DISTRIBUTIONALLY ROBUST POLICY LEARNING UNDER CONCEPT DRIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributionally robust policy learning aims to find a policy that performs well under the worst-case distributional shift, and yet most existing methods for robust policy learning consider the worst-case *joint* distribution of the covariate and the outcome. The joint-modeling strategy can be unnecessarily conservative when we have more information on the source of distributional shifts. This paper studies a more nuanced problem — robust policy learning under the *concept drift*, when only the conditional relationship between the outcome and the covariate changes. To this end, we first provide a doubly-robust estimator for evaluating the worst-case average reward of a given policy under a set of perturbed conditional distributions. We show that the policy value estimator enjoys asymptotic normality even if the nuisance parameters are estimated with a slower-than-root-n rate. We then propose a learning algorithm that outputs the policy maximizing the estimated policy value within a given policy class Π, and show that the sub-optimality gap of the proposed algorithm is of the order $\kappa(\Pi)n^{-1/2}$, with $\kappa(\Pi)$ is the entropy integral of Π under the Hamming distance and n is the sample size. The proposed methods are implemented and evaluated in numerical studies, demonstrating substantial improvement compared with existing benchmarks.

1 INTRODUCTION

In a wide range of fields, the abundance of user-specific historical data provides opportunities for learning efficient individualized policies. Examples include learning the optimal personalized treatment from electronic health record data (Murphy, 2003; Kim et al., 2011; Chan et al., 2012), or obtaining an individualized advertising strategy using past customer behavior data (Bottou et al., 2013; Kallus & Udell, 2016). Driven by such a practical need, a line of works have been devoted to developing efficient policy learning algorithms using historical data — a task often known as *offline policy learning* (Dudík et al., 2011; Zhang et al., 2012; Swaminathan & Joachims, 2015a;b;c; Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Zhou et al., 2023; Zhan et al., 2023; Bibaut et al., 2021; Jin et al., 2021; 2022a).

Most existing methods for offline policy learning deliver performance guarantees under the premise that the target environment remains the same as that from which the historical data is collected. It has been widely observed, however, that such a condition is hardly met in practice (see e.g., Recht et al. (2019); Namkoong et al. (2023); Liu et al. (2023); Jin et al. (2023) and the references therein). Under distribution shift, a policy learned in one environment often shows degraded performance when deployed in another environment. To address this issue, there is an emerging body of research on *robust policy learning*, which aims at finding a policy that still performs well when the target distribution is perturbed. Pioneering works in this area consider the case where the *joint distribution* of the covariates and the outcome is shifted from the training distribution, and propose algorithms that output a policy achieving reliable worst-case performance under the aforementioned shifts Si et al. (2023); Kallus et al. (2022). The joint modeling approach, however, ignores the *type* of distributional shifts, and the resulting worst-case value can be unnecessarily conservative in practice.

Indeed, distributional shifts can be categorized into two classes by their sources: (1) the shift in the covariate X, and/or (2) the shift in the conditional relationship between the outcome Y and the covariate X. The two types of distributional shifts have different implications in differnt applications, and call for distinct treatment (Namkoong et al., 2023; Liu et al., 2023; Jin et al., 2023; Ai & Ren, 2024). For example, when the distribution of covariates changes while that of $Y \mid X$ remains invariant, the distribution shift is identifiable/estimable since the covariates are often accessible in the target in environment. Alternatively, when the $Y \mid X$ distribution changes but the X distribution remains invariant, the distribution shift is no longer identifiable, and we need to account for the worst-case situation. This setting, known as *concept drift*, occurs when the distribution of the unobserved confounder changes over time, or due to sudden external shocks (Widmer & Kubat, 1996; Lu et al., 2018; Gama et al., 2014). For example, in advertising, the customer behavior can evolve over time as the environment changes, while the population remains largely the same. In personalized medicine, treatment may be affecting patients' outcomes through some unmeasured confounders that have different distributions in the training and target cohort, thereby inducing a concept drift.

Our work mainly focuses on robust policy learning under concept drift. Most existing methods for robust policy learning (Si et al., 2023; Kallus et al., 2022) that model the distributional shift jointly without distinguishing the sources, and the corresponding algorithms turn out to be suboptimal. The reason behind their suboptimality is that the worst-case distributions under the two models — the joint-shift model and the concept-drift model — can be substantially different, so it would be a "waste" of our budget to consider adversarial distributions that are not feasible under concept drift. It is worth mentioning that a recent paper by Mu et al. (2022) accounts for the sources of distributional shifts in policy learning; their approach, however, applies only when the covariates take *a finite number of values*, and therefore is limited in its applicability. When the covariate space is infinite, it remains unclear how to efficiently learn a robust policy under concept drift. The current work aims to fill in the gap by answering the following question:

How can we efficiently learn a policy with optimal worst-case average performance under concept drift with minimal assumptions?

We provide a rigorous answer to the above question. Specifically, we assume the covariate distribution remains the same in the training and target environments, while the $Y \mid X$ distribution shift is bounded in KL-divergence by a pre-specified constant δ. Our goal is to find a policy that maximizes the worst-case averaged outcome over all possible target distributions satisfying the previous condition.

1.1 Our contributions

Towards robust policy learning under concept drift, we make the following contributions.

Policy evaluation. Given a policy, we present a doubly-robust estimator for the worst-case policy value under concept drift. We prove that the estimator is asymptotic normal under mild conditions on the estimation rate of the nuisance parameter. Our approach involves first formulating the worst-case policy value under the concept drift model as the optimal objective value of a distributionally robust optimization problem with KL-divergence constraints. The optimization problem is then solved in its dual form. Finally, we plug in the empirical risk optimizer into the dual objective function and take a debiased step to obtain the final estimator.

Policy learning. We propose a robust policy learning algorithm that outputs a policy maximizing the estimated policy value over a policy class Π. Compared with the oracle optimal policy, the policy provided by our algorithm with high probability has a suboptimality gap of the order $\kappa(\Pi) \cdot n^{-1/2}$, where $\kappa(\Pi)$ is a measure quantifying the policy class complexity (to be formalized shortly) and n is the number of samples. Compared with Mu et al. (2022), our algorithm and theory apply to general covariate spaces and potentially infinite policy classes, while their method is restricted to finite covariate space and policy class. Furthermore, the sample dependence of our sub-optimality gap is $O(n^{-1/2})$, which is sharper that the $(n \log n)^{-1/2}$ rate in Mu et al. (2022).

Implementation and empirics. We provide efficient implementation of our robust policy learning algorithm, and compare its empirical performance with existing benchmarks in numerical studies. Our proposed method exhibits substantial improvement.

1.2 Related works

Offline policy learning. There is a long list of works devoted to offline policy learning. Most of them assume no distributional shifts (e.g., Dudík et al. (2011); Zhang et al. (2012); Swaminathan & Joachims (2015a;b;c); Kitagawa & Tetenov (2018); Athey & Wager (2021); Zhou et al. (2023)). Zhan et al. (2023); Jin et al. (2021; 2022a) allow the data to be adaptively collected, but the distribution over the covariate and the (potential) outcomes remain invariant in the training and target environment.

As mentioned earlier, the work of Si et al. (2023); Kallus et al. (2022) study robust policy learning when the joint distribution of (X, Y) ranges in the neighborhood of the training distribution; Mu et al. (2022) consider the case when the covariate shift and $Y \mid X$ shift are specified separately; their method, however, is restricted to finite covariate space, and their sub-optimality gap is logarithmic factors slower than parametric rates. The work of Kallus & Zhou (2021) concerns robust policy learning when the distribution shift is caused by hidden confounders — this is in fact a special type of concept drift — and the corresponding $Y \mid X$ shift is assumed to be bounded uniformly, which is quite different from our f-divergence bound. More recently, Guo et al. (2024) considers a pure covariate shift with a focus on policy evaluation, where the setup and the goal are different from ours.

Distributionally robust optimization. More broadly, our work is also closely related to DRO, where the goal is to learn a model that has good performance under the worst-case distribution (e.g., Bertsimas & Sim (2004); Delage & Ye (2010); Hu & Hong (2013); Duchi et al. (2019); Dudík et al. (2011); Zhang et al. (2023)). The major focus of the aforementioned works involves parameter estimation and prediction in supervised settings; we however take a decision-making perspective and aim at learning a individualized policy with optimal worst-case performance guarantees.

1.3 Notation

We use $[n]$ to denote the discrete set $\{1, 2, \cdots, n\}$ for any $n \in \mathbb{Z}$. We use argmin and argmax to denote the minimizers and maximizers; if the minimzer or the maximizer cannot be attained, we project it back to the feasible set. We denote the usual p-norm as $\|\cdot\|_p$. For any probability measure P defined on the probability space $(\Omega, \sigma(\Omega), P)$. For any function f, we denote the $L_2(P)$-norm of f conventionally as $\|f\|_{L_2(P)} = (\int |f(x)|^2 \, dP(x))^{1/2}$ and $\|f\|_{L_\infty} = \sup_{x\in\mathcal{X}} |f(x)|$. We use $\widehat{P}$ to denote the empirical distribution of P. For any random variables X, Y, we use $X \perp\!\!\!\perp Y$ to denote that X is independent of Y. For a random variable/vector X, we use $\mathbb{E}_X[\cdot]$ to indicate the expectation taken over the distribution of X.

2 Problem formulation

Consider a set of M actions denoted by $[M]$ and let $\mathcal{X} \subseteq \mathbb{R}^d$. Throughout the paper, we follow the potential outcome framework (Imbens & Rubin, 2015), where $Y(a) \in \mathcal{Y}_a \subseteq \mathbb{R}$ denotes the potential outcome had action a been taken for any $a \in [M]$. We posit the underlying data-generating distribution P on the joint covariate-outcome random vector $(X, Y(1), \cdots, Y(M)) \in \mathcal{X} \times \prod_{a=1}^{M} \mathcal{Y}_a$. Consider a data set $\mathcal{D} = \{(X_i, A_i, Y_i)\}_{i\in[n]}$ consisting of n i.i.d. draws of (X, A, Y), where $X_i \in \mathcal{X}$ is the observed contextual vector, $A_i \in [M]$ the action, and $Y_i = Y(A_i)$ the realized reward. The actions are selected by the *behavior policy* π_0, where $\pi_0(a \mid x) := \mathbb{P}(A_i = a \mid X = x)$, for any $a \in [M], x \in \mathcal{X}$. We make the following assumptions for π_0 and P.

Assumption 2.1. The behavior policy π_0 and the joint distribution P satisfy the following.

(1) *Unconfoundedness:* $(Y(1), \cdots, Y(M)) \perp\!\!\!\perp A \mid X$.

(2) *Overlap:* for some $\varepsilon > 0$, $\pi_0(a \mid x) \geq \varepsilon$, for all $(a, x) \in [M] \times \mathcal{X}$.

(3) *Bounded reward support:* there exists $\bar{y} > 0$, such that $0 \leq Y(a) \leq \bar{y}$ for all $a \in [M]$.

The above assumptions are standard in the literature (see e.g., Athey & Wager, 2021; Zhou et al., 2023; Si et al., 2023; Kallus et al., 2022). In particular, the unfoundedness assumption guarantees identifiability, and the overlap assumption ensures sufficient exploration when collecting the training dataset. The bounded reward support is assumed for the ease of exposition, and can be relaxed to the sub-Gaussian reward straightforwardly.

2.1 The KL-distributionally robust formulation

Given the training set $\mathcal{D} = \{(X_i, A_i, Y_i)\}_{i\in[n]}$ and a policy class Π, we aim to learn a policy $\pi \in \Pi$ that achieves high expected reward in a target environment that may deviate from the data-collection environment where $\mathcal{D}$ is collected. While distribution shift can take place in various forms, we focus primarily on the concept drift, where only the conditional reward distribution $Y(a) \mid X$ differs in the training and target environment. The distance between distributions is quantified by the KL divergence.

Definition 2.2 (KL divergence)**.** The KL divergence between two distributions Q and P is defined as $D_{\mathrm{KL}}(Q \,\|\, P) = \mathbb{E}_Q[\log \frac{dQ}{dP}]$, where $\frac{dQ}{dP}$ is the Radon-Nikodym derivative of Q with respect to P.

We define an uncertainty set of neighboring distributions around P, whose conditional outcome distribution is bounded in KL divergence from P. Given a radius $\delta > 0$, the uncertainty set of the conditional distribution is defined as

$$\mathcal{P}(P_{Y\,|\,X}, \delta) := \big\{Q_{Y\,|\,X} : D_{\mathrm{KL}}(Q_{Y\,|\,X} \,\|\, P_{Y\,|\,X}) \le \delta\big\},$$

where $P_{Y\,|\,X}$ and $Q_{Y\,|\,X}$ refers to the distribution of $(Y(1), \ldots, Y(d)) \mid X$ under P and Q respectively. The distributionally robust policy value for any policy π at level δ is defined as

$$\mathcal{V}_\delta(\pi) := \mathbb{E}_{P_X}\left[\inf_{Q_{Y\,|\,X} \in \mathcal{P}(P_{Y\,|\,X}, \delta)} \mathbb{E}_{Q_{Y\,|\,X}}\Big[Y\big(\pi(X)\big) \,\Big|\, X\Big]\right]. \tag{1}$$

The optimal policy in Π is the one that maximizes $\mathcal{V}_\delta(\pi)$, i.e. $\pi^*_\delta := \mathrm{argmax}_{\pi\in\Pi}\ \mathcal{V}_\delta(\pi)$.[1]

Under this formulation, our goal is to learn a "robust" policy with a high value of $\mathcal{V}_\delta(\pi)$ using a dataset drawn from P. The task here is two-fold: we need to (i) estimate the policy value $\mathcal{V}_\delta(\pi)$ for a given policy π, and (ii) find a near-optimal robust policy $\widehat{\pi} \in \Pi$ whose policy value is close to the optimal policy π^*_δ. Here, the performance of a learned policy $\widehat{\pi}$ is measured by the sub-optimality gap (regret), defined as

$$\mathcal{R}_\delta(\widehat{\pi}) := \mathcal{V}_\delta(\pi^*_\delta) - \mathcal{V}_\delta(\widehat{\pi}). \tag{2}$$

In the following sections, we tackle each task sequentially.

2.2 Strong duality

In order to estimate $\mathcal{V}_\delta(\pi)$, we first rewrite the inner optimization problem in Equation (1) in its dual form using standard results in convex optimization. The transformation is formalized in the following lemma, with its proof provided in Appendix B.1.

Lemma 2.3 (Strong Duality)**.** *Given any $\pi \in \Pi$ and any $x \in \mathcal{X}$, the optimal value of inner optimization problem in Equation (1) equals to*

$$-\min_{\alpha\ge 0, \eta\in\mathbb{R}} \mathbb{E}_P\left[\alpha \exp\Big(-\frac{Y(\pi(X)) + \eta}{\alpha} - 1\Big) + \eta + \alpha\delta \,\Big|\, X = x\right]. \tag{3}$$

We note that the optimization problem in (3) depends on x and π — to manifest this dependence, we use $(\alpha^*_\pi(x), \eta^*_\pi(x))$ to denote its optimizer, i.e.,

$$\big(\alpha^*_\pi(x), \eta^*_\pi(x)\big) \in \underset{\alpha\ge 0, \eta\in\mathbb{R}}{\mathrm{argmin}}\ \mathbb{E}_P\left[\alpha \exp\Big(-\frac{Y(\pi(X)) + \eta}{\alpha} - 1\Big) + \eta + \alpha\delta \,\Big|\, X = x\right].$$

With this notation and Lemma 2.3, the robust policy value becomes

$$\mathcal{V}_\delta(\pi) = -\mathbb{E}_P\left[\alpha^*_\pi(X) \exp\Big(-\frac{Y(\pi(X)) + \eta^*_\pi(X)}{\alpha^*_\pi(X)} - 1\Big) + \eta^*_\pi(X) + \alpha^*_\pi(X)\delta\right]. \tag{4}$$

The above formulation has thus translated the original distributionally robust optimization problem into an *empirical risk minimization (ERM)* problem. We note that, unlike the well-studied joint

[1] When the supremum cannot be attained, we can always construct a sequence of policies whose policy values converge to the supremum, and all the arguments go through with a limiting argument.

distributional shift formulation, the above representation admits an optimizer pair $(\alpha^*_\pi(x), \eta^*_\pi(x))$ that is *dependent* on the context x (i.e. α^*_π, η^*_π are functions of x) and the policy π. As we shall see shortly, our proposed policy value estimation procedure employs ERM tools to estimate $(\alpha^*_\pi, \eta^*_\pi)$, and then compute an estimate of $\mathcal{V}_\delta(\pi)$ by plugging $(\alpha^*_\pi, \eta^*_\pi)$ into Equation (4). The remaining challenge in this proposal is the slow estimation rate of the optimizers — if we naïvely plug in the optimizers, the resulting policy value estimator typically has a convergence rate slower than root-n. To overcome this, we incorporate a novel adjustment method to debias the estimator, which allows us to obtain a doubly-robust estimator that achieves root-n rate of convergence even when then nuisance parameters (e.g., $(\alpha^*_\pi, \eta^*_\pi)$) are converging slower than the root-n rate.

We end this section by discussing when $\alpha^*_\pi(x) > 0$. Throughout, we shall make the following mild assumption on the conditional outcome distribution.

Assumption 2.4. For $a \in [M]$ and $x \in \mathcal{X}$, define $\underline{y}(x;a) = \sup\{t : \mathbb{P}(Y(a) < t \mid X = x, A = a) = 0\}$ and $\tilde{p}(x;a) = \mathbb{P}(Y(a) = \underline{y}(x;a) \mid X = x, A = a)$. Let $f^*_{\mathrm{KL}}(x) = e^{x-1}$. It holds that $\tilde{p}(x;a) f^*_{\mathrm{KL}}(1/\tilde{p}(x;a)) + (1 - \tilde{p}(x;a)) f^*_{\mathrm{KL}}(0) > \delta$ for $P_{X|A=a}$-almost all x.

The above assumption requires that $P_{Y|X,A}$ does not posit a large point mass at its essential infimum, which can be satisfied by many commonly used distributions, e.g., all the continuous distributions. Next, the following result from Jin et al. (2022b, Proposition 4), shows that $\alpha^* > 0$ when Assumption 2.4 holds, which ensures that the gradient of the risk function in ERM has a zero mean.

Proposition 2.5 (Jin et al. (2022b))**.** *Under Assumption 2.4, the optimizer α^* of (3) satisfies $\alpha^* > 0$.*

3 Policy value estimation under concept drift

3.1 The estimation procedure

Fixing a policy π, we aim to estimate the policy value $\mathcal{V}_\delta(\pi)$ using the training dataset $\mathcal{D}$. We first split $\mathcal{D}$ into K equally sized disjoint folds, $\mathcal{D}^{(k)}$ for $k \in [K]$,[2] where we slightly abuse the notation and $\mathcal{D}^{(k)}$ to denote the data points or the corresponding indices interchangeably.

For each $k \in [K]$, we use data points in $\mathcal{D}^{(k+1)}$ to obtain the propensity score estimator $\widehat{\pi}_0^{(k)}$ and the optimizers $(\widehat{\alpha}_\pi^{(k)}, \widehat{\eta}_\pi^{(k)})$.[3] Next, we define

$$\widehat{G}_\pi^{(k)}(x, y) := \widehat{\alpha}_\pi^{(k)}(x) \cdot \exp\Big(- \frac{y + \widehat{\eta}_\pi^{(k)}(x)}{\widehat{\alpha}_\pi^{(k)}(x)} - 1 \Big) + \widehat{\eta}_\pi^{(k)}(x) + \widehat{\alpha}_\pi^{(k)}(x) \cdot \delta,$$

and its conditional expectation

$$\bar{g}_\pi^{(k)}(x) := \mathbb{E}_P\Big[\widehat{G}_\pi^{(k)}\big(X, Y(\pi(X))\big) \,\Big|\, X = x \Big].$$

We then use $\mathcal{D}^{(k+2)}$ to obtain $\widehat{g}_\pi^{(k)}$ as an estimator of g_π. The policy value estimator $\widehat{\mathcal{V}}_\delta^{(k)}(\pi)$ for the k-th fold is constructed as

$$\widehat{\mathcal{V}}_\delta^{(k)}(\pi) = \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{\pi(X_i) = A_i\}}{\widehat{\pi}_0^{(k)}(A_i \mid X_i)} \cdot \big(\widehat{G}_\pi^{(k)}(X_i, Y_i) - \widehat{g}_\pi^{(k)}(X_i) \big) + \widehat{g}_\pi^{(k)}(X_i). \tag{5}$$

The final policy value estimator is given by

$$\widehat{\mathcal{V}}_\delta(\pi) := -\frac{1}{K} \sum_{k=1}^{K} \widehat{\mathcal{V}}_\delta^{(k)}(\pi).$$

The complete procedure is summarized in Algorithm 1. A few remarks are in order.

Remark 3.1. The estimation procedure involves three model-fitting steps corresponding to π_0, (α_π, η_π), and $\bar{g}_\pi$, respectively. The propensity score function π_0 can be estimated with off-the-shelf algorithms (e.g., logistic regression, random forest); the conditional mean $g_\pi^{(k)}$ can be obtained by

[2] in practice, we only need a minimum of $K = 3$ folds.

[3] We use the convention that $\mathcal{D}^{(k+j)} = \mathcal{D}^{(k+j \bmod K)}$ for any j, k.

Algorithm 1 Policy estimation under concept drift

```
Input: Dataset $\mathcal{D}$; policy $\pi$; uncertainty set parameter $\delta$; propensity score estimation algorithm $\mathcal{C}$; ERM algorithm $\mathcal{E}$ for obtaining $(\alpha_\pi^*, \eta_\pi^*)$; regression algorithm $\mathcal{R}$ for estimating $\bar{g}_\pi$.

Randomly split $\mathcal{D}$ into $K$ non-overlapping equally-sized folds $\mathcal{D}^{(k)}$, $k \in [K]$;
for $k = 1, \cdots, K$ do
    On $\mathcal{D}^{(k+1)}$: $\widehat{\pi}_0^{(k)} \leftarrow \mathcal{C}(\mathcal{D}^{(k+1)})$, $(\widehat{\alpha}_\pi^{(k)}, \widehat{\eta}_\pi^{(k)}) \leftarrow \mathcal{E}(\mathcal{D}^{(k+1)})$;
    On $\mathcal{D}^{(k+2)}$: $\widehat{\bar{g}}_\pi^{(k)}(\cdot) \leftarrow \mathcal{R}\big(\{X_i, A_i, \widehat{G}_\pi^{(k)}(X_i, Y_i); i \in \mathcal{D}^{(k+2)}\}\big)$;
    On $\mathcal{D}^{(k)}$: compute $\widehat{\mathcal{V}}_\delta^{(k)}(\pi)$ according to Equation (5);
end for
Return: $\widehat{\mathcal{V}}_\delta(\pi) \leftarrow -\frac{1}{K}\sum_{k=1}^{K} \widehat{\mathcal{V}}_\delta^{(k)}(\pi)$.
```

regressing $\widehat{G}_\pi^{(k)}(X_i, Y_i)$ onto X_i for the points such that $A_i = \pi(X_i)$ with standard regression algorithms, e.g., kernel regression (Nadaraya, 1964; Watson, 1964), local polynomial regression (Cleveland, 1979; Cleveland & Devlin, 1988), smoothing spline (Green & Silverman, 1993), regression trees (Loh, 2011) and random forests (Ho et al., 1995). The ERM step is more complex, and will be discussed in detail shortly.

Remark 3.2. The construction of the estimator $\widehat{\mathcal{V}}_\delta(\pi)$ employs two major techniques: cross-fitting and de-biasing. The cross-fitting technique crucially provides the convenient property of independence and the de-biasing technique overcomes the slow rate of estimating the nuisance parameter α_π, η_π, leading to the doubly-robust property of the proposed estimator.

The ERM step. For notational simplicity, we denote $\theta = (\alpha, \eta)$ and write the loss function as

$$\ell(x, y; \theta) = \alpha(x) \exp\Big(-\frac{y + \eta(x)}{\alpha(x)} - 1\Big) + \eta(x) + \alpha(x)\delta. \tag{6}$$

By the notation, $\theta_\pi^*(x) = (\alpha_\pi^*(x), \eta_\pi^*(x))$ is the optimizer of $\mathbb{E}_P[\ell(x, Y(\pi(x)); \theta^*) \,|\, X = x]$. Throughout, we make the following assumption on θ_π^*.

Assumption 3.3. For any policy π, there exist constants $\underline{\alpha}, \bar{\alpha}, \bar{\eta}$ such that

$$0 < \underline{\alpha} \le \alpha_\pi^*(x) \le \bar{\alpha}, \quad \big|\eta_\pi^*(x)\big| \le \bar{\eta}, \quad \text{for all } x \in \mathcal{X}.$$

The above assumption is quite mild. It can be achieved, for example, when $\theta_\pi^*(x)$ is continuous in x and when $\mathcal{X}$ is compact. We refer the readers to Jin et al. (2022b) for a more detailed discussion.

Under the unconfoundedness assumption, it can be seen that θ_π^* is also a minimizer of $\mathbb{E}_P\big[\ell(X, Y; \theta)\mathbb{1}\{A = \pi(X)\}\big]$. We obtain an estimate of θ_π^* by minimizing the empirical risk:

$$\widehat{\theta}_\pi^{(k)} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \frac{1}{|\mathcal{D}^{(k+1)}|} \sum_{i \in D^{(k+1)}} \mathbb{1}\{A_i = \pi(X_i)\} \cdot \ell(X_i, Y_i; \theta) \right\}, \tag{7}$$

where $\Theta \subseteq \{(\alpha, \eta) : \alpha(x) \ge 0, \eta(x) \in \mathbb{R}, \text{ for any } x \in \mathcal{X}\}$ is to be determined. In our implementation, we follow Yadlowsky et al. (2022); Jin et al. (2022b); Sahoo et al. (2022), and adopt the method of sieves (Geman & Hwang, 1982) to solve (7). Specifically, we consider an increasing sequence $\Theta_1 \subset \Theta_2 \subset \cdots$ of spaces of smooth functions, and let $\Theta = \Theta_n$ in Equation (7). For example, Θ_n can be a class of polynomials, splines, or wavelets. It has been shown in Jin et al. (2022b, Section 3.4) that under mild regularity conditions, $\widehat{\theta}_\pi^{(k)}$ converges to θ_π^* at a non-parametric rate. For example, if $\mathcal{X} = \prod_{j=1}^d \mathcal{X}_j \subseteq \mathbb{R}^d$ for some compact intervals $\mathcal{X}_j$ and that θ_π^* belongs to the Hölder class of p-smooth functions, with some other mild regularity conditions, then $\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_2(P_{X \mid A = \pi(X)})} = O_P((\frac{\log n}{n})^{-p/(2p+d)})$ and $\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_\infty} = O_P((\frac{\log n}{n})^{-2p^2/(2p+d)^2})$. We refer the readers to Yadlowsky et al. (2018) and Jin et al. (2022b) for more details.

3.2 Theoretical Guarantees

We are now ready to present the theoretical guarantees for the policy value estimator $\widehat{\mathcal{V}}_\delta(\pi)$. To start, we make the following assumption on the convergence rates of the nuisance parameter estimators.

Assumption 3.4 (Asymptotic estimation rate)**.** For any policy π, assume that for each $k \in [K]$,

(a) The estimators $\widehat{\pi}_0^{(k)}$ and $\widehat{g}_\pi^{(k)}$ satisfy

$$\|\widehat{\pi}_0^{(k)} - \pi_0\|_{L_2(P_{X \mid A=\pi(X)})} = o_P(n^{-\gamma_\pi}),\ \|\widehat{g}_\pi^{(k)} - \bar{g}_\pi^{(k)}\|_{L_2(P_{X \mid A=\pi(X)})} = o_P(n^{-\gamma_g}),$$

for some $\gamma_\pi, \gamma_g \geq 0$ and $\gamma_\pi + \gamma_g \geq \frac{1}{2}$.

(b) The optimizer $\widehat{\theta}_\pi^{(k)}$ satisfies

$$\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_2(P_{X \mid A=\pi(X)})} = o_P(n^{-\frac{1}{4}}),\ \|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_\infty} = o_P(1).$$

Assumption 3.4 (a) requires either the propensity score or the conditional mean of $\widehat{G}_\pi^{(k)}(X, Y)$ is well estimated, and is standard in the double machine learning literature (Chernozhukov et al., 2018; Athey & Wager, 2021; Zhou et al., 2023; Kallus et al., 2019; 2022; Jin et al., 2022b) and can be achieved by various commonly-used machine learning methods discussed in Section 3.1. Assumption 3.4 (b) requires the optimizer $\widehat{\theta}_\pi^{(k)}$ to be estimated at a rate faster than $n^{-1/4}$, and can be achieved by, for example, the estimators discussed in Section 3.1 under mild conditions.

The following theorem states that our estimated policy value $\widehat{\mathcal{V}}_\delta(\pi)$ is consistent for estimating $\mathcal{V}_\delta$ and is asymptotically normal. Its proof is provided in Appendix B.2.

Theorem 3.5 (Asymptotic normality)**.** *Suppose Assumptions 2.1, 2.4, 3.3, and 3.4 hold. For any policy $\pi : \mathcal{X} \mapsto \mathcal{A}$, we have*

$$\sqrt{n} \cdot \big(\widehat{\mathcal{V}}_\delta(\pi) - \mathcal{V}_\delta(\pi)\big) \xrightarrow{d} N(0, \sigma_\pi^2),$$

where

$$\sigma_\pi^2 = \mathrm{Var}\bigg(\frac{\mathbb{1}\{A = \pi(X)\}}{\pi_0(A \mid X)} \cdot \big(G(X,Y) - g(X)\big) + g(X)\bigg);$$
$$G_\pi(x, y) = \ell(x, y; \theta_\pi^*) \text{ and } g_\pi(x) := \mathbb{E}\big[G_\pi(X, Y(\pi(X))) \mid X = x\big].$$

4 Policy learning under concept drift

Building on the results and methodology in Section 3, we turn to the problem of policy learning under concept drift.

Given a policy class Π and an estimated policy value $\widehat{\mathcal{V}}_\delta(\pi)$ for each $\pi \in \Pi$, it is natural to consider optimizing the estimated policy value over Π to find the best policy. The biggest challenge here is that the quantity $\widehat{\theta}_\pi^{(k)}$ in defining $\widehat{\mathcal{V}}_\delta(\pi)$ is not only a function of $x \in \mathcal{X}$, but also a function of $\pi \in \Pi$. The above strategy requires carrying out the ERM step in Section 3.1, for all possible policies $\pi \in \Pi$, posing major computational difficulties.

Instead of solving $\widehat{\theta}_\pi^{(k)}$ for each $\pi \in \Pi$, we propose an alternative strategy that solves a similar ERM problem for each action $a \in [M]$. To see why this is sufficient, note that for any $\pi \in \Pi$,

$$\mathbb{E}\big[\ell(X, Y(\pi(X)); \theta) \mid X = x\big] = \sum_{a=1}^{M} \mathbb{1}\{\pi(X) = a\} \cdot \mathbb{E}[\ell(x, Y(a); \theta) \mid X = x]. \tag{8}$$

Letting $\theta_a^*(x) \in \underset{\theta}{\mathrm{argmin}}\,\big\{\mathbb{E}[\ell(x, Y(a); \theta) \mid X = x]\big\}$, we can see that $\theta_{\pi(x)}^*(x)$ is a minimizer of (8). Then, the policy learning problem reduces to finding $\pi \in \Pi$ that maximizes

$$-\mathbb{E}\Bigg[\alpha_{\pi(X)}^*(X) \cdot \exp\bigg(-\frac{Y(\pi(X)) + \eta_{\pi(X)}^*(X)}{\alpha_{\pi(X)}^*(X)} - 1\bigg) + \eta_{\pi(X)}^*(X) + \alpha_{\pi(X)}^*(X)\delta\Bigg].$$

The following section instantiates this idea and provides a detailed algorithm for policy learning under concept drift.

Algorithm 2 Policy learning under concept drift

```
Input: Dataset $\mathcal{D}$; policy class $\Pi$; uncertainty set parameter $\delta$; propensity score estimation algorithm $\mathcal{C}$; ERM algorithm $\mathcal{E}(\cdot)$ for obtaining $\theta_a^*$; regression algorithm $\mathcal{R}$ for estimating $\bar{g}_a$.

Randomly split $\mathcal{D}$ into $K$ equal-sized folds;
for $k = 1, \ldots, K$ do
    $\widehat{\pi}_0^{(k)} \leftarrow \mathcal{C}(\mathcal{D}^{(k+1)})$,
    for $a = 1, \cdots, M$ do
        $\widehat{\theta}_a^{(k)} \leftarrow \mathcal{E}(\mathcal{D}^{(k+1)})$;
        $\widehat{g}_a^{(k)} \leftarrow \mathcal{R}(X_i, A_i, \widehat{G}_a^{(k)}(X_i, Y_i); i \in \mathcal{D}^{(k+2)})$;
    end for
end for

Return: $\widehat{\pi}_{\mathrm{LN}}$ that maximizes $\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi)$ as in Equation (9).
```

4.1 The learning algorithm

The policy learning algorithm consists of two main steps: (1) solving for θ_a^* for each $a \in [M]$ and constructing the policy value estimator $\widehat{\mathcal{V}}_\delta(\pi)$; (2) learning the optimal policy π_δ^*.

As before, we randomly split the original data set $\mathcal{D}$ into K folds. For each fold $k \in [K]$, we use samples in the $(k+1)$-th data fold $\mathcal{D}^{(k+1)}$ to obtain the propensity estimator $\widehat{\pi}_0^{(k)}(a \mid \cdot)$ (by regression) and the optimizer $\widehat{\theta}_a^{(k)}(\cdot)$ (by ERM) for each $a \in [M]$. Next, for each $a \in [M]$, define

$$G_a(x,y) = \ell(x,y;\theta_a^*),\ \widehat{G}_a^{(k)}(x,y) = \ell(x,y;\widehat{\theta}_a^{(k)}),\ \text{and}\ \bar{g}_a^{(k)}(x) = \mathbb{E}\big[\widehat{G}_a^{(k)}(X, Y(a)) \mid X = x\big].$$

We then obtain an estimator $\widehat{g}_a^{(k)}$ for $\bar{g}_a^{(k)}$ by regressing $\widehat{G}_a^{(k)}(X_i, Y_i)$ onto X_i with $i \in \mathcal{D}^{(k+2)}$. Finally, we obtain the learned policy by maximizing the estimated policy value:

$$\widehat{\pi}_{\mathrm{LN}} = \operatorname*{argmax}_{\pi \in \Pi} \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi) := \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi),\ \text{where}$$
$$\widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi) = \frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i \mid X_i)} \cdot \big(\widehat{G}_{\pi(X_i)}^{(k)}(X_i, Y_i) - \widehat{g}_{\pi(X_i)}^{(k)}(X_i)\big) + \widehat{g}_{\pi(X_i)}^{(k)}(X_i). \tag{9}$$

Above, the optimization problem can be solved by first-order optimization methods or policy tree search as in Zhou et al. (2023); we shall elaborate on the implementation in Section 5. The complete policy learning procedure is summarized in Algorithm 2, in which $\mathcal{D}_a^{(k)} := \{(X_i, A_i, Y_i) \in \mathcal{D}^{(k)} : A_i = a\}$.

4.2 Regret analysis

In this section, we present the regret analysis of $\widehat{\pi}_{\mathrm{LN}}$ obtained by Algorithm 2 (recall that the definition of regret is given in Equation (2)). Before we embark on the formal analysis, we introduce the Hamming entropy integral $\kappa(\Pi)$, which measures the complexity of Π.

Definition 4.1. Given a policy class Π and n data points $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$,

(1) The *Hamming distance* $d_H(\pi, \pi')$ between two policies $\pi, \pi' \in \Pi$ is defined as

$$d_H(\pi, \pi') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\pi(x_i) \neq \pi'(x_i)\}.$$

(2) The *ε-covering number* of $\{x_1, \ldots, x_n\}$, denoted as $\mathcal{C}(\epsilon, \Pi; \{x_1, \ldots, x_n\})$, is the smallest number L of policies $\{\pi_1, \ldots, \pi_L\}$ in Π, such that $\forall\, \pi \in \Pi$, $\exists\, \pi'_\ell$ such that $d_H(\pi, \pi_\ell) \leq \epsilon$.

(3) Denote $N_H(\epsilon, \Pi) := \sup_{n \geq 1} \sup_{x_1, \ldots, x_n} \mathcal{C}(\epsilon, \Pi; \{x_1, \ldots, x_n\})$. The *Hamming entropy integral* of Π is defined as $\kappa(\Pi) := \int_0^1 \sqrt{\log N_H(\epsilon^2, \Pi)}\, d\epsilon$.

Now we present the main result.

Theorem 4.2. *Suppose Assumptions 2.1, 3.3, 3.4 hold. For any $\beta \in (0,1)$, there exists $N \in \mathbb{N}_+$ such that when $n \geq N$, we have with probability at least $1-\beta$ that*

$$\mathcal{R}_\delta(\widehat{\pi}_{\mathrm{LN}}) \leq \frac{5\sqrt{K}C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{n}} \cdot \left(22 + 4\kappa(\Pi) + \sqrt{2\log(K/\beta)}\right),$$

where $C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon) := 6(\bar{\alpha}\cdot\exp(\bar{\eta}/\underline{\alpha}-1)+\bar{\eta}+\bar{\alpha}\delta)/\varepsilon$.

The proof of Theorem 4.2 is deferred to Appendix B. The main idea is to start with the following regret decomposition:

$$\begin{aligned}\mathcal{R}_\delta(\widehat{\pi}_{\mathrm{LN}}) = \mathcal{V}_\delta(\pi^*) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}}) &\leq \mathcal{V}_\delta(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}}) \\ &\leq 2\sup_{\pi\in\Pi}|\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi) - \mathcal{V}_\delta(\pi)|.\end{aligned}$$

We bound the above quantity by establishing uniform convergence results for the policy value estimators.

Remark 4.3. The theorem shows that the dependence of $\mathcal{R}(\widehat{\pi}_{\mathrm{LN}})$ on n is of the order $O(n^{-\frac{1}{2}})$, which outmatches the $O(n^{-\frac{1}{2}}\log n)$ dependence for that of Mu et al. (2022) by a logarithmic factor.

5 Numerical Results

We evaluated our learning algorithm in a simulated setting against the benchmark algorithm SNLN in Si et al. (2023). Our data generating process follows that of the linear boundary example in Si et al. (2023). We let the context set $\mathcal{X} = \{x \in \mathbb{R}^5 : \|x\|_2 \leq 1\}$ to be the closed unit ball of $\mathbb{R}^5$ and let the action set to be $\mathcal{A} = \{1,2,3\}$. We assume the rewards $Y(j)$'s are mutually independent conditioned on X with conditional distribution that follows a Gaussian law. We prepared a training dataset $\mathcal{D}_{\text{train}}$ (used for policy learning tasks) and a testing dataset $\mathcal{D}_{\text{test}}$ (used to empirically derive the underlying true policy value of the learnt policies, as the performance metric). The empirical policy value $\bar{\mathcal{V}}_\delta(\pi)$ of any policy $\pi \in \Pi$ is calculated as the sample average

$$\bar{\mathcal{V}}_\delta(\pi) = -\frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i\in\mathcal{D}_{\text{test}}}\left[\alpha_\pi(X_i)\exp\left(-\frac{Y_i(\pi(X_i))+\eta_\pi(X_i)}{\alpha_\pi(X_i)}-1\right)+\eta_\pi(X_i)+\alpha_\pi(X_i)\delta\right],$$

where the nuisance parameters $\alpha_\pi(X_i), \eta_\pi(X_i)$ are found via optimization:

$$\operatorname*{argmin}_{\alpha\geq 0,\eta\in\mathbb{R}}\left(\frac{1}{|\{j\in\mathcal{D}_{\text{test}} : X_j = X_i\}|}\sum_{j\in\mathcal{D}_{\text{test}}}\left(\alpha\exp\left(-\frac{Y_j(\pi(X_j))+\eta}{\alpha}-1\right)+\eta+\alpha\delta\right)\mathbb{1}\{X_j = X_i\}\right).$$

The testing dataset $\mathcal{D}_{\text{test}}$ realized i.i.d. draws of data tuple $(X, Y(1), \cdots, Y(M))$ so that the empirical policy value $\bar{\mathcal{V}}_\delta(\pi)$ could be computed for all π.

To perform our proposed learning Algorithm 2 combined with the `policytree` R package (Athey & Wager, 2021), we first split the dataset into $K = 3$ folds. Then we use Random Forest Regressor from `scikit-learn` Python library to find both $\widehat{\pi_0}$ and $\widehat{g}$; cubic spline is implemented to approximate θ^* with threshold at 0.001 to guarantee Proposition 2.5. We employ the Nelder-Mead optimization method in `SciPy` Python library (Virtanen et al., 2020) to optimize the coefficients in the spline approximation. For each data point (X_i, A_i, Y_i), using the estimated parameters $\widehat{r}, \widehat{g}, (\widehat{\alpha}_a, \widehat{\eta}_a)_{a\in\mathcal{A}}$ outputted by Algorithm 2, the outcomes $\widehat{Y}_i(a)$ under any action $a \in \mathcal{A}$ for $i \in [n]$ are estimated and stored in an outcome matrix. Then `policytree` finds $\widehat{\pi}_{LN}$ with the outcome matrix.

The benchmark algorithm SNLN is adapted from Si et al. (2023, Algorithm 2). Since Si et al. (2023, Algorithm 2) is designed for joint distribution shift formulation, we revised the original algorithm to fit our concept drift setting. It is well-known that the chain rule of KL-divergence (Cover, 1999) gives

$$D_{\mathrm{KL}}(Q_{X,Y}\,\|\,P_{X,Y}) = D_{\mathrm{KL}}(Q_X\,\|\,P_X) + D_{\mathrm{KL}}(Q_{Y\,|\,X}\,\|\,P_{Y\,|\,X}). \tag{10}$$

Therefore, given any uncertainty set radius δ and known covariate shift (in this experiment, we assume no covariate shift), Si et al. (2023, Algorithm 2) can be used to implement policy learning under concept drift. Note that SNLN admits known propensity scores. As we only consider the case where the propensity scores are unknown, we complement Si et al. (2023, Algorithm 2) with estimated propensity scores from Random Forest Regressor in `scikit-learn`, same as in the implementation of Algorithm 2. The additional setup details are in Appendix A.

Both algorithms are fitted with $K = 3$ folds. We conduct the experiments under three uncertainty set radii $\delta = 0.05, 0.1, 0.2$ and repeat each experiment over 50 random seeds. Table 1 reports the estimated average distribuionally robust values (with 95% confidence intervals) of the learnt policies $\widehat{\pi}_{\text{LN}}$ and $\widehat{\pi}_{\text{SNLN}}$, by Algorithm 2 and Si et al. (2023, Algorithm 2) respectively.

With a higher δ, the distribuionally robust values of $\widehat{\pi}_{\text{LN}}, \widehat{\pi}_{\text{SNLN}}$ are smaller, due to a bigger uncertainty set. Table 1 shows that $\widehat{\pi}_{\text{LN}}$ outperforms the benchmark $\widehat{\pi}_{\text{SNLN}}$ consistently across all tested setups, with higher policy values and smaller 95% confidence intervals. Intuitively, Algorithm 2 admits a subset of the uncertainty set that the benchmark algorithm SNLN considers, as explained in Equation (10). Consequently, $\bar{\mathcal{V}}_{\delta}(\widehat{\pi}_{\text{SNLN}})$ is a lower bound of $\bar{\mathcal{V}}_{\delta}(\widehat{\pi}_{\text{LN}})$ in theory, and by the results in Table 1, in practice. This shows that when only concept drift occurs, Algorithm 2 enjoys a better worst-case performance comparing to the joint distributional shift policy learning benchmark, as the latter is more conservative in this scenario.

In Appendix A, we also provide simulation results of Algorithm 1 for a fixed target policy, which show that Algorithm 1 can estimate the distributionally robust policy value under concept drift efficiently.

Sample Size	$\bar{\mathcal{V}}_{0.05}(\widehat{\pi}_{\text{LN}})$	$\bar{\mathcal{V}}_{0.05}(\widehat{\pi}_{\text{SNLN}})$
7500	$0.2272 \pm 2.2e-3$	$0.0554 \pm 5.9e-3$
13500	$0.2299 \pm 1.8e-3$	$0.0589 \pm 4.5e-3$
16500	$0.2303 \pm 1.7e-3$	$0.0617 \pm 4.2e-3$
19500	$0.2310 \pm 1.8e-3$	$0.0664 \pm 3.9e-3$
	$\bar{\mathcal{V}}_{0.1}(\widehat{\pi}_{\text{LN}})$	$\bar{\mathcal{V}}_{0.1}(\widehat{\pi}_{\text{SNLN}})$
7500	$0.1579 \pm 7.0e-3$	$0.0548 \pm 4.7e-3$
13500	$0.1662 \pm 2.0e-3$	$0.0580 \pm 4.2e-3$
16500	$0.1663 \pm 1.8e-3$	$0.0583 \pm 3.5e-3$
19500	$0.1678 \pm 1.8e-3$	$0.0616 \pm 4.3e-3$
	$\bar{\mathcal{V}}_{0.2}(\widehat{\pi}_{\text{LN}})$	$\bar{\mathcal{V}}_{0.2}(\widehat{\pi}_{\text{SNLN}})$
7500	$0.0781 \pm 2.6e-3$	$0.0182 \pm 3.0e-3$
13500	$0.0802 \pm 2.0e-3$	$0.0183 \pm 3.0e-3$
16500	$0.0804 \pm 2.1e-3$	$0.0200 \pm 3.2e-3$
19500	$0.0831 \pm 2.3e-3$	$0.0219 \pm 3.8e-3$

Table 1: Distributionally robust values of policies $\widehat{\pi}_{\text{LN}}, \widehat{\pi}_{\text{SNLN}}$ found by Algorithm 2 and SNLN respectively. The performance metric is $\bar{\mathcal{V}}_{\delta}(\cdot)$ estimated from the test data. We report the results under cases $\delta = 0.05, 0.1, 0.2$, sequentially in the table below.

References

Jiahao Ai and Zhimei Ren. Not all distributional shifts are equal: Fine-grained robust conformal inference. *arXiv preprint arXiv:2402.13042*, 2024.

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

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

Aurélien Bibaut, Nathan Kallus, Maria Dimakopoulou, Antoine Chambaz, and Mark van Der Laan. Risk minimization from adaptively collected data: Guarantees for supervised and policy learning. *Advances in neural information processing systems*, 34:19261–19273, 2021.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (11), 2013.

Carri W Chan, Vivek F Farias, Nicholas Bambos, and Gabriel J Escobar. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations research*, 60(6):1323–1341, 2012.

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.

William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2(1), 2019.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pp. 401–414, 1982.

Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.

Yihong Guo, Hao Liu, Yisong Yue, and Anqi Liu. Distributionally robust policy evaluation under general covariate shift in contextual bandits. *arXiv preprint arXiv:2401.11353*, 2024.

Tin Kam Ho et al. Proceedings of 3rd international conference on document analysis and recognition. In *Proceedings of 3rd international conference on document analysis and recognition*, 1995.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

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*, 2022a.

Ying Jin, Zhimei Ren, and Zhengyuan Zhou. Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective. *arXiv preprint arXiv:2203.04373*, 2022b.

Ying Jin, Kevin Guo, and Dominik Rothenhäusler. Diagnosing the role of observable distribution shift in scientific replications. *arXiv preprint arXiv:2309.01056*, 2023.

Nathan Kallus and Madeleine Udell. Dynamic assortment personalization in high dimensions. *arXiv preprint arXiv:1610.05604*, 2016.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *arXiv preprint arXiv:1912.12945*, 2019.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pp. 10598–10632. PMLR, 2022.

Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein Jr, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus Jr, Sanjay Gupta, et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

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

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. *arXiv preprint arXiv:2307.05284*, 2023.

Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12):2346–2363, 2018.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Tong Mu, Yash Chandak, Tatsunori B Hashimoto, and Emma Brunskill. Factored dro: Factored distributionally robust policies for contextual bandits. *Advances in Neural Information Processing Systems*, 35:8318–8331, 2022.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Hongseok Namkoong, Steve Yadlowsky, et al. Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*, 2023.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 2023.

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, 2015a.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823. PMLR, 2015b.

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

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *Annals of statistics*, 50(5):2587, 2022.

Ruohan Zhan, Zhimei Ren, Susan Athey, and Zhengyuan Zhou. Policy learning with adaptively collected data. *Management Science*, 2023.

Baqun Zhang, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Zihan Zhang, Wenhao Zhan, Yuxin Chen, Simon S Du, and Jason D Lee. Optimal multi-distribution learning. *arXiv preprint arXiv:2312.05134*, 2023.

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

A EXPERIMENT DETAILS

We adopt the data generating process similar to the linear boundary example in Si et al. (2023). We consider the context set $\mathcal{X} = \{x \in \mathbb{R}^5 : \|x\|_2 \leq 1\}$ is the closed unit ball of $\mathbb{R}^5$, and the action set $\mathcal{A} = \{1, 2, 3\}$. We assume the rewards $Y(j)$'s are mutually independent conditioned on X with conditional distribution $Y(j) \mid X \sim N(\beta_i^\top X, \sigma_j^2)$, for $j = 1, 2, 3$ and vectors $\{\beta_1, \beta_2, \beta_3\} \in \mathbb{R}^5$ and $\{\sigma_1^2, \sigma_2^2, \sigma_3^2\} \in \mathbb{R}_+$. We choose β's and σ's to be

$$\beta_1 = (1, 0, 0, 0, 0), \quad \beta_2 = (-1/2, \sqrt{3}/2, 0, 0, 0), \quad \beta_3 = (-1/2, -\sqrt{3}/2, 0, 0, 0); \quad \sigma = (0.2, 0.5, 0.8).$$

The underlying policy π_0 chooses actions with context x according to the following rules:

$$(\pi_0(1 \,|\, x), \pi_0(2 \,|\, x), \pi_0(3 \,|\, x)) = \begin{cases} (0.5, 0.25, 0.25), & \text{if } \operatorname*{argmax}_{i=1,2,3}\{\beta_i^\top x\} = 1, \\ (0.25, 0.5, 0.25), & \text{if } \operatorname*{argmax}_{i=1,2,3}\{\beta_i^\top x\} = 2, \\ (0.25, 0.25, 0.5), & \text{if } \operatorname*{argmax}_{i=1,2,3}\{\beta_i^\top x\} = 3. \end{cases}$$

We generate $\mathcal{D}_{\text{train}}$ according to the procedure described above as training dataset. We also generate 10,000 samples as our testing dataset $\mathcal{D}_{\text{test}} = \{i \in [10,000] : (X_i, Y_i(1), Y_i(2), Y_i(3))\}$, which we use to estimate the true policy value.

We present the result of the policy estimation experiments in Figure 1, using Algorithm 1 with inputs of the training datasets and the target policy π

$$\pi(x) = \begin{cases} 1, & \text{if } \|x\|_2 \in [0, 1/3], \\ 2, & \text{if } \|x\|_2 \in [1/3, 2/3], \\ 3, & \text{if } \|x\|_2 \in [2/3, 1]. \end{cases}$$

The underlying true policy value is obtained by the testing dataset $\mathcal{D}_{\text{test}}$. Similar to the learning experiment, we repeat the estimation experiment over 50 seeds. Figure 1 shows that as the sample sizes increases, the estimated policy value by Algorithm 1 is more accurate and stable.

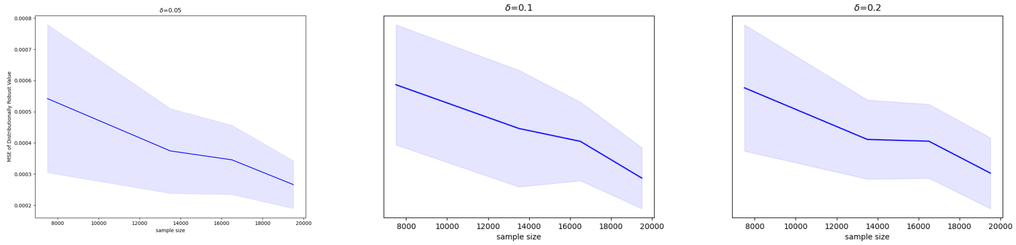

Figure 1: The Mean Square Error (MSE) of the estimated policy value by Algorithm 1. The x-axis is the number of samples used by Algorithm 1, and the y-axis is the mean squared error (MSE) of the policy value estimator.

Implementation details. The experiments were run on the following cloud servers: (i) an Intel Xeon Platinum 8160 @ 2.1 GHz with 766GB RAM and 96 CPU x 2.1 GHz; (ii) an Intel Xeon Platinum 8160 @ 2.1 GHz with 1.5TB RAM and 96 CPU x 2.1 GHz; (iii) an Intel Xeon Gold 6132 @ 2.59 GHz with 768GB RAM and 56 CPU x 2.59 GHz and (iv) an Intel Xeon GPU E5-2697A v4 @ 2.59 GHz with 384GB RAM and 64 CPU x 2.59 GHz.

B Deferred proofs of the main results

B.1 Proof of Lemma 2.3

Fix $\pi \in \Pi$ and $x \in \mathcal{X}$. Letting $L = \frac{dQ_{Y \mid X=x}}{dP_{Y \mid X=x}}$, we can rewrite the inner minimization in Equation (1) as

$$\begin{aligned} \inf_{L \text{ measurable}} \; & \mathbb{E}_{P_{Y \mid X}}[Y(\pi(x))L \mid X=x] \\ \text{s.t. } & \mathbb{E}_{P_{Y \mid X}}[L \mid X=x] = 1, \\ & \mathbb{E}_{P_{Y \mid X}}[f_{\mathrm{KL}}(L) \mid X=x] \le \delta, \end{aligned} \tag{11}$$

where the function $f_{\mathrm{KL}}(x) = x \log x$ represents the KL divergence function. In (11), the first constraint reflects that L is an likelihood ratio, and the second constraint corresponds to the KL divergence bound.

For notational simplicity, let $\mathbb{E}_x$ be the shorthand of $\mathbb{E}_{P_{Y \mid X}}[\cdot \mid X=x]$. By Theorem 8.6.1 of Luenberger (1997), the Slater's condition is satisfied and strong duality holds:

$$\inf_{\substack{\mathbb{E}_x[L]=1, \\ \mathbb{E}_x[f_{\mathrm{KL}}(L)] \le \delta}} \mathbb{E}_x\big[Y(\pi(x))L\big] = \max_{\alpha \ge 0, \eta \in \mathbb{R}} \varphi(\alpha, \eta, x), \tag{12}$$

where

$$\begin{aligned} \varphi(\alpha, \eta, x) &= \inf_{L \ge 0} \mathcal{L}(\alpha, \eta, L, x), \\ \mathcal{L}(\alpha, \eta, L, x) &= \mathbb{E}_x[Y(\pi(x))L] + \eta \cdot \big(\mathbb{E}_x[L] - 1\big) + \alpha \cdot \big(\mathbb{E}_x[f_{\mathrm{KL}}(L)] - \delta\big) \\ &= \mathbb{E}_x\big[Y(\pi(x))L + \eta(L-1) + \alpha(f_{\mathrm{KL}}(L) - \delta)\big]. \end{aligned}$$

We can explicitly work out the minimum of $\mathcal{L}(\alpha, \eta, L, x)$, and we have

$$\varphi(\alpha, \eta, x) = \mathbb{E}_x\left[-\alpha f_{\mathrm{KL}}^*\left(-\frac{Y(\pi(x)) + \eta}{\alpha}\right) - \eta - \alpha\delta\right],$$

where $f_{\mathrm{KL}}^*(y) = \exp(y-1)$ is the conjugate function of f_{KL}. Using Equation (12), we arrive at

$$\inf_{\substack{\mathbb{E}_x[L]=1, \\ \mathbb{E}_x[f_{\mathrm{KL}}(L)] \le \delta}} \mathbb{E}_x\big[Y(\pi(x))L\big] = -\min_{\alpha \ge 0, \eta \in \mathbb{R}} \mathbb{E}_x\left[\alpha \exp\left(-\frac{Y(\pi(x)) + \eta}{\alpha} - 1\right) + \eta + \alpha\delta\right].$$

The proof is thus completed.

B.2 Proof of Theorem 3.5

For notional simplicity, we drop the dependence on P in $\mathbb{E}_P$ when the context is clear. The proof of Theorem 3.5 makes use of the following lemma, which establishes some useful properties of the optimizer θ^*_π. The proof of Lemma B.1 can be found in Appendix C.1.

Lemma B.1. *For any policy π, assume that Assumption 3.3 holds. We have the following properties of the optimizer θ^*_π:*

(1) $\mathbb{E}\big[\nabla_\theta\, \ell(x, Y(\pi(x)); \theta) \,|\, X = x\big] = 0$ at $\theta = \theta^*_\pi(x)$ for any $x \in \mathcal{X}$.

(2) There exists a constant $\xi > 0$ such that for any θ satisfying $\|\theta - \theta^*_\pi\|_{L_\infty} \le \xi$,

$$\big|\ell(x, y; \theta(x)) - \ell(x, y; \theta^*_\pi(x)) - \nabla_\theta \ell(x, y; \theta^*_\pi(x))^\top (\theta(x) - \theta^*_\pi(x))\big| \le \bar{\ell}(x, y) \cdot \big\|\theta(x) - \theta^*_\pi(x)\big\|_2^2,$$

for some function $\bar{\ell}(x, y)$ such that $\sup_{x\in\mathcal{X}} \mathbb{E}[\bar{\ell}(x, Y(\pi(x))) \,|\, X = x] < L$ for some $L > 0$.

(3) There exists a constant $\xi_1 > 0$ such that for any θ satisfying $\|\theta - \theta^*_\pi\|_{L_2(P_{X \,|\, A=\pi(X)})} \le \xi_1$.

$$\big\|\ell(X, Y(\pi(X)); \theta) - \ell(X, Y(\pi(X)); \theta^*_\pi)\big\|_{L_2(P_{X, Y(\pi(X)) \,|\, A=\pi(X)})} \le C_\ell \|\theta - \theta^*_\pi\|_{L_2(P_{X \,|\, A=\pi(X)})},$$

for some constant $C_\ell > 0$.

We proceed to show the asymptotic normality of $\widehat{\theta}_\pi$. For each $k \in [K]$, we first define the following oracle quantity:

$$\mathcal{V}^{*(k)}_\delta(\pi) = \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \frac{\mathbb{1}\{\pi(X_i) = A_i\}}{\pi_0(A_i \,|\, X_i)} \cdot \big(G_\pi(X_i, Y_i) - g_\pi(X_i)\big) + g_\pi(X_i).$$

In the sequel, we shall show that $\widehat{\mathcal{V}}^{(k)}_\delta(\pi) = \mathcal{V}^{*(k)}_\delta(\pi) + o_p(n^{-\frac{1}{2}})$. We begin by decomposing the difference between $\widehat{\mathcal{V}}^{(k)}_\delta(\pi)$ and $\mathcal{V}^{*(k)}_\delta$:

$$\begin{aligned}
&\widehat{\mathcal{V}}^{(k)}_\delta(\pi) - \mathcal{V}^{*(k)}_\delta(\pi)\\
&= \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \left[\frac{\mathbb{1}\{\pi(X_i) = A_i\}}{\widehat{\pi}^{(k)}_0(A_i \,|\, X_i)} \cdot \Big(\widehat{G}^{(k)}_\pi(X_i, Y_i) - \widehat{g}^{(k)}_\pi(X_i)\Big) - \frac{\mathbb{1}\{\pi(X_i) = A_i\}}{\pi_0(A_i \,|\, X_i)} \cdot \Big(G_\pi(X_i, Y_i) - g_\pi(X_i)\Big)\right]\\
&\qquad + \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \big(\widehat{g}^{(k)}_\pi(X_i) - g_\pi(X_i)\big)\\
&= \underbrace{\frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \cdot \big(\widehat{G}^{(k)}_\pi(X_i, Y_i) - G_\pi(X_i, Y_i)\big)}_{\text{(I)}}\\
&\qquad \underbrace{- \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}^{(k)}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \cdot \big(\widehat{g}^{(k)}_\pi(X_i) - \bar{g}^{(k)}_\pi(X_i)\big)}_{\text{(II)}}\\
&\qquad \underbrace{+ \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}^{(k)}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \cdot \big(\widehat{G}^{(k)}_\pi(X_i, Y_i) - \bar{g}^{(k)}_\pi(X_i)\big)}_{\text{(III)}}\\
&\qquad \underbrace{- \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \cdot \big(\widehat{g}^{(k)}_\pi(X_i) - g_\pi(X_i)\big) + \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i\in\mathcal{D}^{(k)}} \big(\widehat{g}^{(k)}_\pi(X_i) - g_\pi(X_i)\big)}_{\text{(IV)}}.
\end{aligned}$$

Bounding term (I). Recall that $(\alpha^*_\pi(x), \eta^*_\pi(x))$ is the minimizer of

$$\mathbb{E}\Big[\ell\big(x, Y(\pi(x)); (\alpha, \eta)\big) \,\Big|\, X = x\Big].$$

By the first-order condition established in part (1) of Lemma B.1, we have

$$\mathbb{E}\Big[\nabla_{\alpha,\eta}\ell\big(x, Y(\pi(x)); (\alpha_\pi^*, \eta_\pi^*)\big) \,\Big|\, X = x\Big] = 0, \tag{13}$$

where we abuse the notation a bit and $\nabla_{\alpha,\eta}\ell(x, y; (\alpha_\pi^*, \eta_\pi^*))$ to denote the gradient of $\ell(x, y; (\alpha, \eta))$ with respect to (α, η) evaluated at $(\alpha_\pi^*(x), \eta_\pi^*(x))$. For any $i \in \mathcal{D}^{(k)}$, by the unconfoundedness condition in Assumption 2.1, we have

$$\begin{aligned}
&\mathbb{E}\bigg[\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \cdot \Big(\widehat{G}_\pi^{(k)}\big(X_i, Y_i\big) - G_\pi\big(X_i, Y_i\big)\Big)\bigg] \\
&= \mathbb{E}\bigg[\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \cdot \Big(\widehat{G}_\pi^{(k)}\big(X_i, Y_i(\pi(X_i))\big) - G_\pi\big(X_i, Y_i(\pi(X_i))\big)\Big)\bigg] \\
&= \mathbb{E}\Big[\widehat{G}_\pi^{(k)}\big(X_i, Y_i(\pi(X_i))\big) - G_\pi\big(X_i, Y_i(\pi(X_i))\big)\Big] \\
&= \mathbb{E}\Big[\ell\big(X_i, Y_i(\pi(X_i)); (\widehat{\alpha}_\pi^{(k)}, \widehat{\eta}_\pi^{(k)})\big) - \ell\big(X_i, Y_i(\pi(X_i)); \alpha_\pi^*, \eta_\pi^*\big) - \nabla_{\alpha,\eta}\ell\big(X_i, Y(\pi(X_i)); (\alpha_\pi^*, \eta_\pi^*)\big)\Big],
\end{aligned}$$

where the last step is due to Equation (13). By Assumption 3.4, $\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_\infty} = o_P(1)$. Therefore, for n sufficiently large, $\|\widehat{\theta}_\pi^{(k)}(x) - \theta_\pi^*(x)\|_2 \le \xi$ for all $x \in \mathcal{X}$. Then by part (2) of Lemma B.1 and Jensen's inequality, we have

$$\begin{aligned}
&\bigg|\mathbb{E}\bigg[\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \cdot \Big(\widehat{G}_\pi^{(k)}\big(X_i, Y_i\big) - G_\pi\big(X_i, Y_i\big)\Big)\bigg]\bigg| \\
&\le \mathbb{E}\bigg[\Big|\ell(X_i, Y_i(\pi(X_i)); (\widehat{\alpha}_\pi^{(k)}, \widehat{\eta}_\pi^{(k)})) - \ell(X_i, Y_i(\pi(X_i)); \alpha_\pi^*, \eta_\pi^*) - \nabla_{\alpha,\eta}\ell\big(X_i, Y(\pi(X_i)); (\alpha_\pi^*, \eta_\pi^*)\big)\Big|\bigg] \\
&\le \mathbb{E}\Big[\bar{\ell}(X_i, Y_i) \cdot \|\widehat{\theta}_\pi^{(k)}(X_i) - \theta_\pi^*(X_i)\|_2^2\Big] \le L\mathbb{E}\big[\|\widehat{\theta}_\pi^{(k)}(X_i) - \theta_\pi^*(X_i)\|_2^2\big] = L\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_2(P_X)}^2.
\end{aligned}$$

By Chebyshev's inequality, we have for any $t > 0$ that

$$\begin{aligned}
&\mathbb{P}\Bigg(\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \cdot (\widehat{G}_\pi^{(k)}(X_i, Y_i) - G_\pi(X_i, Y_i)) \\
&\qquad\qquad - \mathbb{E}\bigg[\frac{\mathbb{1}\{A = \pi(X)\}}{\pi_0(A \,|\, X)} \cdot \Big(\widehat{G}_\pi^{(k)}(X, Y) - G_\pi(X, Y)\Big)\bigg]\bigg| \ge t\Bigg) \\
&\quad \le \frac{1}{|\mathcal{D}^{(k)}|t^2}\mathrm{Var}\bigg(\frac{\mathbb{1}\{A = \pi(X)\}}{\pi_0(A \,|\, X)} \cdot \Big[\widehat{G}_\pi^{(k)}(X, Y) - G_\pi(X, Y)\Big]\bigg) \\
&\quad \le \frac{\big\|\widehat{G}_\pi^{(k)} - G_\pi\big\|_{L_2(P_{X,Y \,|\, A = \pi(X)})}^2}{\varepsilon^2|\mathcal{D}^{(k)}|t^2} \\
&\quad \le \frac{C_\ell\Big(\big\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\big\|_{L_2(P_{X \,|\, A = \pi(X)})}^2\Big)}{\varepsilon^2|\mathcal{D}^{(k)}|t^2},
\end{aligned}$$

where the last step is due to the stability property in Lemma B.1. Therefore, we have that

$$\begin{aligned}
&\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \cdot (\widehat{G}_\pi^{(k)}(X_i, Y_i) - G_\pi(X_i, Y_i)) \\
&\qquad\qquad - \mathbb{E}\bigg[\frac{\mathbb{1}\{A = \pi(X)\}}{\pi_0(A \,|\, X)} \cdot \Big(\widehat{G}_\pi^{(k)}(X, Y) - G_\pi(X, Y)\Big)\bigg]\bigg| \\
&\quad = o_P(n^{-1/2}).
\end{aligned}$$

Combining the above results, we have that

$$\text{term (I)} \le o_P(n^{-1/2}) + L\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_2(P_X)}^2 \le o_P(n^{-1/2}) + \frac{L}{\sqrt{\varepsilon}}\|\widehat{\theta}_\pi^{(k)} - \theta_\pi^*\|_{L_2(P_X)}^2 = o_P(n^{-1/2}),$$

where the second step is due to the overlap and unconfoundedness assumption and the last step is due to Assumption 3.4.

Bounding term (II). Applying the Cauchy-Schwarz inequality to term (II), we have

$$
\begin{aligned}
&\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\left(\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right)\cdot\left(\widehat{g}_\pi^{(k)}(X_i)-\bar{g}_\pi^{(k)}(X_i)\right)\right|\\
\leq&\sqrt{\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\left(\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right)^2}\\
&\qquad\qquad\qquad\times\sqrt{\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\mathbb{1}\{A_i=\pi(X_i)\}\cdot\big(\widehat{g}_\pi^{(k)}(X_i)-\bar{g}_\pi^{(k)}(X_i)\big)^2}\\
=&O_P\left(\epsilon^{-2}\|\widehat{\pi}_0^{(k)}-\pi_0\|_{L_2(P_{X\,|\,A=\pi(X)})}\cdot\|\widehat{g}_\pi^{(k)}-\bar{g}_\pi^{(k)}\|_{L_2(P_{X\,|\,A=\pi(X)})}\right)=o_P(n^{-1/2}),
\end{aligned}
$$

where the next-to-last inequality is due to the lower bound on π_0 and $\widehat{\pi}^{(k)}$; the last equality is due to the given convergence rate of the product estimation error in Assumption 3.4.

Bounding term (III). By Assumption 3.4, for any $\beta\in(0,1)$, there exists $N\in\mathbb{N}_+$ such that for $n\geq N$,

$$
\mathbb{P}\big(\|\widehat{\theta}_\pi^{(k)}-\theta^*\|_{L_\infty}\leq\min(\underline{\alpha},\bar{\eta})/2\big)\geq 1-\beta.
$$

On the event $\|\widehat{\theta}_\pi^{(k)}-\theta^*\|_{L_\infty}\leq\min(\underline{\alpha},\bar{\eta})/2$, we can find a constant L_g such that $|\ell(x,y;\widehat{\theta}_\pi^{(k)})|\leq L_g$ for any $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, which implies that $\mathrm{Var}\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)\,|\,X_i,\mathcal{D}^{(-k)}\big)\leq L_g^2$. Since β is arbitrary, we have that $\mathrm{Var}\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)\,|\,X_i,\mathcal{D}^{(-k)}\big)=O(1)$.

Next, since $\bar{g}_\pi^{(k)}$ is the conditional expectation of $\widehat{G}_\pi^{(k)}$, for any $i\in\mathcal{D}^{(k)}$,

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right)\cdot\left(\widehat{G}_\pi^{(k)}(X_i,Y_i)-\bar{g}_\pi^{(k)}(X_i)\right)\,\middle|\,\mathcal{D}^{(-k)}\right]\\
=&\,\mathbb{E}\Bigg[\mathbb{E}\left[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\,\middle|\,X_i,\mathcal{D}^{(-k)}\right]\\
&\qquad\qquad\times\mathbb{E}\Big[\widehat{G}_\pi^{(k)}(X_i,Y(\pi(X_i)))-\bar{g}_\pi^{(k)}(X_i)\,\Big|\,X_i,\mathcal{D}^{(-k)}\Big]\,\Bigg|\,\mathcal{D}^{(-k)}\Bigg]=0.
\end{aligned}
$$

By Chebyshev's inequality, for any $t>0$,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\left(\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right)\cdot\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)-\bar{g}_\pi^{(k)}(X_i)\big)\right|\geq t\,\middle|\,\mathcal{D}^{(-k)}\right)\\
\leq&\frac{1}{|\mathcal{D}^{(k)}|t^2}\mathrm{Var}\left(\left[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right]\cdot\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)-\bar{g}_\pi^{(k)}(X_i)\big)\,\middle|\,\mathcal{D}^{(-k)}\right)\\
\leq&\frac{1}{|\mathcal{D}^{(k)}|t^2}\mathbb{E}\left[\left[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right]^2\cdot\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)-\bar{g}_\pi^{(k)}(X_i)\big)^2\,\middle|\,\mathcal{D}^{(-k)}\right]\\
=&\frac{1}{|\mathcal{D}^{(k)}|t^2}O\Big(\|\widehat{\pi}_0^{(k)}-\pi_0\|_{L_2(P_{X\,|\,T=\pi(X)})}^2\Big),
\end{aligned}
$$

where the last step is because of the overlap condition and that $\mathrm{Var}\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)\,|\,X_i,\mathcal{D}^{(-k)}\big)=O(1)$. The above inequality implies that term (III) $=O_P\big(\|\widehat{\pi}_0^{(k)}-\pi_0\|_{L_2(P_{X\,|\,A=\pi(X)})}/\sqrt{|\mathcal{D}^{(k)}|}\big)$. By the consistency of $\widehat{\pi}_0^{(k)}$ assumed in Assumption 3.4, term (III) is of rate $o_P(n^{-1/2})$.

Bounding term (IV). We first show that term (IV) is of zero-mean:

$$\mathbb{E}\Bigg[-\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)+\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)\,\Big|\,\mathcal{D}^{(-k)}\Bigg]$$
$$=-\mathbb{E}\Bigg[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)\,\Big|\,\mathcal{D}^{(-k)}\Bigg]+\mathbb{E}\Big[\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\,\big|\,\mathcal{D}^{(-k)}\Big]=0.$$

By Chebyshev's inequality, for any $t>0$,

$$\mathbb{P}\Bigg(\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\frac{\mathbb{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)-\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)\bigg|\geq t\,\bigg|\,\mathcal{D}^{(-k)}\Bigg)$$
$$\leq\frac{1}{|\mathcal{D}^{(k)}|t^2}\mathrm{Var}\Bigg(\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)-\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)\,\bigg|\,\mathcal{D}^{(-k)}\Bigg)$$
$$=\frac{1}{|\mathcal{D}^{(k)}|t^2}\mathbb{E}\Bigg[\frac{(1-\pi_0(\pi(X_i)\,|\,X_i))^2}{\pi_0(\pi(X_i)\,|\,X_i)}\cdot\big(\widehat{g}_\pi^{(k)}(X_i)-g_\pi(X_i)\big)^2\,\Big|\,\mathcal{D}^{(-k)}\Bigg].$$

As a result, term (IV) $=O_P\big(\|\widehat{g}_\pi^{(k)}-g_\pi\|_{L_2(P_X)}/\sqrt{n}\big)$. Note that

$$\begin{aligned}\|\widehat{g}_\pi^{(k)}-g_\pi\|_{L_2(P_X)}&=O(\|\widehat{g}_\pi^{(k)}-g_\pi\|_{L_2(P_{X\,|\,A=\pi(X)})})\\&\leq O\Big(\|\widehat{g}_\pi^{(k)}-\bar{g}_\pi\|_{L_2(P_{X\,|\,A=\pi(X)})}+\|\bar{g}_\pi^{(k)}-g_\pi\|_{L_2(P_{X\,|\,A=\pi(X)})}\Big),\end{aligned}$$

where the first inequality follows from the overlap condition. By Assumption 3.4, $\|\widehat{g}_\pi^{(k)}-\bar{g}_\pi\|_{L_2(P_{X\,|\,A=\pi(X)})}=o_P(1)$. Meanwhile,

$$\begin{aligned}\|\bar{g}_\pi^{(k)}-g_\pi\|^2_{L_2(P_{X\,|\,A=\pi(X)})}&=\mathbb{E}\big[(\bar{g}(X)-g(X))^2\,|\,A=\pi(X)\big]\\&=\mathbb{E}\Big[\Big(\mathbb{E}\big[\ell(X,Y(\pi(X));\widehat{\theta}_\pi^{(k)})-\ell(X,Y(\pi(X));\theta_\pi^*)\,|\,X\big]\Big)^2\,|\,A=\pi(X)\Big]\\&\overset{\text{(i)}}{\leq}\mathbb{E}\Big[\Big(\ell(X,Y(\pi(X));\widehat{\theta}_\pi^{(k)})-\ell(X,Y(\pi(X));\theta_\pi^*)\Big)^2\,|\,A=\pi(X)\Big]\\&\overset{\text{(ii)}}{=}O\Big(\|\widehat{\theta}_\pi^{(k)}-\theta_\pi^*\|^2_{L_2(P_{X\,|\,A=\pi(X)})}\Big)=o_P(1).\end{aligned}$$

Above, we slightly abuse the notation, taking the expectation conditional on $\mathcal{D}^{(-k)}$ without explicitly writing so; step (i) follows from Jensen's inequality and step (ii) from Lemma B.1. Combining everything, we have that term (IV) is of rate $o_P(n^{-1/2})$.

Putting everything together. So far we have shown that for each fold $k\in[K]$, there is

$$\widehat{\mathcal{V}}_\delta^{(k)}(\pi)-\mathcal{V}_\delta^{*(k)}(\pi)=o_P(n^{-1/2}).$$

Averaging over all k folds, we have

$$\begin{aligned}&\sqrt{n}\cdot\big(\widehat{\mathcal{V}}_\delta(\pi)-\mathcal{V}_\delta(\pi)\big)\\&=\frac{1}{\sqrt{n}}\sum_{i\in[n]}\bigg\{\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(G_\pi(X_i,Y_i)-g_\pi(X_i)\big)+g_\pi(X_i)\bigg\}-\mathcal{V}_\delta(\pi)+o_P(1),\end{aligned}$$

By the central limit theorem and Slutsky's theorem.

$$\sqrt{n}\cdot\big(\widehat{\mathcal{V}}_\delta(\pi)-\mathcal{V}_\delta(\pi)\big)\xrightarrow{\text{d.}}\mathcal{N}(0,\sigma^2),$$

where

$$\sigma^2=\mathrm{Var}\bigg(\frac{\mathbb{1}\{A=\pi(X)\}}{\pi_0(A\,|\,X)}\cdot\big(G(X,Y)-g(X)\big)+g(X)\bigg).$$

B.3 Proof of Theorem 4.2

The regret bound of Algorithm 2 builds on the following regret decomposition:

$$\begin{aligned}\mathcal{R}_\delta(\widehat{\pi}_{\mathrm{LN}}) &= \mathcal{V}_\delta(\pi^*) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}})\\ &= \mathcal{V}_\delta(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}})\\ &\le \mathcal{V}_\delta(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}})\\ &\le 2\sup_{\pi\in\Pi}\big|\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi) - \mathcal{V}_\delta(\pi)\big|, \end{aligned} \tag{14}$$

where the second-to-last step is by the choice of $\widehat{\pi}_{\mathrm{LN}}$. For any $\pi \in \Pi$ and any fold $k \in [K]$, we define an intermediate quantity

$$\tilde{\mathcal{V}}_\delta^{(k)} := \frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\cdot\big(G_{\pi(X_i)}(X_i,Y_i) - g_{\pi(X_i)}(X_i)\big) + g_{\pi(X_i)}(X_i).$$

Letting $\tilde{\mathcal{V}}_\delta = \frac{1}{K}\sum_{k=1}^K \tilde{\mathcal{V}}_\delta^{(k)}$, we have

$$\begin{aligned}\big|\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi) - \mathcal{V}_\delta(\pi)\big| &= \bigg|\frac{1}{K}\sum_{k=1}^K \widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi) - \mathcal{V}_\delta(\pi)\bigg|\\ &\le \bigg|\frac{1}{K}\sum_{k=1}^K \widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi) - \tilde{\mathcal{V}}_\delta(\pi)\bigg| + \Big|\tilde{\mathcal{V}}_\delta(\pi) - \mathcal{V}_\delta(\pi)\Big|\\ &\le \sup_{\pi\in\Pi}\bigg|\frac{1}{K}\sum_{k=1}^K \widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi) - \frac{1}{K}\sum_{k=1}^K \tilde{\mathcal{V}}_\delta^{(k)}(\pi)\bigg| + \sup_{\pi\in\Pi}\bigg|\frac{1}{K}\sum_{k=1}^K \tilde{\mathcal{V}}_\delta^{(k)}(\pi) - \mathcal{V}_\delta(\pi)\bigg|.\end{aligned}$$

Taking the supremum over all $\pi \in \Pi$, we have that

$$\sup_{\pi\in\Pi}\big|\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi) - \mathcal{V}_\delta(\pi)\big| \le \sup_{\pi\in\Pi}\Big|\tilde{\mathcal{V}}_\delta(\pi) - \mathcal{V}_\delta(\pi)\Big| + \sup_{\pi\in\Pi}\bigg|\frac{1}{K}\sum_{k=1}^K \widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi) - \frac{1}{K}\sum_{k=1}^K \tilde{\mathcal{V}}_\delta^{(k)}(\pi)\bigg|.$$

We proceed to bound the above two terms separately. The following lemma is essential for establishing the uniform convergence results.

Lemma B.2. *Suppose h is a function of $(x, a, y, \pi(x))$ such that*

(1) $|h| \le C_h$ for some constant $C_h > 0$;

(2) $\mathbb{E}[h(X, A, Y, \pi(X))] = 0$.

Then for any $\beta > 0$, with probability $1 - \beta$, we have that

$$\sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{n=1}^n h\big(X_i, A_i, Y_i, \pi(X_i)\big)\bigg| \le \frac{C_h}{\sqrt{n}}\big(20 + 4\kappa(\Pi) + \sqrt{2\log(1/\beta)}\big).$$

We now focus on the first term. Denote $Z_i = (X_i, A_i, Y_i)$ and take

$$h(Z_i, \pi(X_i)) = \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\cdot\big(G_{\pi(X_i)}(X_i,Y_i) - g_{\pi(X_i)}(X_i)\big) + g_{\pi(X_i)}(X_i) - \mathcal{V}_\delta(\pi).$$

Under the unconfoundedness assumption in Assumption 2.1, $\mathbb{E}[h(Z_i, \pi(X_i))] = 0$. By Assumption 3.3, we have

$$|h(Z_i, \pi(X_i))| \le \frac{6}{\varepsilon}\cdot\bigg(\bar{\alpha}\cdot\exp\Big(\frac{\bar{\eta}}{\underline{\alpha}} - 1\Big) + \bar{\eta} + \bar{\alpha}\delta\bigg) =: C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon).$$

Meanwhile, we have write

$$\sup_{\pi\in\Pi}\bigg|\frac{1}{K}\sum_{k=1}^K \tilde{\mathcal{V}}_\delta^{(k)}(\pi) - \mathcal{V}_\delta(\pi)\bigg| = \sup_{\pi\in\Pi}\bigg|\frac{1}{K}\sum_{k=1}^K \frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}} h(Z_i;\pi)\bigg| = \sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{i=1}^n h(Z_i;\pi)\bigg|.$$

Applying Lemma B.2, for any $\beta \in (0, 1)$, we have with probability at least $1 - \beta$,

$$\sup_{\pi \in \Pi} \big|\tilde{\mathcal{V}}_\delta(\pi) - \mathcal{V}_\delta(\pi)\big| \le \frac{C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)}{\sqrt{n}} \big(20 + 4\kappa(\Pi) + \sqrt{2\log(1/\beta)}\big). \tag{15}$$

We now proceed to the second term. For any $\pi \in \Pi$ and any $k \in [K]$, consider the following decomposition:

$$\begin{aligned}
&\widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi) - \tilde{\mathcal{V}}_\delta^{(k)}(\pi) \\
&= \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i \,|\, X_i)} \big(\widehat{G}^{(k)}_{\pi(X_i)}(X_i, Y_i) - \widehat{g}^{(k)}_{\pi(X_i)}(X_i)\big) + \widehat{g}^{(k)}_{\pi(X_i)}(X_i) \\
&\qquad - \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \big(G_{\pi(X_i)}(X_i, Y_i) - g_{\pi(X_i)}(X_i)\big) - g_{\pi(X_i)}(X_i) \\
&= \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \big(\widehat{G}^{(k)}_{\pi(X_i)}(X_i, Y_i) - \bar{g}^{(k)}_{\pi(X_i)}(X_i)\big) \\
&\qquad + \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \big(\bar{g}^{(k)}_{\pi(X_i)}(X_i) - \widehat{g}^{(k)}_{\pi(X_i)}(X_i)\big) \\
&\qquad + \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \big(\widehat{G}^{(k)}_{\pi(X_i)}(X_i, Y_i) - G_{\pi(X_i)}(X_i, Y_i)\big) \\
&\qquad - \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \big(\widehat{g}^{(k)}_{\pi(X_i)}(X_i) - g_{\pi(X_i)}(X_i)\big) + \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \Big(\widehat{g}^{(k)}_{\pi(X_i)}(X_i) - g_{\pi(X_i)}(X_i)\Big).
\end{aligned}$$

For notational simplicity, we denote

$$\begin{aligned}
K_1(\pi) &:= \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \big(\widehat{G}^{(k)}_{\pi(X_i)}(X_i, Y_i) - \bar{g}^{(k)}_{\pi(X_i)}(X_i)\big), \\
K_2(\pi) &:= \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \big(\bar{g}^{(k)}_{\pi(X_i)}(X_i) - \widehat{g}^{(k)}_{\pi(X_i)}(X_i)\big), \\
K_3(\pi) &:= \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \big(\widehat{G}^{(k)}_{\pi(X_i)}(X_i, Y_i) - G_{\pi(X_i)}(X_i, Y_i)\big), \\
K_4(\pi) &:= -\frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)} \big(\widehat{g}^{(k)}_{\pi(X_i)}(X_i) - g_{\pi(X_i)}(X_i)\big) + \frac{1}{|\mathcal{D}^{(k)}|} \sum_{i \in \mathcal{D}^{(k)}} \Big(\widehat{g}^{(k)}_{\pi(X_i)}(X_i) - g_{\pi(X_i)}(X_i)\Big).
\end{aligned}$$

We proceed to bound each term separately.

Bounding $K_1(\pi)$. Here, we take

$$h_1(Z_i; \pi(X_i)) := \left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \big(\widehat{G}^{(k)}_{\pi(X_i)}(X_i, Y_i) - \bar{g}^{(k)}_{\pi(X_i)}(X_i)\big).$$

Since $\bar{g}^{(k)}_a(X)$ is the conditional expectation of $\widehat{G}^{(k)}_a(X, Y(a))$, we have

$$\begin{aligned}
\mathbb{E}\big[h_1(Z_i, \pi(X_i)) \,|\, \mathcal{D}^{(-k)}\big] &= \mathbb{E}\left[\left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i \,|\, X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i \,|\, X_i)}\right) \big(\widehat{G}^{(k)}_{A_i}(X_i, Y_i) - \bar{g}^{(k)}_{A_i}(X_i)\big) \,\middle|\, \mathcal{D}^{(-k)}\right] \\
&= \mathbb{E}\left[\left(\frac{\pi_0(\pi(X_i))}{\widehat{\pi}_0(A_i \,|\, X_i)} - 1\right) \mathbb{E}\Big[\widehat{G}^{(k)}_{\pi(X_i)}(X_i, Y_i) - \bar{g}^{(k)}_{\pi(X_i)}(X_i) \,\Big|\, X_i, \mathcal{D}^{(-k)}\Big] \,\middle|\, \mathcal{D}^{(-k)}\right] \\
&= 0.
\end{aligned}$$

By Assumption 3.4, there exists $N_1 \in \mathbb{N}_+$, such that when $n \ge N_1$, w. p. at least $1 - \beta$,

$$\max_{a \in [M]} \|\widehat{\theta}^{(k)}_a - \theta^*_a\|_{L_\infty} \le \max(\bar{\alpha}, \underline{\alpha}, \bar{\eta})/2.$$

On the event $\{\max_{a\in[M]} \|\widehat{\theta}_a^{(k)} - \theta_a^*\|_{L_\infty} \le \max(\bar{\alpha}, \underline{\alpha}, \bar{\eta})/2\}$, we have

$$\big|h_i(Z_i, \pi(X_i))\big| \le C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon).$$

We now apply Lemma B.2 to $h_1(Z_i, \pi(X_i))$ on the event $\{\max_{a\in[M]} \|\widehat{\theta}_a^{(k)} - \theta_a^*\|_{L_\infty} \le \max(\bar{\alpha}, \underline{\alpha}, \bar{\eta})/2\}$,

$$\mathbb{P}\bigg(\sup_{\pi\in\Pi} \big|K_1(\pi)\big| \ge \frac{C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\big(20 + 4\kappa(\Pi) + \sqrt{2\log(1/\beta)}\big) \,\Big|\, \mathcal{D}^{(-k)}\bigg) \le \beta.$$

Taking a union bound, with probability at least $1 - 2\beta$, we have that

$$\sup_{\pi\in\Pi} \big|K_1(\pi)\big| \le \frac{C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\big(20 + 4\kappa(\Pi) + \sqrt{2\log(1/\beta)}\big) \tag{16}$$

Bounding $K_2(\pi)$. We first note that by Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i\,|\,X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right)\big(\bar{h}_{A_i}^{(k)}(X_i) - \widehat{g}_{A_i}^{(k)}(X_i)\big)\right| \\
&\le \frac{1}{|\mathcal{D}^{(k)}|\varepsilon^2}\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\big(\widehat{\pi}_0^{(k)}(\pi(X_i)\,|\,X_i) - \pi_0(\pi(X_i)\,|\,X_i)\big)^2}\cdot\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\big(\bar{g}_{\pi(X_i)}^{(k)}(X_i) - \widehat{g}_{\pi(X_i)}^{(k)}(X_i)\big)^2} \\
&\le \frac{1}{|\mathcal{D}^{(k)}|\varepsilon^2}\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\sum_{a=1}^{M}\big(\widehat{\pi}_0^{(k)}(a\,|\,X_i) - \pi_0(a\,|\,X_i)\big)^2}\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\sum_{a=1}^{M}\big(\bar{g}_a^{(k)}(X_i) - \widehat{g}_a^{(k)}(X_i)\big)^2}.
\end{aligned}$$

Then for any $t > 0$, let

$$s = \frac{M}{t\varepsilon^2}\max_{a\in[M]}\Big\{\|\widehat{\pi}_a^{(k)} - \pi_{0,a}^{(k)}\|_{L_2(P_X)}\Big\}\max_{a\in[M]}\Big\{\|\bar{g}_a^{(k)} - \widehat{g}_a^{(k)}\|_{L_2(P_X)}\Big\}.$$

Then

$$\begin{aligned}
&\mathbb{P}\left(\max_{\pi\in\Pi}\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\left(\frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\widehat{\pi}_0(A_i\,|\,X_i)} - \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\right)\big(\widehat{g}_{A_i}^{(k)}(X_i) - \bar{g}_{A_i}^{(k)}(X_i)\big)\right| \ge s \,\middle|\, \mathcal{D}^{(-k)}\right) \\
&\le \mathbb{P}\left(\frac{1}{|\mathcal{D}^{(k)}|\varepsilon^2}\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\sum_{a=1}^{M}\big(\widehat{\pi}_0^{(k)}(a\,|\,X_i) - \pi_0(a\,|\,X_i)\big)^2}\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\sum_{a=1}^{M}\big(\widehat{g}_a^{(k)}(X_i) - \bar{g}_a^{(k)}(X_i)\big)^2} \ge s \,\middle|\, \mathcal{D}^{(-k)}\right) \\
&\le \mathbb{P}\left(\frac{1}{\varepsilon}\sqrt{\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\sum_{a=1}^{M}\big(\widehat{\pi}_0^{(k)}(a\,|\,X_i) - \pi_0(a\,|\,X_i)\big)^2} \ge \frac{\sqrt{M}}{\sqrt{t}\varepsilon}\max_{a\in[M]}\Big\{\|\widehat{\pi}_a^{(k)} - \pi_{0,a}^{(k)}\|_{L_2(P_X)}\Big\} \,\middle|\, \mathcal{D}^{(-k)}\right) \\
&\qquad + \mathbb{P}\left(\frac{1}{\varepsilon}\sqrt{\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\sum_{a=1}^{M}\big(\widehat{g}_a^{(k)}(X_i) - \bar{g}_a^{(k)}(X_i)\big)^2} \ge \frac{\sqrt{M}}{\sqrt{t}\varepsilon}\max_{a\in[M]}\Big\{\|\widehat{g}_a^{(k)} - \widehat{g}_a^{(k)}\|_{L_2(P_X)}\Big\} \,\middle|\, \mathcal{D}^{(-k)}\right) \\
&\le 2t,
\end{aligned}$$

where the last inequality is due to Chebyshev's inequality. Marginalizing over the randomness of $\mathcal{D}^{(-k)}$, for any $\beta \in (0, 1)$, we have with probability at least $1 - \beta$ that

$$\max_{\pi\in\Pi}|K_2(\pi)| < \frac{2M}{\beta\varepsilon^2}\max_{a\in[M]}\Big\{\|\widehat{\pi}_a^{(k)} - \pi_{0,a}^{(k)}\|_{L_2(P_X)}\Big\}\max_{a\in[M]}\Big\{\|\widehat{g}_a^{(k)} - \bar{g}_a^{(k)}\|_{L_2(P_X)}\Big\}.$$

Bounding $K_3(\pi)$. We start by taking

$$h_3(Z_i, \pi(X_i)) = \frac{\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\Big[\widehat{G}_{\pi(X_i)}^{(k)}\big(X_i, Y_i(\pi(X_i))\big) - G_{\pi(X_i)}\big(X_i, Y_i(\pi(X_i))\big)\Big].$$

For any $\pi \in \Pi$,

$$
\begin{aligned}
&\mathbb{E}\big[K_3(\pi)\,|\,\mathcal{D}^{(-k)}\big]\\
=&\,\mathbb{E}\Big[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{G}^{(k)}_{A_i}(X_i,Y_i(\pi(X_i)))-G_{A_i}(X_i,Y_i(\pi(X_i)))\big)\,\big|\,\mathcal{D}^{(-k)}\Big]\\
=&\,\mathbb{E}\Big[\widehat{G}^{(k)}_{\pi(X_i)}(X_i,Y_i(\pi(X_i)))-G_{\pi(X_i)}(X_i,Y_i(\pi(X_i)))\,\big|\,\mathcal{D}^{(-k)}\Big]\\
=&\,\mathbb{E}\Big[\ell\big(X_i,Y_i(\pi(X_i));\theta^{(k)}_{\pi(X_i)}\big)-\ell\big(X_i,Y_i;\theta^*_{\pi(X_i)}\big)-\nabla\ell(X_i,Y_i(\pi(X_i));\theta^*_{\pi(X_i)})^\top\big(\widehat{\theta}^{(k)}_{\pi(X_i)}-\theta^*_{\pi(X_i)}\big)\,\big|\,\mathcal{D}^{(-k)}\Big],
\end{aligned}
$$

where the last step follows from part (1) of Lemma B.1. By Assumption 3.4, for any $\beta \in (0,1)$, there exists $N_3 \in \mathbb{N}$ such that when $n \geq N_3$,

$$
\mathbb{P}\Big(\max_{a\in[M]}\|\widehat{\theta}^{(k)}_a-\theta^*_a\|_{L_\infty}>\min\big(\xi,\bar{\alpha},\underline{\alpha},\bar{\eta}\big)/2\Big)\leq\beta.
$$

On the event $\big\{\max_{a\in[M]}\|\widehat{\theta}^{(k)}_a-\theta^*_a\|_{L_\infty}\leq\min(\xi,\bar{\alpha},\underline{\alpha},\bar{\eta})/2\big\}$, we have

$$
\begin{aligned}
&\Big|\ell\big(X_i,Y_i;\theta^{(k)}_{\pi(X_i)}\big)-\ell\big(X_i,Y_i;\theta^*_{\pi(X_i)}\big)-\nabla\ell(X_i,Y_i;\theta^*_{\pi(X_i)})^\top\big(\widehat{\theta}^{(k)}_{\pi(X_i)}-\theta^*_{\pi(X_i)}\big)\Big|\\
\leq&\,\bar{\ell}(X_i,Y_i)\cdot\sum_{a\in[M]}\big\|\widehat{\theta}_a(X_i)-\theta^*_a(X_i)\big\|_2^2,
\end{aligned}
$$

and

$$
\Big|\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{G}^{(k)}_{A_i}(X_i,Y_i(\pi(X_i)))-G_{A_i}(X_i,Y_i(\pi(X_i)))\big)-\mathbb{E}[K_3(\pi)\,|\,\mathcal{D}^{(-k)}]\Big|\leq C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon).
$$

As a result, on the event $\{\max_{a\in[M]}\|\widehat{\theta}^{(k)}_a-\theta^*_a\|_{L_\infty}\leq\min(\xi,\bar{\alpha},\underline{\alpha},\bar{\eta})/2\}$,

$$
\sup_{\pi\in\Pi}\big|\mathbb{E}[K_3(\pi)\,|\,\mathcal{D}^{(-k)}]\big|\leq L\sum_{a\in[M]}\big\|\widehat{\theta}_a-\theta^*_a\big\|^2_{L_2(P_X)}.
$$

We now take

$$
h_3(Z_i,\pi(X_i))=\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{G}^{(k)}_{A_i}(X_i,Y_i(\pi(X_i)))-G_{A_i}(X_i,Y_i(\pi(X_i)))\big)-\mathbb{E}\big[K_3(\pi)\,|\,\mathcal{D}^{(-k)}\big].
$$

By the previous derivation we have $\mathbb{E}[h_3(Z_i,\pi(X_i))\,|\,\mathcal{D}^{(-k)}]=0$ and $|h_3(Z_i,\pi(X_i))|\leq C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)$ on the event $\{\max_{a\in[M]}\|\widehat{\theta}^{(k)}_a-\theta^*_a\|_{L_\infty}\leq\min(\xi,\bar{\alpha},\underline{\alpha},\bar{\eta})/2\}$. On the same event, applying Lemma B.2, we have

$$
\begin{aligned}
&\mathbb{P}\Big(\max_{\pi\in\Pi}|K_3(\pi)|\geq L\sum_{a\in[M]}\big\|\widehat{\theta}_a-\theta^*_a\big\|^2_{L_2(P_X)}+\frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\big(20+4\kappa(\Pi)+\sqrt{2\log(1/\beta)}\big)\,\Big|\,\mathcal{D}^{(-k)}\Big)\\
\leq&\,\mathbb{P}\Big(\max_{\pi\in\Pi}\Big|K_3(\pi)-\mathbb{E}\big[K_3(\pi)\,|\,\mathcal{D}^{(-k)}\big]\Big|\geq\frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\big(20+4\kappa(\Pi)+\sqrt{2\log(1/\beta)}\big)\,\Big|\,\mathcal{D}^{(-k)}\Big)\\
=&\,\mathbb{P}\Big(\max_{\pi\in\Pi}\Big|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}h_3(Z_i,\pi(X_i))\Big|\geq\frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\big(20+4\kappa(\Pi)+\sqrt{2\log(1/\beta)}\big)\,\Big|\,\mathcal{D}^{(-k)}\Big)\\
\leq&\,\beta.
\end{aligned}
$$

Taking a union bound, with probability at least $1-2\beta$, we have

$$
\max_{\pi\in\Pi}|K_3(\pi)|\leq L\sum_{a\in[M]}\big\|\widehat{\theta}_a-\theta^*_a\big\|^2_{L_2(P_X)}+\frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\big(20+4\kappa(\Pi)+\sqrt{2\log(1/\beta)}\big).
\tag{17}
$$

Bounding $K_4(\pi)$. For $K_4(\pi)$, we take

$$h_4(Z_i,\pi(X_i)) = -\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\big(\widehat{g}^{(k)}_{\pi(X_i)}(X_i)-g_{\pi(X_i)}(X_i)\big)+\big(\widehat{g}^{(k)}_{\pi(X_i)}(X_i)-g_{\pi(X_i)}(X_i)\big).$$

and therefore $K_4(\pi)=\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}^{(k)}}h_4(Z_i,\pi(X_i))$. Again by the unconfoundedness assumption,

$$\mathbb{E}\big[h_4(Z_i,\pi(X_i))\,|\,\mathcal{D}^{(-k)}\big]=0.$$

As the case of bounding $K_3(\pi)$, when $n\geq N_3$, with probability at least $1-\beta$,

$$\max_{a\in[M]}\|\widehat{\theta}^{(k)}_a-\theta^*_a\|_{L_\infty}\leq\max(\xi,\bar{\alpha},\underline{\alpha},\bar{\eta})/2.$$

On the event $\{\max_{a\in[M]}\|\widehat{\theta}^{(k)}_a-\theta^*_a\|_{L_\infty}\leq\max(\xi,\bar{\alpha},\underline{\alpha},\bar{\eta})/2\}$, we have

$$\big|h_4(Z_i,\pi(X_i))\big|\leq 2C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon).$$

Applying Lemma B.2 to $h_4(Z_i,\pi(X_i))$ and taking a union bound, we have with probability at least $1-2\beta$ that

$$\max_{\pi\in\Pi}\big|K_4(\pi)\big|\leq\frac{2C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\big(20+4\kappa(\Pi)+\sqrt{2\log(1/\beta)}\big). \tag{18}$$

Combining (15)-(18) and taking a union bound over $k\in[K]$, when $n\geq\max(N_1,N_3)$ we have that with probability at least $1-8\beta$,

$$\begin{aligned}\sup_{\pi\in\Pi}\big|\widehat{\mathcal{V}}^{\mathrm{LN}}_\delta(\pi)-\mathcal{V}_\delta(\pi)\big|\leq{}&\frac{5\sqrt{K}C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{n}}\big(20+4\kappa(\Pi)+\sqrt{2\log(K/\beta)}\big)+L\sum_{a\in[M]}\big\|\widehat{\theta}_a-\theta^*_a\big\|^2_{L_2(P_X)}\\&+\frac{2M}{\beta\varepsilon^2}\max_{a\in[M]}\Big\{\|\widehat{\pi}^{(k)}_a-\pi^{(k)}_{0,a}\|_{L_2(P_X)}\Big\}\max_{a\in[M]}\Big\{\|\widehat{g}^{(k)}_a-\bar{g}^{(k)}_a\|_{L_2(P_X)}\Big\}.\end{aligned}$$

By Assumption 3.4, there exists $N_4\in\mathbb{N}_+$, such that when $n\geq N_4$,

$$\mathbb{P}\bigg(\max_{a\in[M]}\|\widehat{\theta}_a-\theta^*_a\|^2_{L_2(P_X)}\geq\frac{1}{L\sqrt{n}},\ \max_{a\in[M]}\Big\{\|\widehat{\pi}^{(k)}_a-\pi^{(k)}_{0,a}\|_{L_2(P_X)}\Big\}\max_{a\in[M]}\Big\{\|\widehat{g}^{(k)}_a-\bar{g}^{(k)}_a\|_{L_2(P_X)}\Big\}\geq\frac{\beta\varepsilon^2}{2M\sqrt{n}}\bigg)\leq\beta.$$

Taking a union bound, with probability at least $1-9\beta$, we have that

$$\sup_{\pi\in\Pi}\big|\widehat{\mathcal{V}}^{\mathrm{LN}}_\delta(\pi)-\mathcal{V}_\delta(\pi)\big|\leq\frac{5\sqrt{K}C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{n}}\big(22+4\kappa(\Pi)+\sqrt{2\log(K/\beta)}\big).$$

We have thus completed the proof of Theorem 4.2.

C Proof of technical lemmas

C.1 Proof of Lemma B.1

Recall that our loss function is

$$\ell(x,y;\theta)=\alpha\cdot e^{-\frac{y+\eta}{\alpha}-1}+\eta+\alpha\delta.$$

By the strong duality, $\mathbb{E}[\ell(X,Y(\pi(X));\theta)\,|\,X]$ is convex in θ; by Proposition 2.5, the first-order condition of convex optimization implies

$$\nabla_\theta\mathbb{E}\big[\ell(x,Y(\pi(x));\theta^*)\,\big|\,X=x\big]=0.$$

For any $x\in\mathcal{X}$ and $\theta\in\Theta$,

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ell(x,y;\theta)&=\Big(1+\frac{y+\eta}{\alpha}\Big)\cdot\exp\Big(-\frac{y+\eta}{\alpha}-1\Big)+\delta,\\\frac{\partial}{\partial_\eta}\ell(x,y;\theta)&=1-\exp\Big(-\frac{y+\eta}{\alpha}-1\Big).\end{aligned}\tag{19}$$

For some a such that $|a - \alpha^*| \leq \alpha^*/2$, by the monotonicity of $\frac{\partial}{\partial \alpha}\ell(x, y; (\alpha, \eta^*))$ in α, we have

$$\left|\frac{\partial}{\partial \alpha}\ell(x, y; (a, \eta^*))\right| \leq \max\left\{\left|\frac{\partial}{\partial \alpha}\ell(x, y; (3\alpha^*/2, \eta^*))\right|, \left|\frac{\partial}{\partial \alpha}\ell(x, y; (\alpha^*/2, \eta^*))\right|\right\},$$

where the right-hand side is integrable. By dominated convergence theorem,

$$\mathbb{E}\Big[\frac{\partial}{\partial \alpha}\ell(x, Y(\pi(x)); \theta^*) \,\big|\, X = x\Big] = \frac{\partial}{\partial \alpha}\mathbb{E}\Big[\ell(x, Y(\pi(x)); \theta^*) \,\big|\, X = x\Big] = 0.$$

Similarly, since $\frac{\partial}{\partial_\eta}\ell(x, y; (\alpha^*, \eta))$ is non-decreasing in η, for $|\eta - \eta^*| \leq 1$,

$$\left|\frac{\partial}{\partial_\eta}\ell(x, y; (\alpha^*, \eta))\right| \leq \max\left\{\left|\frac{\partial}{\partial_\eta}\ell(x, y; (\alpha^*, \eta^* + 1))\right|, \left|\frac{\partial}{\partial_\eta}\ell(x, y; (\alpha^*, \eta^* - 1))\right|\right\},$$

with the right-hand side being integrable. By dominated convergence theorem,

$$\mathbb{E}\Big[\frac{\partial}{\partial \eta}\ell(x, Y(\pi(x)); \theta^*) \,\big|\, X = x\Big] = \frac{\partial}{\partial \eta}\mathbb{E}\Big[\ell(x, Y(\pi(x)); \theta^*) \,\big|\, X = x\Big] = 0.$$

We have thus completed the proof of Lemma B.1.

Next, for any $x \in \mathcal{X}$ and $\theta \in \Theta$,

$$\begin{aligned}
\frac{\partial^2}{\partial \alpha^2}\ell(x, y; \theta) &= \frac{(y+\eta)^2}{\alpha^3}\exp\Big(-\frac{y+\eta}{\alpha} - 1\Big),\\
\frac{\partial^2}{\partial \alpha \partial \eta}\ell(x, y; \theta) &= -\frac{y+\eta}{\alpha^2}\exp\Big(-\frac{y+\eta}{\alpha} - 1\Big),\\
\frac{\partial^2}{\partial \eta^2}\ell(x, y; \theta) &= \frac{1}{\alpha}\exp\Big(-\frac{y+\eta}{\alpha} - 1\Big).
\end{aligned}$$

By the Taylor expansion,

$$\begin{aligned}
&\ell(x, y; \theta) - \ell(x, y; \theta^*) = \nabla \ell(x, y; \theta^*)^\top(\theta - \theta^*) + \frac{1}{2}(\theta - \theta^*)^\top \nabla^2 \ell(x, y; \tilde{\theta})(\theta - \theta^*),\\
\Rightarrow\ & \big|\ell(x, y; \theta) - \ell(x, y; \theta^*) - \nabla \ell(x, y; \theta^*)^\top(\theta - \theta^*)\big|\\
&\qquad \leq \frac{1}{2}\Big(\frac{(y+\tilde{\eta})^2}{\tilde{\alpha}^3} + \frac{1}{\tilde{\alpha}}\Big)\exp\Big(-\frac{y+\tilde{\eta}}{\tilde{\alpha}} - 1\Big)\|\theta - \theta^*\|_2^2,
\end{aligned}$$

where $\tilde{\theta} = t\theta + (1-t)\theta^*$ for some $t \in [0, 1]$ and α and η implicitly depend on x. To emphasize the dependence on x, we write $\alpha(x)$ and $\eta(x)$ in the following. Letting $\xi = \min(\underline{\alpha}, |\underline{\eta}|, |\overline{\eta}|)/2$, consider $\tilde{\theta} = (\tilde{\alpha}, \tilde{\eta})$ such that $|\tilde{\alpha}(x) - \alpha(x)| \leq \xi$ and $|\tilde{\eta}(x) - \eta(x)| \leq \xi$, for all $x \in \mathcal{X}$. Then,

$$\begin{aligned}
&\frac{1}{2}\Big(\frac{(y+\tilde{\eta}(x))^2}{\tilde{\alpha}(x)^3} + \frac{1}{\tilde{\alpha}(x)}\Big)\exp\Big(-\frac{y+\tilde{\eta}(x)}{\tilde{\alpha}(x)} - 1\Big)\cdot\|\tilde{\theta}(x) - \theta^*(x)\|_2^2\\
&\leq \Big(\frac{8\bar{y}^2 + 8\bar{\eta}^2}{\underline{\alpha}^3} + \frac{2}{\underline{\alpha}}\Big)\cdot\exp\Big(-\frac{\underline{y} + \underline{\eta}}{\bar{\alpha}} - 1\Big)\cdot\|\theta(x) - \theta^*(x)\|_2^2.
\end{aligned}$$

We have thus completed the proof of (2).

We proceed to prove (3). Again by the Taylor expansion,

$$\ell(x, y; \theta) - \ell(x, y; \theta^*) = \nabla \ell(x, y; \tilde{\theta})^\top(\theta(x) - \theta^*(x)),$$

where $\tilde{\theta} = t\theta + (1-t)\theta^*$ for some $t \in [0, 1]$. Plugging the expressions of the gradient in Equation (19), we have

$$\begin{aligned}
&\big[\ell(x, y; \theta) - \ell(x, y; \theta^*)\big]^2 = \big[\nabla \ell(x, y; \tilde{\theta})^\top(\theta(x) - \theta^*(x))\big]^2\\
&= \big(\theta(x) - \theta^*(x)\big)^\top \nabla \ell(x, y; \tilde{\theta}) \nabla \ell(x, y; \tilde{\theta})^\top \big(\theta(x) - \theta^*(x)\big)\\
&\leq \bigg\{\Big[\Big(1 + \frac{y+\eta(x)}{\alpha(x)}\Big)\exp\Big(-\frac{y+\eta(x)}{\alpha(x)} - 1\Big) + \delta\Big]^2 + \Big[1 - \exp\Big(-\frac{y+\eta(x)}{\alpha(x)} - 1\Big)\Big]^2\bigg\}\cdot\big\|\theta(x) - \theta^*(x)\big\|_2^2\\
&\leq C(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta)\cdot\big\|\theta(x) - \theta^*(x)\big\|_2^2,
\end{aligned}$$

where $C(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta)$ is a function of $\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta$. Taking the expectation over X, we have

$$\left\|\ell(x, Y; \theta) - \ell(x, Y; \theta^*)\right\|_{L_2(P_{X,Y \mid A=\pi(X)})} \leq C(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta) \cdot \left\|\theta - \theta^*\right\|_{L_2(P_{X,Y \mid A=\pi(X)})},$$

completing the proof of (3).

C.2 Proof of Lemma B.2

For any $i \in [n]$, let $z_i = (x_i, a_i, y_i)$ and $z_i' = (x_i', a_i', y_i')$. Define

$$f(z_1, \ldots, z_n; \pi) = \frac{1}{n} \sum_{i=1}^{n} h(z_i, \pi(x_i)).$$

We can check that for any $\pi \in \Pi$ and any $j \in [n]$,

$$\begin{aligned}
&\left|f(z_1, \ldots, z_j, \ldots, z_n; \pi)\right| - \sup_{\pi' \in \Pi} \left|f(z_1, \ldots, z_j', \ldots, z_n; \pi')\right| \\
\leq &\left|f(z_1, \ldots, z_j, \ldots, z_n; \pi)\right| - \left|f(z_1, \ldots, z_j', \ldots, z_n; \pi)\right| \\
\leq &\sup_{\pi \in \Pi} \left|f(z_1, \ldots, z_j, \ldots, z_n; \pi) - f(z_1, \ldots, z_j', \ldots, z_n; \pi)\right| \\
= &\sup_{\pi \in \Pi} \frac{1}{n} \left|h(z_j; \pi) - h(z_j'; \pi)\right| \leq 2C_h/n.
\end{aligned} \tag{20}$$

Above, the first inequality is because of the definition of sup and the second is due to the triangle inequality; the last step is due to the boundedness of h. Taking the supremum over all $\pi \in \Pi$ in (20), we have that

$$\sup_{\pi \in \Pi} \left|f(z_1, \ldots, z_j, \ldots, z_n; \pi)\right| - \sup_{\pi \in \Pi} \left|f(z_1, \ldots, z_j', \ldots, z_n; \pi)\right| \leq 2C_h/n.$$

By the bounded difference inequality (Wainwright, 2019, Corollary 2.21), for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\pi \in \Pi} \left|\frac{1}{n} h\big(Z_i, \pi(X_i)\big)\right| - \mathbb{E}\left[\sup_{\pi \in \Pi} \left|\frac{1}{n} h\big(Z_i, \pi(X_i)\big)\right|\right] \geq t\right) \\
=&\mathbb{P}\left(\sup_{\pi \in \Pi} \left|f\big(\{Z_i\}_{i \in [n]}; \pi\big)\right| - \mathbb{E}\left[\sup_{\pi \in \Pi} \left|f\big(\{Z_i\}_{i \in [n]}; \pi\big)\right|\right] \geq t\right) \leq e^{-\frac{nt^2}{2C_h^2}}.
\end{aligned}$$

Take $t = C_h \sqrt{\frac{2}{n} \log\left(\frac{1}{\beta}\right)}$. Then with probability at least $1 - \beta$,

$$\sup_{\pi \in \Pi} \left|\frac{1}{n} h\big(Z_i, \pi(X_i)\big)\right| < \mathbb{E}\left[\sup_{\pi \in \Pi} \left|\frac{1}{n} h\big(Z_i, \pi(X_i)\big)\right|\right] + C_h \sqrt{\frac{2}{n} \log\left(\frac{1}{\beta}\right)}.$$

It remains to bound the expectation term. Let $Z'_1, \dots, Z'_n$ be an i.i.d. copy of $Z_1, \dots, Z_n$, and $\epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{\pm 1\})$. Then

$$\begin{aligned}
&\mathbb{E}\bigg[\sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{i\in[n]} h(Z_i,\pi(X_i)) - \mathbb{E}\Big[h(Z_i,\pi(X_i))\Big]\bigg|\bigg] \\
=&\mathbb{E}\bigg[\sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{i\in[n]} h(Z_i,\pi(X_i)) - \mathbb{E}_{Z'}\Big[\frac{1}{n}\sum_{i\in[n]} h(Z'_i,\pi(X'_i))\Big]\bigg|\bigg] \\
\overset{\text{(i)}}{\le}&\mathbb{E}\bigg[\sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{i\in[n]} h(Z_i,\pi(X_i)) - \frac{1}{n}\sum_{i\in[n]} h(Z'_i,\pi(X_i))\bigg|\bigg] \\
\overset{\text{(ii)}}{=}&\mathbb{E}\bigg[\sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{i\in[n]} \epsilon_i\big(h(Z_i,\pi(X_i)) - h(Z'_i,\pi(X_i))\big)\bigg|\bigg], \\
\le& 2\mathbb{E}\bigg[\sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{i\in[n]} \epsilon_i h(Z_i,\pi(X_i))\bigg|\bigg] \\
=& 2\mathbb{E}\bigg[\mathbb{E}_\epsilon\Big[\sup_{\pi\in\Pi}\bigg|\frac{1}{n}\sum_{i\in[n]} \epsilon_i h(Z_i,\pi(X_i))\bigg|\Big]\bigg],
\end{aligned} \tag{21}$$

step (i) is by Jensen's inequality and step (ii) is because of the symmetry of (Z_i, Z'_i). Before proceeding, we introduce the ℓ_2 distance on the policy space Π, as well as the corresponding covering number.

Definition C.1. Given a function h and a set of realized data $z_1, \dots, z_n$,

(1) the ℓ_2 distance between two policies $\pi_1, \pi_2 \in \Pi$ with respect to $\{z_1, \dots, z_n\}$ is defined as

$$\ell_2(\pi_1, \pi_2; \{z_1, \dots, z_n\}) = \frac{1}{2C_h}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(h(z_i, \pi(x_i)\big) - h(z_i; \pi'(x_i))\big)^2}.$$

(2) $N_2(\gamma, \Pi; \{z_1, \dots, z_n\})$ is the minimum number of policies needed to γ-cover Π under ℓ_2 with respect $\{z_1, \dots, z_n\}$.

Under the ℓ_2 distance, we define a sequence of approximation operators $A_j : \Pi \mapsto \Pi$ for $j \in [J]$, where $J = \lceil \log_2 n \rceil$. Specifically, for any $j = 0, 1, \dots, J$, let S_j be the set of policies that 2^{-j}-covers Π and satisfies $|S_j| = N_2(2^{-j}, \Pi; \{Z_1, \dots, Z_n\})$. Specially, $S_0 = \{\bar{\pi}\}$, with π_0 is an arbitrary policy in Π — this is a valid choice since for any $\pi \in \Pi$,

$$\ell_2(\pi, \bar{\pi}; \{Z_1, \dots, Z_n\}) = \frac{1}{2C_h}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(h(Z_i, \pi(X_i)) - h(Z_i, \bar{\pi}(X_i))\big)^2} \le 1.$$

The approximation operators are defined in a backward manner: for any $\pi \in \Pi$,

(1) define $A_J[\pi] = \underset{\pi' \in S_J}{\text{argmin}}\ \ell_2\big(\pi, \pi'; \{Z_1, \dots, Z_n\}\big)$;

(2) for $j = J-1, \dots, 0$, define

$$A_j[\pi] = \underset{\pi' \in S_j}{\text{argmin}}\ \ell_2\big(A_{j+1}[\pi], \pi'; \{Z_1, \dots, Z_n\}\big).$$

Using the sequential approximation operators, we decompose the inner expectation term in (21) (Rademacher complexity) as

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i h(Z_i,\pi(X_i))\right|\right]\\
\leq\,&\mathbb{E}_\epsilon\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i\big[h(Z_i,\pi(X_i))-h(Z_i,A_J[\pi](X_i))\big]\right|\right]\\
&\quad+\mathbb{E}_\epsilon\left[\sup_{\pi\in\Pi}\left|\sum_{j=1}^{J}\frac{1}{n}\sum_{i\in[n]}\epsilon_i\big[h(Z_i,A_j[\pi](X_i))-h(Z_i,A_{j-1}[\pi](X_i))\big]\right|\right]\\
&\quad+\mathbb{E}_\epsilon\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i h(Z_i,A_0[\pi](X_i))\right|\right]\\
=:\,&\Xi_1+\Xi_2+\Xi_3.
\end{aligned}$$

For any $\pi\in\Pi$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i\big[h(Z_i,\pi(X_i))-h(Z_i,A_J[\pi](X_i))\big]\right| &\leq \frac{1}{n}\sqrt{n\sum_{i\in[n]}\big(h\big(Z_i,\pi(X_i)\big)-h(Z_i,A_J[\pi](X_i))\big)^2}\\
&=2C_h\cdot\ell_2(\pi,A_J(\pi);\{Z_1,\ldots,Z_n\})\\
&\leq 2C_h 2^{-J}\leq\frac{2C_h}{n},
\end{aligned}$$

where the second-to-last step is because $A_J(\pi)$ is 2^{-J}-close to π and the last step is by the choice of J. As a result the above derivation, $\Xi_1\leq 2C_h/n$.

Next, for any $j=1,\ldots,J$ we use P_j to denote the projection of π to S_j, i.e., $A_{j-1}[\pi]=P_{j-1}[A_j[\pi]]$. For any $s>0$,

$$\begin{aligned}
&\mathbb{P}_\epsilon\left(\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i\big[h(Z_i,A_j[\pi](X_i))-h(Z_i;A_{j-1}[\pi](X_i))\big]\right|\geq s\right)\\
\leq&\sum_{\pi'\in S_j}\mathbb{P}_\epsilon\left(\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i\Big[h(Z_i,\pi'(X_i))-h(Z_i,P_{j-1}[\pi'](X_i))\Big]\right|\geq s\right)\\
\leq&\sum_{\pi'\in S_j}2\cdot\exp\left(-\frac{2ns^2}{\sum_{i=1}^n[h(Z_i,\pi'(X_i))-h(Z_i,P_{j-1}[\pi'](X_i))]^2/n}\right)\\
=&\sum_{\pi'\in S_j}2\cdot\exp\left(-\frac{ns^2}{2C_h^2\ell_2(\pi',P_{j-1}(\pi');Z)^2}\right)\\
\leq&\,2N_2(2^{-j},\Pi;Z)\cdot\exp\Big(-\frac{ns^2}{C_h^2 2^{-2j+1}}\Big),
\end{aligned}$$

we Z is a shorthand for $\{Z_1,\ldots,Z_n\}$. For any $j=1,\ldots,J$ and $m\in\mathbb{N}$, take

$$s_{j,m}=\frac{C_h}{2^{j-1/2}}\sqrt{\frac{1}{n}\log\big(N_2(2^{-j},\Pi;Z)\cdot 2^{m+1}\big)}.$$

For a fixed m, with a union bound over $j=1,\ldots,J$ we have that

$$\begin{aligned}
&P_\epsilon\left(\sup_{\pi\in\Pi}\left|\sum_{j=1}^{J}\frac{1}{n}\sum_{i\in[n]}\epsilon_i\big[h(Z_i,A_j[\pi](X_i))-h(Z_i,A_{j-1}[\pi](X_i))\big]\right|\geq\sum_{j=1}^{J}s_{j,m}\right)\\
\leq&\sum_{j=1}^{J}P_\epsilon\left(\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i\big[h(Z_i,A_j[\pi](X_i))-h(Z_i,A_{j-1}[\pi](X_i))\big]\right|\geq s_{j,m}\right)\leq\sum_{j=1}^{J}\frac{1}{j^2 2^m}\leq\frac{1}{2^{m-1}}.
\end{aligned}$$

To proceed, we shall use the following lemma, whose proof is deferred to Appendix C.3.

Lemma C.2. *For any realization $z_1, \dots, z_n$ and $\gamma > 0$, there is $N_2(\gamma, \Pi; z_1, \dots, z_n) \leq N_H(\gamma^2, \Pi)$.*

By Lemma C.2, for any $m \in \mathbb{N}_+$,

$$\begin{aligned}
\sum_{j=1}^{J} s_{j,m} &= \sum_{j=1}^{J} \frac{C_h}{2^{j-1/2}\sqrt{n}} \sqrt{\log\left(N_2(2^{-j}, \Pi; Z) \cdot 2^{m+1}\right)} \\
&\leq \sum_{j=1}^{J} \frac{C_h}{2^{j-1/2}\sqrt{n}} \sqrt{\log(N_H(2^{-2j}, \Pi)) + (m+1)\log 2} \\
&\overset{\text{(i)}}{\leq} \frac{2C_h}{\sqrt{n}} \sum_{j=1}^{J} 2^{-j} \cdot \left(\sqrt{\log(N_H(2^{-2j}, \Pi))} + \sqrt{m+1}\right) \\
&\overset{(ii)}{\leq} \frac{4C_h}{\sqrt{n}} \left(\kappa(\Pi) + \sqrt{m+1}\right) =: u_m,
\end{aligned}$$

where step (i) uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$; step (ii) uses the definition of $\kappa(\Pi)$. Then

$$\begin{aligned}
\Xi_2 &= \mathbb{E}_\epsilon\left[\sup_{\pi \in \Pi}\left|\sum_{j=1}^{J} \frac{1}{n} \sum_{i \in [n]} \epsilon_i \Big[h(Z_i, A_j[\pi](X_i)) - h(Z_i, A_{j-1}[\pi](X_i))\Big]\right|\right] \\
&= \int_0^\infty \mathbb{P}_\epsilon\left(\sup_{\pi \in \Pi}\left|\sum_{j=1}^{J} \frac{1}{n} \sum_{i \in [n]} \epsilon_i \Big[h(Z_i, A_j[\pi](X_i)) - h(Z_i, A_{j-1}[\pi](X_i))\Big]\right| > s\right) ds \\
&\leq u_1 + \sum_{k=1}^{\infty} (u_{k+1} - u_k) \cdot 2^{-k+1} \\
&= \frac{C(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \underline{\eta})}{\sqrt{n\varepsilon}} \cdot \left(\kappa(\Pi) + \sqrt{2} + \sum_{k=1}^{\infty} (\sqrt{k+2} - \sqrt{k+1}) \cdot 2^{-k+1}\right) \leq \frac{4C_h}{\sqrt{n}} \cdot \big(\kappa(\Pi) + 4\big).
\end{aligned}$$

Finally, we consider Ξ_3. Recall that $S_0 = \{\bar{\pi}\}$, and therefore

$$\Xi_3 = \mathbb{E}_\epsilon\left[\left|\frac{1}{n}\sum_{i \in [n]} \epsilon_i h(Z_i, \bar{\pi}(X_i))\right|\right] \leq \sqrt{\mathbb{E}_\epsilon\left[\left(\frac{1}{n}\sum_{i \in [n]} \epsilon_i h(Z_i, \bar{\pi}(X_i))\right)^2\right]} \leq \frac{2C_h}{\sqrt{n}}.$$

Putting everything together, we have with probability $1 - \beta$ that

$$\sup_{\pi \in \Pi}\left|\frac{1}{n}\sum_{i=1}^{n} h(Z_i, \pi(X_i))\right| \leq \frac{C_h}{\sqrt{n}}\left(20 + 4\kappa(\Pi) + \sqrt{2\log\left(\frac{1}{\beta}\right)}\right).$$

C.3 Proof of Lemma C.2

Fix $\gamma > 0$. If $N_H(\gamma^2, \Pi) = \infty$, the lemma is trivially true. Otherwise, let $N_0 = N_H(\gamma^2; \Pi)$. For any realization $z_1, \dots, z_n$, define

$$(\pi^*_{i,1}, \pi^*_{i,2}) = \operatorname*{argmax}_{\pi_1, \pi_2}\left\{|h(z_i, \pi_1(x_i)) - h(z_i, \pi_2(x_i))|\right\}.$$

Implicitly, $(\pi^*_{i,1}, \pi^*_{i,2})$ depends on z_i. For an arbitrary positive integer m and $i \in [n]$, we define

$$n_i = \left\lceil \frac{m}{4C_h^2 n}\left\{h(z_i, \pi^*_{i,1}(x_i)) - h(z_i, \pi^*_{i,2}(x_i))\right\}^2 \right\rceil.$$

We then construct a new set of data

$$\{\tilde{z}_1, \dots, \tilde{z}_N\} = \{z_1, \dots, z_1, z_2, \dots, z_2, \dots, z_n, \dots, z_n\},$$

where z_i appears n_i times and

$$N = \sum_{i=1}^{n} n_i = \sum_{i=1}^{n} \left\lceil \frac{m}{4C_h^2 n}\left\{h(z_i, \pi^*_{i,1}(x_i)) - h(z_i, \pi^*_{i,2}(x_i))\right\}^2 \right\rceil \leq m + n.$$

By definition, there exists a policy set S_0 to be a γ^2-cover of Π the Hamming distance with respect to $\tilde{x} := \{\tilde{x}_1, \ldots, \tilde{x}_N\}$ such that $|S_0| = N_0$. As a result, for any $\pi \in \Pi$, there exists $\pi' \in S_0$ such that $H(\pi, \pi'; \tilde{x}) \leq \gamma^2$. On the other hand,

$$\begin{aligned}
H(\pi, \pi'; \tilde{x}) &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\pi(\tilde{x}_i) \neq \pi'(\tilde{x}_i)\} \\
&\overset{(i)}{=} \frac{1}{N} \sum_{i=1}^{n} n_i \mathbb{1}\{\pi(x_i) \neq \pi'(x_i)\} \\
&\geq \frac{1}{N} \sum_{i=1}^{n} \frac{m}{4C_h^2 n} \big\{h(z_i, \pi^*_{i,1}(x_i)) - h(z_i, \pi^*_{i,2}(x_i))\big\}^2 \cdot \mathbb{1}\{\pi(x_i) \neq \pi'(x_i)\} \\
&\overset{\text{(ii)}}{\geq} \frac{1}{N} \sum_{i=1}^{n} \frac{m}{4C_h^2 n} \big\{h(z_i, \pi(x_i)) - h(z_i, \pi'(x_i))\big\}^2 \cdot \mathbb{1}\{\pi(x_i) \neq \pi'(x_i)\} \\
&\overset{\text{(iii)}}{=} \frac{1}{N} \sum_{i=1}^{n} \frac{m}{4C_h^2 n} \big\{h(z_i, \pi(x_i)) - h(z_i, \pi'(x_i))\big\}^2.
\end{aligned}$$

Above, step (i) and (ii) follow from the choice of $\tilde{z}$ and $(\pi^*_{i,1}, \pi^*_{i,2})$, respectively; step (iii) is because when $\pi(x_i) = \pi'(x_i)$, $h(z_i, \pi(x_i)) = h(z_i, \pi(x'_i))$. By the definition of the ℓ_2 distance and that $N \leq m + n$, we further have

$$\gamma^2 \geq H(\pi, \pi'; \tilde{x}) \geq \frac{m}{(m+n)} \ell^2(\pi, \pi'; z).$$

Since m is arbitrary, we take m to infinity and have $\ell_2(\pi, \pi'; z) \leq \gamma$. By definition, S_0 is a γ-cover of Π under ℓ_2 with respect to $z_1, \ldots, z_n$, and therefore $N_2(\gamma, \Pi; z_1, \ldots, z_n) \leq N_H(\gamma, \Pi)$.