{n}) := \max_{\lambda \in \Lambda} \mathfrak{s}(\lambda \| \bar{n}).$$

6.2 Mixture-Agnostic Regret Bound

The following result captures a root-n finite-sample bound for the mixture-agnostic regret that parallels the optimal regret bounds typically seen in the offline policy learning literature.

Theorem 1 (Mixture-Agnostic Regret Bound). *Suppose Assumptions 1, 2, and 3 hold. For any $\epsilon > 0$, let Λ_ϵ denote a minimal ϵ-covering set of Λ under the ℓ_1 distance. Set $T = (n/\mathfrak{s}(\Lambda\|\bar{n}))^{1+\alpha}$ for any choice of $\alpha > 0$. Then, for any $\epsilon > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, the EG-OPO policy $\hat{\pi} = \pi^{(T)}$ achieves the regret bound*

$$R_{\lambda'}(\hat{\pi}) \leqslant \min_{\pi \in \Pi} \max_{\lambda \in \Lambda} R_\lambda(\pi) + 2B\epsilon + \xi_{\epsilon,\delta}(n; \Pi, \Lambda)$$

for any $\lambda' \in \Lambda$, where

$$\xi_{\epsilon,\delta}(n;\Pi,\Lambda) := C_{\epsilon,\delta}\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right),$$

where the constant $B = \max_{s \in \mathcal{S}} B_s$ is a uniform upper bound on all the potential outcomes across sources, $C_{\epsilon,\delta} = \sqrt{C \log(|\Lambda_\epsilon|/\delta)}$ with universal constant C, and $V = \max_{s \in \mathcal{S}} \max_{\pi \in \Pi} \mathbb{E}_{\bar{\mathcal{D}}_s}[\Gamma^s(\pi(X^s))^2]$ is the worst-case AIPW score variance across sources.

This result shows that the EG-OPO algorithm achieves minimal worst-case mixture regret uniformly across any valid mixture distribution, with a vanishing rate proportional to $\mathcal{O}_p(\kappa(\Pi)\sqrt{V \cdot \mathfrak{s}(\Lambda\|\bar{n})/n})$. This ensures the policy's robustness across a broad range of target settings. Importantly, the trained policy retains strong performance on each individual source distribution while generalizing to new settings represented by a mixture of source distributions, broadening the applicability of policy learning.

Note that if we set the discretization granularity to be $\epsilon = o(\sqrt{\mathfrak{s}(\Lambda\|\bar{n})/n})$, then the $2B\epsilon$ term in the above bound can be absorbed into the vanishing error term $o_p(\sqrt{\mathfrak{s}(\Lambda\|\bar{n})/n})$. However, depending on the corresponding size of $|\Lambda_\epsilon|$ and other terms in the above bound, more favorable choices may be possible.

Additionally, the root-n vanishing rate is moderated by the skewness which can also scale with the total sample size. For example, if $\Lambda = \{\bar{n}\}$ then $\mathfrak{s}(\Lambda\|\bar{n})/n = 1/n$, and if $\Lambda = \{(1, 0, \dots, 0)\}$ then $\mathfrak{s}(\Lambda\|\bar{n})/n = 1/n_1$. Thus, this skewness-moderated rate generalizes and smoothly interpolates between the rates one expects from the uniform weighted model and the single source model. Indeed, when sources are identical and $\Lambda = \{\bar{n}\}$, we effectively recover the best known rates from standard offline policy learning (Zhou et al., 2023).

6.3 Proof Sketch

In this section, we provide a high-level sketch of the proof of Theorem 1. Refer to the appendix for the full proofs and discussions.

First, we show the mixture regret $R_{\lambda'}(\hat{\pi})$ can be decomposed into the *empirical mixture regret* $\hat{R}_{\lambda'}(\hat{\pi})$ and the *mixture empirical process* $\sup_{\pi_a,\pi_b} |\Delta_\lambda(\pi_a, \pi_b) - \hat{\Delta}_\lambda(\pi_a, \pi_b)|$, where $\Delta_\lambda(\pi_a, \pi_b) = Q_\lambda(\pi_a) - Q_\lambda(\pi_b)$ and $\hat{\Delta}_\lambda(\pi_a, \pi_b) = \hat{Q}_\lambda(\pi_a) - \hat{Q}_\lambda(\pi_b)$ are the *true* and *empirical mixture policy value differences* between any two policies $\pi_a, \pi_b \in \Pi$.

The empirical mixture regret can be bounded by the suboptimality bound of the EG-OPO algorithm using the following lemma which we demonstrate follows from the results of (Freund and Schapire, 1996).

Lemma 1. *For any T and any $\lambda' \in \Lambda_\epsilon$,*

$$\hat{R}_{\lambda'}(\hat{\pi}) \leqslant \min_{\pi\in\Pi} \max_{\lambda\in\Lambda} \hat{R}_\lambda(\pi) + 2\hat{B}\sqrt{\frac{\log|\Lambda_\epsilon|}{T}}$$

where $\hat{B}$ is a uniform bound on $\hat{R}_{\lambda'}(\pi)$.

Next, the mixture empirical process can be further decomposed into *oracle* and *approximate mixture empirical processes*. For the oracle empirical process, we make use of Talagrand concentration inequalities and symmetrization arguments (Koltchinskii, 2011) to establish tight high-probability bounds based on the *weighted Rademacher complexity* (Mohri et al., 2019) for the policy value function class. We then follow a Dudley chaining argument (Dudley, 1967) to show that this weighted Rademacher complexity can be bounded by measures of policy class complexity, source mixture set skewness, and worst-case source AIPW score variance as they appear in Theorem 1.

The approximate empirical process can simply be bounded directly by decomposing the differences and relating them to the assumptions on the data-generating process and the nuisance parameter estimation error stated in Assumptions 1, 2, 3. For robustness across valid mixtures, for each of these empirical process bounds we take a union bound over the valid mixture weights set.

Moreover, the resulting bound on the mixture empirical process allows us to relate the minimax empirical mixture regret that appears in Lemma 1 to the minimax mixture regret that finally appears in Theorem 1. Lastly, the additional ϵ term in the mixture-agnostic regret bound is due to the discretization error over the minimal cover Λ_ϵ of the valid mixture weights set Λ.

6.4 Target Regret Bound

The following result captures a regret bound on a target distribution that is not captured by the family of valid mixture distributions. Refer to Appendix D for a proof of this result.

Theorem 2 (Target Regret Bound)**.** *Let $R(\hat{\pi})$ denote the target regret of $\hat{\pi}$ under a target distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}^d$. This is bounded as*

$$R(\hat{\pi}) \leqslant 2 \cdot \mathrm{disc}(\mathcal{D}_\Lambda, \mathcal{D}) + \max_{\lambda\in\Lambda} R_\lambda(\hat{\pi}),$$

where

$$\mathrm{disc}(\mathcal{D}_\Lambda, \mathcal{D}) = \min_{\lambda\in\Lambda} \max_{\pi\in\Pi} |Q_\lambda(\pi) - Q(\pi)|$$

is the minimax mixture policy value discrepancy. Moreover, if the same boundedness condition also holds for $\mathcal{D}$, then this discrepancy can be further bounded by

$$\mathrm{disc}(\mathcal{D}_\Lambda, \mathcal{D}) \leqslant B \cdot \min_{\lambda\in\Lambda} \mathrm{TV}(\mathcal{D}_\lambda, \mathcal{D}),$$

where TV *is the total variation distance.*

This result immediately follows from a decomposition of the target regret into the mixture regret and its associated discrepancy. Note that the worst-case mixture regret here can be bounded using the result of Theorem 1.

7 EXPERIMENTS

7.1 Setup

In this section, we describe our experimental setup. We consider the source set $\mathcal{S} = [k]$ with $k = 3$ sources, a binary action set $\mathcal{A} = \{a_1, a_2\}$, and the context space $\mathcal{X} = [-1, 1]^p$ with $p = 2 \times q$ where $q = 4$. For every data source $s \in \mathcal{S}$, we construct the following data-generating process:

- $A^s \sim \text{Uniform}(\mathcal{A})$,
- $X^s \sim \text{Uniform}(\mathcal{X})$,
- $Y^s(a)|X^s \sim \text{Normal}\left(\mu_s(X^s; a), \sigma_s^2\right)$ for all $a \in \mathcal{A}$,

where the source reward mean is $\mu_s(x; a) = x_a^T \theta_s$ given source parameter $\theta_s \sim \text{Normal}(\vec{0}, \sigma^2 I_q)$ with $\sigma^2 = 5$ and the source reward variance is $\sigma_s^2 = 1$. For a given total sample size $n \in \mathbb{N}$, each source $s \in \mathcal{S}$ is allocated a local sample sample size determined by the following increasing function $n_s = \nu_s(n) = n/k$. By construction, this entire data-generating process satisfies our data assumptions stated in Section 3.3.

For the set of valid mixtures, we choose $\Lambda = \{e_s\}_{s \in \mathcal{S}}$ where e_s is the one-hot vector of source $s \in \mathcal{S}$.

For the policy class Π, we choose the class of fixed depth-2 decision trees, and for the OPO oracle, we use the PolicyTree algorithm (Sverdrup et al., 2020). Refer to Appendix E for additional details on the algorithm implementation, namely on the cross-fitting strategy for nuisance parameter and AIPW score estimation.

To assess the robustness of our approach, we will compare the empirical regret of different policies under two settings: a fixed source distribution (source $s = 1$) and a nearby mixture distribution (mixture weights $\lambda = [0.9, 0.05, 0.05]$). Under both settings, we will compare the *EG-OPO policy* against two baseline policies: the *aggregate policy* trained on data aggregated from all sources and the *source policy* trained on an equivalent amount of source data, each trained with a single call to the PolicyTree OPO oracle. The analysis considers a range of training sample sizes up to 500. Our plots will display the rolling mean and standard deviation bands of empirical regrets over three seed runs, with regrets computed relative to corresponding optimal models trained on 2,000 similarly sampled data points.

7.2 Performance on Source Distribution

We begin by comparing our algorithm against our baselines on the source distribution for a fixed source $s = 1$. In Figure 1, we plot the source regret $R_1(\pi)$ for the aggregate policy $\hat{\pi}_{\mathcal{S}}$, source policy $\hat{\pi}_1$, and EG-OPO policy $\hat{\pi}_{EG}$. We see that the aggregate policy performs poorly due to distribution shift in the data from all other sources. For a similar reason, the EG-OPO policy does not perform as well as the source policy, but it still shows improvement as the sample size increases, consistent with our theoretical results.

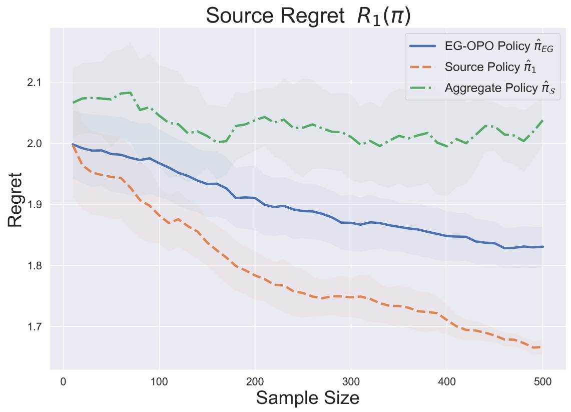

Figure 1: Empirical source regret $R_1(\pi)$.

7.3 Performance on Mixture Distribution

Next, we compare performances on a mixture distribution close to the source distribution $\mathcal{D}_1$, namely $\mathcal{D}_\lambda$ where $\lambda = [0.9, 0.05, 0.05]$. In Figure 2, we plot the mixture regret $R_\lambda(\pi)$ for the same policies. The aggregate policy again performs poorly, as the mixture distribution, although it accounts for other sources, is not close to the uniform distribution. Similarly, the local policy struggles due to the slight distribution shift, despite the mixture weights being close to the source degenerate weights, and performs even worse than the aggregate policy. In contrast, the EG-OPO policy remains robust and performs well in this setting, further supporting the consistent performance of our approach across different distributions uniformly at once.

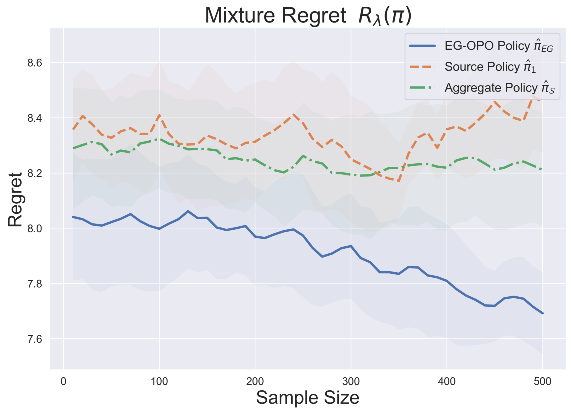

Figure 2: Empirical mixture regret $R_\lambda(\pi)$.

8 DISCUSSION

An ideal generalizable policy should perform well across a range of test distributions, not just the empirical distribution of its training data. If the test distribution exactly matches the uniformly combined distribution of all sources, the optimal strategy might be to train a policy on this aggregated dataset. However, such a policy may lack robustness to even slight deviations from this assumption. For example, our experiments above demonstrate that when the test distribution corresponds to a single source, a policy trained on aggregate heterogeneous data can significantly underperform. In contrast, our proposed algorithm explicitly accounts for distributional heterogeneity, enabling it to generalize across all valid test distributions simultaneously.

Our theoretical results further illustrate that a key factor influencing generalization is the number of sources and the amount of data and heterogeneity they introduce. When all sources are identical, the optimal mixture aligns with the empirical mixture, and increasing the number of sources simply provides more data, tightening regret bounds on the same test distribution. However, when sources differ, the optimal mixture deviates from the empirical mixture, creating a tradeoff between the benefits of data volume and diversity and the challenges posed by source heterogeneity. Our theoretical regret bound quantifies this tradeoff through total data size and skewness relative to the empirical mixture. In practice, a data source should be incorporated into training only if the additional samples it provides outweigh the increase in skewness it introduces subject to sufficient coverage to the test distributions considered.

While our primary focus is theoretical, our empirical results suggest that incorporating sufficiently similar sample-rich sources can significantly improve performance on test distributions corresponding to sample-poor sources. However, as heterogeneity increases, this benefit diminishes and can even become detrimental in highly heterogeneous settings if the test distribution is not covered by a source mixture. This underscores a fundamental tradeoff between data quantity and source heterogeneity, particularly in source-specific test scenarios. A deeper investigation into this tradeoff is an important direction for future work.

9 CONCLUSION

We presented an approach for robust offline policy learning using observational data from multiple sources, formulated as minimax regret optimization to ensure low regret across target distributions modeled as source mixtures. Our algorithm, combining doubly robust evaluation and no-regret learning strategies, achieves minimal worst-case regret bounds up to a root-n vanishing rate of the total data. Our theoretical analysis and experiments confirm its effectiveness in learning robust generalizable policies across different environments.

References

Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory*, pages 2704–2729. PMLR, 2022.

Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 687–696, 2017.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47 (2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709.

Alina Beygelzimer, John Langford, and Bianca Zadrozny. Machine learning techniques—reductions between prediction quality metrics. *Performance Modeling and Engineering*, pages 3–28, 2008.

Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 247–262. Springer, 2009.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *The Journal of Machine Learning Research*, 22(1):5928–5976, 2021.

Aldo Gael Carranza, Sanath Kumar Krishnamurthy, and Susan Athey. Flexible and efficient contextual bandits with heterogeneous treatment effect oracles. *arXiv preprint arXiv:2203.16668*, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Yoav Freund and Robert E Schapire. Game theory, online prediction and boosting. In *Proceedings of the* *ninth annual conference on Computational learning theory*, pages 325–332, 1996.

Li He, Long Xia, Wei Zeng, Zhi-Ming Ma, Yihong Zhao, and Dawei Yin. Off-policy learning for multiple loggers. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1184–1193, 2019.

Joey Hong, Branislav Kveton, Manzil Zaheer, Sumeet Katariya, and Mohammad Ghavamzadeh. Multi-task off-policy learning from bandit feedback. In *International Conference on Machine Learning*, pages 13157–13173. PMLR, 2023.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning" without"overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Nathan Kallus. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31, 2018.

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

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

Yishay Mansour, Mehryar Mohri, Jae Ro, Ananda Theertha Suresh, and Ke Wu. A theory of multiple-source adaptation with limited target labeled data. In *International Conference on Artificial Intelligence and Statistics*, pages 2332–2340. PMLR, 2021.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Erik Sverdrup, Ayush Kanodia, Zhengyuan Zhou, Susan Athey, and Stefan Wager. policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232, 2020.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Ruohan Zhan, Zhimei Ren, Susan Athey, and Zhengyuan Zhou. Policy learning with adaptively collected data. *arXiv preprint arXiv:2105.02344*, 2021.

Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]

(c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]

   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]

   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A AUXILIARY RESULTS

The following known results will be used in our regret bound proofs. See Chapter 2 of Koltchinskii (2011) for discussions of these results.

Lemma 2 (Hoeffding's inequality). *Let $Z_1, \dots, Z_n$ be independent random variables with $Z_i \in [a_i, b_i]$ almost surely. For all $t > 0$, the following inequality holds*

$$\mathbb{P}\left(\left|\sum_{i=1}^n Z_i - \mathbb{E}\left[Z_i\right]\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 3 (Talagrand's inequality). *Let $Z_1, \dots, Z_n$ be independent random variables in $\mathcal{Z}$. For any class of real-valued functions $\mathcal{H}$ on $\mathcal{Z}$ that is uniformly bounded by a constant $U > 0$ and for all $t > 0$, the following inequality holds*

$$\mathbb{P}\left(\left|\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n h(Z_i)\right| - \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n h(Z_i)\right|\right]\right| \geqslant t\right) \leqslant C\exp\left(-\frac{t}{CU}\log\left(1 + \frac{Ut}{D}\right)\right),$$

where $C > 0$ is a universal constant and $D \geqslant \mathbb{E}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n h^2(Z_i)\right]$.

Lemma 4 (Ledoux-Talagrand contraction inequality). *Let $Z_1, \dots, Z_n$ be independent random variables in $\mathcal{Z}$. For any class of real-valued functions $\mathcal{H}$ on $\mathcal{Z}$ and any L-Lipschitz function φ, the following inequality holds*

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \varepsilon_i(\varphi\circ h)(Z_i)\right|\right] \leqslant 2L\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \varepsilon_i h(Z_i)\right|\right],$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent Rademacher random variables.

Lastly, we state an auxiliary inequality that serves as a typical candidate for the quantity denoted by D above in Talagrand's inequality. This result follows as a corollary of the Ledoux-Talagrand contraction inequality and a symmetrization argument. We provide a proof for completeness.

Lemma 5. *Let $Z_1, \dots, Z_n$ be independent random variables in $\mathcal{Z}$. For any class of real-valued functions $\mathcal{H}$ on $\mathcal{Z}$ and any L-Lipschitz function φ, the following inequality holds*

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n(\varphi\circ h)(Z_i)\right] \leqslant \sup_{h\in\mathcal{H}}\sum_{i=1}^n \mathbb{E}\left[(\varphi\circ h)(Z_i)\right] + 4L\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \varepsilon_i h(Z_i)\right|\right],$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent Rademacher random variables.

Proof. We have that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n(\varphi\circ h)(Z_i)\right] - \sup_{h\in\mathcal{H}}\sum_{i=1}^n \mathbb{E}\left[(\varphi\circ h)(Z_i)\right] && (1)\\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n(\varphi\circ h)(Z_i) - \sup_{h\in\mathcal{H}}\sum_{i=1}^n \mathbb{E}\left[(\varphi\circ h)(Z_i)\right]\right] && (2)\\
&\leqslant \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n(\varphi\circ h)(Z_i) - \sum_{i=1}^n \mathbb{E}\left[(\varphi\circ h)(Z_i)\right]\right|\right] && (3)\\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \Big((\varphi\circ h)(Z_i) - \mathbb{E}\left[(\varphi\circ h)(Z_i)\right]\Big)\right|\right] && (4)\\
&\leqslant 2\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \varepsilon_i(\varphi\circ h)(Z_i)\right|\right] && (5)\\
&\leqslant 4L\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \varepsilon_i h(Z_i)\right|\right]. && (6)
\end{aligned}$$

Inequality (3) follows from the triangle inequality, inequality (5) follows from a standard symmetrization argument (see Koltchinskii (2011)), and inequality (6) follows from the Ledoux-Talagrand contraction inequality (see Lemma 4). The result follows by moving the second term in Equation (1) to the right-hand side in the last inequality. □

B COMPLEXITY AND HETEROGENEITY MEASURES

In this section, we introduce important quantities of policy class complexity and source heterogeneity that appear in our analysis. All throughout, we let $n = \sum_{s \in \mathcal{S}} n_s$ be the total sample size across sources, $n_{\mathcal{S}} = (n_s)_{s \in \mathcal{S}}$ the vector of sample sizes across sources, and $\bar{n} = (n_s/n)_{s \in \mathcal{S}}$ the empirical distribution of samples across sources.

B.1 Weighted Rademacher Complexity

Our learning bounds will rely on the following weighted multiple-source generalization of Rademacher complexity introduced in Mohri et al. (2019).

Definition 7 (Weighted Rademacher complexity)**.** Suppose there is a set of sources $\mathcal{S}$, where each source $s \in \mathcal{S}$ possesses a data-generating distribution $\mathcal{D}_s$ defined over a common space $\mathcal{Z}$. For each source $s \in \mathcal{S}$, consider a collection of independently sampled random variables $Z_1^s, \ldots, Z_{n_s}^s \sim \mathcal{D}_s$, and let $Z = \{Z_i^s \mid s \in \mathcal{S}, i \in [n_s]\}$ denote the set of samples across all sources. Additionally, let $\varepsilon = \{\varepsilon_i^s \mid s \in \mathcal{S}, i \in [n_s]\}$ be a corresponding set of independent Rademacher random variables.

The *empirical weighted Rademacher complexity* of a function class $\mathcal{F}$ on $\mathcal{Z}$ given multi-source data Z under fixed mixture weights $\lambda \in \Delta_{\mathcal{S}}$ and sample sizes $n_{\mathcal{S}}$ is

$$\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\mathcal{F}; Z) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s f(Z_i^s) \right| \;\middle|\; Z \right],$$

where the expectation is taken with respect to the Rademacher random variables ε. Additionally, the *weighted Rademacher complexity* of $\mathcal{F}$ under fixed mixture weights $\lambda \in \Delta_{\mathcal{S}}$ and sample sizes $n_{\mathcal{S}}$ is

$$\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s f(Z_i^s) \right| \right],$$

where the expectation is taken with respect to the multi-source random variables Z and the independent Rademacher random variables ε.

B.2 Hamming Distance & Entropy Integral

We provide additional details on the definition of the entropy integral introduced in Section 6.1.

Definition 8 (Hamming distance, covering number, and entropy integral)**.** Consider a policy class Π and a multi-source covariate set $x = \{x_i^s \mid s \in \mathcal{S}, i \in [n_s]\} \subset \mathcal{X}$ across sources $\mathcal{S}$ with source sample sizes $n_{\mathcal{S}}$. We define the following:

(a) the Hamming distance between any two policies $\pi_1, \pi_2 \in \Pi$ given multi-source covariate set x is

$$\mathrm{H}(\pi_1, \pi_2; x) := \frac{1}{\sum_{s \in \mathcal{S}} n_s} \sum_{s \in \mathcal{S}} \sum_{i=1}^{n_s} \mathbf{1}\{\pi_1(x_i^s) \neq \pi_2(x_i^s)\};$$

(b) an ϵ-cover of Π under the Hamming distance given covariate set x is any policy set $\Pi_\epsilon \subset \Pi$ such that for any $\pi \in \Pi$ there exists some $\pi' \in \Pi_\epsilon$ such that $\mathrm{H}(\pi, \pi'; x) \leqslant \epsilon$;

(c) the ϵ-covering number of Π under the Hamming distance given covariate set x is

$$N_{\mathrm{H}}(\epsilon, \Pi; x) := \min\{|\Pi_\epsilon| \mid \Pi_\epsilon \text{ is an } \epsilon\text{-cover of } \Pi \text{ w.r.t. } \mathrm{H}(\cdot, \cdot; x)\};$$

(d) the ϵ-covering number of Π under the Hamming distance is

$$N_{\mathrm{H}}(\epsilon, \Pi) := \sup\{N_{\mathrm{H}}(\epsilon, \Pi; x) \mid x \in \mathcal{X}_{\mathcal{S}}\},$$

where $\mathcal{X}_{\mathcal{S}}$ is the set of all covariate sets in $\mathcal{X}$ across sources $\mathcal{S}$ with arbitrary sample sizes;

(e) the entropy integral of Π is

$$\kappa(\Pi) := \int_0^1 \sqrt{\log N_{\mathrm{H}}(\epsilon^2, \Pi)} d\epsilon.$$

B.3 ℓ_λ Distance

Throughout our analysis, we will consider the function class

$$\mathcal{F}_\Pi := \{Q(\cdot, \pi) : \mathcal{Z} \to \mathbb{R} \mid \pi \in \Pi\},$$

where

$$Q(z_i^s; \pi) := \gamma_i^s(\pi(x_i^s))$$

for any covariate-score vector $z_i^s = (x_i^s, \gamma_i^s) \in \mathcal{Z} = \mathcal{X} \times \mathbb{R}^d$ and $\pi \in \Pi$, where $\gamma_i^s(a)$ is the a-th coordinate of the score vector γ_i^s.

Definition 9 (ℓ_λ distance and covering number)**.** Consider a policy class Π, function class $\mathcal{F}_\Pi$, and a multi-source covariate-score set $z = \{z_i^s \mid s \in \mathcal{S}, i \in [n_s]\} \subset \mathcal{Z}$ across sources $\mathcal{S}$ with source sample sizes $n_{\mathcal{S}}$ and source mixture weights λ. We define:

(a) the ℓ_λ distance with respect to function class $\mathcal{F}_\Pi$ between any two policies $\pi_1, \pi_2 \in \Pi$ given covariate-score set z is

$$\ell_\lambda(\pi_1, \pi_2; z) = \sqrt{\frac{\sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2} \big(Q(z_i^s; \pi_1) - Q(z_i^s; \pi_2)\big)^2}{\sup_{\pi_a, \pi_b \in \Pi} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2} \big(Q(z_i^s; \pi_a) - Q(z_i^s; \pi_b)\big)^2}};$$

(b) an ϵ-cover of Π under the ℓ_λ distance given covariate-score set z is any policy set Π_ϵ such that for any $\pi \in \Pi$ there exists some $\pi' \in \Pi_\epsilon$ such that $\ell_\lambda(\pi, \pi'; z) \leqslant \epsilon$;

(c) the ϵ-covering number of Π under the ℓ_λ distance given covariate-score set z is

$$N_{\ell_\lambda}(\epsilon, \Pi; z) := \min\{|\Pi_\epsilon| \mid \Pi_\epsilon \text{ is an } \epsilon\text{-cover of } \Pi \text{ w.r.t. } \ell_\lambda(\cdot, \cdot; z)\}.$$

The following lemma relates the covering numbers of the two policy distances we have defined.

Lemma 6. *Let $z = \{z_i^s \mid s \in \mathcal{S}, i \in [n_s]\} \subset \mathcal{Z}$ be a multi-source covariate-score set across sources $\mathcal{S}$ with source sample sizes $n_{\mathcal{S}}$ and source mixture weights λ. For any $\epsilon > 0$,*

$$N_{\ell_\lambda}(\epsilon, \Pi; z) \leqslant N_H(\epsilon^2, \Pi).$$

Proof. Fix $\epsilon > 0$. Without loss of generality, we assume $N_{\mathrm{H}}(\epsilon^2, \Pi) < \infty$, otherwise the result trivially holds. Let $\Pi_\epsilon = \{\pi_1, \ldots, \pi_{N_0}\}$ be a corresponding Hamming ϵ^2-cover of Π.

Consider any arbitrary $\pi \in \Pi$. By definition, there exists a $\pi' \in \Pi_\epsilon$ such that for any multi-source covariate set $\tilde{x} = \{\tilde{x}_i^s \mid s \in \mathcal{S}, i \in [\tilde{n}_c]\}$ with any given sample sizes $\tilde{n}_c > 0$ the following holds:

$$\mathrm{H}(\pi, \pi'; \tilde{x}) = \frac{1}{\tilde{n}} \sum_{s\in\mathcal{S}} \sum_{i=1}^{\tilde{n}_c} \mathbf{1}\{\pi(\tilde{x}_i^s) \neq \pi'(\tilde{x}_i^s)\} \leqslant \epsilon^2,$$

where $\tilde{n} = \sum_{s\in\mathcal{S}} \tilde{n}_c$. Using this pair of policies π, π' we generate an augmented data set $\tilde{z}$ from z as follows. Let m be a positive integer and define $\tilde{z}$ to be a collection of multiple copies of all covariate-score tuples $z_i^s \in z$, where each z_i^s appears

$$\tilde{n}_i^s = \left\lceil \frac{m \cdot \frac{\lambda_s^2}{n_s^2} \big(Q(z_i^s; \pi) - Q(z_i^s; \pi')\big)^2}{\sup_{\pi_a, \pi_b} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2} \big(Q(z_i^s; \pi_a) - Q(z_i^s; \pi_b)\big)^2} \right\rceil$$

times in $\tilde{z}$. Therefore, the source sample sizes in this augmented data set are $\tilde{n}_c = \sum_{i=1}^{n_s} \tilde{n}_i^s$ and the total sample size is $\tilde{n} = \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \tilde{n}_i^s$. The total sample size is bounded as

$$\begin{aligned}
\tilde{n} &= \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \left\lceil \frac{m \cdot \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi) - Q(z_i^s;\pi')\big)^2}{\sup_{\pi_a,\pi_b} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi_a) - Q(z_i^s;\pi_b)\big)^2} \right\rceil \\
&\leqslant \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \left(\frac{m \cdot \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi) - Q(z_i^s;\pi')\big)^2}{\sup_{\pi_a,\pi_b} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi_a) - Q(z_i^s;\pi_b)\big)^2} + 1 \right) \\
&\leqslant \frac{m \cdot \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi) - Q(z_i^s;\pi')\big)^2}{\sup_{\pi_a,\pi_b} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi_a) - Q(z_i^s;\pi_b)\big)^2} + n \leqslant m + n.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\mathrm{H}(\pi,\pi';\tilde{z}) &= \frac{1}{\tilde{n}} \sum_{s\in\mathcal{S}} \sum_{i=1}^{\tilde{n}_c} \mathbf{1}\{\pi(x_i^s) \neq \pi'(x_i^s)\} \\
&= \frac{1}{\tilde{n}} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \tilde{n}_i^s \cdot \mathbf{1}\{\pi(x_i^s) \neq \pi'(x_i^s)\} \\
&\geqslant \frac{1}{\tilde{n}} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{m \cdot \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi) - Q(z_i^s;\pi')\big)^2}{\sup_{\pi_a,\pi_b} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi_a) - Q(z_i^s;\pi_b)\big)^2} \mathbf{1}\{\pi(x_i^s) \neq \pi'(x_i^s)\} \\
&= \frac{m}{\tilde{n}} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi) - Q(z_i^s;\pi')\big)^2}{\sup_{\pi_a,\pi_b} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\big(Q(z_i^s;\pi_a) - Q(z_i^s;\pi_b)\big)^2} \\
&\geqslant \frac{m}{m+n} \ell_\lambda^2(\pi,\pi';z).
\end{aligned}$$

This implies that

$$\ell_\lambda(\pi,\pi';z) \leqslant \sqrt{\frac{m+n}{m}\mathrm{H}(\pi,\pi';\tilde{z})} \leqslant \sqrt{1 + \frac{n}{m}} \cdot \epsilon.$$

Letting $m \to \infty$ yields $\ell_\lambda(\pi,\pi';z) \leqslant \epsilon$. This establishes that for any $\pi \in \Pi$, there exists a $\pi' \in \Pi_\epsilon$ such that $\ell_\lambda(\pi,\pi';z) \leqslant \epsilon$, and thus $N_{\ell_\lambda}(\epsilon,\Pi;z) \leqslant N_{\mathrm{H}}(\epsilon^2,\Pi)$. □

B.4 Mixture Skewness

An important quantity that arises in our analysis is

$$\sum_{s\in\mathcal{S}} \frac{\lambda_s^2}{\bar{n}_s} = \mathop{\mathbb{E}}_{s\sim\lambda}\left[\frac{\lambda_s}{\bar{n}_s}\right],$$

which captures a measure of the imbalance of the mixture weight distribution λ relative to the empirical distribution of samples $\bar{n}$. The following result makes this interpretation more clear:

$$\begin{aligned}
\sum_{s\in\mathcal{S}} \frac{\lambda_s^2}{\bar{n}_s} &= \sum_{s\in\mathcal{S}} \frac{\lambda_s^2}{\bar{n}_s} + \sum_{s\in\mathcal{S}} \bar{n}_s - 2\sum_{s\in\mathcal{S}} \lambda_s + 1 \\
&= \sum_{s\in\mathcal{S}} \left(\frac{\lambda_s^2}{\bar{n}_s} + \frac{\bar{n}_s^2}{\bar{n}_s} - \frac{2\lambda_s \bar{n}_s}{\bar{n}_s} \right) + 1 \\
&= \sum_{s\in\mathcal{S}} \frac{(\lambda_s - \bar{n}_s)^2}{\bar{n}_s} + 1 \\
&= \chi^2(\lambda || \bar{n}) + 1,
\end{aligned}$$

where $\chi^2(\lambda||\bar{n})$ is the chi-squared divergence from $\bar{n}$ to λ. Following Mohri et al. (2019), this quantity is referred to as skewness.

Definition 10 (Skewness)**.** The *skewness* of a given set of mixture weights λ relative to the empirical distribution of samples $\bar{n} := (n_s/n)_{s\in\mathcal{S}}$ is

$$\mathfrak{s}(\lambda\|\bar{n}) := 1 + \chi^2(\lambda||\bar{n}),$$

where $\chi^2(\lambda||\bar{n})$ is the chi-squared divergence of λ from $\bar{n}$. Additionally, the *mixture-agnostic skewness* is

$$\mathfrak{s}(\Lambda\|\bar{n}) := \max_{\lambda\in\Lambda} \mathfrak{s}(\lambda\|\bar{n}).$$

C BOUNDING MIXTURE REGRET

C.1 Preliminaries

C.1.1 Function Classes

As mentioned in Appendix B.3, the function class we will be considering in our analysis is

$$\mathcal{F}_\Pi := \{Q(\cdot;\pi) : \mathcal{Z} \to \mathbb{R} \mid \pi \in \Pi\}, \tag{7}$$

where

$$Q(z_i^s;\pi) := \gamma_i^s(\pi(x_i^s)), \tag{8}$$

for any covariate-score vector $z_i^s = (x_i^s, \gamma_i^s) \in \mathcal{Z} = \mathcal{X} \times \mathbb{R}^d$ and $\pi \in \Pi$, where $\gamma_i^s(a)$ is the a-th coordinate of the score vector γ_i^s. It will also be useful to consider the Minkowski difference of $\mathcal{F}_\Pi$ with itself,

$$\Delta\mathcal{F}_\Pi := \mathcal{F}_\Pi - \mathcal{F}_\Pi = \{\Delta(\cdot;\pi_a,\pi_b) : \mathcal{Z} \to \mathbb{R} \mid \pi_a,\pi_b \in \Pi\}, \tag{9}$$

where

$$\Delta(z_i^s;\pi_a,\pi_b) := Q(z_i^s;\pi_a) - Q(z_i^s;\pi_b) = \gamma_i^s(\pi_a(x_i^s)) - \gamma_i^s(\pi_b(x_i^s)), \tag{10}$$

for any $z_i^s = (x_i^s, \gamma_i^s) \in \mathcal{Z}$ and $\pi_a, \pi_b \in \Pi$.

C.1.2 Policy Value Estimators

Augmented Inverse Propensity Weighted Scores As discussed in Section 4.2, we used propensity-weighted scores to estimate the policy values. We considered the construction of the oracle local AIPW scores

$$\Gamma_i^s(a) = \mu_s(X_i^s;a) + \big(Y_i^s(a) - \mu_s(X_i^s;a)\big)\omega_s(X_i^s;a)\mathbf{1}\{A_i^s = a\}$$

for each $a \in \mathcal{A}$, constructed from the available observable samples (X_i^s, A_i^s, Y_i^s) taken from the partially observable counterfactual samples $(X_i^s, A_i^s, Y_i^s(a_1), \ldots, Y_i^s(a_d)) \sim \bar{\mathcal{D}}_s$ for $i \in [n_s]$. Similarly, we discussed the construction of the approximate local AIPW scores

$$\hat{\Gamma}_i^s(a) = \hat{\mu}_s(X_i^s;a) + \big(Y_i^s - \hat{\mu}_s(X_i^s;a)\big)\hat{\omega}_s(X_i^s;a)\mathbf{1}\{A_i^s = a\}$$

for each $a \in \mathcal{A}$, given fixed estimates $\hat{\mu}_s$ and $\hat{\omega}_s$ of μ_s and ω_s, respectively. In practice, we use cross-fitting to make the estimates fixed and independent relative to the data on which they are evaluated (see further discussion in Appendix E). Note that only this second set of AIPW scores can be constructed from the observed data. The first set is only "constructed" for analytic purposes for our proofs.

Policy Value Estimates and Policy Value Difference Estimates Using the source data and the constructed AIPW scores, we let $Z_i^s = (X_i^s, \Gamma_i^s(a_1), \ldots, \Gamma_i^s(a_d))$ and $\hat{Z}_i^s = (X_i^s, \hat{\Gamma}_i^s(a_1), \ldots, \hat{\Gamma}_i^s(a_d))$ for each $i \in [n_s]$. Using these constructed vectors, we define the oracle and approximate mixture policy value estimates of $Q_\lambda(\pi)$, respectively, as

$$\tilde{Q}_\lambda(\pi) := \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Gamma_i^s(\pi(X_i^s)) = \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} Q(Z_i^s;\pi),$$

$$\hat{Q}_\lambda(\pi) := \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \hat{\Gamma}_i^s(\pi(X_i^s)) = \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} Q(\hat{Z}_i^s;\pi),$$

for any $\pi \in \Pi$, where we use the function class defined in Equation 8 for these representations, which we will use throughout our proofs for notational convenience. It will also be very useful to define the following corresponding *policy value difference* quantities:

$$\begin{aligned}
\Delta_\lambda(\pi_a, \pi_b) &:= Q_\lambda(\pi_a) - Q_\lambda(\pi_b), \\
\tilde{\Delta}_\lambda(\pi_a, \pi_b) &:= \tilde{Q}_\lambda(\pi_a) - \tilde{Q}_\lambda(\pi_b) = \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Delta(Z_i^s; \pi_a, \pi_b), \\
\hat{\Delta}_\lambda(\pi_a, \pi_b) &:= \hat{Q}_\lambda(\pi_a) - \hat{Q}_\lambda(\pi_b) = \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Delta(\hat{Z}_i^s; \pi_a, \pi_b),
\end{aligned}$$

for any $\pi_a, \pi_b \in \Pi$, where we use the function class defined in Equation (10) for these representations, which we will use throughout our proofs for notational convenience.

Unbiased Estimates The following result can be used to readily show that the oracle estimator for the mixture policy value is unbiased.

Lemma 7. *Suppose Assumption 1 holds. For any $\pi \in \Pi$,*

$$\mathop{\mathbb{E}}_{Z^s \sim \bar{\mathcal{D}}_s} [\Gamma^s(\pi(X^s))] = Q_s(\pi) \quad \textit{and} \quad \mathop{\mathbb{E}}_{s \sim \lambda} \mathop{\mathbb{E}}_{Z^s \sim \bar{\mathcal{D}}_s} [\Gamma^s(\pi(X^s))] = Q_\lambda(\pi),$$

where $Z^s = (X^s, A^s, Y^s(a_1), \ldots, Y^s(a_d)) \sim \bar{\mathcal{D}}_s$.

Proof. First, observe that for any $a \in \mathcal{A}$,

$$\begin{aligned}
\Gamma^s(a) &= \mu_s(X^s; a) + (Y^s(A^s) - \mu_s(X^s; a))\, \omega_s(X^s; a) \mathbf{1}\{A^s = a\} \\
&= \mu_s(X^s; a) + (Y^s(a) - \mu_s(X^s; a))\, \omega_s(X^s; a) \mathbf{1}\{A^s = a\}
\end{aligned}$$

and so for any $(X^s, A^s, \vec{Y}^s) = (X^s, A^s, Y^s(a_1), \ldots, Y^s(a_d)) \sim \bar{\mathcal{D}}_s$,

$$\begin{aligned}
\mathbb{E}_{A^s, \vec{Y}^s}\left[\Gamma^s(a) \mid X^s\right] &= \mu_s(X^s; a) + \mathbb{E}_{A^s, \vec{Y}^s}\left[(Y^s(a) - \mu_s(X^s; a))\, \omega_s(X^s; a) \mathbf{1}\{A^s = a\} \mid X^s\right] \\
&= \mu_s(X^s; a) + \mathbb{E}_{\vec{Y}^s}\left[Y^s(a) - \mu_s(X^s; a) \mid X^s\right] \cdot \mathbb{E}_{A^s}\left[\omega_s(X^s; a) \mathbf{1}\{A^s = a\} \mid X^s\right] \\
&= \mu_s(X^s; a) + \mathbb{E}_{\vec{Y}^s}\left[Y^s(a) - \mu_s(X^s; a) \mid X^s\right] \cdot \omega_s(X^s; a) e_s(X^s; a) \\
&= \mu_s(X^s; a) + \mathbb{E}_{\vec{Y}^s}\left[Y^s(a) \mid X^s\right] - \mu_s(X^s; a) \\
&= \mathbb{E}_{\vec{Y}^s}\left[Y^s(a) \mid X^s\right].
\end{aligned}$$

The second equality follows from the unconfoundedness assumption stated in Assumption 1. This immediately implies that

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z^s \sim \bar{\mathcal{D}}_s} [\Gamma^s(\pi(X^s))] &= \mathop{\mathbb{E}}_{Z^s \sim \bar{\mathcal{D}}_s} \left[\sum_{a \in \mathcal{A}} \mathbf{1}\{\pi(X^s) = a\} \Gamma^s(a)\right] \\
&= \mathbb{E}_{X^s}\left[\sum_{a \in \mathcal{A}} \mathbf{1}\{\pi(X^s) = a\}\, \mathbb{E}_{A^s, \vec{Y}^s}\left[\Gamma^s(a) \mid X^s\right]\right] \\
&= \mathbb{E}_{X^s}\left[\sum_{a \in \mathcal{A}} \mathbf{1}\{\pi(X^s) = a\}\, \mathbb{E}_{A^s, \vec{Y}^s}\left[Y^s(a) \mid X^s\right]\right] \\
&= \mathbb{E}_{X^s}\left[\mathbb{E}_{A^s, \vec{Y}^s}\left[Y^s(\pi(X^s)) \mid X^s\right]\right] \\
&= \mathop{\mathbb{E}}_{Z^s \sim \bar{\mathcal{D}}_s}\left[Y^s(\pi(X^s)\right] \\
&= Q_s(\pi),
\end{aligned}$$

and thus,

$$\mathop{\mathbb{E}}_{s \sim \lambda} \mathop{\mathbb{E}}_{Z^s \sim \bar{\mathcal{D}}_s} [\Gamma^s(\pi(X^s))] = \mathop{\mathbb{E}}_{s \sim \lambda} \mathop{\mathbb{E}}_{Z^s \sim \mathcal{D}_s} \left[Y^s(\pi(X^s))\right] = Q_\lambda(\pi).$$

□

C.1.3 Data-generating Distributions and Sufficient Statistics

Note that the contexts and AIPW scores are *sufficient statistics* for the corresponding oracle and approximate estimators of the policy values. Moreover, our results will mostly directly depend on properties of the sufficient statistics. Therefore, it will sometimes be simpler for our analysis to work with the distribution of the sufficient statistics directly. For any $(X^s, A^s, Y^s(a_1), \ldots, Y^s(a_d)) \sim \bar{\mathcal{D}}_s$, let $(X^s, \Gamma^s(a_1), \ldots, \Gamma^s(a_d))$ be the sufficient statistic of contexts and oracle AIPW scores, and denote its induced distribution as $\tilde{\mathcal{D}}_s$ defined over $\mathcal{Z} = \mathcal{X} \times \mathbb{R}^d$.

For simplicity and without loss of generality, when proving results that only involve the contexts and AIPW scores, we will assume the data is sampled from the distributions of the sufficient statistics, e.g., $Z^s \sim \tilde{\mathcal{D}}_s$. When we have a discussion involving constructing the AIPW scores from the observable data, we will be more careful about the source distributions and typically assume the data is sampled from the complete data-generating distributions, e.g., $Z^s \sim \bar{\mathcal{D}}_s$.

C.2 Bounding Weighted Rademacher Complexity

First, to simplify our analysis, we can easily bound the weighted Rademacher complexity of $\Delta\mathcal{F}_\Pi$ by that of $\mathcal{F}_\Pi$ as follows.

Lemma 8.

$$\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\Delta\mathcal{F}_\Pi) \leqslant 2\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\mathcal{F}_\Pi).$$

Proof.

$$\begin{aligned}
\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\Delta\mathcal{F}_\Pi) &= \mathbb{E}\left[\sup_{\pi_a, \pi_b \in \Pi} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s \Delta(Z_i^s; \pi_a, \pi_b) \right|\right] \\
&= \mathbb{E}\left[\sup_{\pi_a, \pi_b \in \Pi} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s \Big(Q(Z_i^s; \pi_a) - Q(Z_i^s; \pi_b)\Big) \right|\right] \\
&\leqslant \mathbb{E}\left[\sup_{\pi_a, \pi_b \in \Pi} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s Q(Z_i^s; \pi_a) \right| + \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s Q(Z_i^s; \pi_b) \right|\right] \\
&= 2\,\mathbb{E}\left[\sup_{\pi \in \Pi} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s Q(Z_i^s; \pi) \right|\right] \\
&= 2\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\mathcal{F}_\Pi).
\end{aligned}$$

□

Therefore, we can simply focus on bounding the weighted Rademacher complexity of $\mathcal{F}_\Pi$.

Proposition 1. *Suppose Assumptions 1 and 2 hold. Then,*

$$\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\mathcal{F}_\Pi) \leqslant (14 + 6\kappa(\Pi)) \sqrt{\frac{V_{\lambda, n_{\mathcal{S}}}}{n}} + o\left(\sqrt{\frac{\mathfrak{s}(\lambda \| \bar{n})}{n}}\right),$$

where

$$V_{\lambda, n_{\mathcal{S}}} = \sup_{\pi_a, \pi_b \in \Pi} \sum_{s \in \mathcal{S}} \frac{\lambda_s^2}{\bar{n}_s} \mathop{\mathbb{E}}_{Z^s \sim \tilde{\mathcal{D}}_s} [\Delta^2(Z^s; \pi_a, \pi_b)].$$

Proof. We follow a chaining argument to bound the weighted Rademacher complexity of $\mathcal{F}_\Pi$.

Constructing the policy approximation chain. First, for each source $s \in \mathcal{S}$, let $Z_1^c, \ldots, Z_{n_s}^s$ be n_s independent random variables sampled from $\tilde{\mathcal{D}}_s$, where each $Z_i^s = (X_i^s, \vec{\Gamma}_i^s) \in \mathcal{Z} = \mathcal{X} \times \mathbb{R}^d$. Additionally, let $Z = \{Z_i^s \mid s \in \mathcal{S}, i \in [n_c]\}$ represent the corresponding set of samples across all sources.

Next, set $K = \lceil \log_2 n \rceil$. We will construct a sequence $\{\Psi_k : \Pi \to \Pi\}_{k=0}^K$ of policy approximation operators that satisfies the following properties. For any $k = 0, \ldots, K$, (P1) $\max_{\pi \in \Pi} \ell_\lambda(\Psi_{k+1}(\pi), \Psi_k(\pi); Z) \leqslant \epsilon_k := 2^{-k}$

(P2) $|\{\Psi_k(\pi) \mid \pi \in \Pi\}| \leqslant N_{\ell_\lambda}(\epsilon_k, \Pi; Z)$

We use the notational shorthand that $\Psi_{K+1}(\pi) = \pi$ for any $\pi \in \Pi$. We will construct the policy approximation chain via a backward recursion scheme. First, let Π_k denote the smallest ϵ_k-covering set of Π under the ℓ_λ distance given data Z. Note, in particular, that $|\Pi_0| = 1$ since the ℓ_λ distance is never more than 1 and so any single policy is enough to 1-cover all policies in Π. Then, the backward recursion is as follows: for any $\pi \in \Pi$,

1. define $\Psi_K(\pi) = \arg\min_{\pi' \in \Pi_K} \ell_\lambda(\pi, \pi'; Z)$;
2. for each $k = K-1, \ldots, 1$, define $\Psi_k(\pi) = \arg\min_{\pi' \in \Pi_k} \ell_\lambda(\Psi_{k+1}(\pi), \pi'; Z)$;
3. define $\Psi_0(\pi) \equiv 0$.

Note that although $\Psi_0(\pi)$ is not in Π, it can still serve as a 1-cover of Π since the ℓ_λ distance is always bounded by 1. Before proceeding, we check that each of the stated desired properties of the constructed operator chain is satisfied:

(P1) Pick any $\pi \in \Pi$. Clearly, $\Psi_{k+1}(\pi) \in \Pi$. Then, by construction of Π_k, there exists a $\pi' \in \Pi_k$ such that $\ell_\lambda(\Psi_{k+1}(\pi), \pi'; Z) \leqslant \epsilon_k$. Therefore, by construction of $\Psi_k(\pi)$, we have $\ell_\lambda(\Psi_{k+1}(\pi), \Psi_k(\pi); Z) \leqslant \ell_\lambda(\Psi_{k+1}(\pi), \pi'; Z) \leqslant \epsilon_k$.

(P2) By construction of Ψ_k, we have that $\Psi_k(\pi) \in \Pi_k$ for every $\pi \in \Pi$. Therefore, $|\{\Psi_k(\pi) \mid \pi \in \Pi\}| \leqslant |\Pi_k| = N_{\ell_\lambda}(\epsilon_k, \Pi; Z)$.

Thus, the constructed chain satisfies the desired properties. Next, we observe that since $\Psi_0(\pi) \equiv 0$, we have that $Q(Z_i^s; \Psi_0(\pi)) = 0$ and

$$\begin{aligned}
Q(Z_i^s; \pi) &= Q(Z_i^s; \pi) - Q(Z_i^s; \Psi_0(\pi)) \\
&= Q(Z_i^s; \pi) - Q(Z_i^s; \Psi_K(\pi)) + \sum_{k=1}^{K} Q(Z_i^s; \Psi_k(\pi) - Q(Z_i^s; \Psi_{k-1}(\pi)) \\
&= \Delta(Z_i^s; \pi, \Psi_K(\pi)) + \sum_{k=1}^{K} \Delta(Z_i^s; \Psi_k(\pi), \Psi_{k-1}(\pi))
\end{aligned}$$

Therefore, we can decompose the weighted Rademacher complexity of $\mathcal{F}_\Pi$ as follows:

$$\begin{aligned}
\mathfrak{R}_{\lambda, n_{\mathcal{S}}}(\mathcal{F}_\Pi) \leqslant{}& \mathbb{E}\left[\sup_{\pi \in \Pi} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s \Delta(Z_i^s; \pi, \Psi_K(\pi)) \right| \right] \\
&+ \mathbb{E}\left[\sup_{\pi \in \Pi} \left| \sum_{s \in \mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s \left(\sum_{k=1}^{K} \Delta(Z_i^s; \Psi_k(\pi), \Psi_{k-1}(\pi)) \right) \right| \right]
\end{aligned}$$

We will obtain bounds separately for these two terms, which we refer to as the *negligible regime* term and the *effective regime* term, respectively.

Bounding the negligible regime. First, we define the following quantities:

$$B_{\lambda, n_{\mathcal{S}}}(Z) := \sup_{\pi_a, \pi_b \in \Pi} \sum_{s \in \mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2} \Delta^2(Z_i^s; \pi_a, \pi_b) \quad \text{and} \quad B_{\lambda, n_{\mathcal{S}}} := \mathbb{E}\left[B_{\lambda, n_{\mathcal{S}}}(Z)\right]$$

Given any realization of independent Rademacher random variables $\varepsilon = \{\varepsilon_i^s \mid s \in \mathcal{S}, i \in [n_s]\}$ and multi-source data Z, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Delta(Z_i^s;\pi,\Psi_K(\pi)\right| &\leqslant \sqrt{\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}(\varepsilon_i^s)^2}\cdot\sqrt{\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\frac{\lambda_s^2}{n_s^2}\Delta^2(Z_i^s;\pi,\Psi_K(\pi)}\\ &= \sqrt{n}\cdot\sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\ell_2(\pi,\Psi_K(\pi);Z)\\ &\leqslant \sqrt{nB_{\lambda,n_{\mathcal{S}}}(Z)}\epsilon_K\\ &\leqslant \sqrt{\frac{B_{\lambda,n_{\mathcal{S}}}(Z)}{n}}.\end{aligned}$$

Then, by Jensen's inequality,

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Delta(Z_i^s;\pi,\Psi_K(\pi)\right|\right]\leqslant\mathbb{E}\left[\sqrt{\frac{B_{\lambda,n_{\mathcal{S}}}(Z)}{n}}\right]\leqslant\sqrt{\frac{B_{\lambda,n_{\mathcal{S}}}}{n}}.$$

Bounding the effective regime. For any $k\in[K]$, let

$$t_{k,\delta}=\sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\epsilon_k\tau_{k,\delta}$$

where $\tau_{k,\delta}>0$ is some constant to be specified later. By Hoeffding's inequality (in Lemma 2),

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\right|>t_{k,\delta}\;\middle|\;Z\right)\\ &\leqslant 2\exp\left(-\frac{t_{k,\delta}^2}{2\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\frac{\lambda_s^2}{n_s^2}\Delta^2(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))}\right)\\ &= 2\exp\left(-\frac{t_{k,\delta}^2}{2B_{\lambda,n_{\mathcal{S}}}(Z)\ell_2^2(\Psi_k(\pi),\Psi_{k-1}(\pi);Z)}\right)\\ &\leqslant 2\exp\left(-\frac{t_{k,\delta}^2}{2B_{\lambda,n_{\mathcal{S}}}(Z)\epsilon_{k-1}^2}\right)\\ &= 2\exp\left(-\frac{t_{k,\delta}^2}{8B_{\lambda,n_{\mathcal{S}}}(Z)\epsilon_k^2}\right)\\ &= 2\exp\left(-\frac{\tau_{k,\delta}^2}{8}\right).\end{aligned}$$

Here, we used the fact that $\epsilon_{k-1}=2\epsilon_k$. Setting

$$\tau_{k,\delta}=\sqrt{8\log\left(\frac{\pi^2k^2}{3\delta}N_{\ell_\lambda}(\epsilon_k,\Pi;Z)\right)}$$

and applying a union bound over the policy space, we obtain

$$\begin{aligned}&\mathbb{P}\left(\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\right|>t_{k,\delta}\;\middle|\;Z\right)\\ &\leqslant 2\left|\Pi_k\right|\cdot\exp\left(-\frac{\tau_{k,\delta}^2}{8}\right)\\ &\leqslant 2N_{\ell_2}(\epsilon_k,\Pi;Z)\cdot\exp\left(-\frac{\tau_{k,\delta}^2}{8}\right)\\ &= \frac{6\delta}{\pi^2k^2}.\end{aligned}$$

By a further union bound over $k \in [K]$, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Big(\sum_{k=1}^{K}\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\Big)\right| > \sum_{k=1}^{K} t_{k,\delta}\ \Big|\ Z\right)\\
&\leq \sum_{k=1}^{K}\mathbb{P}\left(\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\right| > t_{k,\delta}\ \Big|\ Z\right)\\
&\leq \sum_{k=1}^{K}\frac{6\delta}{\pi^2k^2} \leq \delta.
\end{aligned}$$

Therefore, given multi-source data Z, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Big(\sum_{k=1}^{K}\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\Big)\right|\\
&\leq \sum_{k=1}^{K} t_{k,\delta}\\
&= \sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\sum_{k=1}^{K}\epsilon_k\sqrt{8\log\left(\frac{\pi^2k^2}{3\delta}N_{\ell_2}(\epsilon_k,\Pi;Z)\right)}\\
&= \sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\sum_{k=1}^{K}\epsilon_k\left(\sqrt{8\log\frac{\pi^2}{3\delta}+16\log k+8\log N_{\ell_2}(\epsilon_k,\Pi;Z)}\right)\\
&\leq \sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\sum_{k=1}^{K}\epsilon_k\left(\sqrt{8\log\frac{\pi^2}{3\delta}}+\sqrt{16\log k}+\sqrt{8\log N_{\ell_2}(\epsilon_k,\Pi;Z)}\right)\\
&\leq \sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\sum_{k=1}^{K}\epsilon_k\left(\sqrt{8\log(4/\delta)}+\sqrt{16\log k}+\sqrt{8\log N_{\mathrm{H}}(\epsilon_k,\Pi)}\right)\\
&\leq \sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\left(\sqrt{8\log(4/\delta)}+2+\sqrt{8}\sum_{k=1}^{\infty}\epsilon_k\sqrt{\log N_{\mathrm{H}}(\epsilon_k,\Pi)}\right)\\
&\leq \sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\left(\sqrt{8\log(4/\delta)}+2+\sqrt{8}\kappa(\Pi)\right).
\end{aligned}$$

Next, we turn this high-probability bound into a bound on the conditional expectation. First, let $F_R(\cdot \mid Z)$ be the cumulative distribution of the random variable

$$R \coloneqq \sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Big(\sum_{k=1}^{K}\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\Big)\right|$$

conditional on Z. Above, we have shown that

$$1-F_R\left(\sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\left(\sqrt{8\log(4/\delta)}+2+\sqrt{8}\kappa(\Pi)\right)\ \Big|\ Z\right)\leq\delta.$$

For any non-negative integer l, let $\Delta_l = \sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\cdot\left(\sqrt{8\log(4/\delta_l)}+2+\sqrt{8}\kappa(\Pi)\right)$ where $\delta_l = 2^{-l}$. Since R is non-negative, we can compute and upper bound the conditional expectation of R given Z as follows:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Bigg(\sum_{k=1}^{K}\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\Bigg)\right|\;\Bigg|\;Z\right]\\
&=\int_0^\infty(1-F_R(r|Z))\,dr\\
&\leqslant\sum_{l=0}^{\infty}(1-F_R(\Delta_l|Z))\cdot\Delta_l\\
&\leqslant\sum_{l=0}^{\infty}\delta_l\cdot\Delta_l\\
&=\sum_{l=0}^{\infty}2^{-l}\cdot\sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\left(\sqrt{8(l+2)\log 2}+2+\sqrt{8}\kappa(\Pi)\right)\\
&\leqslant\sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\left(4\sqrt{8\log 2}+4+2\sqrt{8}\kappa(\Pi)\right)\\
&\leqslant\sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\left(14+6\kappa(\Pi)\right).
\end{aligned}$$

Taking the expectation with respect to Z and using Jensen's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Bigg(\sum_{k=1}^{K}\Delta(Z_i^s;\Psi_k(\pi),\Psi_{k-1}(\pi))\Bigg)\right|\right]\\
&\leqslant(14+6\kappa(\Pi))\,\mathbb{E}\left[\sqrt{B_{\lambda,n_{\mathcal{S}}}(Z)}\right]\\
&\leqslant(14+6\kappa(\Pi))\sqrt{B_{\lambda,n_{\mathcal{S}}}}.
\end{aligned}$$

Refining the upper bound. We see that the quantity $B_{\lambda,n_{\mathcal{S}}}$ appears in the bounds for the negligible and effective regimes. Therefore, the task is to find an appropriate bound for this quanitity. One could easily bound $B_{\lambda,n_{\mathcal{S}}}$ using worst-case bounds on the AIPW element. Instead, we use Lemma 5 to get a more refined bound on $B_{\lambda,n_{\mathcal{S}}}$.

To make use this result, we first identify the set of independent random variables $\tilde{Z}_i^s = T(Z_i^s) = (X_i^s, \frac{\lambda_s}{n_s}\vec{\Gamma}_i^s)$ for each $s \in \mathcal{S}$ and $i \in [n_s]$ and the function class $\mathcal{H} = \{\Delta(\cdot;\pi_a,\pi_b) \mid \pi_a,\pi_b \in \Pi\}$. We also identify the Lipschitz function $\varphi : u \mapsto u^2$ defined over the set $\mathcal{U} \subset \mathbb{R}$ containing all possible outputs of any $\Delta(\cdot;\pi_a,\pi_b)$ given any realization of $\tilde{Z}_i^s$ for any $s \in \mathcal{S}$ as input. To further capture this domain, note that by the boundedness and overlap assumptions in Assumption 1, it is easy to verify that there exists some $U > 0$ for all $s \in \mathcal{S}$ such that $|\Gamma_i^s(a)| \leqslant U$ for any $a \in \mathcal{A}$ and any realization of Z_i^s. This implies that

$$|\Delta(\tilde{Z}_i^s;\pi_a,\pi_b)| = \frac{\lambda_s}{n_s}\left|\Gamma_i^s(\pi_a(X_i^s)) - \Gamma_i^s(\pi_b(X_i^s))\right| \leqslant 2U\frac{\lambda_s}{n_s},$$

for any realization of $\tilde{Z}_i^s$ and any $\pi_a,\pi_b \in \Pi$. Moreover, note that

$$\frac{\lambda_s}{n_s} \leqslant \sqrt{\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{n_s^2}} \leqslant \frac{1}{\sqrt{\min_{s\in\mathcal{S}} n_s}}\sqrt{\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{n_s}} = \frac{1}{\sqrt{\min_{s\in\mathcal{S}} n_s}}\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}} =: s_{\lambda,n_{\mathcal{S}}} \tag{11}$$

Therefore, $\mathcal{U} \subset [-2Us_{\lambda,n_{\mathcal{S}}}, 2Us_{\lambda,n_{\mathcal{S}}}]$, and thus, for any $u, v \in \mathcal{U}$, we have that

$$|\varphi(u) - \varphi(v)| = |u^2 - v^2| = |u + v| \cdot |u - v| \leqslant 4Us_{\lambda,n_{\mathcal{S}}}\,|u - v|\,.$$

Therefore, $L = 4Us_{\lambda,n_{\mathcal{S}}}$ is a valid Lipschitz constant for φ. Then, through these identifications, Lemma 5 guarantees the following upper bound:

$$\begin{aligned}
B_{\lambda,n_{\mathcal{S}}} &= \mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\frac{\lambda_s^2}{n_s^2}\Delta^2(Z_i^s;\pi_a,\pi_b)\right] \\
&= \mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\varphi\circ\Delta(\tilde{Z}_i^s;\pi_a,\pi_b)\right] \\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\mathbb{E}\left[\varphi\circ\Delta(\tilde{Z}_i^s;\pi_a,\pi_b)\right] + 16Us_{\lambda,n_{\mathcal{S}}}\,\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\varepsilon_i^s\Delta(\tilde{Z}_i^s;\pi_a,\pi_b)\right|\right] \\
&= \sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\frac{\lambda_s^2}{n_s^2}\mathbb{E}\left[\Delta^2(Z_i^s;\pi_a,\pi_b)\right] + 16Us_{\lambda,n_{\mathcal{S}}}\,\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^s\Delta(Z_i^s;\pi_a,\pi_b)\right|\right] \\
&= \sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{n_s}\mathbb{E}\left[\Delta^2(Z_i^s;\pi_a,\pi_b)\right] + 16U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\Delta\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}} \\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{n_s}\mathbb{E}\left[\Delta^2(Z_i^s;\pi_a,\pi_b)\right] + 32U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}} \\
&= \frac{V_{\lambda,n_{\mathcal{S}}}}{n} + 32U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}}.
\end{aligned}$$

Establishing the skewness bound. Before proceeding, note that by the local data size scaling assumption stated in Assumption 2, $n_s = \Omega(\nu_c(n))$ for some increasing function ν_c for any $s \in \mathcal{S}$. This immediately implies that $s_{\lambda,n_{\mathcal{S}}}$ is dominated as

$$s_{\lambda,n_{\mathcal{S}}} = \frac{1}{\sqrt{\min_{s\in\mathcal{S}} n_s}}\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}} \leqslant o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).$$

Combining results. Thus, combining the bounds for the negligible and effective regime and including the refined bound and the skewness bound, we have

$$\begin{aligned}
&\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) \\
&\leqslant \sqrt{\frac{B_{\lambda,n_{\mathcal{S}}}}{n}} + (14+6\kappa(\Pi))\sqrt{B_{\lambda,n_{\mathcal{S}}}} \\
&\leqslant \sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n^2} + 32U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)\frac{s_{\lambda,n_{\mathcal{S}}}}{n}} + (14+6\kappa(\Pi))\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n} + 32U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}}} \\
&\leqslant \sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n^2}} + \sqrt{32U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)\frac{s_{\lambda,n_{\mathcal{S}}}}{n}} + (14+6\kappa(\Pi))\left(\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + \sqrt{32U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}}}\right) \\
&\leqslant (14+6\kappa(\Pi))\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + \sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n^2}} + \mathcal{O}\left(\sqrt{\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}}}\right) \qquad (12)
\end{aligned}$$

This gives an upper bound on $\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)$ in terms of itself. To decouple this dependence, we express

$$\begin{aligned}
\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) &\leqslant \mathcal{O}\left(\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}}\right) + \mathcal{O}\left(\sqrt{\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}}}\right) \\
&\leqslant A_1\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + A_2\sqrt{\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}}} \qquad (13)
\end{aligned}$$

for some constants A_1, A_2, and we split this inequality into the following two exhaustive cases.

Case 1: $A_2\sqrt{s_{\lambda,n_{\mathcal{S}}}} \leqslant \frac{1}{2}\sqrt{\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)}$

In this case, we can bound the second term in the right-hand side of inequality (13) to get

$$\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) \leqslant A_1\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + \frac{\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)}{2},$$

and so

$$\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) \leqslant 2A_1\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}}. \tag{14}$$

Moreover,

$$\begin{aligned} V_{\lambda,n_{\mathcal{S}}} &= \sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{\bar{n}_s}\mathop{\mathbb{E}}_{Z^s\sim\tilde{\mathcal{D}}_s}[\Delta^2(Z^s;\pi_a,\pi_b)] \\ &\leqslant \sup_{\pi_a,\pi_b\in\Pi}\max_{s\in\mathcal{S}}\mathop{\mathbb{E}}_{Z^s\sim\tilde{\mathcal{D}}_s}[\Delta^2(Z^s;\pi_a,\pi_b)]\cdot\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{\bar{n}_s} = \bar{V}\mathfrak{s}(\lambda\|\bar{n}), \end{aligned}$$

where $\bar{V} = \sup_{\pi_a,\pi_b\in\Pi}\max_{s\in\mathcal{S}}\mathbb{E}_{Z^s\sim\tilde{\mathcal{D}}_s}[\Delta(Z^s;\pi_a,\pi_b)]$, which is a constant value. Note that the last equality holds by the skewness identity in established in Appendix B.4. Plugging this into inequality (14), we get

$$\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) \leqslant 2A_1\sqrt{\frac{\bar{V}\mathfrak{s}(\lambda\|\bar{n})}{n}} \leqslant \mathcal{O}\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).$$

Case 2: $A_2\sqrt{s_{\lambda,n_{\mathcal{S}}}} > \frac{1}{2}\sqrt{\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)}$

In this case, one can easily rearrange terms to get that

$$\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) < 4A_2^2 s_{\lambda,n_{\mathcal{S}}} \leqslant o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).$$

Therefore, in either case, we have that at least $\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) \leqslant \mathcal{O}\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)$. We can plug this asymptotic bound into inequality (12) to arrive at the desired result:

$$\begin{aligned} &\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi) \\ &\leqslant (14+6\kappa(\Pi))\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + \sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n^2}} + \mathcal{O}\left(\sqrt{\mathcal{O}\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)s_{\lambda,n_{\mathcal{S}}}}\right) \\ &\leqslant (14+6\kappa(\Pi))\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + \sqrt{\frac{\bar{V}\mathfrak{s}(\lambda\|\bar{n})}{n^2}} + \mathcal{O}\left(\sqrt{\mathcal{O}\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)}\right) \\ &\leqslant (14+6\kappa(\Pi))\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) + o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) \\ &\leqslant (14+6\kappa(\Pi))\sqrt{\frac{V_{\lambda,n_{\mathcal{S}}}}{n}} + o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right). \end{aligned}$$

□

C.3 Bounding Oracle Mixture Empirical Process

Proposition 2. *Suppose Assumptions 1 and 2 hold. Fix $\lambda \in \Lambda$. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{\pi_a,\pi_b\in\Pi}|\Delta_\lambda(\pi_a,\pi_b) - \tilde{\Delta}_\lambda(\pi_a,\pi_b)| \leqslant C\kappa(\Pi)\sqrt{V\cdot\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}\cdot\log\frac{1}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right),$$

where C is a universal constant and

$$V = \max_{s\in\mathcal{S}}\sup_{\pi\in\Pi}\mathop{\mathbb{E}}_{Z^s\sim\bar{\mathcal{D}}_s}[\Gamma^s(\pi(X^s))^2].$$

Proof. First, for each source $s \in \mathcal{S}$, let $Z_1^c, \dots, Z_{n_s}^s$ be n_s independent random variables sampled from $\tilde{\mathcal{D}}_s$, where each $Z_i^s = (X_i^s, \vec{\Gamma}_i^s) \in \mathcal{Z} = \mathcal{X} \times \mathbb{R}^d$. Additionally, let $Z = \{Z_i^s \mid s \in \mathcal{S}, i \in [n_c]\}$ represent the corresponding set of samples across all sources.

In Lemma 7, we showed that $\mathbb{E}_{Z^s \sim \tilde{\mathcal{D}}_s}[Q(Z^s;\pi)] = Q_s(\pi)$. This implies that

$$\mathbb{E}[\tilde{Q}_\lambda(\pi)] = \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \mathbb{E}\left[Q(Z_i^s;\pi)\right] = \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} Q_s(\pi) = \sum_{s\in\mathcal{S}} \lambda_s Q_s(\pi) = Q_\lambda(\pi).$$

Additionally,

$$\mathbb{E}\left[\tilde{\Delta}_\lambda(\pi_a,\pi_b)\right] = \mathbb{E}\left[\tilde{Q}_\lambda(\pi_a)\right] - \mathbb{E}\left[\tilde{Q}_\lambda(\pi_b)\right] = Q_\lambda(\pi_a) - Q_\lambda(\pi_b) = \Delta_\lambda(\pi_a,\pi_b).$$

Therefore, we can follow a symmetrization argument to upper bound the expected oracle regret in terms of a Rademacher complexity, namely the weighted Rademacher complexity. Let Z' be an independent copy of Z and let $\varepsilon = \{\varepsilon_i^s \mid s \in \mathcal{S}, i \in [n_s]\}$ be a set of independent Rademacher random variables. Then,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi} |\Delta_\lambda(\pi_a,\pi_b) - \tilde{\Delta}_\lambda(\pi_a,\pi_b)|\right] \\
&= \mathbb{E}_Z\left[\sup_{\pi_a,\pi_b\in\Pi} \left| \mathbb{E}_{Z'}\left[\sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Delta(Z_i'^s;\pi_a,\pi_b)\right] - \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Delta(Z_i^s;\pi_a,\pi_b)\right|\right] \\
&= \mathbb{E}_Z\left[\sup_{\pi_a,\pi_b\in\Pi} \left| \mathbb{E}_{Z'}\left[\sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Delta(Z_i'^s;\pi_a,\pi_b) - \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Delta(Z_i^s;\pi_a,\pi_b)\right]\right|\right] \\
&\leqslant \mathbb{E}_Z\left[\mathbb{E}_{Z'}\left[\sup_{\pi_a,\pi_b\in\Pi} \left| \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \Big(\Delta(Z_i'^s;\pi_a,\pi_b) - \Delta(Z_i^s;\pi_a,\pi_b)\Big)\right|\right]\right] \\
&= \mathbb{E}_{Z,Z',\varepsilon}\left[\sup_{\pi_a,\pi_b\in\Pi} \left| \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s \Big(\Delta(Z_i'^s;\pi_a,\pi_b) - \Delta(Z_i^s;\pi_a,\pi_b)\Big)\right|\right] \\
&\leqslant 2\,\mathbb{E}_{Z,\varepsilon}\left[\sup_{\pi_a,\pi_b\in\Pi} \left| \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} \varepsilon_i^s \Delta(Z_i^s;\pi_a,\pi_b)\right|\right] \\
&= 2\mathfrak{R}_{\lambda,\bar{n}}(\Delta\mathcal{F}_\Pi) \\
&\leqslant 4\mathfrak{R}_{\lambda,\bar{n}}(\mathcal{F}_\Pi).
\end{aligned}$$

The first equalities and inequalities follow from standard symmetrization arguments, and the last inequality follows from Lemma 8. Next, we use this bound on the expected oracle regret and Talagrand's inequality (Lemma 3), to establish a high-probability bound on the oracle regret. In particular, we identify the set of independent random variables $\tilde{Z} = \{\tilde{Z}_i^s = (X_i^s, \frac{\lambda_s}{n_s}\Gamma_i^s) \mid s \in \mathcal{S}, i \in [n_s]\}$ and the function class $\mathcal{H} = \{h(\cdot;\pi_a,\pi_b) \mid \pi_a,\pi_b \in \Pi\}$ where

$$h(\tilde{Z}_i^s;\pi_a,\pi_b) = \mathbb{E}[\Delta(\tilde{Z}_i^s;\pi_a,\pi_b)] - \Delta(\tilde{Z}_i^s;\pi_a,\pi_b), \tag{15}$$

which is uniformly bounded for any $s \in \mathcal{S}$ and $i \in [n_s]$ by

$$\begin{aligned}
|h(\tilde{Z}_i^s;\pi_a,\pi_b)| &= \frac{\lambda_s}{n_s}\left|\mathbb{E}\left[\Gamma_i^s(\pi_a(X_i^s)) - \Gamma_i^s(\pi_b(X_i^s))\right] - \left(\Gamma_i^s(\pi_a(X_i^s)) - \Gamma_i^s(\pi_b(X_i^s))\right)\right| \\
&\leqslant \frac{\lambda_s}{n_s} 4U \\
&\leqslant 4U s_{\lambda,n_\mathcal{S}} =: U_{\lambda,n_\mathcal{S}},
\end{aligned}$$

where $U > 0$ is a uniform upper bound on $|\Gamma_i^s(a)|$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$ guaranteed by Assumption 1, and where the last inequality follows from Inequality (11), Additionally, we have

$$s_{\lambda,n_\mathcal{S}} = \frac{1}{\sqrt{\min_{s\in\mathcal{S}} n_c}} \sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}} \leqslant o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right),$$

as discussed in the proof of Proposition 1. Lastly, to use Talagrand's inequality, we set the constant D (specified in Lemma 3) to be

$$D = \sup_{\pi_a,\pi_b\in\Pi} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \mathbb{E}\left[h^2(\tilde{Z}_i^s;\pi_a,\pi_b)\right] + 8U_{\lambda,n_{\mathcal{S}}}\,\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\varepsilon_i^s h(\tilde{Z}_i^s;\pi_a,\pi_b)\right|\right].$$

By Lemma 5, this choice of D meets the required condition to use in Talagrand's inequality. In particular, we identify $\varphi : u \mapsto u^2$ defined over the set $\mathcal{U} \subset \mathbb{R}$ containing all possible outputs of any function in $\mathcal{H}$ given any realization of $\tilde{Z}_i^s$ for any $s \in \mathcal{S}$ as input. The uniform bound established above on realizable outputs of h given input $\tilde{Z}_i^s$ implies that $\mathcal{U} \subset [-U_{\lambda,n_{\mathcal{S}}}, U_{\lambda,n_{\mathcal{S}}}]$, and therefore, the Lipschitz constant of φ is $L = 2U_{\lambda,n_{\mathcal{S}}}$, as required.

Next, after setting t to be the positive solution of

$$\frac{t^2}{CD + CU_{\lambda,n_{\mathcal{S}}}t} = \log(C/\delta),$$

Talagrand's inequality guarantees

$$\begin{aligned}
&\mathbb{P}\left(\left|\sup_{\pi_a,\pi_b\in\Pi}\left|\Delta_\lambda(\pi_a,\pi_b) - \tilde{\Delta}_\lambda(\pi_a,\pi_b)\right| - \mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\Delta_\lambda(\pi_a,\pi_b) - \tilde{\Delta}_\lambda(\pi_a,\pi_b)\right|\right]\right| \geqslant t\right)\\
&= \mathbb{P}\left(\left|\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s} h(\tilde{Z}_i^s;\pi_a,\pi_b)\right| - \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s} h(\tilde{Z}_i^s;\pi_a,\pi_b)\right|\right]\right| \geqslant t\right)\\
&\leqslant C\exp\left(-\frac{t}{CU_{\lambda,n_{\mathcal{S}}}}\log\left(1 + \frac{U_{\lambda,n_{\mathcal{S}}}t}{D}\right)\right)\\
&\leqslant C\exp\left(-\frac{t^2}{CD + CU_{\lambda,n_{\mathcal{S}}}t}\right) = \delta.
\end{aligned}$$

Here, we used the inequality $\log(1+x) \geqslant \frac{x}{1+x}$ for any $x \geqslant 0$. Observe that, by construction,

$$\begin{aligned}
t &= \frac{1}{2}CU_{\lambda,n_{\mathcal{S}}}\log(C/\delta) + \sqrt{\frac{1}{4}C^2U_{\lambda,n_{\mathcal{S}}}^2\log^2(C/\delta) + CD\log(C/\delta)}\\
&\leqslant CU_{\lambda,n_{\mathcal{S}}}\log(C/\delta) + \sqrt{CD\log(C/\delta)}
\end{aligned}$$

and

$$\begin{aligned}
D &= \sup_{\pi_a,\pi_b\in\Pi} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \mathbb{E}\left[h^2(\tilde{Z}_i^s;\pi_a,\pi_b)\right] + 8U_{\lambda,\bar{n}}\,\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\varepsilon_i^s h(\tilde{Z}_i^s;\pi_a,\pi_b)\right|\right]\\
&= \sup_{\pi_a,\pi_b\in\Pi} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\mathbb{E}\left[\big(\mathbb{E}\left[\Delta(Z_i^s;\pi_a,\pi_b)\right] - \Delta(Z_i^s;\pi_a,\pi_b)\big)^2\right]\\
&\quad + 8U_{\lambda,n_{\mathcal{S}}}\,\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\frac{\lambda_s}{n_s}\varepsilon_i^s\big(\mathbb{E}\left[\Delta(Z_i^s;\pi_a,\pi_b)\right] - \Delta(Z_i^s;\pi_a,\pi_b)\big)\right|\right]\\
&= \sup_{\pi_a,\pi_b\in\Pi} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\Big(\mathbb{E}\left[\Delta^2(Z_i^s;\pi_a,\pi_b)\right] - \mathbb{E}\left[\Delta(Z_i^s;\pi_a,\pi_b)\right]^2\Big)\\
&\quad + 8U_{\lambda,n_{\mathcal{S}}}\,\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\frac{\lambda_s}{n_s}\varepsilon_i^s\big(\mathbb{E}\left[\Delta(Z_i^s;\pi_a,\pi_b)\right] - \Delta(Z_i^s;\pi_a,\pi_b)\big)\right|\right]\\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} \frac{\lambda_s^2}{n_s^2}\mathbb{E}\left[\Delta^2(Z_i^s;\pi_a,\pi_b)\right] + 16U_{\lambda,n_{\mathcal{S}}}\,\mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\sum_{i=1}^{n_s}\frac{\lambda_s}{n_s}\varepsilon_i^s\Delta(Z_i^s;\pi_a,\pi_b)\right|\right]\\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi} \sum_{s\in\mathcal{S}} \frac{\lambda_s^2}{n_s}\mathbb{E}\left[\Delta^2(Z_i^s;\pi_a,\pi_b)\right] + 16U_{\lambda,n_{\mathcal{S}}}\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\Delta\mathcal{F}_\Pi)\\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi} \sum_{s\in\mathcal{S}} \frac{\lambda_s^2}{n_s}\mathbb{E}\left[\Delta^2(Z_i^s;\pi_a,\pi_b)\right] + 32U_{\lambda,n_{\mathcal{S}}}\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)\\
&= \frac{V_{\lambda,n_{\mathcal{S}}}}{n} + 128U\mathfrak{R}_{\lambda,n_{\mathcal{S}}}(\mathcal{F}_\Pi)s_{\lambda,n_{\mathcal{S}}}.
\end{aligned}$$

Therefore, with this setup, Talagrand's inequality guarantees that with probability at least $1-\delta$

$$
\begin{aligned}
&\sup_{\pi_a,\pi_b\in\Pi}\left|\Delta_\lambda(\pi_a,\pi_b)-\tilde{\Delta}_\lambda(\pi_a,\pi_b)\right| \\
&\leqslant \mathbb{E}\left[\sup_{\pi_a,\pi_b\in\Pi}\left|\Delta_\lambda(\pi_a,\pi_b)-\tilde{\Delta}_\lambda(\pi_a,\pi_b)\right|\right]+t \\
&= 4\mathfrak{R}_{\lambda,n_\mathcal{S}}(\mathcal{F}_\Pi)+\sqrt{CD\log(C/\delta)}+CU_{\lambda,n_\mathcal{S}}\log(C/\delta) \\
&\leqslant 4\mathfrak{R}_{\lambda,n_\mathcal{S}}(\mathcal{F}_\Pi)+\sqrt{C\left(\frac{V_{\lambda,n_\mathcal{S}}}{n}+128U\mathfrak{R}_{\lambda,n_\mathcal{S}}(\mathcal{F}_\Pi)s_{\lambda,n_\mathcal{S}}\right)\log\left(C/\delta\right)}+4CUs_{\lambda,n_\mathcal{S}}\log\left(C/\delta\right) \\
&\leqslant 4\mathfrak{R}_{\lambda,n_\mathcal{S}}(\mathcal{F}_\Pi)+\sqrt{C\log(C/\delta)\frac{V_{\lambda,n_\mathcal{S}}}{n}}+\sqrt{128UC\log(C/\delta)\mathfrak{R}_{\lambda,n_\mathcal{S}}(\mathcal{F}_\Pi)s_{\lambda,n_\mathcal{S}}}+4UC\log(C/\delta)s_{\lambda,n_\mathcal{S}} \\
&\leqslant \left((56+24\kappa(\Pi))\sqrt{\frac{V_{\lambda,n_\mathcal{S}}}{n}}+o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)\right)+\sqrt{C\log(C/\delta)\frac{V_{\lambda,n_\mathcal{S}}}{n}} \\
&\qquad+\sqrt{\mathcal{O}_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)}+o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) \\
&\leqslant \left(56+24\kappa(\Pi)+\sqrt{C\log(C/\delta)}\right)\sqrt{\frac{V_{\lambda,n_\mathcal{S}}}{n}}+o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).
\end{aligned}
$$

It holds that the constant in Talagrand's inequality satisfies $C\geqslant 1$. Moreover, without significant loss of generality, we can consider $\kappa(\Pi)\geqslant 1$. Lastly, we can map $\delta/C\mapsto\delta$ to get

$$
\sup_{\pi_a,\pi_b\in\Pi}\left|\Delta_\lambda(\pi_a,\pi_b)-\tilde{\Delta}_\lambda(\pi_a,\pi_b)\right|\leqslant C'\kappa(\Pi)\sqrt{\frac{V_{\lambda,n_\mathcal{S}}}{n}\log\left(\frac{1}{\delta}\right)}+o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)
$$

where we set $C'=56+25+\sqrt{C}$. Here, we used the bounds previously established in the proof of Proposition 1 that $\mathfrak{R}_{\lambda,n_\mathcal{S}}(\mathcal{F}_\Pi)\leqslant\mathcal{O}\big(\sqrt{\mathfrak{s}(\lambda\|\bar{n})/n}\,\big)$ and $s_{\lambda,n_\mathcal{S}}\leqslant o\big(\sqrt{\mathfrak{s}(\lambda\|\bar{n})/n}\,\big)$.

Lastly, we decompose the weighted variance term $V_{\lambda,n_\mathcal{S}}$ as

$$
\begin{aligned}
V_{\lambda,n_\mathcal{S}} &= \sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{\bar{n}_s}\mathop{\mathbb{E}}_{Z^s\sim\tilde{\mathcal{D}}_s}\left[\Delta^2(Z^s;\pi_a,\pi_b)\right] \\
&= \sup_{\pi_a,\pi_b\in\Pi}\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{\bar{n}_s}\mathop{\mathbb{E}}_{Z^s\sim\bar{\mathcal{D}}_s}\left[\big(\Gamma^s(\pi_a(X^s))-\Gamma^s(\pi_b(X^s))\big)^2\right] \\
&\leqslant \max_{s\in\mathcal{S}}\sup_{\pi_a,\pi_b\in\Pi}\mathop{\mathbb{E}}_{Z^s\sim\bar{\mathcal{D}}_s}\left[\big(\Gamma^s(\pi_a(X^s))-\Gamma^s(\pi_b(X^s))\big)^2\right]\cdot\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{\bar{n}_s} \\
&\leqslant 4\cdot\max_{s\in\mathcal{S}}\sup_{\pi\in\Pi}\mathop{\mathbb{E}}_{Z^s\sim\bar{\mathcal{D}}_s}\left[\Gamma^s(\pi(X^s))^2\right]\cdot\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{\bar{n}_s} \\
&= 4V\cdot\mathfrak{s}(\lambda\|\bar{n}).
\end{aligned}
$$

We absorb the factor of $\sqrt{4}$ into the universal constant to get the desired result.

□

C.4 Bounding Approximate Mixture Empirical Process

First, we state a more general form of Assumption 3 that is sufficient for bounding the approximate mixture empirical process.

Assumption 4. For any source $s\in\mathcal{S}$, the source estimates $\hat{\mu}_s$ and $\hat{\omega}_s$ of the nuisance parameters μ_s and ω_s, respectively, trained on n_s source data points satisfy the following squared error bounds:

$$\mathbb{E}_{\mathcal{D}_s}\big[\,\|\hat{\mu}_s(X^s) - \mu_s(X^s)\|_2^2\,\big] \leqslant \frac{o(1)}{n_s^{\zeta_\mu}},$$
$$\mathbb{E}_{\mathcal{D}_s}\big[\,\|\hat{\omega}_s(X^s) - \omega_s(X^s)\|_2^2\,\big] \leqslant \frac{o(1)}{n_s^{\zeta_w}},$$

for some $0 < \zeta_\mu, \zeta_w < 1$ with $\zeta_\mu + \zeta_w \geqslant 1$.

This assumption subsumes Assumption 3 and allows flexibility in the estimation of nuisance parameters. It allows less accurate estimation of the inverse conditional propensities if we can estimate the response response functions at a faster rate, and vice versa. Now, we can return to bounding the approximate mixture empirical process with this more general assumption on the error rates of the nuisance parameter estimates.

Proposition 3. *Suppose Assumptions 1, 2, and 3 hold. Fix $\lambda \in \Lambda$. Then,*

$$\sup_{\pi_a,\pi_b\in\Pi} |\tilde{\Delta}_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| \leqslant o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).$$

Proof. Recall that $\{(X_i^s, A_i^s, Y_i^s)\}_{i=1}^{n_s}$ is the data collected by source $s \in \mathcal{S}$ as described in Section 3.3. We assume each source estimates the local nuisance parameters using a cross-fitting strategy, as discussed in Algorithm 2. Under this strategy, each source $s \in \mathcal{S}$ divides their local dataset into K folds, and for each fold k, the source estimates μ_s and ω_s using the rest $K-1$ folds. Let $k_c : [n_s] \to [K]$ denote the surjective mapping that maps a data point index to its corresponding fold containing the data point. We let $\hat{\mu}_s^{-k_s(i)}$ and $\hat{\omega}_s^{-k_s(i)}$ denote the estimators of μ_s and ω_s fitted on the $K-1$ folds of source s other than $k_s(i)$.

As discussed Section 4, recall the oracle AIPW scores

$$\Gamma_i^s(a) = \mu_s(X_i^s;a) + \big(Y_i^s(a) - \mu_s(X_i^s;a)\big)\omega_s(X_i^s;a)\mathbf{1}\{A_i^s = a\}$$

and approximate AIPW scores

$$\hat{\Gamma}_i^s(a) = \hat{\mu}_s^{-k_s(i)}(X_i^s;a) + \big(Y_i^s - \hat{\mu}_s^{-k_s(i)}(X_i^s;a)\big)\hat{\omega}_s^{-k_s(i)}(X_i^s;a)\mathbf{1}\{A_i^s = a\}$$

for any $a \in \mathcal{A}$, where $k_s(i)$ is the fold corresponding to data point i of source s. One can verify that the difference between the oracle and approximate AIPW scores can be expressed as

$$\hat{\Gamma}_i^s(a) - \Gamma_i^s(a) = \Gamma_i^{s\prime}(a) + \Gamma_i^{s\prime\prime}(a) + \Gamma_i^{s\prime\prime\prime}(a),$$

where

$$\begin{aligned}
\Gamma_i^{s\prime}(a) &= \Big(\hat{\mu}_s^{-k_s(i)}(X_i^s;a) - \mu_s(X_i^s;a)\Big)\big(1 - \omega_s(X_i^s;a)\mathbf{1}\{A_i^s = a\}\big),\\
\Gamma_i^{s\prime\prime}(a) &= \big(Y_i^s(a) - \mu_s(X_i^s;a)\big)\Big(\hat{\omega}_s^{-k_s(i)}(X_i^s;a) - \omega_s(X_i^s;a)\Big)\mathbf{1}\{A_i^s = a\},\\
\Gamma_i^{s\prime\prime\prime}(a) &= \Big(\mu_s(X_i^s;a) - \hat{\mu}_s^{-k_s(i)}(X_i^s;a)\Big)\Big(\hat{\omega}_s^{-k_s(i)}(X_i^s;a) - \omega_s(X_i^s;a)\Big)\mathbf{1}\{A_i^s = a\}.
\end{aligned}$$

This induces the following decomposition of the approximate regret:

$$\hat{\Delta}_\lambda(\pi_a,\pi_b) - \tilde{\Delta}_\lambda(\pi_a,\pi_b) = S_1(\pi_a,\pi_b) + S_2(\pi_a,\pi_b) + S_3(\pi_a,\pi_b),$$

where

$$\begin{aligned}
S_1(\pi_a,\pi_b) &= \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s}\sum_{i=1}^{n_s} \Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime}(\pi_b(X_i^s)),\\
S_2(\pi_a,\pi_b) &= \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s}\sum_{i=1}^{n_s} \Gamma_i^{s\prime\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime\prime}(\pi_b(X_i^s)),\\
S_3(\pi_a,\pi_b) &= \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s}\sum_{i=1}^{n_s} \Gamma_i^{s\prime\prime\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime\prime\prime}(\pi_b(X_i^s)).
\end{aligned}$$

We further decompose S_1 and S_2 by folds as follows:

$$S_1(\pi_a, \pi_b) = \sum_{k=1}^{K} S_1^k(\pi_a, \pi_b),$$
$$S_2(\pi_a, \pi_b) = \sum_{k=1}^{K} S_2^k(\pi_a, \pi_b),$$

where

$$S_1^k(\pi_a, \pi_b) = \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{\{i|k_s(i)=k\}} \Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime}(\pi_b(X_i^s)),$$
$$S_2^k(\pi_a, \pi_b) = \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s} \sum_{\{i|k_s(i)=k\}} \Gamma_i^{s\prime\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime\prime}(\pi_b(X_i^s)),$$

for each $k \in [K]$. To determine a bound on the approximate regret, we will establish high probability bounds for the worst-case absolute value over policies of each term in this decomposition. For convenience, for any policy π, we will denote $\pi(x; a) = \mathbf{1}\{\pi(x) = a\}$.

Bounding S_1. We wish to bound the quantity $\sup_{\pi_a,\pi_b\in\Pi} |S_1(\pi_a, \pi_b)|$. We first bound the quantity $\sup_{\pi_a,\pi_b\in\Pi} \left|S_1^k(\pi_a, \pi_b)\right|$ for any $k \in [K]$.

First, note that since $\hat{\mu}_s^{-k_s(i)}$ is estimated using data outside fold $k_s(i)$, when we condition on the data outside fold $k_s(i)$, $\hat{\mu}_s^{-k_s(i)}$ is fixed and each term in $S_1(\pi_a, \pi_b)$ is independent. This allows us to compute

$$\begin{aligned}
&\mathbb{E}\left[\Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s}(\pi_b(X_i^s))\right] \\
&= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\left(\pi_a(X_i^s; a) - \pi_b(X_i^s; a)\right)\left(\hat{\mu}_s^{-k_s(i)}(X_i^s; a) - \mu_s(X_i^s; a)\right)\left(1 - \omega_s(X_i^s; a)\mathbf{1}\{A_i^s = a\}\right)\right] \\
&= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\mathbb{E}\left[\left(\pi_a(X_i^s; a) - \pi_b(X_i^s; a)\right)\left(\hat{\mu}_s^{-k_s(i)}(X_i^s; a) - \mu_s(X_i^s; a)\right)\left(1 - \omega_s(X_i^s; a)\mathbf{1}\{A_i^s = a\}\right) \;\middle|\; X_i^s\right]\right] \\
&= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\left(\pi_a(X_i^s; a) - \pi_b(X_i^s; a)\right)\left(\hat{\mu}_s^{-k_s(i)}(X_i^s; a) - \mu_s(X_i^s; a)\right)\mathbb{E}\left[1 - \omega_s(X_i^s; a)\mathbf{1}\{A_i^s = a\} \;\middle|\; X_i^s\right]\right] = 0
\end{aligned}$$

Therefore,

$$\begin{aligned}
&K \sup_{\pi_a,\pi_b\in\Pi} \left|S_1^k(\pi_a, \pi_b)\right| \\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi} \left| \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s/K} \sum_{\{i|k_s(i)=k\}} \Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime}(\pi_b(X_i^s)) \right| \\
&= \sup_{\pi_a,\pi_b\in\Pi} \left| \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s/K} \sum_{\{i|k_s(i)=k\}} \left(\Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime}(\pi_b(X_i^s))\right) - \mathbb{E}\left[\Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime}(\pi_b(X_i^s))\right] \right|.
\end{aligned}$$

Identifying $\Gamma_i^{s\prime}$ with Γ_i^s and sample sizes $n_{\mathcal{S}}/K$ with $n_{\mathcal{S}}$, the right-hand side in the above inequality is effectively an oracle regret and so we can apply Proposition 2 to obtain that with probability at least $1 - \delta$,

$$\begin{aligned}
&K \sup_{\pi_a,\pi_b\in\Pi} \left|S_1^k(\pi_a, \pi_b)\right| \\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi} \left| \sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s/K} \sum_{\{j|k_i(j)=k\}} \left(\Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime}(\pi_b(X_i^s))\right) - \mathbb{E}\left[\Gamma_i^{s\prime}(\pi_a(X_i^s)) - \Gamma_i^{s\prime}(\pi_b(X_i^s))\right] \right| \\
&\leqslant C_{\Pi,\delta} \sqrt{\max_{s\in\mathcal{S}} \sup_{\pi\in\Pi} \mathbb{E}\left[\Gamma_i^{s\prime}(\pi(X_i^s))^2 \mid \hat{\mu}_s^{-k_s(i)}\right] \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n/K}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n/K}}\right) \\
&\leqslant C_{\Pi,\delta}\,(1/\eta - 1) \sqrt{K \max_{s\in\mathcal{S}} \mathbb{E}\left[||\hat{\mu}_s^{-k_s(i)}(X_i^s) - \mu_s(X_i^s)||_2^2 \mid \hat{\mu}_s^{-k_s(i)}\right] \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right),
\end{aligned}$$

where $C_{\Pi,\delta} = C\kappa(\Pi)\sqrt{\log(1/\delta)}$ for some universal constant C, and $\eta = \min_{s\in\mathcal{S}} \eta_s$ for η_s in the overlap assumption stated in in Assumption 1. The last inequality follows from a uniform bound on $\Gamma_i^{s\prime}(\pi(X_i^s))$ and the overlap assumption.

By the assumption on finite sample error bounds for the nuisance functions stated above, for every $s \in \mathcal{S}$

$$\mathbb{E}\left[||\hat{\mu}_s^{-k_s(i)}(X_i^s) - \mu_s(X_i^s)||^2 \mid \hat{\mu}_s^{-k_s(i)}\right] \leqslant \frac{g_s\left(\alpha_K n_s\right)}{\left(\alpha_K n_s\right)^{\zeta_\mu}},$$

where $\alpha_K = 1 - K^{-1}$, g_s is some decreasing function, and $0 < \zeta_\mu < 1$. Then,

$$\begin{aligned}\max_{s\in\mathcal{S}} \mathbb{E}\left[||\hat{\mu}_s^{-k_s(i)}(X_i^s) - \mu_s(X_i^s)||^2 \mid \hat{\mu}_s^{-k_s(i)}\right] &\leqslant \max_{s\in\mathcal{S}} \frac{g_s(\alpha_K n_s)}{(\alpha_K n_s)^{\zeta_\mu}} \\ &\leqslant \frac{\max_{s\in\mathcal{S}} g_s(\alpha_K n_s)}{\alpha_K^{\zeta_\mu} \cdot \min_{s\in\mathcal{S}} n_s^{\zeta_\mu}} \\ &\leqslant \frac{\max_{s\in\mathcal{S}} g_s(\alpha_K n_s)}{\alpha_K^{\zeta_\mu} \cdot \min_{s\in\mathcal{S}} n_s^{\zeta_\mu}}.\end{aligned}$$

By the local data size scaling assumption in Assumption 2, for any $s \in \mathcal{S}$, we have that $n_s = \Omega(\nu_c(n))$ where ν_c is an increasing function. In other words, there exists a constant $\tau > 0$ such that $n_s \geqslant \tau\nu_s(n)$ for sufficiently large n. Then, since g_s is decreasing, $g_s(\alpha_K n_c) < g_s(\tau\alpha_K\nu_s(n))$ for sufficiently large n. Moreover, since ν_s is increasing and $\tau\alpha_K > 0$, $\tilde{\nu}_c = \tau\alpha_K\nu_s$ is also increasing, and since g_s is decreasing, the composition $\tilde{g}_s = g_s \circ \tilde{\nu}_c$ is decreasing. Therefore, $g_s(\alpha_K n_s)$ is asymptotically bounded by a decreasing function $\tilde{g}_s$ of n. This observation and the fact that the maximum of a set of decreasing functions is itself decreasing imply that $\max_{s\in\mathcal{S}} g_s(\alpha_K n_s)$ is asymptotically bounded by the decreasing function $\tilde{g}$ defined by $\tilde{g}(n) = \max_{s\in\mathcal{S}} \tilde{g}_s(n)$. In other words,

$$\max_{s\in\mathcal{S}} g_s(\alpha_K n_s) \leqslant \tilde{g}(n) \leqslant o(1).$$

Additionally, since $n_s = \Omega(\nu_s(n))$ and $\zeta_\mu > 0$, we also have that

$$\frac{1}{\min_{s\in\mathcal{S}} n_s^{\zeta_\mu}} \leqslant o(1).$$

These two observations imply

$$\max_{s\in\mathcal{S}} \mathbb{E}\left[||\hat{\mu}_s^{-k_s(i)}(X_i^s) - \mu_s(X_i^s)||^2 \mid \hat{\mu}_s^{-k_s(i)}\right] \leqslant \frac{\max_{s\in\mathcal{S}} g_s(\alpha_K n_s)}{\alpha_K^{\zeta_\mu} \cdot \min_{s\in\mathcal{S}} n_s^{\zeta_\mu}} \leqslant o\left(1\right).$$

Therefore,

$$\begin{aligned}&\sup_{\pi_a,\pi_b\in\Pi} \left|S_1^k(\pi_a, \pi_b)\right| \\ &\leqslant C_{\Pi,\delta}\left(1/\eta - 1\right)\sqrt{\frac{1}{K}\max_{s\in\mathcal{S}} \mathbb{E}\left[||\hat{\mu}_s^{-k_s(i)}(X_i^s) - \mu_s(X_i^s)||_2^2 \mid \hat{\mu}_s^{-k_s(i)}\right] \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) \\ &\leqslant C_{\Pi,\delta}\left(1/\eta - 1\right)\sqrt{\frac{1}{K} \cdot o\left(\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}\right)} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) \\ &\leqslant o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right),\end{aligned}$$

and

$$\sup_{\pi_a,\pi_b\in\Pi} \left|S_1(\pi_a, \pi_b)\right| \leqslant \sum_{k=1}^{K} \sup_{\pi_a,\pi_b\in\Pi} \left|S_1^k(\pi_a, \pi_b)\right| \leqslant o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).$$

Bounding S_2. The bound for $\sup_{\pi_a,\pi_b\in\Pi} |S_2(\pi_a,\pi_b)|$ follows the same argument as that of S_1. We first bound $\sup_{\pi_a,\pi_b\in\Pi} \left|S_2^k(\pi_a,\pi_b)\right|$ for any $k \in [K]$.

First, note that since $\hat{\omega}_s^{-k_s(i)}$ is estimated using data outside fold $k_s(i)$, when we condition on the data outside fold $k_s(i)$, $\hat{\omega}_s^{-k_s(i)}$ is fixed and each term in $S_2(\pi_a,\pi_b)$ is independent. This allows us to compute

$$
\begin{aligned}
&\mathbb{E}\left[\Gamma_i^{s''}(\pi_a(X_i^s)) - \Gamma_i^{s''}(\pi_b(X_i^s))\right] \\
&= \mathbb{E}\left[\sum_{a\in\mathcal{A}} (\pi_a(X_i^s;a) - \pi_b(X_i^s;a))\,(Y_i^s(a) - \mu_s(X_i^s;a))\left(\hat{\omega}_s^{-k_s(i)}(X_i^s;a) - \omega_s(X_i^s;a)\right)\mathbf{1}\{A_i^s = a\}\right] \\
&= \mathbb{E}\left[(\pi_a(X_i^s;A_i^s) - \pi_b(X_i^s;A_i^s))\,(Y_i^s(A_i^s) - \mu_s(X_i^s;A_i^s))\left(\hat{\omega}_s^{-k_s(i)}(X_i^s;a) - \omega_s(X_i^s;a)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(\pi_a(X_i^s;A_i^s) - \pi_b(X_i^s;A_i^s))\,(Y_i^s(A_i^s) - \mu_s(X_i^s;A_i^s))\left(\hat{\omega}_s^{-k_s(i)}(X_i^s;a) - \omega_s(X_i^s;a)\right) \;\middle|\; X_i^s, A_i^s\right]\right] \\
&= \mathbb{E}\left[(\pi_a(X_i^s;A_i^s) - \pi_b(X_i^s;A_i^s))\,\mathbb{E}\left[Y_i^s(A_i^s) - \mu_s(X_i^s;A_i^s) \mid X_i^s, A_i^s\right]\left(\hat{\omega}_s^{-k_s(i)}(X_i^s;a) - \omega_s(X_i^s;a)\right)\right] = 0
\end{aligned}
$$

Therefore, we can follow the exact same argument as above, eliciting Proposition 2, to obtain that with probability at least $1-\delta$,

$$
\begin{aligned}
&K \sup_{\pi_a,\pi_b\in\Pi} \left|S_2^k(\pi_a,\pi_b)\right| \\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s/K} \sum_{\{i|k_s(i)=k\}} \Gamma_i^{s''}(\pi_a(X_i^s)) - \Gamma_i^{s''}(\pi_b(X_i^s))\right| \\
&= \sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s/K} \sum_{\{i|k_s(i)=k\}} \left(\Gamma_i^{s''}(\pi_a(X_i^s)) - \Gamma_i^{s''}(\pi_b(X_i^s))\right) - \mathbb{E}\left[\Gamma_i^{s''}(\pi_a(X_i^s)) - \Gamma_i^{s''}(\pi_b(X_i^s))\right]\right| \\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}} \frac{\lambda_s}{n_s/K} \sum_{\{j|k_i(j)=k\}} \left(\Gamma_i^{s''}(\pi_a(X_i^s)) - \Gamma_i^{s''}(\pi_b(X_i^s))\right) - \mathbb{E}\left[\Gamma_i^{s''}(\pi_a(X_i^s)) - \Gamma_i^{s''}(\pi_b(X_i^s))\right]\right| \\
&\leqslant C_{\Pi,\delta}\sqrt{\max_{s\in\mathcal{S}} \mathbb{E}\left[\Gamma_i^{s''}(\pi(X_i^s))^2 \mid \hat{\omega}_s^{-k_s(i)}\right]\cdot\frac{\mathfrak{s}(\lambda\|\bar{n})}{n/K}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n/K}}\right) \\
&\leqslant C_{\Pi,\delta}\sqrt{2BK\max_{s\in\mathcal{S}} \mathbb{E}\left[||\hat{\omega}_s^{-k_s(i)}(X_i^s) - \omega_s(X_i^s)||_2^2 \mid \hat{\omega}_s^{-k_s(i)}\right]\cdot\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right),
\end{aligned}
$$

where $C_{\Pi,\delta} = C\kappa(\Pi)\sqrt{\log(1/\delta)}$ for some universal constant C, and $B = \max_{s\in\mathcal{S}} B_s$ for the bounds B_s on the outcomes defined in Assumption 1. The last inequality follows from a uniform bound on $\Gamma_i^{s''}(\pi(X_i^s))$.

We follow the exact same argument as above to get

$$
\max_{s\in\mathcal{S}} \mathbb{E}\left[||\hat{\omega}_s^{-k_s(i)}(X_i^s) - \omega_s(X_i^s)||^2 \mid \hat{\omega}_s^{-k_s(i)}\right] \leqslant o\left(\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}\right).
$$

Therefore,

$$
\begin{aligned}
&\sup_{\pi_a,\pi_b\in\Pi} \left|S_2^k(\pi_a,\pi_b)\right| \\
&\leqslant C_{\Pi,\delta}\sqrt{\frac{2B}{K}\max_{s\in\mathcal{S}} \mathbb{E}\left[||\hat{\omega}_s^{-k_s(i)}(X_i^s) - \omega_s(X_i^s)||_2^2 \mid \hat{\omega}_s^{-k_s(i)}\right]} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) \\
&\leqslant C_{\Pi,\delta}\sqrt{\frac{2B}{K}\cdot o\left(\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}\right)} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) \\
&\leqslant o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right),
\end{aligned}
$$

and

$$\sup_{\pi_a,\pi_b\in\Pi}\left|S_2(\pi_a,\pi_b)\right| \leqslant \sum_{k=1}^{K}\sup_{\pi_a,\pi_b\in\Pi}\left|S_2^k(\pi_a,\pi_b)\right| \leqslant o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).$$

Bounding S_3. Next, we bound the contribution from S_3. We have that

$$\begin{aligned}
&\sup_{\pi_a,\pi_b\in\Pi}\left|S_3(\pi_a,\pi_b)\right| \\
&= \sup_{\pi_a,\pi_b\in\Pi}\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\Gamma_i^{s'''}(\pi_a(X_i^s)) - \Gamma_i^{s'''}(\pi_b(X_i^s))\right| \\
&\leqslant 2\left|\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\sum_{a\in\mathcal{A}}\left(\mu_s(X_i^s;a) - \hat{\mu}_s^{-k_s(i)}(X_i^s;a)\right)\left(\hat{\omega}_s^{-k_s(i)}(X_i^s;a) - \omega_s(X_i^s;a)\right)\right| \\
&\leqslant 2\sqrt{\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\left\|\mu_s(X_i^s) - \hat{\mu}_s^{-k_s(i)}(X_i^s)\right\|_2^2}\sqrt{\sum_{s\in\mathcal{S}}\frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}\left\|\hat{\omega}_s^{-k_s(i)}(X_i^s) - \omega_s(X_i^s)\right\|_2^2} \\
&\leqslant 2\sqrt{\sum_{s\in\mathcal{S}}\lambda_s\frac{g_s(\alpha_K n_s)}{(\alpha_K n_s)^{\zeta_\mu}}}\sqrt{\sum_{s\in\mathcal{S}}\lambda_s\frac{g_s(\alpha_K n_s)}{(\alpha_K n_s)^{\zeta_w}}} \\
&\leqslant \frac{2}{\alpha_K^{(\zeta_\mu+\zeta_w)/2}}\sqrt{\max_{s\in\mathcal{S}}\frac{\lambda_s}{n_s^{\zeta_\mu}}\sum_{s\in\mathcal{S}}g_s(\alpha_K n_s)}\sqrt{\max_{s\in\mathcal{S}}\frac{\lambda_s}{n_s^{\zeta_w}}\sum_{s\in\mathcal{S}}g_s(\alpha_K n_s)} \\
&= \frac{2}{\alpha_K^{(\zeta_\mu+\zeta_w)/2}}\sum_{s\in\mathcal{S}}g_s(\alpha_K n_s)\sqrt{\max_{s\in\mathcal{S}}\frac{\lambda_s^2}{n_s^{\zeta_\mu+\zeta_w}}} \\
&\leqslant \frac{2}{\alpha_K^{(\zeta_\mu+\zeta_w)/2}}\sum_{s\in\mathcal{S}}g_s(\alpha_K n_s)\sqrt{\max_{s\in\mathcal{S}}\frac{\lambda_s^2}{n_s}} \\
&\leqslant \frac{2}{\alpha_K^{(\zeta_\mu+\zeta_w)/2}}\sum_{s\in\mathcal{S}}g_s(\alpha_K n_s)\sqrt{\sum_{s\in\mathcal{S}}\frac{\lambda_s^2}{n_s}} \\
&\leqslant \frac{2}{\alpha_K^{(\zeta_\mu+\zeta_w)/2}}\sum_{s\in\mathcal{S}}g_s(\alpha_K n_s)\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}.
\end{aligned}$$

As discussed earlier, $g_s(\alpha_K n_s)$ is asymptotically bounded by a decreasing function of n. Since the sum of decreasing functions is decreasing, $\sum_{s\in\mathcal{S}} g_s(\alpha_K n_s)$ is asymptotically bounded by a decreasing function $\tilde{g}$ in n. In other words, $\sum_{s\in\mathcal{S}} g_s(\alpha_K n_s) \leqslant \tilde{g}(n) \leqslant o(1)$. Therefore,

$$\sup_{\pi_a,\pi_b\in\Pi}\left|S_3(\pi_a,\pi_b)\right| \leqslant \frac{2}{\alpha_K^{(\zeta_\mu+\zeta_w)/2}}\cdot o(1)\cdot\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}} \leqslant o\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).$$

Combine results. Putting all the above bounds together, we have

$$\begin{aligned}
\sup_{\pi_a,\pi_b\in\Pi}\left|\tilde{\Delta}_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)\right| &\leqslant \sup_{\pi_a,\pi_b\in\Pi}\left|S_1(\pi_a,\pi_b) + S_2(\pi_a,\pi_b) + S_3(\pi_a,\pi_b)\right| \\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi}\left|S_1(\pi_a,\pi_b)\right| + \sup_{\pi_a,\pi_b\in\Pi}\left|S_2(\pi_a,\pi_b)\right| + \sup_{\pi_a,\pi_b\in\Pi}\left|S_3(\pi_a,\pi_b)\right| \\
&\leqslant o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right).
\end{aligned}$$

□

C.5 Bounding Mixture Empirical Process

Proposition 4. *Suppose Assumptions 1, 2, and 3 hold. For any $\epsilon > 0$, let Λ_ϵ denote the minimal ϵ-covering set of Λ under the ℓ_1 distance. For any $\epsilon > 0$, any $\delta \in (0,1)$, and any $\lambda \in \Lambda_\epsilon$, with probability at least $1-\delta$,*

$$\sup_{\pi_a,\pi_b \in \Pi} |\Delta_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| \leqslant \xi_{\epsilon,\delta}(n;\Pi,\Lambda),$$

where

$$\xi_{\epsilon,\delta}(n;\Pi,\Lambda) := C\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\Lambda\|\bar{n})}{n} \cdot \log\frac{|\Lambda_\epsilon|}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right),$$

with C being a universal constant and

$$V = \max_{s\in\mathcal{S}}\sup_{\pi\in\Pi}\mathbb{E}_{\tilde{\mathcal{D}}_s}[\Gamma^s(\pi(X^s))^2].$$

Proof. By Propositions 2 and 3, for any fixed choice of $\lambda \in \Lambda_\epsilon$, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{\pi_a,\pi_b \in \Pi} |\Delta_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| \\ &\leqslant \sup_{\pi_a,\pi_b \in \Pi} |\Delta_\lambda(\pi_a,\pi_b) - \tilde{\Delta}_\lambda(\pi_a,\pi_b)| + \sup_{\pi_a,\pi_b \in \Pi} |\tilde{\Delta}_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| \\ &\leqslant C\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n} \cdot \log\frac{1}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right),\end{aligned}$$

where C is a universal constant. In other words, for any fixed $\lambda \in \Lambda_\epsilon$,

$$\mathbb{P}\left(\sup_{\pi_a,\pi_b \in \Pi} |\Delta_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| > C\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n} \cdot \log\frac{1}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)\right) \leqslant \delta.$$

Then, by a union bound over Λ_ϵ, it follows that

$$\begin{aligned}&\mathbb{P}\left(\forall\lambda \in \Lambda_\epsilon, \sup_{\pi_a,\pi_b \in \Pi} |\Delta_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| > C\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n} \cdot \log\frac{1}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)\right) \\ &\leqslant \sum_{\lambda\in\Lambda_\epsilon}\mathbb{P}\left(\sup_{\pi_a,\pi_b \in \Pi} |\Delta_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| > C\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n} \cdot \log\frac{1}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right)\right) \\ &\leqslant |\Lambda_\epsilon|\delta.\end{aligned}$$

Mapping $\delta \mapsto \delta/|\Lambda_\epsilon|$, this implies that for every $\lambda \in \Lambda_\epsilon$, with probability at least $1-\delta$,

$$\begin{aligned}\sup_{\pi_a,\pi_b \in \Pi} |\Delta_\lambda(\pi_a,\pi_b) - \hat{\Delta}_\lambda(\pi_a,\pi_b)| &\leqslant C\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\lambda\|\bar{n})}{n} \cdot \log\frac{|\Lambda_\epsilon|}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\lambda\|\bar{n})}{n}}\right) \\ &\leqslant C\kappa(\Pi)\sqrt{V \cdot \frac{\mathfrak{s}(\Lambda\|\bar{n})}{n} \cdot \log\frac{|\Lambda_\epsilon|}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right)\end{aligned}$$

□

C.6 EG-OPO Suboptimality Bound

Lemma 1. *For any T and any $\lambda' \in \Lambda_\epsilon$,*

$$\hat{R}_{\lambda'}(\hat{\pi}) \leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda}\hat{R}_\lambda(\pi) + 2\hat{B}\sqrt{\frac{\log|\Lambda_\epsilon|}{T}}$$

where $\hat{B}$ is a uniform bound on $\hat{R}_{\lambda'}(\pi)$.

Proof. The regret guarantee of the exponentiated gradient algorithm, specifically Corollary 2.14 of (Shalev-Shwartz et al., 2012), states that for any $\lambda' \in \Lambda_\epsilon$

$$-\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\lambda\sim\rho_t}\left[\frac{R_\lambda(\pi_t)}{\hat{B}}\right] \leqslant -\sum_{t=1}^{T} \frac{\hat{R}_{\lambda'}(\pi_t)}{\hat{B}} + \frac{\log|\Lambda_\epsilon|}{\eta\hat{B}} + \eta\hat{B}T.$$

Substituting $\eta\hat{B} = \sqrt{\log|\Lambda_\epsilon|/T}$ gives

$$-\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\lambda\sim\rho_t}\left[\frac{R_\lambda(\pi)}{\hat{B}}\right] \leqslant -\sum_{t=1}^{T} \frac{\hat{R}_{\lambda'}(\pi)}{\hat{B}} + 2\sqrt{T\log|\Lambda_\epsilon|}.$$

Recalling that $P_t = \text{Uniform}(\pi_1, \ldots, \pi_t)$, following the proof techniques of Freund and Schapire (1996) gives that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\pi\sim P_T}\left[\frac{\hat{R}_{\lambda'}(\pi)}{\hat{B}}\right] &= \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{R}_{\lambda'}(\pi_t)}{\hat{B}} \\
&\leqslant \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\lambda\sim\rho_t}\left[\frac{R_\lambda(\pi_t)}{\hat{B}}\right] + 2\sqrt{\frac{\log|\Lambda_\epsilon|}{T}} \\
&= \frac{1}{T}\sum_{t=1}^{T}\min_{\pi\in\Pi}\mathop{\mathbb{E}}_{\lambda\sim\rho_t}\left[\frac{\hat{R}_\lambda(\pi)}{\hat{B}}\right] + 2\sqrt{\frac{\log|\Lambda_\epsilon|}{T}} \\
&\leqslant \min_{\pi\in\Pi}\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\lambda\sim\rho_t}\left[\frac{R_\lambda(\pi)}{\hat{B}}\right] + 2\sqrt{\frac{\log|\Lambda_\epsilon|}{T}} \\
&\leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda}\frac{\hat{R}_\lambda(\pi)}{\hat{B}} + 2\sqrt{\frac{\log|\Lambda_\epsilon|}{T}}.
\end{aligned}$$

Multiplying through by $\hat{B}$ gives

$$\mathop{\mathbb{E}}_{\pi\sim P_T}\left[\hat{R}_{\lambda'}(\pi)\right] \leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda}\hat{R}_\lambda(\pi) + 2\hat{B}\sqrt{\frac{\log|\Lambda_\epsilon|}{T}}.$$

□

Proposition 5. *For any T and any $\lambda' \in \Lambda_\epsilon$,*

$$\mathop{\mathbb{E}}_{\pi\sim P_T}\left[\hat{R}_{\lambda'}(\pi)\right] \leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda}R_\lambda(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda) + 2\hat{B}\sqrt{\frac{\log|\Lambda_\epsilon|}{T}}$$

where $\hat{B}$ is a uniform bound on $\hat{R}_{\lambda'}(\pi)$.

Proof. By Proposition 4,

$$\begin{aligned}
\hat{R}_{\lambda'}(\pi) &= \hat{\Delta}_{\lambda'}(\hat{\pi}_{\lambda'}, \pi) \\
&\leqslant \Delta_{\lambda'}(\hat{\pi}_{\lambda'}, \pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda) \\
&\leqslant \Delta_{\lambda'}(\pi^*_{\lambda'}, \pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda) \\
&= R_{\lambda'}(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda).
\end{aligned}$$

Therefore, using the result of 1, it follows that for any $\lambda' \in \Lambda_\epsilon$

$$\begin{aligned}
\mathop{\mathbb{E}}_{\pi\sim P_T}\left[\hat{R}_{\lambda'}(\pi)\right] &\leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda}\hat{R}_\lambda(\pi) + 2\hat{B}\sqrt{\frac{\log|\Lambda_\epsilon|}{T}} \\
&\leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda}R_\lambda(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda) + 2\hat{B}\sqrt{\frac{\log|\Lambda_\epsilon|}{T}}.
\end{aligned}$$

□

Next, we show that choosing $T = (n/\mathfrak{s}(\Lambda\|\bar{n}))^{1+\alpha}$ for some $\alpha > 0$ suffices to ensure that the optimization error is no larger than the statistical error.

Corollary 1. *For any $T = \Omega(n/\mathfrak{s}(\Lambda\|\bar{n}))$ and any $\lambda' \in \Lambda_\epsilon$, the distribution P_T satisfies*

$$\mathop{\mathbb{E}}_{\pi\sim P_T}\left[\hat{R}_{\lambda'}(\pi)\right] \leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda).$$

Proof. Since $B \geqslant |Y^s(a)|$, for any $\lambda' \in \Lambda_\epsilon$,

$$\begin{aligned}
\hat{R}_{\lambda'}(\pi) &\leqslant R_{\lambda'}(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda)\\
&\leqslant Q_{\lambda'}(\pi^*_{\lambda'}) - Q_{\lambda'}(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda)\\
&\leqslant 2B + \xi_{\epsilon,\delta}(n;\Pi,\Lambda)\\
&\leqslant 2B + \mathcal{O}_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right).
\end{aligned}$$

This last inequality gives a candidate bound for $\hat{R}_{\lambda'}(\pi)$ to apply Proposition 5. Then, if we set $T = (n/\mathfrak{s}(\Lambda\|\bar{n}))^{1+\alpha}$ for any choice of $\alpha > 0$, we have that

$$2\cdot\left(2B + \mathcal{O}_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right)\right)\cdot\sqrt{\frac{\log|\Lambda_\epsilon|}{T}} \leqslant o_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right),$$

and so

$$\mathop{\mathbb{E}}_{\pi\sim P_T}\left[\hat{R}_{\lambda'}(\pi)\right] \leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda).$$

□

C.7 Proof of Theorem 1

Theorem 1 (Mixture-Agnostic Regret Bound)**.** *Suppose Assumptions 1, 2, and 3 hold. For any $\epsilon > 0$, let Λ_ϵ denote a minimal ϵ-covering set of Λ under the ℓ_1 distance. Set $T = (n/\mathfrak{s}(\Lambda\|\bar{n}))^{1+\alpha}$ for any choice of $\alpha > 0$. Then, for any $\epsilon > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, the EG-OPO policy $\hat{\pi} = \pi^{(T)}$ achieves the regret bound*

$$R_{\lambda'}(\hat{\pi}) \leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + 2B\epsilon + \xi_{\epsilon,\delta}(n;\Pi,\Lambda)$$

for any $\lambda' \in \Lambda$, where

$$\xi_{\epsilon,\delta}(n;\Pi,\Lambda) := C_{\epsilon,\delta}\kappa(\Pi)\sqrt{V\cdot\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right),$$

where the constant $B = \max_{s\in\mathcal{S}} B_s$ is a uniform upper bound on all the potential outcomes across sources, $C_{\epsilon,\delta} = \sqrt{C\log(|\Lambda_\epsilon|/\delta)}$ with universal constant C, and $V = \max_{s\in\mathcal{S}}\max_{\pi\in\Pi}\mathbb{E}_{\bar{\mathcal{D}}_s}[\Gamma^s(\pi(X^s))^2]$ is the worst-case AIPW score variance across sources.

Proof. Fix $\epsilon > 0$. For any $\lambda' \in \Lambda_\epsilon$,

$$\begin{aligned}
R_{\lambda'}(\hat{\pi}) &= \Delta_{\lambda'}(\pi^*_{\lambda'}, \hat{\pi})\\
&= \Delta_{\lambda'}(\pi^*_{\lambda'}, \hat{\pi}) - \hat{\Delta}_{\lambda'}(\pi^*_{\lambda'}, \hat{\pi}) + \hat{\Delta}_{\lambda'}(\pi^*_{\lambda'}, \hat{\pi})\\
&\leqslant \sup_{\pi_a,\pi_b\in\Pi}|\Delta_\lambda(\pi_a,\pi_b) - \hat{\Delta}_{\lambda'}(\pi_a,\pi_b)| + \hat{\Delta}_{\lambda'}(\pi^*_{\lambda'}, \hat{\pi}).
\end{aligned}$$

By Proposition 4, the first term is bounded as

$$\sup_{\pi_a,\pi_b\in\Pi}|\Delta_{\lambda'}(\pi_a,\pi_b) - \hat{\Delta}_{\lambda'}(\pi_a,\pi_b)| \leqslant \xi_{\epsilon,\delta}(n;\Pi,\Lambda).$$

For the second term, we have

$$\begin{aligned}\hat{\Delta}_{\lambda'}(\pi^*_{\lambda'}, \hat{\pi}) &= \hat{Q}_{\lambda'}(\pi^*_{\lambda}) - \hat{Q}_{\lambda'}(\hat{\pi}) \\ &\leqslant \hat{Q}_{\lambda'}(\hat{\pi}_{\lambda'}) - \hat{Q}_{\lambda'}(\hat{\pi}) \\ &= \hat{R}_{\lambda'}(\hat{\pi}) \\ &= \mathop{\mathbb{E}}_{\pi\sim P_T}[\hat{R}_{\lambda'}(\pi)] \\ &\leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + \xi_{\epsilon,\delta}(n;\Pi,\Lambda),\end{aligned}$$

where the last inequality follows from the suboptimality bound of the HedgeOPO algorithm established in Corollary 1.

Putting it all together, we have that for any $\lambda' \in \Lambda_\epsilon$,

$$R_{\lambda'}(\hat{\pi}) \leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + 2\xi_{\epsilon,\delta}(n;\Pi,\Lambda).$$

Lastly, note that for any $\lambda'' \in \Lambda$, by construction there exists a $\lambda' \in \Lambda_\epsilon$ such that $\|\lambda'' - \lambda'\|_1 \leqslant \varepsilon$. Therefore, for any $\lambda'' \in \Lambda$,

$$\begin{aligned}R_{\lambda''}(\hat{\pi}) &= Q_{\lambda''}(\pi^*_{\lambda''}) - Q_{\lambda''}(\hat{\pi}) \\ &= \sum_{s\in\mathcal{S}} \lambda''_s \left(Q_s(\pi^*_{\lambda''}) - Q_s(\hat{\pi})\right) \\ &= \sum_{s\in\mathcal{S}} (\lambda''_s - \lambda'_s)\left(Q_s(\pi^*_{\lambda''}) - Q_s(\hat{\pi})\right) + \sum_{s\in\mathcal{S}} \lambda'_s \left(Q_s(\pi^*_{\lambda''}) - Q_s(\hat{\pi})\right) \\ &\leqslant 2B\sum_{s\in\mathcal{S}} (\lambda''_s - \lambda'_s) + \left(Q_{\lambda'}(\pi^*_{\lambda''}) - Q_{\lambda'}(\hat{\pi})\right) \\ &\leqslant 2B\|\lambda'' - \lambda'\|_1 + \left(Q_{\lambda'}(\pi^*_{\lambda'}) - Q_{\lambda'}(\hat{\pi})\right) \\ &\leqslant 2B\epsilon + R_{\lambda'}(\hat{\pi}) \\ &\leqslant 2B\epsilon + \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + 2\xi_{\epsilon,\delta}(n;\Pi,\Lambda).\end{aligned}$$

Thus,

$$\begin{aligned}\max_{\lambda\in\Lambda} R_\lambda(\hat{\pi}) &\leqslant \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + 2B\epsilon + 2\xi_{\epsilon,\delta}(n;\Pi,\Lambda) \\ &= \min_{\pi\in\Pi}\max_{\lambda\in\Lambda} R_\lambda(\pi) + 2B\epsilon + 2C\kappa(\Pi)\sqrt{V\cdot\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}\cdot\log\frac{|\Lambda_\epsilon|}{\delta}} + o_p\left(\sqrt{\frac{\mathfrak{s}(\Lambda\|\bar{n})}{n}}\right).\end{aligned}$$

Redefining $C \mapsto C/2$ gives the result. □

D BOUNDING TARGET REGRET

D.1 Proof of Theorem 2

Theorem 2 (Target Regret Bound)**.** *Let $R(\hat{\pi})$ denote the target regret of $\hat{\pi}$ under a target distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}^d$. This is bounded as*

$$R(\hat{\pi}) \leqslant 2\cdot\mathrm{disc}(\mathcal{D}_\Lambda, \mathcal{D}) + \max_{\lambda\in\Lambda} R_\lambda(\hat{\pi}),$$

where

$$\mathrm{disc}(\mathcal{D}_\Lambda, \mathcal{D}) = \min_{\lambda\in\Lambda}\max_{\pi\in\Pi} |Q_\lambda(\pi) - Q(\pi)|$$

is the minimax mixture policy value discrepancy. Moreover, if the same boundedness condition also holds for $\mathcal{D}$, then this discrepancy can be further bounded by

$$\mathrm{disc}(\mathcal{D}_\Lambda, \mathcal{D}) \leqslant B\cdot\min_{\lambda\in\Lambda}\mathrm{TV}(\mathcal{D}_\lambda, \mathcal{D}),$$

where TV *is the total variation distance.*

Proof. First, we have that

$$\begin{aligned}
R(\hat{\pi}) &= Q(\pi^*) - Q(\hat{\pi}) \\
&= Q(\pi^*) - Q(\hat{\pi}) + Q_\lambda(\pi^*) - Q_\lambda(\pi^*) + Q_\lambda(\hat{\pi}) - Q_\lambda(\hat{\pi}) \\
&= \big(Q(\pi^*) - Q_\lambda(\pi^*)\big) + \big(Q_\lambda(\hat{\pi}) - Q(\hat{\pi})\big) + Q_\lambda(\pi^*) - Q_\lambda(\pi) \\
&\leqslant 2 \max_{\pi \in \Pi} |Q(\pi) - Q_\lambda(\pi)| + Q_\lambda(\pi^*) - Q_\lambda(\hat{\pi}) \\
&\leqslant 2 \max_{\pi \in \Pi} |Q(\pi) - Q_\lambda(\pi)| + Q_\lambda(\pi^*_\lambda) - Q_\lambda(\hat{\pi}) \\
&= 2 \max_{\pi \in \Pi} |Q(\pi) - Q_\lambda(\pi)| + R_\lambda(\hat{\pi}).
\end{aligned}$$

Since $R(\hat{\pi})$ does not depend on λ, we can freely take the minimum over $\lambda \in \Lambda$ on the first term on the right-hand side of this last inequality and the maximum over $\lambda \in \Lambda$ on the second term. This gives

$$R(\hat{\pi}) \leqslant 2 \min_{\lambda \in \Lambda} \max_{\pi \in \Pi} |Q(\pi) - Q_\lambda(\pi)| + \max_{\lambda \in \Lambda} R_\lambda(\hat{\pi}).$$

Next, let $\mathcal{F}_\infty^B$ be the space of functions uniformly bounded by B. Since we assume the potential outcomes under the source distributions and the target distribution are bounded by B, it follows by the definition of integral probability metric distances under uniformly bounded test functions (Sriperumbudur et al., 2009) that

$$\begin{aligned}
\max_{\pi \in \Pi} |Q(\pi) - Q_\lambda(\pi)| &= \max_{\pi \in \Pi} \left| \mathop{\mathbb{E}}_{Z \sim \mathcal{D}} [Y(\pi(X))] - \mathop{\mathbb{E}}_{Z \sim \mathcal{D}_\lambda} [Y(\pi(X))] \right| \\
&\leqslant \max_{f \in \mathcal{F}_\infty^B} \left| \mathop{\mathbb{E}}_{Z \sim \mathcal{D}} [f(Z)] - \mathop{\mathbb{E}}_{Z \sim \mathcal{D}_\lambda} [f(Z)] \right| \\
&\leqslant B \cdot \mathrm{TV}(\mathcal{D}, \mathcal{D}_\lambda),
\end{aligned}$$

and thus,

$$\mathrm{disc}(\mathcal{D}, \mathcal{D}_\Lambda) \leqslant B \cdot \min_{\lambda \in \Lambda} \mathrm{TV}(\mathcal{D}, \mathcal{D}_\lambda) = B \cdot \mathrm{TV}(\mathcal{D}, \mathcal{D}_\Lambda).$$

□

E ADDITIONAL ALGORITHM DETAILS

E.1 Oracle Implementation Details & Complexity

At each time step, the EG-OPO algorithm requires $O(|\Lambda_\epsilon|)$ calls to the OPO oracle, and the update at each time can be done in $O(|\Lambda_\epsilon|)$ time. Therefore, the complexity of algorithm is $O(|\Lambda_\epsilon|^2 \cdot T \cdot C_{\mathrm{OPO}})$ where C_{OPO} is the complexity of the OPO oracle.

We used the policy class Π of depth-2 decision trees and we used the PolicyTree method (Sverdrup et al., 2020) as the OPO oracle for this class. This method does exact standard offline policy learning for this class of policies. It fits a fixed-depth decision tree by exhaustive search, given the set of rewards (AIPW scores) for all actions and the associated feature vectors. PolicyTree runs in $O(p^k \cdot n^k \cdot (\log n + d) + p \cdot n \cdot \log n)$ time, where k is the depth of the decision trees, p is the context dimension, d is the number of actions, and n is the number of samples.

A single 4.05 GHz Apple M3 Max CPU was used to run the experiments.

E.2 Nuisance Parameter Estimation

There is an additional upfront cost of policy value estimation. Our results rely on efficient estimation of $Q_\lambda(\pi)$ for any policy π, which in turn relies on efficient estimation of $Q_s(\pi)$. We leverage ideas of double machine learning (Chernozhukov et al., 2018) to guarantee efficient policy value estimation given only high-level conditions on the predictive accuracy of machine learning methods on estimating the nuisance parameters of doubly robust policy value estimators. In this work, we use machine learning and cross-fitting strategies to estimate the nuisance parameters locally. The nuisance parameter estimates must satisfy the conditions of Assumption 3. Under these conditions, extensions of the results of (Chernozhukov et al., 2018; Athey and Wager, 2021) would imply that the doubly robust local policy value estimates $\hat{Q}_s(\pi)$ for any policy π are asymptotically efficient for estimating $Q_s(\pi)$.

The conditions and estimators that guarantee these error assumptions have been extensively studied in the estimation literature. These include parametric or smoothness assumptions for non-parametric estimation. The conditional response function $\mu_s(x;a) = \mathbb{E}_{\bar{\mathcal{D}}_s}[Y^s(a)|X^s = x]$ can be estimated by regressing observed rewards on observed contexts. The inverse conditional propensity function $\omega_s(x;a) = 1/\mathbb{P}_{\bar{\mathcal{D}}_s}(A^s = a|X^s = x)$ can be estimated by estimating the conditional propensity function $e_s(x;a) = \mathbb{P}_{\bar{\mathcal{D}}_s}(A^s = a|X^s = x)$ and then taking the inverse. Under sufficient regularity and overlap assumptions, this gives accurate estimates. We can take any flexible approach to estimate these nuisance parameters. We could use standard parametric estimation methods like logistic regression and linear regression, or we could use non-parametric methods like classification and regression forests to make more conservative assumptions on the true models. Lastly, we note that if it is known that some sources have the same data-generating distribution, it should be possible to learn the nuisance parameters across similar sources.

In our experiments, we estimate the nuisance parameters with the non-parametric method of generalized random forests and we make use of the `grf` package (Athey et al., 2019).

E.3 Cross-fitted AIPW Estimation

Once the nuisance parameters are estimated, they can be used for estimating AIPW scores. Refer to Algorithm 2 for the pseudocode on how we conduct the cross-fitting strategy for AIPW score estimation. Under this strategy, each source $s \in \mathcal{S}$ divides their local dataset into K folds, and for each fold k, the source estimates μ_s and ω_s using the rest $K-1$ folds. During AIPW estimation for a single data point, the nuisance parameter estimate that is used in the AIPW estimate is the one that was not trained on the fold that contained that data point. This cross-fitting estimation strategy is described in additional detail in (Zhou et al., 2023).

Algorithm 2 Cross-fitted AIPW: source-Side

```
Require: local data $\{(X_i^s, A_i^s, Y_i^s)\}_{i=1}^{n_s}$, nuisance parameter estimates $\hat{\mu}_s$ and $\hat{\omega}_s$, number of folds $K$
Partition local data into $K$ folds
Define surjective mapping $k_c : [n_s] \to [K]$ of point index to corresponding fold index
for $k = 1, \dots, K$ do
    Fit estimators $\hat{\mu}_s^{-k}$ and $\hat{\omega}_s^{-k}$ using rest of data not in fold $k$
end for
for $i = 1, \dots, n_s$ do
    for $a \in \mathcal{A}$ do
        $\hat{\Gamma}_i^s(a) \leftarrow \hat{\mu}_s^{-k_s(i)}(X_i^s; a) + \big(Y_i^s - \hat{\mu}_s^{-k_s(i)}(X_i^s; a)\big) \cdot \hat{\omega}_s^{-k_s(i)}(X_i^s; a) \cdot \mathbf{1}\{A_i^s = a\}$
    end for
end for
```

F ADDITIONAL DISCUSSION

F.1 Limitations & Future Work

Our work has several limitations that warrant further exploration. First, we make certain assumptions about the data-generating process that may not always hold. While we have discussed potential relaxations, such as relaxing the boundedness and uniform overlap assumptions in the data-generating distributions, further investigation is needed to fully understand the impact of these adjustments. In particular, interesting question arises regarding the pessimism principle in overcoming the uniform overlap assumption (Jin et al., 2022). Specifically, we wonder if it would be necessary to have coverage under the locally optimal policy for each data source, which could influence the robustness of our approach.

Moreover, in this work, we estimate nuisance parameters separately for each data source. When sources share the same data-generating distribution, there is an opportunity to improve efficiency by learning nuisance parameters across similar sources, rather than treating them independently.

Finally, we have not fully established whether the regret bounds we provide are optimal. Prior work (Mohri et al., 2019) suggests that skewness-based bounds for distributed supervised learning are optimal, and our results reduce to regret-optimal results from (Athey and Wager, 2021; Zhou et al., 2023) when the sources are identical. This provides strong indications of optimality, but establishing lower bounds in our setting is an open area for future research.