Distributionally Robust Policy Learning under Concept Drifts

Jingyuan Wang [1] **Zhimei Ren** [2] **Ruohan Zhan** [3] **Zhengyuan Zhou** [1 4]

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 $\kappa(\Pi)n^{-1/2}$, where $\kappa(\Pi)$ is the entropy integral of Π under the Hamming distance and n is the sample size. A matching lower bound is provided to show the optimality of the rate. The proposed methods are implemented and evaluated in numerical studies, demonstrating substantial improvement compared with existing benchmarks.

[1]Stern School of Business, New York University [2]Department of Statistics and Data Science, University of Pennsylvania [3]IEDA, Hong Kong University of Science and Technology [4]Arena Technologies. Correspondence to: Zhengyuan Zhou <zhengyuanzhou24@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

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 researchers devise 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 are different in nature, have different implications on the objectives, and call for distinct treatment (Namkoong et al., 2023; Liu et al., 2023; Jin et al.,

	Distribution shift	Unknown π_0	General $\mathcal{X}$	Upper bound	Lower bound
Athey & Wager (2021)	✗	✓	✓	—	—
Zhou et al. (2023)	✗	✓	✓	—	—
Si et al. (2023)	Joint	✗	✓	—	—
Kallus et al. (2022)	Joint	✓	✓	—	—
Mu et al. (2022)	Separate	✓	✗	$O\big(\sqrt{\frac{\log n \log(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert)}{n}}\big)$	✗
This work	Separate	✓	✓	$O\big(\frac{\kappa(\Pi)}{\sqrt{n}}\big)$	$\Omega\big(\sqrt{\frac{\mathrm{Ndim}(\Pi)}{n}}\big)$

Table 1. Comparison of results in the offline policy learning literature. "Unknown π_0" refers to whether an algorithm assumes knowledge of the behavior policy π_0. "General $\mathcal{X}$" refers to whether an algorithm allows for general types of covariates. Athey & Wager (2021); Zhou et al. (2023); Si et al. (2023); Kallus et al. (2022) have the regret upper and lower bounds for the specific problems they consider that are not directly comparable to ours, so we do not include them in the table. $|\mathcal{X}|$ refers to the cardinality of the covariate support (if finite) and $|\mathcal{A}|$ to that of the action set. $\kappa(\Pi)$ and $\mathrm{Ndim}(\Pi)$ are the entropy integral under Hamming distance and the Natarajan dimension of a policy class Π, with the relation $\kappa(\Pi) = O(\log(d)\mathrm{Ndim}(\Pi))$, where d is the dimension of the covariate space.

2023; Ai & Ren, 2024). To be concrete, imagine that the distribution of covariates changes while that of $Y \mid X$ remains invariant — in this case, the distribution shift is identifiable/estimable since the covariates are often accessible in the target environment. As a result, it is often unnecessary to account for the worst-case covariate shift rather than directly correcting for it. Alternatively, when the $Y \mid X$ distribution changes but the X distribution remains invariant, the distribution shift is no longer identifiable, where we can instead apply the worst-case consideration to guarantee performance. This latter setting, known as *concept drift*, occurs 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 product recommendation, similar population segments in developed and emerging markets may prefer different product features. In these applications, with the one extra bit of information that the shift is only in the conditional reward distribution, can we obtain a more efficient policy learning algorithm?

Motivated by the above situations, we study robust policy learning under concept drift. Most existing methods for robust policy learning (Si et al., 2023; Kallus et al., 2022) 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 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,[1] 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 solving the dual form of a distributionally robust optimization problem and taking a de-biased step to deal with the slow convergence of the optimizer, thereby obtaining an estimator with root-n convergence rate.

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 regret/suboptimality gap of the order $\kappa(\Pi)/\sqrt{n}$, where $\kappa(\Pi)$ is a measure quantifying the

[1] Otherwise, the covariate shift can be easily adjusted by covariate matching discussed earlier.

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. We complement the upper bound with a matching lower bound, thus establishing the minimax optimality of our proposed algorithm. We summarize the comparison between our result and prior work in Table 1 for better demonstration.

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, 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. 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 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 an individualized policy with optimal worst-case performance guarantees.

2. Preliminaries

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 $\pi_0(a \mid x) := \mathbb{P}(A_i = a \mid X = x)$ is the *propensity score*, 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 unconfoundedness 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 environments. 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 $\mathcal{P}(P_{Y \mid X}, \delta) := \{Q_{Y \mid X} : D_{\mathrm{KL}}(Q_{Y \mid X} \,\|\, P_{Y \mid X}) \leq \delta\}$, where $P_{Y \mid X}$ and $Q_{Y \mid X}$ refers to the distribution of $(Y(1), \ldots, Y(M)) \mid X$ under P and Q respectively. The distributionally robust policy value for any policy π at level δ is defined as

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

The optimal policy in Π is the one that maximizes $\mathcal{V}_\delta(\pi)$, i.e. $\pi^*_\delta := \mathrm{argmax}_{\pi\in\Pi}\ \mathcal{V}_\delta(\pi)$.[2]

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):

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

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 (see e.g., Luenberger (1997)). The transformation is formalized in the following lemma, with its proof provided in Appendix D.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*

$$-\min_{\alpha\geq 0,\eta\in\mathbb{R}} \mathbb{E}_P\left[V_\delta(\alpha,\eta;\pi)\,\middle|\, X=x\right]. \tag{3}$$

where $V_\delta(\alpha,\eta;\pi) = \alpha\exp(-\frac{Y(\pi(X))+\eta}{\alpha} - 1) + \eta + \alpha\delta$.

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

$$\big(\boldsymbol{\alpha}^*_\pi(x), \boldsymbol{\eta}^*_\pi(x)\big) \in \operatorname*{argmin}_{\alpha\geq 0,\eta\in\mathbb{R}} \mathbb{E}_P\left[V_\delta(\alpha,\eta;\pi)\,\middle|\, X=x\right].$$

With this notation and Lemma 2.3, the robust policy value

$$\mathcal{V}_\delta(\pi) = -\mathbb{E}_P\left[V_\delta(\boldsymbol{\alpha}^*_\pi(X), \boldsymbol{\eta}^*_\pi(X);\pi)\right]. \tag{4}$$

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 distributional shift formulation, the above representation admits an optimizer pair $(\boldsymbol{\alpha}^*_\pi(x), \boldsymbol{\eta}^*_\pi(x))$ that is *dependent* on the context x (i.e. $\boldsymbol{\alpha}^*_\pi, \boldsymbol{\eta}^*_\pi$ are functions of x) and the policy π. As we shall see shortly, our proposed policy value estimation procedure employs ERM tools to estimate $(\boldsymbol{\alpha}^*_\pi, \boldsymbol{\eta}^*_\pi)$, and then compute an estimate of $\mathcal{V}_\delta(\pi)$ by plugging $(\boldsymbol{\alpha}^*_\pi, \boldsymbol{\eta}^*_\pi)$ 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 convergence rate even when then nuisance parameters (e.g., $(\boldsymbol{\alpha}^*_\pi, \boldsymbol{\eta}^*_\pi)$) are converging slower than the root-n rate.

We end this section by providing a sufficient condition to ensure $\boldsymbol{\alpha}^*_\pi(x) > 0$, which we make throughout the paper.

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 $\tilde{p}(x;a) = \mathbb{P}(Y(a) = \underline{y}(x;a) \mid X = x, A = a)$. It holds that $\log(1/\tilde{p}(x;a)) > \delta$ for $P_{X|A=a}$-almost all x.

The above assumption is mild and can be satisfied by many commonly used distributions, e.g., all the continuous distributions; it requires that $P_{Y\mid X,A}$ does not posit a large point mass at its essential infimum. The following result from Jin et al. (2022b, Proposition 4), shows that $\alpha^* > 0$ when Assumption 2.4 holds, ensuring 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 $k \in [K]$,[3] where we slightly abuse the notation and use $\mathcal{D}^{(k)}$ to denote the data points or the corresponding indices interchangeably.

For each $k \in [K]$, we first use data points in $\mathcal{D}^{(k+1)}$ to obtain the propensity score estimator $\widehat{\pi}_0^{(k)}$ and the optimizers $(\widehat{\boldsymbol{\alpha}}^{(k)}_\pi, \widehat{\boldsymbol{\eta}}^{(k)}_\pi)$.[4] We then define

$$\begin{aligned}\widehat{G}^{(k)}_\pi(x,y) :=& \widehat{\boldsymbol{\alpha}}^{(k)}_\pi(x)\cdot e^{-\frac{y+\widehat{\boldsymbol{\eta}}^{(k)}_\pi(x)}{\widehat{\boldsymbol{\alpha}}^{(k)}_\pi(x)}-1} + \widehat{\boldsymbol{\eta}}^{(k)}_\pi(x) + \widehat{\boldsymbol{\alpha}}^{(k)}_\pi(x)\cdot\delta,\\ \bar{g}^{(k)}_\pi(x) :=& \mathbb{E}_P\Big[\widehat{G}^{(k)}_\pi\big(X, Y(\pi(X))\big)\,\Big|\, X=x\Big].\end{aligned}$$

We next use $\mathcal{D}^{(k+2)}$ to obtain $\widehat{g}^{(k)}_\pi$ as an estimator of $\bar{g}^{(k)}_\pi$.

[2] 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.

[3] We assume without loss of generality that n is divisible by K. In practice, we only need a minimum of $K = 3$ folds.

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

Algorithm 1 Policy estimation under concept drift

Input: Dataset $\mathcal{D}$; policy π; uncertainty set parameter δ; propensity score estimation algorithm $\mathcal{C}$; ERM algorithm $\mathcal{E}$ for obtaining $(\boldsymbol{\alpha}_\pi^*, \boldsymbol{\eta}_\pi^*)$; 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**

$\widehat{\pi}_0^{(k)} \leftarrow \mathcal{C}(\mathcal{D}^{(k+1)})$, $(\widehat{\boldsymbol{\alpha}}_\pi^{(k)}, \widehat{\boldsymbol{\eta}}_\pi^{(k)}) \leftarrow \mathcal{E}(\mathcal{D}^{(k+1)})$;

$\widehat{g}_\pi^{(k)} \leftarrow \mathcal{R}\big(\{X_i, A_i, \widehat{G}_\pi^{(k)}(X_i, Y_i); i \in \mathcal{D}^{(k+2)}\}\big)$;

Compute $\widehat{\mathcal{V}}_\delta^{(k)}(\pi)$ according to Equation (5);

end for

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

The policy value estimator $\widehat{\mathcal{V}}_\delta^{(k)}(\pi)$ for the k-th fold is

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

The policy value estimator is $\widehat{\mathcal{V}}_\delta(\pi) := -\frac{1}{K}\sum_{k=1}^K \widehat{\mathcal{V}}_\delta^{(k)}(\pi)$. The complete procedure is summarized in Algorithm 1.

Remark 3.1. The estimation procedure involves three model-fitting steps corresponding to π_0, $(\boldsymbol{\alpha}_\pi^*, \boldsymbol{\eta}_\pi^*)$, 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 $\bar{g}_\pi^{(k)}$ can be obtained by 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 details 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 splits the training dataset $\mathcal{D}$ into K folds equally and fits the nuisance parameters on the off-fold data. This crucially provides the convenient property of independence between the fitted nuisance parameters and policy value estimators. In contrast to the naïve plug-in estimator whose convergence rate is largely affected by the slow estimation rate of the nuisance parameters, the de-biasing technique addresses this limitation, thereby 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 from Lemma 2.3 as

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

By the notation, $\boldsymbol{\theta}_\pi^*(x) = (\boldsymbol{\alpha}_\pi^*(x), \boldsymbol{\eta}_\pi^*(x))$ is the optimizer of $\mathbb{E}_P[\ell(x, Y(\pi(x)); \theta) \mid X = x]$ with respect to θ. Throughout, we make the following assumption on θ_π^*.

Assumption 3.3. For any policy π, $\exists \underline{\alpha}, \bar{\alpha}, \bar{\eta}$ such that $0 < \underline{\alpha} \le \boldsymbol{\alpha}_\pi^*(x) \le \bar{\alpha}$, $\big|\boldsymbol{\eta}_\pi^*(x)\big| \le \bar{\eta}$, $\forall x \in \mathcal{X}$.

The above assumption is mild and can be achieved, for example, when $\boldsymbol{\theta}_\pi^*(x)$ 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, $\boldsymbol{\theta}_\pi^*$ is also a minimizer of $\mathbb{E}_P\big[\ell(X, Y; \boldsymbol{\theta}(X))\mathbb{1}\{A = \pi(X)\}\big]$. We can thus estimate $\boldsymbol{\theta}_\pi^*$ by minimizing the empirical risk:

$$\widehat{\boldsymbol{\theta}}_\pi^{(k)} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \Big\{\widehat{\mathbb{E}}_{\mathcal{D}^{(k+1)}}[\ell(X, Y; \boldsymbol{\theta}(X))\mathbb{1}\{A = \pi(X)\}]\Big\}, \quad (7)$$

where $\Theta \subseteq \{(\boldsymbol{\alpha}, \boldsymbol{\eta}) \mid \boldsymbol{\alpha} : \mathcal{X} \mapsto \mathbb{R}_{\ge 0},\ \boldsymbol{\eta} : \mathcal{X} \mapsto \mathbb{R}\}$ 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, $\widehat{\boldsymbol{\theta}}_\pi^{(k)}$ converges to $\boldsymbol{\theta}_\pi^*$ at a reasonably fast 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 $\boldsymbol{\theta}_\pi^*$ belongs to the Hölder class of p-smooth functions — with some other mild regularity conditions — then $\|\widehat{\boldsymbol{\theta}}_\pi^{(k)} - \boldsymbol{\theta}_\pi^*\|_{L_2(P_{X \mid A = \pi(X)})} = O_P((\frac{\log n}{n})^{-p/(2p+d)})$ and $\|\widehat{\boldsymbol{\theta}}_\pi^{(k)} - \boldsymbol{\theta}_\pi^*\|_{L_\infty} = O_P((\frac{\log n}{n})^{-2p^2/(2p+d)^2})$. The solution details are given in Appendix B, and we also refer the readers to Yadlowsky et al. (2018) and Jin et al. (2022b).

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 assume the following for the convergence rates of the nuisance parameter estimators.

Assumption 3.4 (Asymptotic estimation rate)**.** For any policy π, assume that for each $k \in [K]$, the estimators $\widehat{\pi}_0^{(k)}$, $\widehat{g}_\pi^{(k)}$ and the empirical risk optimizer $\widehat{\boldsymbol{\theta}}_\pi^{(k)}$ satisfy

$$\|\widehat{\pi}_0^{(k)} - \pi_0\|_{L_2(P_{X \mid A = \pi(X)})} = o_P(n^{-\gamma_1}),$$
$$\|\widehat{g}_\pi^{(k)} - \bar{g}_\pi^{(k)}\|_{L_2(P_{X \mid A = \pi(X)})} = o_P(n^{-\gamma_2}),$$
$$\|\widehat{\boldsymbol{\theta}}_\pi^{(k)} - \boldsymbol{\theta}_\pi^*\|_{L_2(P_{X \mid A = \pi(X)})} = o_P(n^{-\frac{1}{4}}),$$
$$\|\widehat{\boldsymbol{\theta}}_\pi^{(k)} - \boldsymbol{\theta}_\pi^*\|_{L_\infty} = o_P(1),$$

for some $\gamma_1, \gamma_2 \ge 0$ and $\gamma_1 + \gamma_2 \ge \frac{1}{2}$.

Assumption 3.4 (1) requires either the propensity score π_0 or the conditional mean of $\widehat{G}_\pi^{(k)}(X,Y)$ is well estimated. This is a standard assumption 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 (2) requires the optimizer $\widehat{\boldsymbol{\theta}}_\pi^{(k)}$ 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 empirical sensitivity analysis in Jin et al. (2022b) also provides some justification for Assumption 3.4.

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 D.2.

Theorem 3.5 (Asymptotic normality)**.** *Suppose Assumptions 2.1, 2.4, 3.3, and 3.4 hold. For any policy $\pi:\mathcal{X}\mapsto[M]$, we have $\sqrt{n}\cdot\big(\widehat{\mathcal{V}}_\delta(\pi)-\mathcal{V}_\delta(\pi)\big)\xrightarrow{\text{d}} N(0,\sigma_\pi^2)$, where*

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

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 nuisance parameter $\widehat{\boldsymbol{\theta}}_\pi^{(k)}$ 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 $\widehat{\boldsymbol{\theta}}_\pi^{(k)}$ for each $\pi\in\Pi$, we propose a computational shortcut that solves a similar ERM problem for each action $a\in[M]$. To see why this is sufficient, note that for any $\pi\in\Pi$,

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

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

$$-\mathbb{E}\bigg[\ell(X,Y(\pi(X));\boldsymbol{\theta}_{\pi(X)}^*(X))\bigg].$$

Remark 4.1 (Computational efficiency of the proposed shortcut). Constructing $\widehat{\boldsymbol{\theta}}_\pi(x)$ with $\widehat{\boldsymbol{\theta}}_{\pi(x)}(x)$ substantially reduces the computational complexity of the policy learning task. It is virtually infeasible to estimate $\widehat{\boldsymbol{\theta}}_\pi(x)$ for each π in a policy class Π with infinite number of policies. Alternatively, solving for $\widehat{\boldsymbol{\theta}}_{\pi(x)}(x)$ transforms the infeasible task of computing a class of infinite nuisance parameters $\{\boldsymbol{\theta}_\pi(x):\pi\in\Pi\}$ to the feasible task of computing a finite one $\{\boldsymbol{\theta}_a(x):a\in[M]\}$. It remains an interesting future direction to extend this trick to continuous action spaces.

4.1. The Learning Algorithm

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

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 $\widehat{\pi}_0^{(k)}(a\,|\,\cdot)$ (by regression) and the optimizer $\boldsymbol{\theta}_a^*$ (by ERM) for each $a\in[M]$. Next, for each $a\in[M]$, define

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

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:

$$\widehat{\pi}_{\mathrm{LN}}=\underset{\pi\in\Pi}{\mathrm{argmax}}\,\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi):=-\frac{1}{K}\sum_{k=1}^K\widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi);\tag{9}$$
$$\begin{aligned}\widehat{\mathcal{V}}_\delta^{\mathrm{LN},(k)}(\pi)&=\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i\,|\,X_i)}\\&\cdot\big(\widehat{G}_{\pi(X_i)}^{(k)}(X_i,Y_i)-\widehat{g}_{\pi(X_i)}^{(k)}(X_i)\big)+\widehat{g}_{\pi(X_i)}^{(k)}(X_i).\end{aligned}$$

The above optimization problem can be solved efficiently 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.

4.2. Regret Upper Bound

In this section, we present the regret analysis of $\widehat{\pi}_{\mathrm{LN}}$ obtained by Algorithm 2 (recall the definition of regret 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 4.2 (Hamming entropy integral)**.** Given a policy class Π and n data points $\{x_1,\ldots,x_n\}\subseteq\mathcal{X}$,

Algorithm 2 Policy learning under concept drift

Input: Dataset $\mathcal{D}$; policy class Π; uncertainty set parameter δ; propensity score estimation algorithm $\mathcal{C}$; ERM algorithm $\mathcal{E}(\cdot)$ for obtaining $\boldsymbol{\theta}_a^*$; 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**
    $\widehat{\boldsymbol{\theta}}_a^{(k)} \leftarrow \mathcal{E}(\mathcal{D}^{(k+1)})$;
    $\widehat{g}_a^{(k)} \leftarrow \mathcal{R}(X_i, A_i, \widehat{G}_a^{(k)}(X_i, Y_i); i \in \mathcal{D}^{(k+2)})$;
  end for
end for
Return: $\widehat{\pi}_{\mathrm{LN}}$ that maximizes $\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi)$ as in Equation (9).

(1) The *Hamming distance* between two policies $\pi, \pi' \in \Pi$ is $d_H(\pi, \pi') := \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\pi(x_i) \neq \pi'(x_i)\}$.

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

The following theorem provides a regret upper bound for the policy learned by Algorithm 2.

Theorem 4.3. *Suppose Assumptions 2.1, 2.4, 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*

$$\mathcal{R}_\delta(\widehat{\pi}_{\mathrm{LN}}) \leq \frac{C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)}{\sqrt{n}}\big(65 + 8\kappa(\Pi) + \sqrt{\log(1/\beta)}\big),$$

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 4.3 is deferred to Appendix D. At a high level, we decompose the regret and upper bound it with the supremum of the estimation error of policy values:

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

where the last step uses the choice of $\widehat{\pi}_{\mathrm{LN}}$. We bound the above quantity by establishing uniform convergence results for the policy value estimators. Through a careful chaining argument, we show that the dependence of $\mathcal{R}(\widehat{\pi}_{\mathrm{LN}})$ on n is of the order $O(n^{-\frac{1}{2}})$, which is sharper than the $O(n^{-\frac{1}{2}}\log n)$ dependence for that of Mu et al. (2022) by a logarithmic factor. We also note that both regrets are asymptotic in n and hold for sufficiently large n.

4.3. Regret Lower Bound

In this section, we complement the regret upper bound in Theorem 4.3 with a minimax lower bound that characterizes the fundamental difficulty of policy learning under concept drift. Our lower bound is stated in terms of the Natarajan dimension (Natarajan, 1989), defined as follows.

Definition 4.4 (Natarajan dimension)**.** Given an M-action policy class Π, we say a set of m points $\{x_1, \ldots, x_m\}$ is *shattered* by Π if there exist two functions $f_{-1}, f_1 : \{x_1, \ldots, x_m\} \mapsto [M]$ such that $f_{-1}(x_j) \neq f_1(x_j)$ for all $j \in [m]$ and for any $\sigma \in \{\pm 1\}^m$, there exists a policy $\pi \in \Pi$ such that for any $j \in [m]$, $\pi(x_j) = f_{\sigma_j}(x_j)$. The *Natarajan dimension* $\mathrm{Ndim}(\Pi)$ of Π is the size of the largest set shattered by Π.

Remark 4.5 (Connection to other complexity measures)*.* As can be seen from the definition, the Natarajan dimension generalizes the Vapnik-Chervonenkis (VC) dimension (Vapnik & Chervonenkis, 2015) to the multi-class classification setting. The Natarajan dimension is also closely related to the Hamming entropy integral $\kappa(\Pi)$ in our upper bound, as $\kappa(\Pi) = O(\sqrt{\log(d)\mathrm{Ndim}(\Pi)})$ (Cai et al., 2020).

Theorem 4.6 (Regret lower bound)**.** *Let $\mathcal{P}$ denote the set of all distributions of $(X, A, Y(1), \ldots, Y(M))$ that satisfy Assumption 2.1, 2.4, 3.3, and 3.4.*[5] *Suppose that $\delta \leq 0.2$, $n \geq \mathrm{Ndim}(\Pi)^2$, and $\mathrm{Ndim}(\Pi) \geq 4/(9\varepsilon)$. For any policy leaning algorithm that outputs $\widehat{\pi}$ as a function of $\{(X_i, A_i, Y_i)\}_{i=1}^n$, there is $\sup_{P\in\mathcal{P}} \mathbb{E}_{P^n}[\mathcal{R}(\widehat{\pi})] \geq \frac{1}{120}\sqrt{\frac{\mathrm{Ndim}(\Pi)}{n\varepsilon}}$.*

The proof of Theorem 4.6 is provided in Appendix D.4. Theorem 4.6 implies that for any learning algorithm, there exists a problem instance such that the regret scales as $\Omega(\sqrt{\mathrm{Ndim}(\Pi)/n})$.

Remark 4.7 (Optimality of Algorithm 2)*.* Recalling the relationship between the Natarajan dimension and the Hamming entropy integral in the remark above, we see that our proposed algorithm achieves the minimax rate in the sample size and the policy class complexity.

5. Numerical Results

We evaluated our learning algorithm in two settings: a simulated and a real-world dataset, against the benchmark algorithm SNLN in Si et al. (2023, Algorithm 2).

Simulated Dataset. Our data generating process follows that of the linear boundary example in Si et al. (2023). We let the context set $\mathcal{X}$ to be the closed unit ball of $\mathbb{R}^5$ and let the action set to be $\{1, 2, 3\}$; the rewards $Y(a)$'s are mutually independent conditioned on X with $Y(a) \mid X$

[5]When we say a distribution P satisfies Assumption 3.4, we mean that under P there exist $\widehat{\boldsymbol{\theta}}$, $\widehat{\pi}_0$, and $\widehat{g}$ that satisfy the convergence rates in Assumption 3.4.

METRIC	δ	POLICY	n =7500	n =13500	n =16500	n =19500
$\bar{\mathcal{V}}_\delta$	0.05	$\widehat{\pi}_{\text{LN}}$	0.2272±0.002	0.2299±0.001	0.2303±0.001	0.2310±0.001
		$\widehat{\pi}_{\text{SNLN}}$	0.0554±0.005	0.0589±0.004	0.0617±0.004	0.0664±0.003
	0.1	$\widehat{\pi}_{\text{LN}}$	0.1579±0.007	0.1662±0.002	0.1663±0.002	0.1678±0.002
		$\widehat{\pi}_{\text{SNLN}}$	0.0548±0.004	0.0580±0.004	0.0583±0.003	0.0616±0.004
	0.2	$\widehat{\pi}_{\text{LN}}$	0.0781±0.003	0.0802±0.002	0.0804±0.002	0.0831±0.002
		$\widehat{\pi}_{\text{SNLN}}$	0.0182±0.003	0.0183±0.003	0.0200±0.003	0.0219±0.003

Table 2. Empirical robust policy value $\bar{\mathcal{V}}_\delta$ of policies $\widehat{\pi}_{\text{LN}}, \widehat{\pi}_{\text{SNLN}}$ on the simulated dataset and the corresponding, over 50 seeds.

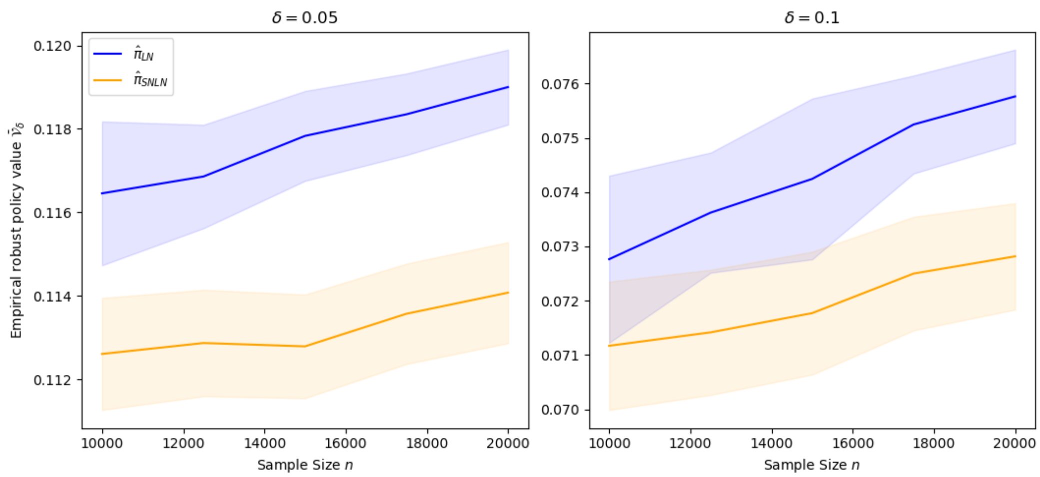

Figure 1. Empirical robust policy value $\bar{\mathcal{V}}_\delta$ of policies $\widehat{\pi}_{\text{LN}}, \widehat{\pi}_{\text{SNLN}}$ on the real-world dataset, over 50 seeds. Shading corresponds to 95% confidence intervals.

being Gaussian, for $a \in [3]$. The training datasets $\mathcal{D}_{\text{train}}$ are generated with an unknown given behavior policy π_0 over 50 random seeds. We similarly generate 100 testing datasets $\mathcal{D}_{\text{test}}$ of size 10,000. The details are given in Appendix C.

Real-world Dataset. We consider the dataset of a large-scale randomized experiment comparing assistance programs offered to French unemployed individuals provided in Behaghel et al. (2014). The decision maker is trying to learn a personalized policy that decides whether to provide: (i) an intensive counseling program run by a public agency $(A = 0)$; or (ii) a similar program run by private agencies $(A = 1)$, to an unemployed individual. The reward Y is binary and indicates reemployment within six months. The processed dataset is provided in Kallus (2023).

Implementation. In our implementation, the number of splits is taken to be $K = 3$. We use the Random Forest regressor from the `scikit-learn` Python library to estimate $\widehat{\pi}_0$ and $\widehat{g}$. For estimating $\boldsymbol{\theta}^*$, we adopt the cubic spline method and employ the Nelder-Mead optimization method in `SciPy` Python library (Virtanen et al., 2020) to optimize the coefficients in the spline approximation, where the obtained estimator has threshold at 0.001 to guarantee Proposition 2.5. Finally, we optimize and find $\widehat{\pi}_{\text{LN}}$ with `policytree` (Sverdrup et al., 2020).[6]

The benchmark algorithm SNLN is adapted from Si et al. (2023, Algorithm 2) as in Kallus et al. (2022).[7] 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. The well-known KL-divergence chain rule (Cover, 1999) gives

$$\begin{aligned} &D_{\text{KL}}(Q_{X,Y} \,\|\, P_{X,Y}) \\ &= D_{\text{KL}}(Q_X \,\|\, P_X) + D_{\text{KL}}(Q_{Y\,|\,X} \,\|\, P_{Y\,|\,X}). \end{aligned} \tag{10}$$

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`,

[6]A working example on the real-world dataset is given in https://github.com/off-policy-learning/concept-drift-robust-learning.

[7]The implemented code of SNLN benchmark is provided by Kallus et al. (2022).

Metric	Policy	$n = 7500$	$n = 13500$	$n = 16500$	$n = 19500$
$\tilde{\mathcal{V}}_{0.1}^{\min}$	$\widehat{\pi}_{\text{LN}}$	0.2075 ± 0.015	0.2139 ± 0.005	0.2149 ± 0.007	0.2167 ± 0.003
	$\widehat{\pi}_{\text{SNLN}}$	0.1884 ± 0.007	0.2009 ± 0.008	0.2017 ± 0.006	0.2020 ± 0.004

Table 3. Empirical worst case policy reward on the KL-sphere $\tilde{\mathcal{V}}_{\delta}^{\min}$ of policies $\widehat{\pi}_{\text{LN}}, \widehat{\pi}_{\text{SNLN}}$, over 20 seeds.

the same way as in the implementation of Algorithm 2.

Evaluation. For a learned policy $\widehat{\pi}$, we evaluate its performance by the following performance metrics. (i) We use the testing dataset $\mathcal{D}_{\text{test}}$ to estimate the robust policy value $\mathcal{V}_{\delta}^{*}(\widehat{\pi})$ by the empirical robust policy value

$$\bar{\mathcal{V}}_{\delta}(\widehat{\pi}) = -\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} \ell(X_i, Y_i(\pi(X_1)); \boldsymbol{\theta}_{\widehat{\pi}}(X_i)),$$

where the nuisance parameters $\boldsymbol{\theta}_{\widehat{\pi}}(X)$ are found via cubic spline method and employ the Nelder-Mead optimization method using the testing dataset according to policy $\widehat{\pi}$: $\mathcal{D}_{\text{test},\widehat{\pi}} = \{(X_i, Y_i(\widehat{\pi}(X_i)))\}$. (ii) We also perturb the simulated dataset to mimic possible real-world distributional shift. For each testing dataset j containing 10000 data points of the total 100 testing datasets, we generate a new testing dataset, with each reward distribution $(\tilde{Y}_i^{(j)}(1), \tilde{Y}_i^{(j)}(2), \tilde{Y}_i^{(j)}(3))$ randomly sampled on the KL-sphere centered at the reward distribution $(Y_i^{(j)}(1), Y_i^{(j)}(2), Y_i^{(j)}(3))$ of the testing data point with a radius δ. Then we evaluate $\widehat{\pi}$ using

$$\tilde{\mathcal{V}}_{\delta}^{\min}(\widehat{\pi}) := \min_{j \in [100]} \left\{ \frac{1}{10000} \sum_{i=1}^{10000} \tilde{Y}_i^{(j)}\left(\widehat{\pi}(X_i^{(j)})\right) \right\}.$$

This simulates a more realistic scenario where the policy performance is measured by test datasets with concept drifts.

Results. Table 2 and 1 reports the values $\bar{\mathcal{V}}_{\delta}$ of the learned policies $\widehat{\pi}_{\text{LN}}$ and $\widehat{\pi}_{\text{SNLN}}$ on the simulated and the real-wrold dataset, respectively. Table 3 provides the result of $\tilde{\mathcal{V}}_{\delta}^{\min}$. All results are reported with 95% confidence intervals. Table 2 shows that $\widehat{\pi}_{\text{LN}}$ outperforms the benchmark $\widehat{\pi}_{\text{SNLN}}$ consistently, with higher policy values and similar 95% confidence intervals. In Figure 1, $\widehat{\pi}_{\text{LN}}$ continues to show this advantage over $\widehat{\pi}_{\text{SNLN}}$ on the real-world dataset. With a higher δ, the policy values of $\widehat{\pi}_{\text{LN}}, \widehat{\pi}_{\text{SNLN}}$ are smaller, due to a bigger uncertainty set. Table 3 shows that $\widehat{\pi}_{\text{LN}}$ achieves higher worst-case rewards than $\widehat{\pi}_{\text{SNLN}}$ does, in a more realistic setting with concept drift testing datasets. Together, we see that $\widehat{\pi}_{\text{LN}}$ succeeds in finding a better policy under concept drift; while the performance of $\widehat{\pi}_{\text{SNLN}}$ is comprised by its conservative policy learning process, in which it considers joint distributional shifts even though it is given the information that no covariate shifts took place.

The results align with the intuition that Algorithm 2 admits a subset of the uncertainty set that SNLN considers, as explained in Equation (10). Consequently, $\mathcal{V}_{\delta}(\widehat{\pi}_{\text{SNLN}})$ is a lower bound of $\mathcal{V}_{\delta}(\widehat{\pi}_{\text{LN}})$ in theory, and by the results in Table 2, in practice. In real-world applications, knowing the source of the distribution shift effectively shrinks the uncertainty set, thereby yielding less conservative results. Since it is fairly easy to identify covariate shifts (comparing to detecting concept drift), when the decision maker observes none or little covariate shifts and would like to hedge against the risk of concept drift, it is suitable to apply our method which outperforms existing method designed for learning under joint distributional shifts.

One limitation of our methodology (as well as in other DRO works) is the choice of δ. The parameter δ controls the size of the uncertainty set considered and thus controls the degree of robustness in our model — the larger δ, the more robust the output. The empirical performance of the algorithm substantially depends on the selection of δ. A small δ leads to negligible robustification effect and the algorithm would learn an over-aggressive policy; a large δ tends to yield more conservative results. A more detailed discussion can be found in Si et al. (2023).

In Appendix C, 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.

6. Discussion

In this paper, we study the policy learning problem under concept drift, where we propose a doubly-robust policy value estimator that is consistent and asymptotically normal, and then develop a minimax optimal policy learning algorithm, whose regret is $O_p(\kappa(\Pi)n^{-1/2})$ with a matching lower bound. Numerical results show that our learning algorithm outperforms the benchmark algorithm under concept drift. We also note that our results on pure concept drift could be extended to a more general setting where the concept drift adopts the same form as ours, but there is in addition an identifiable covariate shift as in Jin et al. (2022b). The details can be found in Appendix E.

Acknowledgments

This work is generously supported by the ONR grant 13983263 and the NSF grant CCF-2312205. The authors would like to thank Miao Lu and Wenhao Yang for pointing out an error in an earlier draft of this manuscript and their helpful comments. Z. R. is supported by Wharton Analytics. R. Z. is supported by RGC grant ECS-26210324.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

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

Behaghel, L., Crépon, B., and Gurgand, M. Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American economic journal: applied economics*, 6(4):142–174, 2014.

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

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

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

Cai, F., Qu, Z., Xia, L., and Zhou, Z. Multi-action offline policy learning with bayesian optimization. *arXiv preprint arXiv:2003.07545v1*, 2020.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pp. 1–66, 2024.

Chan, C. W., Farias, V. F., Bambos, N., and Escobar, G. J. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations research*, 60(6):1323–1341, 2012.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

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

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

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

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

Duchi, J. C., Hashimoto, T., and Namkoong, H. Distributionally robust losses against mixture covariate shifts. *Under review*, 2(1), 2019.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

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

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

Guo, Y., Liu, H., Yue, Y., and Liu, A. Distributionally robust policy evaluation under general covariate shift in contextual bandits. *arXiv preprint arXiv:2401.11353*, 2024.

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

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

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

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

Jin, Y., Ren, Z., Yang, Z., and Wang, Z. Policy learning" without"overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022a.

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

Jin, Y., Guo, K., and Rothenhäusler, D. Diagnosing the role of observable distribution shift in scientific replications. *arXiv preprint arXiv:2309.01056*, 2023.

Kallus, N. Treatment effect risk: Bounds and inference. *Management Science*, 69(8):4579–4590, 2023.

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

Kallus, N., Mao, X., and Uehara, M. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *arXiv preprint arXiv:1912.12945*, 2019.

Kallus, N., Mao, X., Wang, K., and Zhou, Z. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pp. 10598–10632. PMLR, 2022.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein Jr, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus Jr, J., Gupta, S., et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1): 44–53, 2011.

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

Liu, J., Wang, T., Cui, P., and Namkoong, H. On the need for a language describing distribution shifts: Illustrations on tabular datasets. *arXiv preprint arXiv:2307.05284*, 2023.

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

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12):2346–2363, 2018.

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

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

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

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

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

Natarajan, B. K. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Sahoo, R., Lei, L., and Wager, S. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. Distributionally robust batch contextual bandits. *Management Science*, 2023.

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., and Wager, S. policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232, 2020.

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

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

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

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pp. 11–30. Springer, 2015.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

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

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

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

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. Bounds on the conditional and average treatment effect with unobserved confounding factors. *Annals of statistics*, 50(5):2587, 2022.

Yang, W., Zhang, L., and Zhang, Z. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

Zhan, R., Ren, Z., Athey, S., and Zhou, Z. Policy learning with adaptively collected data. *Management Science*, 2023.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Zhang, Z., Zhan, W., Chen, Y., Du, S. S., and Lee, J. D. Optimal multi-distribution learning. *arXiv preprint arXiv:2312.05134*, 2023.

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

A. 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 minimizer or the maximizer cannot be attained, we project it back to the feasible set. We denote the usual p-norm as $\|\cdot\|_p$. Denote P to be any probability measure defined on the probability space $(\Omega, \sigma(\Omega), P)$ and $\widehat{P}$ to be the empirical distribution of 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 $\|f\|_{L_\infty} = \sup_{x \in \mathcal{X}} |f(x)|$. 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.

B. The ERM Solution

To solve the ERM problem, we follow Geman & Hwang (1982); Yadlowsky et al. (2022); Jin et al. (2022b) and adopt the method of sieves: consider an increasing sequence $\Theta_1 \subset \Theta_2 \subset \cdots$ of spaces of smooth functions, and let $(\widehat{\alpha}_\pi^{(k)}, \widehat{\eta}_\pi^{(k)}) = \operatorname*{argmin}_{\theta \in \Theta_n} \mathbb{E}_n[\ell(X, Y(\pi(X)); \theta)]$. In our case, we consider the following classes of sufficiently smooth functions. For $q_1 = \lceil q \rceil - 1$ and $q_2 = q - q_1$ (where q is the smoothness parameter), define the following function class for η^*:

$$\Theta_c^q(\mathcal{X}) = \left\{ h \in C^{q_1}(\mathcal{X}) : \sup_{\substack{x \in \mathcal{X} \\ \sum_{l=1}^p \gamma_l < q_1}} |D^\gamma h(x)| + \sup_{\substack{x \neq x' \in \mathcal{X} \\ \sum_{l=1}^p \beta_l = q_1}} \frac{|D^\beta h(x) - D^\beta h(x')|}{\|x - x'\|^{q_2}} \leq c \right\},$$

where we denote the derivative $D^d = \sum_{l=1}^p \frac{\partial^{d_l} D}{\partial x_l}$. To ensure the non-negativeness of α_π^* in Proposition 2.5, we define the truncated function class $\Theta_c^q(\mathcal{X}, \epsilon) := \{x \mapsto \max\{h(x), \epsilon\} : h \in \Theta_c^q(\mathcal{X})\}$ for the search of α_π. Consequently, the function class we consider is $\Theta = \Theta_c^q(\mathcal{X}) \times \Theta_c^q(\mathcal{X}, \epsilon)$.

optimizers well, we need the true optimizer $(\alpha_\pi^*, \eta_\pi^*)$ to be sufficiently smooth in x. Convexity and stability are also desirable property of the loss function for learning the optimizers. In the next assumption, we present the regularity condition on the conditional distribution $\mathbb{P}_{Y \mid X}$ to ensure smoothness of optimizers.

Assumption B.1 (Smooth conditional reward distribution)**.** The conditional reward distribution $\mathbb{P}_{Y(a) \mid X=x}$ is smooth in x, i.e. for some $h \in \mathcal{X}$, $\mathbb{P}_{Y(a) \mid X=x+th} = \mathbb{P}_{Y(a) \mid X=x} + t \cdot \mathbb{P}_h$ for some measure $\mathbb{P}_h$ on $\mathcal{Y}$.

Taking practical terms into consideration, Assumption B.1 is reasonable as we assume the conditional distributions of $Y(a)$ are close for similar covariates, for any action $a \in \mathcal{A}$. Jin et al. (2022b, Appendix B.2) presents detailed discussion to justify smoothness of the optimizers under Assumption B.1.

Next, we would like to discuss the regularity conditions of the loss function ℓ in Equation (6) and its conditional expectation $\mathbb{E}[\ell(x, Y; \theta) \mid X = x, A = \pi(x)]$. In particular, we require stability of the loss function and its conditional expectation so that plugging in estimators of the optimizers will not cause large errors, which is a mild condition that can be satisfied under the first-order Taylor expansion condition. Readers can refer to (Van der Vaart, 2000; Jin et al., 2022b) for a more detailed discussion. Later, Lemma D.1 summarises the regularity conditions in formal terms and indicates that in our case, with KL-divergence and the loss function defined as in Equation (6), all the above regularity conditions are satisfied.

C. Experiment Details

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 [3]; the rewards $Y(a)$'s are mutually independent conditioned on X with $Y(a) \mid X \sim N(\beta_a^\top X, \sigma_a^2)$, for $a \in [3]$. We choose β's and σ's to be

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

The training dataset $\mathcal{D}_{\text{train}} = \{(X_i, \pi_0(X_i), Y_i(\pi_0(X_i)))\}_{i=1}^n$ is generated with a given behavior policy π_0 (unknown to policy learning algorithms), which chooses actions conditioned on context x according to the following rules:

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

We also generate 100 testing datasets, each with sample size 10,000. Each testing dataset $\mathcal{D}_{\text{test}}$ consists of i.i.d. draws of data tuple $\{(X_i, Y_i(1), Y_i(2), Y_i(3))\}_{i=1}^n$, and is generated similarly to the procedure described above.

We present the result of the policy estimation experiments in Figure 2, 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 2 shows that as the sample size increases, the estimated policy value by Algorithm 1 is more accurate and stable.

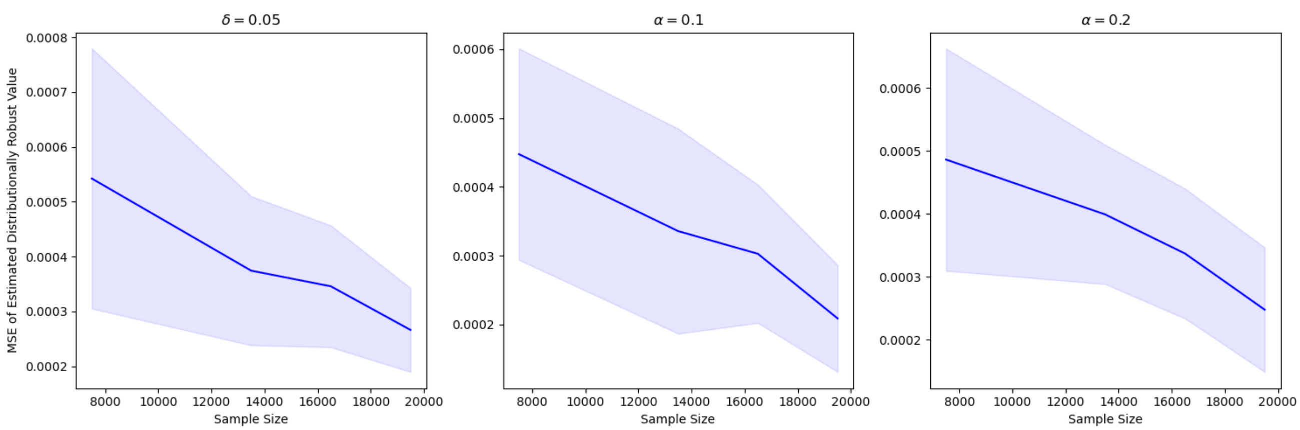

Figure 2. The Mean Square Error (MSE) of the estimated policy value by Algorithm 1. The x-axis is the number of samples used by Algorithm 1, and the y-axis is the mean squared error (MSE) of the policy value estimator.

Computation 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.

D. Deferred Proofs of the Main Results

D.1. Proof of Lemma 2.3

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

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

where the function $f_{\text{KL}}(x) = x \log x$ 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:

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

where

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

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

$$\varphi(\alpha,\eta,x) = \mathbb{E}_x\bigg[-\alpha f^*_{\text{KL}}\Big(-\frac{Y(\pi(x))+\eta}{\alpha}\Big) - \eta - \alpha\delta\bigg],$$

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

$$\inf_{\substack{\mathbb{E}_x[L]=1,\\ \mathbb{E}_x[f_{\text{KL}}(L)]\leq\delta}} \mathbb{E}_x\big[Y(\pi(x))L\big] = -\min_{\alpha\geq 0,\eta\in\mathbb{R}} \mathbb{E}_x\bigg[\alpha\exp\bigg(-\frac{Y(\pi(x))+\eta}{\alpha}-1\bigg)+\eta+\alpha\delta\bigg].$$

The proof is thus completed.

D.2. Proof of Theorem 3.5

For notational 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 $\boldsymbol{\theta}^*_\pi$. The proof of Lemma D.1 can be found in Appendix F.1.

Lemma D.1. *For any policy π, assume that Assumption 3.3 holds. We have the following properties of the optimizer $\boldsymbol{\theta}^*_\pi$.*

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

(2) There exists a constant $\xi > 0$ such that for any x and θ satisfying $\|\theta - \boldsymbol{\theta}^*_\pi(x)\|_2 \leq \xi$,

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

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

(3) There exists a constant $\xi_1 > 0$ such that for any $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_\pi\|_{L_\infty} \leq \xi_1$.

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

for some constant $C_\ell > 0$.

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

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

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

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

Bounding Term (I). Recall that $\boldsymbol{\theta}_\pi^*(x)$ is the minimizer of

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

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

$$
\mathbb{E}\Big[\nabla_\theta\ell\big(x,Y(\pi(x));\boldsymbol{\theta}(x)\big)\,\Big|\,X=x\Big]=0. \tag{13}
$$

For any $i\in\mathcal{D}^{(k)}$, by the unconfoundedness condition in Assumption 2.1, we have

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

where the last step is due to Equation (13). By Assumption 3.4, $\|\widehat{\boldsymbol{\theta}}_\pi^{(k)}-\boldsymbol{\theta}_\pi^*\|_{L_\infty}=o_P(1)$. Therefore, for any $\beta\in(0,1)$, there exists $N\in\mathbb{N}_+$ such that for $n\geq N$, $\|\widehat{\theta}_\pi^{(k)}-\theta_\pi^*\|_{L_\infty}\leq\min(\xi,\xi_1)$. On the event that $\|\widehat{\theta}_\pi^{(k)}(x)-\theta_\pi^*(x)\|_{L_\infty}\leq\min(\xi,\xi_1)$

by part (2) of Lemma D.1 and Jensen's inequality, we have

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

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

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

where the last step is due to part (3) of Lemma D.1. Combining the above results, we have that

$$\text{term (I)}=O_P\big(n^{-1/2}\cdot\|\widehat{\boldsymbol{\theta}}_\pi^{(k)}-\boldsymbol{\theta}_\pi^*\|_{L_2(P_X)}+\|\widehat{\boldsymbol{\theta}}_\pi^{(k)}-\boldsymbol{\theta}_\pi^*\|_{L_2(P_X)}^2\big)=o_P(n^{-1/2}),$$

where the last step is due to Assumption 3.4.

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

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

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 $N_1\in\mathbb{N}_+$ such that for $n\ge N_1$,

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\theta}}_\pi^{(k)}-\boldsymbol{\theta}^*\|_{L_\infty}\le\min(\underline{\alpha},\bar\eta)/2\big)\ge1-\beta.$$

On the event $\|\widehat{\boldsymbol{\theta}}_\pi^{(k)}-\boldsymbol{\theta}^*\|_{L_\infty}\le\min(\underline{\alpha},\bar\eta)/2$,

$$\big|\widehat{G}_\pi^{(k)}(x,y)\big|=\big|\ell(x,y;\widehat{\boldsymbol{\theta}}_\pi^{(k)})\big|\le\bar\alpha\exp\Big(\frac{\bar y+\bar\eta}{\underline{\alpha}}-1\Big)+\bar\eta+\bar\alpha\delta=:L_g.$$

Next, for any $i \in \mathcal{D}^{(k)}$,

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

where the first step is by the unconfoundedness assumption and the second step is due to the fact that $\bar{g}_\pi^{(k)}$ is the conditional expectation of $\widehat{G}_\pi^{(k)}$.

On the event $\{\|\widehat{\boldsymbol{\theta}}_\pi^{(k)}-\boldsymbol{\theta}_\pi^*\|_{L_\infty}\le\min(\underline{\alpha},\bar{\eta})\}$. By Chebyshev's inequality, for any $t>0$,

$$\begin{aligned}
&\mathbb{P}\Bigg(\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\bigg(\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i \mid X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i \mid X_i)}\bigg)\cdot\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)-\bar{g}_\pi^{(k)}(X_i)\big)\bigg|\ge t\,\bigg|\,\mathcal{D}^{(-k)}\Bigg)\\
\le&\,\frac{1}{|\mathcal{D}^{(k)}|t^2}\mathrm{Var}\Bigg(\bigg[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i \mid X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i \mid X_i)}\bigg]\cdot\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)-\bar{g}_\pi^{(k)}(X_i)\big)\,\bigg|\,\mathcal{D}^{(-k)}\Bigg)\\
\le&\,\frac{1}{|\mathcal{D}^{(k)}|t^2}\mathbb{E}\Bigg[\bigg[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\widehat{\pi}_0^{(k)}(A_i \mid X_i)}-\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i \mid X_i)}\bigg]^2\cdot\big(\widehat{G}_\pi^{(k)}(X_i,Y_i)-\bar{g}_\pi^{(k)}(X_i)\big)^2\,\bigg|\,\mathcal{D}^{(-k)}\Bigg]\\
\le&\,\frac{4L_g^2}{|\mathcal{D}^{(k)}|\varepsilon^4t^2}\|\widehat{\pi}_0^{(k)}-\pi_0\|^2_{L_2(P_{X\,\mid\,T=\pi(X)})}.
\end{aligned}$$

The above inequality along with a union bound implies that

$$\text{term (III)} = O_P\big(\|\widehat{\pi}_0^{(k)}-\pi_0\|_{L_2(P_{X\,\mid\,A=\pi(X)})}/\sqrt{|\mathcal{D}^{(k)}|}\big)=o_P(n^{-1/2}),$$

where the last step is by the consistency of $\widehat{\pi}_0^{(k)}$ assumed in Assumption 3.4.

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

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

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

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

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

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

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

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

Above, step (i) follows from Jensen's inequality and step (ii) from part (3) of Lemma D.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

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

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

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

D.3. Proof of Theorem 4.3

By Assumption 3.3, taking $\pi(x) \equiv a$ for any $a \in [M]$, there exist constants $\bar{\alpha}_a, \underline{\alpha}_a, \bar{\eta}_a$ such that

$$
0 < \underline{\alpha}_a \leq \boldsymbol{\alpha}_a^*(x) \leq \bar{\alpha}_a, \quad |\boldsymbol{\eta}_a(x)| \leq \bar{\eta}_a, \quad \forall x \in \mathcal{X}.
$$

Letting $\underline{\alpha} = \min_{a \in [M]} \underline{\alpha}_a$, $\bar{\alpha} = \max_{a \in [M]} \bar{\alpha}_a$, $\bar{\eta} = \max_{a \in [M]} \bar{\eta}_a$, it follows that

$$
0 < \underline{\alpha} \leq \boldsymbol{\alpha}_a^*(x) \leq \bar{\alpha}, \quad |\boldsymbol{\eta}_a(x)| \leq \bar{\eta}, \quad \forall x \in \mathcal{X}, \forall a \in [M]. \tag{14}
$$

For any $a \in [M]$, if we take $\pi(x) \equiv a$, then by (1) of Lemma D.1,

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

By (2) of Lemma D.1, for any $a \in [M]$, there exists a constant $\xi_a > 0$ such that for any $\|\theta - \boldsymbol{\theta}_a^*(x)\|_2 \leq \xi_a$

$$
|\ell(x, y; \theta) - \ell(x, y; \boldsymbol{\theta}_a^*(x)) - \nabla_\theta \ell(x, y; \boldsymbol{\theta}_a^*(x))^\top (\theta - \boldsymbol{\theta}_a^*(x))| \leq \bar{\ell}_a(x, y) \|\theta - \boldsymbol{\theta}_a^*(x)\|_2^2,
$$

for some function $\bar{\ell}_a(x, y) \leq L_a$ for some constant L_a. Similarly, we shall take $\xi = \min_{a \in [M]} \xi_a$, $\bar{\ell}(x, y) = \max_a \bar{\ell}_a(x, y)$, and $L = \sum_{a \in [M]} L_a$.

By (3) of Lemma D.1, for any $a \in [M]$, there exists a constant $\xi_{1,a} > 0$ such that for any $\|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_{L_\infty} \leq \xi_{1,a}$,

$$
\big\|\ell(X, Y(a); \boldsymbol{\theta}(X)) - \ell(X, Y(a); \boldsymbol{\theta}_a^*(X))\big\|_{L_2(P_{X, Y(a) \,|\, A=a})} \leq C_{\ell,a} \|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_{L_2(P_{X \,|\, A=a})}.
$$

Taking $\xi_1 = \min_{a \in [M]} \xi_{1,a}$ and $C_\ell = \sum_{a \in [M]} C_{\ell,a}$, the above inequality holds for any $a \in [M]$ and any $\|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_{L_\infty} \leq \xi_1$.

D.3.1. Regret decomposition

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

$$\begin{aligned}
\mathcal{R}_\delta(\widehat{\pi}_{\mathrm{LN}}) &= \mathcal{V}_\delta(\pi^*) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}}) \\
&= \mathcal{V}_\delta(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}}) \\
&\le \mathcal{V}_\delta(\pi^*) - \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi^*) + \widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\widehat{\pi}_{\mathrm{LN}}) - \mathcal{V}_\delta(\widehat{\pi}_{\mathrm{LN}}) \\
&\le 2 \sup_{\pi\in\Pi} \big|\widehat{\mathcal{V}}_\delta^{\mathrm{LN}}(\pi) - \mathcal{V}_\delta(\pi)\big|,
\end{aligned} \tag{15}$$

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

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

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

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

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

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

We shall show that the first term above is $O_P(n^{-1/2})$ and the second term is $o_P(n^{-1/2})$. In the following, we refer to the two terms as the effective term and the negligible term, respectively. The following lemma is essential for establishing the uniform convergence results.

Lemma D.2. *Suppose h is a function of $(x, a, y, \pi(x))$. Given a set of data $\{z_i = (x_i, a_i, y_i)\}_{i=1}^n$, suppose that $|h(z_i, \pi(x_i))| \le c_i(z_i)$. Then the Rademacher complexity*

$$\mathbb{E}_\epsilon\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i h\big(x_i, a_i, y_i, \pi(x_i)\big)\Big|\Big] \le \frac{\sqrt{\sum_{i=1}^n c_i(z_i)^2}}{n}\cdot\big(32 + 4\kappa(\Pi)\big),$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{\pm 1\}$ are i.i.d. Rademacher random variables and $\mathbb{E}_\epsilon$ means the expectation over ϵ.

D.3.2. The effective term

Denote $Z_i = (X_i, A_i, Y_i)$ and take

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

Under the unconfoundedness assumption in Assumption 2.1, $\mathbb{E}[h(Z_i, \pi(X_i))] = 0$. By Equation (14), we have

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

Meanwhile, we have write

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

Next, we define

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

Consider two arbitrary data sets $\{z_i\}_{i=1}^n$ and $\{z_i'\}_{i=1}^n$. We can check that for any $\pi\in\Pi$ and any $j\in[n]$,

$$\begin{aligned}
&\big|f(z_1,\dots,z_j,\dots,z_n;\pi)\big|-\sup_{\pi'\in\Pi}\big|f(z_1,\dots,z_j',\dots,z_n;\pi')\big|\\
\le&\big|f(z_1,\dots,z_j,\dots,z_n;\pi)\big|-\big|f(z_1,\dots,z_j',\dots,z_n;\pi)\big|\\
\le&\sup_{\pi\in\Pi}\big|f(z_1,\dots,z_j,\dots,z_n;\pi)-f(z_1,\dots,z_j',\dots,z_n;\pi)\big|\\
=&\sup_{\pi\in\Pi}\frac{1}{n}\big|h(z_j;\pi)-h(z_j';\pi)\big|\le C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)/n.
\end{aligned}\tag{16}$$

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 (16), we have that

$$\sup_{\pi\in\Pi}\big|f(z_1,\dots,z_j,\dots,z_n;\pi)\big|-\sup_{\pi\in\Pi}\big|f(z_1,\dots,z_j',\dots,z_n;\pi)\big|\le C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)/n.$$

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\pi\in\Pi}\left|\frac{1}{n}h\big(Z_i,\pi(X_i)\big)\right|-\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}h\big(Z_i,\pi(X_i)\big)\right|\right]\ge t\right)\\
=&\mathbb{P}\left(\sup_{\pi\in\Pi}\big|f\big(\{Z_i\}_{i\in[n]};\pi\big)\big|-\mathbb{E}\left[\sup_{\pi\in\Pi}\big|f\big(\{Z_i\}_{i\in[n]};\pi\big)\big|\right]\ge t\right)\le e^{-\frac{2nt^2}{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)^2}}.
\end{aligned}$$

Take $t=C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)\sqrt{\frac{1}{2n}\log\big(\frac{1}{\beta}\big)}$. Then with probability at least $1-\beta$,

$$\sup_{\pi\in\Pi}\left|\frac{1}{n}h\big(Z_i,\pi(X_i)\big)\right|<\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}h\big(Z_i,\pi(X_i)\big)\right|\right]+C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)\sqrt{\frac{1}{2n}\log\big(\frac{1}{\beta}\big)}.$$

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 let $\epsilon_i\overset{\text{i.i.d.}}{\sim}\text{Unif}(\{\pm1\})$. Then

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}h(Z_i,\pi(X_i))-\mathbb{E}\Big[h(Z_i,\pi(X_i))\Big]\right|\right]\\
=&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}h(Z_i,\pi(X_i))-\mathbb{E}_{Z'}\left[\frac{1}{n}\sum_{i\in[n]}h(Z_i',\pi(X_i'))\right]\right|\right]\\
\overset{\text{(i)}}{\le}&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}h(Z_i,\pi(X_i))-\frac{1}{n}\sum_{i\in[n]}h(Z_i',\pi(X_i'))\right|\right]\\
\overset{\text{(ii)}}{=}&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i\big(h(Z_i,\pi(X_i))-h(Z_i',\pi(X_i'))\big)\right|\right],\\
\le&2\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i h(Z_i,\pi(X_i))\right|\right]\\
=&2\mathbb{E}\left[\mathbb{E}_{\epsilon}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i h(Z_i,\pi(X_i))\right|\right]\right],
\end{aligned}\tag{17}$$

step (i) is by Jensen's inequality and step (ii) is because of the symmetry of (Z_i, Z_i').

Applying Lemma D.2,

$$\mathbb{E}_\epsilon\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i h(Z_i,\pi(X_i))\right|\right] \le \frac{2C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{n}}(32+4\kappa(\Pi)).$$

Combining the above, for any $\beta\in(0,1)$, we have with probability at least $1-\beta$,

$$\sup_{\pi\in\Pi}\big|\tilde{\mathcal{V}}_\delta(\pi)-\mathcal{V}_\delta(\pi)\big| \le \frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{n}}\big(64+8\kappa(\Pi)+\sqrt{\log(1/\beta)}\big). \tag{18}$$

D.3.3. Bounding the negligible term

We now proceed to the negligible 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. To ease the presentation, we shall write $\mathbb{E}_k$ and $\mathbb{P}_k$ as the expectation and probability conditioned on $\mathcal{D}^{(-k)}$, respectively.

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

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

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

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

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

$$|\ell(x, y, ; \widehat{\boldsymbol{\theta}}_a(x))| \leq 2\bar{\alpha}\exp\left(\frac{2\bar{y} + 4\bar{\eta}}{\underline{\alpha}} - 1\right) + 2\bar{\eta} + 2\bar{\alpha}\delta.$$

Letting $C_1(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon) = 4\bar{\alpha}\exp\left(\frac{2\bar{y}+4\bar{\eta}}{\underline{\alpha}} - 1\right) + 4\bar{\eta} + 4\bar{\alpha}\delta)/\varepsilon^2$, We can then check that

$$\begin{aligned}
|h_1(Z_i; \pi(X_i))| &\leq 2C_1(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon) \cdot \big|\widehat{\pi}_0(\pi(X_i) \,|\, X_i) - \pi_0(\pi(X_i) \,|\, X_i)\big| \\
&\leq 2C_1(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon) \cdot \max_{a \in [M]} \big|\widehat{\pi}_0(a \,|\, X_i) - \pi_0(a \,|\, X_i)\big| =: c_1(X_i).
\end{aligned}$$

The upper bound is a constant conditional on X_i's and $\mathcal{D}^{(-k)}$. We now apply the bounded difference inequality conditional on $X = \{X_i\}_{i \in [n]}$:

$$\begin{aligned}
&\mathbb{P}_k\left(\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(-k)}|}\sum_{i \in \mathcal{D}^{(k)}} h_1(Z_i, \pi(X_i))\bigg| - \mathbb{E}_k\left[\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} h_1(Z_i, \pi(X_i))\bigg| \,\bigg|\, X\right] \geq t \,\bigg|\, X\right) \\
&\leq \exp\left(-\frac{2|\mathcal{D}^{(k)}|^2 t^2}{\sum_{i \in \mathcal{D}^{(k)}} c_1(X_i)^2}\right).
\end{aligned}$$

Taking $t = \sqrt{\sum_{i \in \mathcal{D}^{(k)}} c_1(X_i)^2 \log(1/\beta)}/|\mathcal{D}^{(k)}|$, we have with probability at least $1 - \beta$,

$$\begin{aligned}
\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(-k)}|}\sum_{i \in \mathcal{D}^{(k)}} h_1(Z_i, \pi(X_i))\bigg| &\leq \mathbb{E}_k\left[\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} h_1(Z_i, \pi(X_i))\bigg| \,\bigg|\, X\right] \\
&\quad + \frac{\sqrt{\sum_{i \in \mathcal{D}^{(k)}} c_1(X_i)^2}}{|\mathcal{D}^{(k)}|}\sqrt{\log(1/\beta)}.
\end{aligned}$$

For each $i \in \mathcal{D}^{(k)}$, we take A_i' and Y_i' as i.i.d. copies of A_i and Y_i conditional on X_i, respectively. By a similar symmetrization argument as in the proof for the effective term, we have

$$\begin{aligned}
&\mathbb{E}_k\left[\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} h_1\big(Z_i, \pi(X_i)\big)\bigg| \,\bigg|\, X\right] \\
=&\mathbb{E}_k\left[\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} h_1\big(X_i, A_i, Y_i, \pi(X_i)\big) - \mathbb{E}_{A', Z'}\Big[\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} h\big(X_i, A_i', Y_i', \pi(X_i)\big)\Big]\bigg| \,\bigg|\, X\right] \\
\leq&\mathbb{E}_k\left[\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} h_1(X_i, A_i, Y_i, \pi(X_i)) - \frac{1}{|\mathcal{D}^{(-k)}|}\sum_{i \in \mathcal{D}^{(k)}} h(X_i, A_i', Y_i', \pi(X_i))\bigg| \,\bigg|\, X\right] \\
\leq&\mathbb{E}_k\left[\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} \epsilon_i\Big(h_1(X_i, A_i, Y_i, \pi(X_i)) - h(X_i, A_i', Y_i', \pi(X_i))\Big)\bigg| \,\bigg|\, X\right] \\
\leq&2\mathbb{E}_k\Big[\sup_{\pi \in \Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i \in \mathcal{D}^{(k)}} \epsilon_i h_1(X_i, A_i, Y_i, \pi(X_i))\bigg| \,\bigg|\, X\Big].
\end{aligned}$$

Applying Lemma D.2 with $c_i = c_1(X_i)$, we have that

$$\mathbb{E}_\epsilon\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\epsilon_i h_1(X_i,A_i,Y_i,\pi(X_i))\Big|\,\Big|\,X\Big] \le \frac{2\sqrt{\sum_{i\in\mathcal{D}^{(k)}} c_1(X_i)^2}}{|\mathcal{D}^{(k)}|}(32+4\kappa(\Pi)).$$

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

$$\mathbb{P}_k\bigg(\sup_{\pi\in\Pi}\big|K_1(\pi)\big| \ge \frac{\sqrt{\sum_{i\in\mathcal{D}^{(k)}} c_1(X_i)^2}}{|\mathcal{D}^{(k)}|}\big(64+8\kappa(\Pi)+\sqrt{\log(1/\beta)}\big)\,\Big|\,X\bigg) \le \beta.$$

Since $\big|\widehat{\pi}_0(a\,|\,X)-\pi_0(a\,|\,X)\big|^2 \le 1$,

$$\begin{aligned}
&\mathbb{P}_k\bigg(\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\max_{a\in[M]}\big(\widehat{\pi}_0(a\,|\,X)-\pi_0(a\,|\,X)\big)^2 - \sum_{a\in[M]}\mathbb{E}\big[(\widehat{\pi}_0(a\,|\,X)-\pi_0(a\,|\,X))^2\big] \ge t\bigg)\\
\le{}&\mathbb{P}_k\bigg(\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\sum_{a\in[M]}\big(\widehat{\pi}_0(a\,|\,X)-\pi_0(a\,|\,X)\big)^2 - \sum_{a\in[M]}\mathbb{E}\big[(\widehat{\pi}_0(a\,|\,X)-\pi_0(a\,|\,X))^2\big] \ge t\bigg)\\
\le{}&\sum_{a\in[M]}\mathbb{P}_k\bigg(\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\big(\widehat{\pi}_0(a\,|\,X)-\pi_0(a\,|\,X)\big)^2 - \mathbb{E}\big[(\widehat{\pi}_0(a\,|\,X)-\pi_0(a\,|\,X))^2\big] \ge t\bigg)\\
\le{}&M\exp\big(-2|\mathcal{D}^{(k)}|t^2\big).
\end{aligned}$$

Taking a union bound, with probability at least $1-3\beta$, we have that

$$\begin{aligned}
\sup_{\pi\in\Pi}\big|K_1(\pi)\big| \le{}& \frac{2C_1(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{\sqrt{|\mathcal{D}^{(k)}|}}\Big(20+4\kappa(\Pi)+\sqrt{2\log(1/\beta)}\Big)\\
&\times\bigg(\sum_{a\in[M]}\|\widehat{\pi}_0-\pi_0\|_{L_2(P_{X\,|\,A=a})} + \Big(\frac{1}{2n}\log(M/\beta)\Big)^{1/4}\bigg).
\end{aligned}$$

Since $\sum_{a\in[M]}\|\widehat{\pi}_0-\pi_0\|_{L_2(P_{X\,|\,A=a})} = o_P(1)$, there exists $N_1' \ge N_1$ such that when $n \ge N_1'$, with probability at least $1-\beta/(4K)$,

$$\sup_{\pi\in\Pi}\big|K_1(\pi)\big| \le \frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{4\sqrt{n}}. \tag{19}$$

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

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

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

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

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

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

By Assumption 3.4, there exists $N_2'\in\mathbb{N}_+$ such that when $n\geq N_2'$, with probability at least $1-\beta/(4K)$,

$$\sup_{\pi\in\Pi}|K_2(\pi)|\leq\frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{4\sqrt{n}}. \tag{20}$$

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

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

For any $\pi\in\Pi$,

$$\begin{aligned}
&\mathbb{E}_k\big[h_3(Z_i,\pi(X_i))\,|\,X_i\big]\\
=&\,\mathbb{E}_k\Big[\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\big(\widehat{G}^{(k)}_{A_i}(X_i,Y_i(\pi(X_i)))-G_{A_i}(X_i,Y_i(\pi(X_i)))\big)\,\big|\,X_i\Big]\\
=&\,\mathbb{E}_k\Big[\widehat{G}^{(k)}_{\pi(X_i)}(X_i,Y_i(\pi(X_i)))-G_{\pi(X_i)}(X_i,Y_i(\pi(X_i)))\,|\,X_i\Big]\\
=&\,\mathbb{E}_k\Big[\ell\big(X_i,Y_i(\pi(X_i));\boldsymbol{\theta}^{(k)}_{\pi(X_i)}(X_i)\big)-\ell(X_i,Y_i;\boldsymbol{\theta}^*_{\pi(X_i)}(X_i))\\
&\qquad-\nabla\ell(X_i,Y_i(\pi(X_i));\boldsymbol{\theta}^*_{\pi(X_i)}(X_i))^\top\big(\widehat{\boldsymbol{\theta}}^{(k)}_{\pi(X_i)}(X_i)-\boldsymbol{\theta}^*_{\pi(X_i)}(X_i)\big)\,\big|\,X_i\Big],
\end{aligned}$$

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

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

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

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

As a result,

$$\sup_{\pi\in\Pi}\left|\mathbb{E}_k[K_3(\pi)\,|\,X]\right| \le \sup_{\pi\in\Pi}\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\mathbb{E}_k\big[h_3(Z_i,\pi(X_i))\,|\,X_i\big]$$
$$\le \frac{L}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\sum_{a\in[M]}\big\|\widehat{\boldsymbol{\theta}}_a(X_i)-\boldsymbol{\theta}^*_a(X_i)\big\|_2^2.$$

On the same event,

$$\begin{aligned}
\big|h_3(Z_i,\pi(X_i))\big| &= \left|\frac{\mathbb{1}\{A_i=\pi(X_i)\}}{\pi_0(A_i\,|\,X_i)}\cdot\Big\{\ell\big(X_i,Y_i(\pi(X_i));\boldsymbol{\theta}^{(k)}_{\pi(X_i)}(X_i)\big)-\ell(X_i,Y_i;\boldsymbol{\theta}^*_{\pi(X_i)}(X_i))\Big\}\right|\\
&\le \frac{1}{\varepsilon}\Big|\nabla\ell\big(X_i,Y_i(\pi(X_i));\tilde{\boldsymbol{\theta}}_{\pi(X_i)}(X_i)\big)^\top(\boldsymbol{\theta}^{(k)}_{\pi(X_i)}(X_i)-\boldsymbol{\theta}^*_{\pi(X_i)}(X_i))\Big|\\
&\le \frac{1}{\varepsilon}\big\|\nabla\ell\big(X_i,Y_i(\pi(X_i));\tilde{\boldsymbol{\theta}}_{\pi(X_i)}(X_i)\big)\big\|_2\big\|\widehat{\boldsymbol{\theta}}^{(k)}_{\pi(X_i)}(X_i)-\boldsymbol{\theta}^*_{\pi(X_i)}(X_i)\big\|_2\\
&\le C_2(\bar\alpha,\underline{\alpha},\bar\eta,\delta,\varepsilon)\max_{a\in[M]}\big\|\widehat{\boldsymbol{\theta}}^{(k)}_a(X_i)-\boldsymbol{\theta}^*_a(X_i)\big\|_2,
\end{aligned}$$

where $C_2(\bar\alpha,\underline{\alpha},\bar\eta,\delta,\varepsilon)=(1+(\bar y+\bar\eta)/\underline{\alpha})e^{(\bar y+\bar\eta)/\underline{\alpha}-1}+\delta+1$ is a constant. Let $\bar h_3(Z_i,\pi(X_i))=h_3(Z_i,\pi(X_i))-\mathbb{E}_k[h_3(Z_i,\pi(X_i))\,|\,X_i]$, and we have that

$$|\bar h_3(Z_i,\pi(X_i))|\le 2C_2(\bar\alpha,\underline{\alpha},\bar\eta,\delta,\varepsilon)\max_{a\in[M]}\big\|\widehat{\boldsymbol{\theta}}^{(k)}_a(X_i)-\boldsymbol{\theta}^*_a(X_i)\big\|_2.$$

Next, we apply the bounded difference theorem conditional on X_i's:

$$\mathbb{P}_k\left(\sup_{\pi\in\Pi}\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\bar h_3(Z_i,\pi(X_i))\right|-\mathbb{E}_k\left[\sup_{\pi\in\Pi}\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\bar h_3(Z_i,\pi(X_i))\right|\,\middle|\,X\right]\ge t\,\middle|\,X\right)$$
$$\le\exp\left(-\frac{|\mathcal{D}^{(k)}|^2t^2}{2C_2(\bar\alpha,\underline{\alpha},\bar\eta,\delta,\varepsilon)^2\sum_{i\in\mathcal{D}^{(k)}}\max_{a\in[M]}\big\|\widehat{\boldsymbol{\theta}}^{(k)}_a(X_i)-\boldsymbol{\theta}^*_a(X_i)\big\|_2^2}\right),$$

for any $t>0$. Taking $t=C_2(\bar\alpha,\underline{\alpha},\bar\eta,\delta,\varepsilon)\sqrt{2\sum_{i\in\mathcal{D}^{(k)}}\max_{a\in[M]}\big\|\widehat{\boldsymbol{\theta}}^{(k)}_a(X_i)-\boldsymbol{\theta}^*_a(X_i)\big\|_2^2}/|\mathcal{D}^{(k)}|$, we have with probability at least $1-\beta$ that

$$\sup_{\pi\in\Pi}\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\bar h_3(Z_i,\pi(X_i))\right|\le\mathbb{E}_k\left[\sup_{\pi\in\Pi}\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\bar h_3(Z_i,\pi(X_i))\right|\,\middle|\,X\right]$$
$$+\frac{C_2(\bar\alpha,\underline{\alpha},\bar\eta,\delta,\varepsilon)}{|\mathcal{D}^{(k)}|}\sqrt{2\sum_{i\in\mathcal{D}^{(k)}}\sum_{a\in[M]}\big\|\widehat{\boldsymbol{\theta}}^{(k)}_a(X_i)-\boldsymbol{\theta}^*_a(X_i)\big\|_2^2}.$$

For the expectation term,the same symmetrization argument as in the proof for $K_1(\pi)$ leads to

$$\mathbb{E}_k\left[\sup_{\pi\in\Pi}\left|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\bar h_3(Z_i,\pi(X_i))\right|\,\middle|\,X\right]\le 2\mathbb{E}\left[\sup_{\pi\in\Pi}\Big|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\Big|\epsilon_i\bar h_3(Z_i,\pi(X_i))\Big|\,\Big|\,X\right].$$

Then by Lemma D.2, we have

$$\mathbb{E}_\epsilon\left[\sup_{\pi\in\Pi}\Big|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\Big|\epsilon_i\bar h_3(Z_i,\pi(X_i))\Big|\right]$$
$$\le\frac{2C_2(\bar\alpha,\underline{\alpha},\bar\eta,\delta,\varepsilon)}{|\mathcal{D}^{(k)}|}(32+\kappa(\Pi))\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\sum_{a\in[M]}\|\widehat{\boldsymbol{\theta}}^{(k)}_a(X_i)-\boldsymbol{\theta}^*_a(X_i)\|_2^2}.$$

By Hoeffding's inequality, we have that

$$\mathbb{P}_k\bigg(\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\|\widehat{\boldsymbol{\theta}}_a^{(k)}(X_i)-\boldsymbol{\theta}_a^*(X_i)\|_2^2-\|\boldsymbol{\theta}_a^{(k)}-\boldsymbol{\theta}_a\|_{L_2(P_X)}^2\geq\|\boldsymbol{\theta}_a^{(k)}-\boldsymbol{\theta}_a\|_{L_\infty}^2\sqrt{\frac{1}{2n}\log\Big(\frac{1}{\beta}\Big)}\bigg)\leq\beta.$$

Taking a union bound, with probability at least $1-3\beta$, we have that

$$\begin{aligned}&\sup_{\pi\in\Pi}\big|K_3(\pi)\big|\\ \leq&\frac{C_2(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)(130+4\kappa(\Pi))}{\sqrt{|\mathcal{D}^{(k)}|}}\Big(\sum_{a\in[M]}\|\widehat{\boldsymbol{\theta}}_a^{(k)}-\boldsymbol{\theta}_a\|_{L_2(P_X)}+\sqrt{M(\bar{\alpha}+\bar{\eta})}\Big(\frac{1}{|\mathcal{D}^{(k)}|}\log\Big(\frac{M}{\beta}\Big)\Big)^{1/4}\Big)\\ &\quad+L\Big(\sum_{a\in[M]}\|\widehat{\boldsymbol{\theta}}_a^{(k)}-\boldsymbol{\theta}_a\|_{L_2(P_X)}^2+\frac{M\|\widehat{\boldsymbol{\theta}}_a-\boldsymbol{\theta}_a\|_{L_\infty}\sqrt{\log(M/\beta)}}{\sqrt{|\mathcal{D}^{(k)}|}}\Big).\end{aligned}$$

By Assumption 3.4, $\|\widehat{\boldsymbol{\theta}}_a^{(k)}-\boldsymbol{\theta}_a\|_{L_2(P_X)}=o_P(n^{-1/4})$ and $\|\widehat{\boldsymbol{\theta}}_a^{(k)}-\boldsymbol{\theta}_a\|_{L_\infty}=o_P(1)$, so there exists $N_3'\geq N_3$ such that when $n\geq N_3'$, with probability at least $1-\beta/(4K)$,

$$\sup_{\pi\in\Pi}\big|K_3(\pi)\big|\leq\frac{C_0(\bar{\alpha},\underline{\alpha},\bar{\eta},\delta,\varepsilon)}{4\sqrt{n}}.\tag{21}$$

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

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

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,

$$\mathbb{E}_k\big[h_4(Z_i,\pi(X_i))\big]=0.$$

Due to the overlap condition, we further have that

$$|h_4(Z_i,\pi(X_i))|\leq\frac{2}{\varepsilon}\big|\widehat{g}_{\pi(X_i)}^{(k)}(X_i)-g_{\pi(X_i)}(X_i)\big|\leq\frac{2}{\varepsilon}\max_{a\in[M]}\big|\widehat{g}_a^{(k)}(X_i)-g_a(X_i)\big|.$$

As before, we apply the bounded difference theorem conditional on X_i's and the symmetrization argument to obtain

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{\pi\in\Pi}\big|K_4(\pi)\big|-2\mathbb{E}_k\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\epsilon_i h_4(Z_i,\pi(X_i))\Big|\,\Big|\,X\Big]\geq t\,\Big|\,X\bigg)\\ \leq&\,\mathbb{P}\bigg(\sup_{\pi\in\Pi}\big|K_4(\pi)\big|-\mathbb{E}_k\Big[\sup_{\pi\in\Pi}\big|K_4(\pi)\big|\,\big|\,X\Big]\geq t\,|\,X)\\ \leq&\exp\bigg(-\frac{\varepsilon^2|\mathcal{D}^{(k)}|^2t^2}{2\sum_{i\in\mathcal{D}^{(k)}}\max_{a\in[M]}(\widehat{g}_a^{(k)}(X_i)-g_a(X_i))^2}\bigg).\end{aligned}$$

We now apply Lemma D.2:

$$\mathbb{E}_\epsilon\Bigg[\sup_{\pi\in\Pi}\bigg|\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}\epsilon_i h_4(Z_i,\pi(X_i))\bigg|\,\bigg|\,X\Bigg]\leq\frac{\sqrt{\sum_{i\in\mathcal{D}^{(k)}}\max_{a\in[M]}\big(\widehat{g}_a(X_i)-g_a(X_i)\big)^2}}{|\mathcal{D}^{(k)}|\varepsilon}(64+8\kappa(\Pi)).$$

By Assumption 3.4, there exists $N_4\in\mathbb{N}_+$, such that when $n\geq N_4$, with probability at least $1-\beta$,

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

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

$$\big|\ell(x, y; \widehat{\boldsymbol{\theta}}_a(x))\big| \le 2C_1(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon).$$

On the same event, by Hoeffding's inequality, we have that

$$\mathbb{P}_k\bigg(\frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}}(\widehat{g}_a(X_i) - g_a(X_i))^2 - \|\widehat{g}_a^{(k)} - g_a\|_{L_2(P_X)}^2 \ge t\bigg) \le \exp\Big(-\frac{t^2|\mathcal{D}^{(k)}|}{8C_1(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)^2}\Big).$$

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

$$\begin{aligned}\max_{\pi\in\Pi}\big|K_4(\pi)\big| \le &\frac{1}{\varepsilon\sqrt{|\mathcal{D}^{(k)}|}}\Big(128 + 16\kappa(\Pi) + \sqrt{2\log(1/\beta)}\Big)\\ &\times\Big(\sum_{a\in[M]}\|\widehat{g}_a^{(k)} - g_a^*\|_{L_2(P_X)} + 2M\sqrt{C_1(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)}(\log(M/\beta)/n)^{1/4}\Big).\end{aligned}$$

By Assumption 3.4, $\sum_{a\in[M]}\|\widehat{g}_a^{(k)} - g_a^*\|_{L_2(P_X)} = o_P(1)$, so there exists $N_4' \ge N_4$ such that when $n \ge N_4'$, with probability at least $1 - \beta/(4K)$,

$$\sup_{\pi\in\Pi}\big|K_4(\pi)\big| \le \frac{C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)}{4\sqrt{n}}. \tag{22}$$

Combining (18)-(22) and taking a union bound over $k \in [K]$, when $n \ge \max(N_1, N_2, N_3, N_4)$ we have that with probability at least $1 - \beta$,

$$\sup_{\pi\in\Pi}\big|\widehat{\mathcal{V}}_\delta^{\text{LN}}(\pi) - \tilde{\mathcal{V}}_\delta(\pi)\big| \le \frac{C_0(\bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta, \varepsilon)}{\sqrt{n}}.$$

We have thus completed the proof of Theorem 4.3.

D.4. Proof of Theorem 4.6

We first state somes results from (Si et al., 2023) that will be used in the proof. For any $p, q \in [0, 1]$, define

$$D(p\,\|\,q) = p\log\Big(\frac{p}{q}\Big) + (1-p)\log\Big(\frac{1-p}{1-q}\Big), \text{ and } g_\delta(q) = \inf_{p:D_{\text{KL}}(p\,\|\,q)\le\delta} p,$$

Lemma D.3 (Adapted from Lemma A17 of Si et al. (2023))**.** *For $\delta \le 0.2$, $g_\delta(q)$ is differentiable and $g_\delta'(q) \ge 1/2$ for $q \in [0.4, 0.6]$.*

Note that our definition of $g_\delta(q)$ is slightly different from that in (Si et al., 2023), so we include the proof of Lemma D.3 in Appendix F.4 for completeness.

For notational simplicity, we use d to denote the Natarajan dimension of the policy class Π. By the definition of Natarajan dimension, there exists a set of d data points $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ shattered by Π: there exist two functions $f_{-1}, f_1 : \{x_1, \dots, x_d\} \mapsto [M]$ such that $f_{-1}(x_j) \ne f_1(x_j)$ for any $j \in [d]$ and for any $\sigma \in \{-1.1\}^d$, there exists $\pi \in \Pi$, such that $\pi(x_j) = f_{\sigma_j}(x_j)$ for all $j \in [d]$.

Next, we construct a class of distributions indexed by $\sigma \in \{\pm 1\}^d$ that are "hard instances" for the learning problem. Fix any $\sigma \in \{\pm 1\}^d$, we construct distribution P_σ as follows. First, the covariate are drawn uniformly from $\{x_1, \dots, x_d\}$, i.e.,

$$X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\big(\{x_1, \dots, x_d\}\big).$$

Given X_i, the action A_i is chosen according to the behavior policy π_0, where for any $j \in [d]$,

$$\pi_0(f_1(x_j)\,|\,x_j) = \pi_0(f_{-1}(x_j)\,|\,x_j) = \frac{\varepsilon}{2}, \text{ and } \pi_0(a\,|\,x_j) = \frac{1-\varepsilon}{K-2} \text{ for all } a \ne f_1(x_j), f_{-1}(x_j).$$

The potential outcomes are generated as follows:

$$Y_i(f_1(x_j)) \,|\, X_i = x_j \sim \bar{y} \cdot \text{Bern}\Big(\frac{1+\sigma_j\Delta}{2}\Big),\ Y_i(f_{-1}(x_j)) \,|\, X_i = x_j \sim \bar{y} \cdot \text{Bern}\Big(\frac{1-\sigma_j\Delta}{2}\Big),$$
$$\text{and } Y_i(a) = \bar{y} \cdot \text{Bern}(1/4) \text{ for all } a \neq f_1(x_j), f_{-1}(x_j),$$

where $\Delta \in (0, 0.1)$ is some constant to be determined later. Note that the distribution of (X_i, A_i) does not depend on σ. By construction, it is clear that the data-generating process satisfies Assumption 2.1. For any $p \in \{(1+\Delta)/2, (1-\Delta)/2, 1/4\}$, $\log(1/(1-p)) > \delta$. Therefore, the data-generating process also satisfies Assumption 2.4. As for Assumption 3.3, it suffices to check the Bernoulli distributions with parameters $(1+\Delta)/2, (1-\Delta)/2, 1/4$, and can be verified. Since $n \geq d^2$, we can obtain $\widehat{\boldsymbol{\theta}}$, $\widehat{g}$, and $\widehat{\pi}_0$ that converges at rate $O_P(n^{-1/4})$ (by stratifying on X), thereby satisfying Assumption 3.4.

We now proceed to establish the lower bound. For any policy learning algorithm that returns $\widehat{\pi}$, the worst-case regret is lower bounded by the average regret over the class of hard instances we have constructed above:

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}[\mathcal{R}(\widehat{\pi})] \geq \frac{1}{2^d} \sum_{\sigma \in \{\pm 1\}^d} \mathbb{E}_{P_\sigma^n}[\mathcal{R}_\delta(\widehat{\pi})].$$

We now focus on the right-hand side above. Fix $\sigma \in \{\pm 1\}^d$. Recall that $\mathcal{R}_\delta(\widehat{\pi}) = \mathcal{V}_\delta(\pi^*) - \mathcal{V}_\delta(\widehat{\pi})$. For the optimal policy value, there is

$$\begin{aligned}
\mathcal{V}_\delta(\pi^*) &= \max_{\pi \in \Pi} \mathbb{E}_{P_{\sigma,X}}\Big[\inf_{Q_{Y\,|\,X} \in \mathcal{P}(P_{\sigma,Y\,|\,X},\delta)} \mathbb{E}_{Q_{Y\,|\,X}}\big[Y(\pi(X)) \,|\, X\big]\Big] \\
&\overset{\text{(i)}}{=} \max_{\pi \in \Pi} \max_{\boldsymbol{\alpha},\boldsymbol{\eta}} \mathbb{E}_{P_\sigma}\Big[-\boldsymbol{\alpha}(X)\exp\Big(-\frac{Y(\pi(X))+\boldsymbol{\eta}(X)}{\boldsymbol{\alpha}(X)} - 1\Big) - \boldsymbol{\eta}(X) - \boldsymbol{\alpha}(X)\delta\Big] \\
&= \max_{\boldsymbol{\alpha},\boldsymbol{\eta}} \max_{\pi \in \Pi} \mathbb{E}_{P_\sigma}\Big[-\boldsymbol{\alpha}(X)\exp\Big(-\frac{Y(\pi(X))+\boldsymbol{\eta}(X)}{\boldsymbol{\alpha}(X)} - 1\Big) - \boldsymbol{\eta}(X) - \boldsymbol{\alpha}(X)\delta\Big],
\end{aligned} \tag{23}$$

where the step (i) follows from the duality result in Proposition 2.3. We now take a closer look at the expectation above: by the construction of P_σ,

$$\begin{aligned}
&\mathbb{E}_{P_\sigma}\Big[-\boldsymbol{\alpha}(X)\exp\Big(-\frac{Y(\pi(X))+\boldsymbol{\eta}(X)}{\boldsymbol{\alpha}(X)} - 1\Big) - \boldsymbol{\eta}(X) - \boldsymbol{\alpha}(X)\delta\Big] \\
=\ & \frac{1}{d}\sum_{j=1}^d \mathbb{E}_{P_\sigma}\Big[-\boldsymbol{\alpha}(x_j)\exp\Big(-\frac{Y(\pi(x_j))+\boldsymbol{\eta}(x_j)}{\boldsymbol{\alpha}(x_j)} - 1\Big) - \boldsymbol{\eta}(x_j) - \boldsymbol{\alpha}(x_j)\delta \,\Big|\, X = x_j\Big] \\
=\ & \frac{1}{d}\sum_{j=1}^d -\boldsymbol{\alpha}(x_j)\exp\Big(-\frac{\boldsymbol{\eta}(x_j)}{\boldsymbol{\alpha}(x_j)} - 1\Big) \cdot \mathbb{E}_P\Big[\exp\Big(-\frac{Y(\pi(x_j))}{\boldsymbol{\alpha}(x_j)}\Big) \,\Big|\, X = x_j\Big] - \boldsymbol{\eta}(x_j) - \boldsymbol{\alpha}(x_j)\delta.
\end{aligned}$$

Letting $p_j = \mathbb{P}(Y(\pi(x_j)) = 1 \,|\, X = x_j)$, we have

$$\mathbb{E}_P\Big[\exp\Big(-\frac{Y(\pi(x_j))}{\boldsymbol{\alpha}(x_j)}\Big) \,\Big|\, X = x_j\Big] = p_j \exp(-1/\boldsymbol{\alpha}(x_j)) + 1 - p_j,$$

which is decreasing in p_j and is minimized when $\pi(x_j) = f_{\sigma_j}(x_j)$. By construction, such a policy π is in Π. As a result,

$$\begin{aligned}
(23) &= \max_{\boldsymbol{\alpha},\boldsymbol{\eta}} \frac{1}{d}\sum_{j=1}^d \mathbb{E}_{P_\sigma}\Big[-\boldsymbol{\alpha}(x_j)\exp\Big(-\frac{Y(f_{\sigma_j}(x_j))+\boldsymbol{\eta}(x_j)}{\boldsymbol{\alpha}(x_j)} - 1\Big) - \boldsymbol{\eta}(x_j) - \boldsymbol{\alpha}(x_j)\delta \,\Big|\, X = x_j\Big] \\
&= \frac{1}{d}\sum_{j=1}^d \max_{\alpha,\eta} \mathbb{E}_{P_\sigma}\Big[-\alpha\exp\Big(-\frac{Y(f_{\sigma_j}(x_j))+\eta}{\alpha} - 1\Big) - \eta - \alpha\delta \,\Big|\, X = x_j\Big] \\
&= \frac{1}{d}\sum_{j=1}^d \inf_{Q_{Y\,|\,X} \in \mathcal{P}(P_{\sigma,Y\,|\,X=x_j},\delta)} \mathbb{E}_{Q_{Y\,|\,X}}\big[Y(f_{\sigma_j}(x_j)) \,|\, X = x_j\big] = g\Big(\frac{1+\Delta}{2}\Big).
\end{aligned}$$

The last step is because $Y(f_{\sigma_j}(x_j)) \,|\, X = x_j \sim \text{Bern}((1+\Delta)/2)$. Similarly, for $\mathcal{V}(\widehat{\pi})$, we have

$$\begin{aligned}
\mathcal{V}_\delta(\widehat{\pi}) &= \mathbb{E}_{P_{\sigma,X}}\left[\inf_{Q_{Y\,|\,X}\in\mathcal{P}(P_{\sigma,Y\,|\,X},\delta)} \mathbb{E}_{Q_{Y\,|\,X}}\big[Y\big(\widehat{\pi}(X)\big)\,|\,X\big]\right] \\
&= \frac{1}{d}\sum_{j=1}^d \inf_{Q_{Y\,|\,X}\in\mathcal{P}(P_{\sigma,Y\,|\,X=x_j},\delta)} \mathbb{E}_{Q_{Y\,|\,X}}\big[Y(\widehat{\pi}(x_j))\,|\,X=x_j\big] \\
&= \frac{1}{d}\sum_{j=1}^d \mathbb{1}\big\{\widehat{\pi}(x_j)=f_{\sigma_j}(x_j)\big\} g\Big(\frac{1+\Delta}{2}\Big) + \mathbb{1}\big\{\widehat{\pi}(x_j)=f_{-\sigma_j}(x_j)\big\} g\Big(\frac{1-\Delta}{2}\Big) \\
&\qquad + \mathbb{1}\big\{\widehat{\pi}(x_j)\neq f_{\sigma_j}(x_j), f_{-\sigma_j}(x_j)\big\} g(1/4).
\end{aligned}$$

Combining the calculation above, we have

$$\begin{aligned}
\mathcal{R}(\widehat{\pi}) &= \frac{1}{d}\sum_{j=1}^d \mathbb{1}\{\widehat{\pi}(x_j)=f_{-\sigma_j}(x_j)\}\cdot\Big\{g\Big(\frac{1+\Delta}{2}\Big)-g\Big(\frac{1-\Delta}{2}\Big)\Big\} \\
&\qquad + \mathbb{1}\{\widehat{\pi}(x_j)\neq f_{\sigma_j}(x_j), f_{-\sigma_j}(x_j)\}\cdot\Big\{g\Big(\frac{1+\Delta}{2}\Big)-g(1/4)\Big\} \\
&\overset{\text{(i)}}{\geq} \frac{1}{d}\sum_{j=1}^d +\mathbb{1}\{\widehat{\pi}(x_j)\neq f_{\sigma_j}(x_j)\}\cdot\Big\{g\Big(\frac{1+\Delta}{2}\Big)-g(1/4)\Big\} \\
&\overset{\text{(ii)}}{\geq} \frac{1}{d}\sum_{j=1}^d \mathbb{1}\{\widehat{\pi}(x_j)\neq f_{\sigma_j}(x_j)\}\cdot g'(\xi)\Delta \overset{\text{(iii)}}{\geq} \frac{\Delta}{2d}\sum_{j=1}^d \mathbb{1}\{\widehat{\pi}(x_j)\neq f_{\sigma_j}(x_j)\},
\end{aligned}$$

where step (i) uses that g is non-decreasing (c.f. Cauchois et al. (2024, Proposition 1)); in step (ii), $\xi\in((1-\Delta)/2,(1+\Delta)/2)$, and step (iii) follows from Lemma D.3.

Next, we denote $\sigma[j]$ to be the vector σ with the j-th element flipped. Then, we have

$$\begin{aligned}
\frac{1}{2^d}\sum_{\sigma\in\{\pm1\}^d}\mathbb{E}_{P_\sigma^n}[\mathcal{R}(\widehat{\pi})] &\geq \frac{1}{2^d}\sum_{\sigma\in\{\pm1\}^d}\mathbb{E}_{P_\sigma^n}\left[\frac{\Delta}{2d}\sum_{j=1}^d\mathbb{1}\{\widehat{\pi}(x_j)\neq f_{\sigma_j}(x_j)\}\right] \\
&= \frac{\Delta}{d2^{d+1}}\sum_{j=1}^d\sum_{\sigma:\sigma_j=1}\Big\{\mathbb{P}_{P_\sigma^n}\big(\widehat{\pi}(x_j)\neq f_1(x_j)\big)+\mathbb{P}_{P_{\sigma[j]}^n}\big(\widehat{\pi}(x_j)\neq f_{-1}(x_j)\big)\Big\} \\
&\geq \frac{\Delta}{d2^{d+1}}\sum_{j=1}^d\sum_{\sigma:\sigma_j=1}\mathbb{P}_{P_\sigma^n}\big(\widehat{\pi}(x_j)\neq f_1(x_j)\big)+\mathbb{P}_{P_{\sigma[j]}^n}\big(\widehat{\pi}(x_j)= f_1(x_j)\big) \\
&\geq \frac{\Delta}{d2^{d+1}}\sum_{j=1}^d\sum_{\sigma:\sigma_j=1}\big(1-D_{\text{TV}}(P_\sigma^n,P_{\sigma[j]}^n)\big),
\end{aligned} \tag{24}$$

where the last step follows from the definition of the TV distance. By Pinsker's inequality, there is

$$\begin{aligned}
D_{\text{TV}}^2\big(P_\sigma^n,P_{\sigma[j]}^n\big) &\leq \frac{1}{2}D_{\text{KL}}\big(P_\sigma^n\,\|\,P_{\sigma[j]}^n\big) \\
&= \frac{1}{2}\sum_{i=1}^n\mathbb{E}_{P_\sigma}\left[\log\Big(\frac{\mathrm{d}P_\sigma}{\mathrm{d}P_{\sigma[j]}}(X_i,A_i,Y_i)\Big)\right] \\
&= \frac{1}{2}\sum_{i=1}^n\mathbb{E}_{P_\sigma}\left[\mathbb{1}\big\{X_i=x_j,A_i=f_{\pm1}(x_j)\big\}\cdot\Delta\log\Big(\frac{1+\Delta}{1-\Delta}\Big)\right] \\
&\leq \frac{3n\varepsilon}{2d}\Delta^2,
\end{aligned}$$

where the last step follows from $x\log(\frac{1+x}{1-x}) \leq 3x^2$, for $x \in (0, 1/3)$. Take $\Delta = \frac{1}{15}\sqrt{\frac{d}{n\varepsilon}}$ — this is possible since $n \geq d^2$ and $d \geq 4/(9\varepsilon)$ and then

$$(24) \geq \sqrt{\frac{d}{n}}\frac{d2^{d-1}}{15d2^{d+2}} = \frac{1}{120} \times \sqrt{\frac{d}{n\varepsilon}}.$$

E. Generalization to Identifiable Covariate Distribution Shift

We now extend our methodology to handle situations where shift in X and $Y \mid X$ distributions are both present. We note that, in most practical cases, the decision maker has access to the covariates in the target environment, making the shift covariate distribution identifiable and estimable — it is therefore unnecessary to guard against the worst-case shift.

Method. Formally, suppose we have access to a training dataset $\mathcal{D}$ collected in an environment P, and aims at learning a policy that behaves well in the environment Q. Here, we assume that $Q_{Y \mid X} \in \mathcal{P}(P_{Y \mid X}, \delta)$ for some given radius $\delta > 0$ but do not impose any constraints on X distribution shift except that Q_X is absolutely continuous with respect to P_X and that the likelihood ratio is bounded. Letting $r(x) = \frac{dQ_X}{dP_X}(x)$, we can estimate r with X from P and Q using standard tools (by directly estimating the density ratio or by means of classification algorithms). For any policy π, the distributional robust policy value can be written as

$$\mathcal{V}_\delta(\pi) := \mathbb{E}_{Q_X}\Big[\inf_{Q_{Y \mid X} \in \mathcal{P}(P_{Y \mid X}, \delta)} \mathbb{E}_{Q_{Y \mid X}}\big[Y(\pi(X)) \mid X\big]\Big]. \tag{25}$$

Note that the inner expectation of (25) is the same as in (1). By Lemma 2.3, there is

$$\begin{aligned}(25) &= \mathbb{E}_{Q_X \times P_{Y \mid X}}\left[\boldsymbol{\alpha}^*_\pi(X)\exp\Big(-\frac{Y(\pi(X)) + \boldsymbol{\eta}^*_\pi(X)}{\boldsymbol{\alpha}^*_\pi(X)}\Big) + \boldsymbol{\eta}^*_\pi(X) + \boldsymbol{\alpha}^*_\pi(X)\delta\right] \\ &= \mathbb{E}_{Q_X \times P_{Y \mid X}}\left[\frac{\mathbb{1}\{A = \pi(X)\}}{\pi_0(A \mid X)}\left(\boldsymbol{\alpha}^*_\pi(X)\exp\Big(-\frac{Y(A) + \boldsymbol{\eta}^*_\pi(X)}{\boldsymbol{\alpha}^*_\pi(X)}\Big) + \boldsymbol{\eta}^*_\pi(X) + \boldsymbol{\alpha}^*_\pi(X)\delta\right)\right] \\ &= \mathbb{E}_P\left[r(X)\frac{\mathbb{1}\{A = \pi(X)\}}{\pi_0(A \mid X)}\left(\boldsymbol{\alpha}^*_\pi(X)\exp\Big(-\frac{Y(A) + \boldsymbol{\eta}^*_\pi(X)}{\boldsymbol{\alpha}^*_\pi(X)}\Big) + \boldsymbol{\eta}^*_\pi(X) + \boldsymbol{\alpha}^*_\pi(X)\delta\right)\right].\end{aligned}$$

The above expression ensures that the robust policy value is identifiable with the data accessible to the decision maker. With this representation, subsequent policy evaluation and learning are similar to the pure concept drift case, and we provide the adaptation below.

In addition to $\mathcal{D}$, we let $\tilde{\mathcal{D}} = \{X_i\}_{i=1}^m$ denote the covariates from environment Q, i.e., $X_i \overset{\text{i.i.d.}}{\sim} Q_X$. Assume that $\lim_{m,n\to\infty} m/n = \gamma$. As before, we adopt a K-fold cross-fitting scheme, where we split both $\mathcal{D}$ and $\tilde{\mathcal{D}}$ into K non-overlapping equally-sized folds. For $k \in [K]$, we use $\mathcal{D}^{(k+1)}$ and $\tilde{\mathcal{D}}^{(k+1)}$ to obtain $\widehat{\pi}_0^{(k)}$, $\widehat{r}^{(k)}$, and $(\widehat{\boldsymbol{\alpha}}_\pi^{(k)}, \widehat{\boldsymbol{\eta}}_\pi^{(k)})$ as estimates of π_0, r, and $(\boldsymbol{\alpha}^*_\pi, \boldsymbol{\eta}^*_\pi)$, respectively; we then use $\mathcal{D}^{(k+2)}$ and $\tilde{\mathcal{D}}^{(k+2)}$ to obtain $\widehat{g}_\pi^{(k)}$ as an estimate for $\bar{g}_\pi^{(k)}$. The estimator of the k-th fold is

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

and the final robust policy value estimator is $\widehat{\mathcal{V}}_\delta(\pi) = -\frac{1}{K}\sum_{k=1}^K \widehat{\mathcal{V}}_\delta^{(k)}(\pi)$. We then obtain the learned policy via

$$\widehat{\pi}_{\text{LN}} = \operatorname*{argmax}_{\pi \in \Pi} \widehat{\mathcal{V}}(\pi),$$

where we apply the same computational trick as in the pure concept shift case.

Theoretical Guarantees. We now extend the theoretical guarantees to the general case. Since the proof is similar to the pure concept drift case, the proof sketch is provided. As a prerequisite, we modify Assumption 3.4 to be:

Assumption E.1. For any policy π, assume that for each $k \in [K]$, the estimators $\widehat{\pi}_0^{(k)}$, $\widehat{r}^{(k)}$, $\widehat{g}_\pi^{(k)}$, and the empirical risk optimizer $\widehat{\boldsymbol{\theta}}_\pi^{(k)}$ satisfy

$$\|\widehat{r}^{(k)}/\widehat{\pi}_0^{(k)} - r/\pi_0\|_{L_2(P_{X\,|\,A=\pi(X)})} = o_P(n^{-\gamma_1}),\ \|\widehat{g}_\pi^{(k)} - \bar{g}_\pi^{(k)}\|_{L_2(P_{X\,|\,A=\pi(X)})} = o_P(n^{-\gamma_2}),$$
$$\|\widehat{\boldsymbol{\theta}}_\pi^{(k)} - \boldsymbol{\theta}_\pi^*\|_{L_2(P_{X\,|\,A=\pi(X)})} = o_P(n^{-\frac{1}{4}}),\ \|\widehat{\boldsymbol{\theta}}_\pi^{(k)} - \boldsymbol{\theta}_\pi^*\|_{L_\infty} = o_P(1),$$

for some $\gamma_1, \gamma_2 \geq 0$ and $\gamma_1 + \gamma_2 \geq \frac{1}{2}$.

Theorems E.2 and E.3 establish the asymptotic normality of the policy value estimator and the regret upper bound.

Theorem E.2. *Suppose Assumptions 2.1, 2.4, 3.3, and E.1 hold. Additionally assume that $dQ_X/dP_X \leq C$, a.s., for some constants $C > 0$, and that $\lim_{m,n\to\infty} m/n = \gamma$. For any policy $\pi : \mathcal{X} \mapsto [M]$, we have $\sqrt{n}\cdot\big(\widehat{\mathcal{V}}_\delta(\pi) - \mathcal{V}_\delta(\pi)\big) \xrightarrow{\text{d}} N(0, \sigma_\pi^2)$, where*

$$\sigma_\pi^2 = \mathrm{Var}\bigg(\frac{r(X)\mathbb{1}\{A = \pi(X)\}}{\pi_0(A\,|\,X)} \cdot \big(G_\pi(X,Y) - g_\pi(X)\big)\bigg) + \gamma \cdot \mathrm{Var}(g_\pi(X)).$$

Theorem E.3. *Suppose Assumptions 2.1, 2.4, 3.3, E.1 hold. Additionally assume that $dQ_X/dP_X \leq C$, a.s., for some constants $C > 0$, and that $\lim_{m,n\to\infty} m/n = \gamma$. 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*

$$\mathcal{R}_\delta(\widehat{\pi}_{\mathrm{LN}}) \leq \frac{C(\kappa(\Pi) + \sqrt{\log(1/\beta)}),}{\sqrt{n}}$$

where $C > 0$ is a constant independent of n and Π.

Proof Sketch. Recall that

$$G_\pi(x,y) = \boldsymbol{\alpha}_\pi^*(x)\exp\Big(-\frac{y + \boldsymbol{\eta}_\pi^*(x)}{\boldsymbol{\alpha}_\pi^*(x)}\Big) + \boldsymbol{\eta}_\pi^*(x) + \boldsymbol{\alpha}_\pi^*(x)\delta \ \text{ and } \ g_\pi(x) = \mathbb{E}\big[G_\pi(X, Y(\pi(X)))\,|\,X = x\big]$$

It can be checked that

$$\begin{aligned}(25) &= \mathbb{E}_P\Big[\frac{r(X)\mathbb{1}\{A = \pi(X)\}}{\pi_0(A\,|\,X)} G_\pi(X,Y)\Big] \\ &= \mathbb{E}_P\Big[\frac{r(X)\mathbb{1}\{A = \pi(X)\}}{\pi_0(A\,|\,X)} \cdot \big(G_\pi(X,Y) - g_\pi(X)\big)\Big] + \mathbb{E}_{Q_X}\big[g_\pi(X)\big].\end{aligned}$$

So if we define $\bar{\mathcal{V}}_\delta(\pi) = -\frac{1}{K}\sum_{k=1}^K \bar{\mathcal{V}}_\delta^{(k)}(\pi)$, with

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

we have by the central limit theorem that

$$\sqrt{n}\Big(\bar{\mathcal{V}}_\delta(\pi) - \mathcal{V}_\delta(\pi)\Big) \xrightarrow{\text{d}} N(0, \sigma_\pi^2).$$

As in the proof of Theorem 3.5, we can decompose the difference between $\bar{\mathcal{V}}_\delta^{(k)}(\pi)$ and $\widehat{\mathcal{V}}_\delta^{(k)}(\pi)$ as follows:

$$\begin{aligned}\widehat{\mathcal{V}}_\delta^{(k)}(\pi) - \bar{\mathcal{V}}_\delta^{(k)}(\pi) =& \frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}} \frac{r(X_i)\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(X_i)}\big(\widehat{G}_\pi^{(k)}(X_i,Y_i) - G_\pi(X_i,Y_i)\big) \\ &+ \frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}} \mathbb{1}\{A_i = \pi(X_i)\}\Big(\frac{\widehat{r}(X_i)}{\widehat{\pi}_0(X_i)} - \frac{r(X_i)}{\pi_0(X_i)}\Big)\big(\widehat{G}_\pi^{(k)}(X_i,Y_i) - \bar{g}_\pi^{(k)}(X_i)\big) \\ &+ \frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}} \mathbb{1}\{A_i = \pi(X_i)\}\Big(\frac{\widehat{r}(X_i)}{\widehat{\pi}_0(X_i)} - \frac{r(X_i)}{\pi_0(X_i)}\Big)\big(\bar{g}_\pi^{(k)}(X_i) - \widehat{g}_\pi(X_i)\big) \\ &- \frac{1}{|\mathcal{D}^{(k)}|}\sum_{i\in\mathcal{D}^{(k)}} \frac{r(X_i)\mathbb{1}\{A_i = \pi(X_i)\}}{\pi_0(X_i)}(\widehat{g}_\pi^{(k)}(X_i) - g_\pi(X_i)) + \frac{1}{|\tilde{\mathcal{D}}^{(k)}|}\sum_{i\in\tilde{\mathcal{D}}^{(k)}}(\widehat{g}_\pi^{(k)}(X_i) - g_\pi(X_i)).\end{aligned} \tag{26}$$

Following almost the same steps in the proof of Theorem 3.5, we can show that

$$
\begin{aligned}
(26) =& O_P\Big(\|\widehat{\boldsymbol{\theta}}_\pi^{(k)} - \boldsymbol{\theta}_\pi^*\|^2_{L_2(P_{X\,|\,A=\pi(X)})}\Big) + O_P\Big(\Big\|\frac{\widehat{r}^{(k)}}{\widehat{\pi}^{(k)}} - \frac{r}{\pi}\Big\|_{L_2(P_{X\,|\,A=\pi(X)})} \cdot n^{-1/2}\Big) \\
&+ O_P\Big(\Big\|\frac{\widehat{r}^{(k)}}{\widehat{\pi}_0^{(k)}} - \frac{r}{\pi}\Big\|_{L_2(P_{X\,|\,A=\pi(X)})} \cdot \|\widehat{g}_\pi^{(k)} - g_\pi\|_{L_2(P_{X\,|\,A=\pi(X)})}\Big) + O_P\big(\|\widehat{g}_\pi^{(k)} - g_\pi\|_{L_2(P_{X\,|\,A=\pi(X)})} \cdot n^{-1/2}\big) \\
=& o_P(n^{-1/2}).
\end{aligned}
$$

Taking a union bound over $k \in [K]$, we can conclude $\widehat{\mathcal{V}}_\delta(\pi) = \bar{\mathcal{V}}_\delta(\pi) + o_P(n^{-1/2})$, and therefore $\widehat{\mathcal{V}}_\delta(\pi)$ is asymptotically normal. The proof of Theorem E.3 follows similarly.

F. Proof of technical lemmas

F.1. Proof of Lemma D.1

Proof of (1). Given θ, recall that our loss function is

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

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 problem implies

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

Meanwhile, we can compute the gradient of $\ell(x,y;\theta)$ as

$$
\begin{aligned}
&\frac{\partial}{\partial\alpha}\ell(x,y;\theta) = \Big(1 + \frac{y+\eta}{\alpha}\Big)\cdot \exp\Big(-\frac{y+\eta}{\alpha} - 1\Big) + \delta, \\
&\frac{\partial}{\partial_\eta}\ell(x,y;\theta) = 1 - \exp\Big(-\frac{y+\eta}{\alpha} - 1\Big).
\end{aligned} \tag{27}
$$

For any a such that $|a - \boldsymbol{\alpha}_\pi^*(x)| \le \boldsymbol{\alpha}_\pi^*(x)$, we have

$$
\Big|\frac{\partial}{\partial\alpha}\ell(x,y;(a,\boldsymbol{\eta}_\pi^*(x)))\Big| \le \Big(1 + \frac{2(\bar{y}+\bar{\eta})}{\underline{\alpha}}\Big)\cdot \exp\Big(\frac{2(\bar{y}+\bar{\eta})}{\underline{\alpha}} - 1\Big) + \delta < \infty.
$$

By the mean value theorem and the dominated convergence theorem, we can change the order of expectation and taking limits and therefore

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

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

$$
\Big|\frac{\partial}{\partial_\eta}\ell(x,y;(\boldsymbol{\alpha}_\pi^*(x),\eta))\Big| \le \max\bigg\{\Big|\frac{\partial}{\partial_\eta}\ell(x,y;(\boldsymbol{\alpha}_\pi^*(x),\boldsymbol{\eta}_\pi^*(x)+1))\Big|, \Big|\frac{\partial}{\partial_\eta}\ell(x,y;(\boldsymbol{\alpha}_\pi^*(x),\boldsymbol{\eta}_\pi^*(x)-1))\Big|\bigg\},
$$

with the right-hand side being integrable under $P_{Y\,|\,X}$. Agian by the mean-value theorem and the dominated convergence theorem,

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

We have thus completed the proof part (1) of Lemma D.1.

Proof of (2). We now compute the Hessian of $\ell(x, y; \theta)$:

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

By the Taylor expansion,

$$\ell(x,y;\theta) - \ell(x,y;\boldsymbol{\theta}_\pi^*(x)) = \nabla\ell(x,y;\boldsymbol{\theta}_\pi^*(x))^\top(\theta - \theta_\pi^*(x)) + \frac{1}{2}(\theta - \boldsymbol{\theta}_\pi^*(x))^\top \nabla^2\ell(x,y;\tilde{\theta})(\theta - \boldsymbol{\theta}_\pi^*(x)),$$
$$\Rightarrow \big|\ell(x,y;\theta) - \ell(x,y;\boldsymbol{\theta}_\pi^*(x)) - \nabla\ell(x,y;\boldsymbol{\theta}_\pi^*(x))^\top(\theta - \boldsymbol{\theta}_\pi^*(x))\big|$$
$$\le \frac{1}{2}\Big(\frac{(y+\tilde{\eta})^2}{\tilde{\alpha}^3} + \frac{1}{\tilde{\alpha}}\Big)\exp\Big(-\frac{y+\tilde{\eta}}{\tilde{\alpha}}-1\Big)\|\theta - \boldsymbol{\theta}_\pi^*(x)\|_2^2,$$

where $\tilde{\theta} = t\theta + (1-t)\boldsymbol{\theta}_\pi^*(x)$ for some $t \in [0,1]$ and the last step is because

$$\big\|\nabla^2\ell(x,y;\tilde{\theta})\big\|_{\text{op}} \le \Big(\frac{(y+\tilde{\eta})^2}{\tilde{\alpha}^3} + \frac{1}{\tilde{\alpha}}\Big)\exp\Big(-\frac{y+\tilde{\eta}}{\tilde{\alpha}}-1\Big)$$

Let $\xi = \min(\underline{\alpha}, \bar{\eta})/2$. For any θ such that $\|\theta - \boldsymbol{\theta}_\pi^*(x)\|_2 \le \xi$, we also have $|\tilde{\alpha} - \boldsymbol{\alpha}_\pi^*(x)| \le \xi$ and $|\tilde{\eta} - \boldsymbol{\eta}_\pi^*(x)| \le \xi$. Then

$$\frac{1}{2}\Big(\frac{(y+\tilde{\eta})^2}{\tilde{\alpha}^3} + \frac{1}{\tilde{\alpha}}\Big)\exp\Big(-\frac{y+\tilde{\eta}}{\tilde{\alpha}}-1\Big) \le \Big(\frac{8\bar{y}^2 + 8\bar{\eta}^2}{\underline{\alpha}^3} + \frac{2}{\underline{\alpha}}\Big)\cdot\exp\Big(\frac{2\bar{y}+4\bar{\eta}}{\underline{\alpha}}-1\Big).$$

Letting the right-hand side be $\bar{\ell}(x,y)$, we have thus completed the proof of (2).

Proof of (3). By the Taylor expansion,

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

where $\tilde{\boldsymbol{\theta}} = t\boldsymbol{\theta}(x) + (1-t)\boldsymbol{\theta}_\pi^*$ for some $t \in [0,1]$. Let $\xi_1 = \min(\underline{\alpha}, \bar{\eta})/2$. When $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\pi^*\|_{L_\infty} \le \xi_1$, we have $|\tilde{\alpha} - \boldsymbol{\alpha}_\pi^*(x)| \le \xi_1$ and $|\tilde{\eta} - \boldsymbol{\eta}_\pi^*(x)| \le \xi_1$. Plugging the expressions of the gradient in Equation (27), we have

$$\big[\ell(x,y;\boldsymbol{\theta}(x)) - \ell(x,y;\boldsymbol{\theta}_\pi^*(x))\big]^2 = \big[\nabla\ell(x,y;\tilde{\boldsymbol{\theta}}(x))^\top(\boldsymbol{\theta}(x) - \boldsymbol{\theta}_\pi^*(x))\big]^2$$
$$\le \Bigg\{\Bigg[\Big(1 + \frac{y+\tilde{\boldsymbol{\eta}}(x)}{\tilde{\boldsymbol{\alpha}}(x)}\Big)\exp\Big(-\frac{y+\tilde{\boldsymbol{\eta}}(x)}{\tilde{\boldsymbol{\alpha}}(x)}-1\Big) + \delta\Bigg]^2 + \Bigg[1 - \exp\Big(-\frac{y+\tilde{\boldsymbol{\eta}}(x)}{\tilde{\boldsymbol{\alpha}}(x)}-1\Big)\Bigg]^2\Bigg\}\cdot\big\|\boldsymbol{\theta}(x) - \boldsymbol{\theta}_\pi^*(x)\big\|_2^2$$
$$\le C(\bar{y}, \bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta)\cdot\big\|\boldsymbol{\theta}(x) - \boldsymbol{\theta}_\pi^*(x)\big\|_2^2,$$

where $C(\bar{y}, \bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta)$ is a function of $(\bar{y}, \bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta)$. Taking the expectation over $P_{X,Y \mid A=\pi(X)}$, we have

$$\big\|\ell(X,Y;\boldsymbol{\theta}(X)) - \ell(X,Y;\boldsymbol{\theta}_\pi^*(X))\big\|_{L_2(P_{X,Y \mid A=\pi(X)})} \le C(\bar{y}, \bar{\alpha}, \underline{\alpha}, \bar{\eta}, \delta)\cdot\big\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\big\|_{L_2(P_{X \mid A=\pi(X)})},$$

completing the proof of (3).

F.2. Proof of Lemma D.2

We first introduce the ℓ_2 distance on the policy space Π, as well as the corresponding covering number.

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

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

$$\ell_2\big(\pi_1, \pi_2; \{z_1, \ldots, z_n\}\big) = \sqrt{\frac{\sum_{i=1}^n \big(h(z_i, \pi_1(x_i)) - h(z_i; \pi_2(x_i))\big)^2}{4\sum_{i=1}^n c_i(z_i)^2}}.$$

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

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 $j = 0, 1, \ldots, J$, let S_j be the set of policies that 2^{-j}-covers Π and satisfies $|S_j| = N_2(2^{-j}, \Pi; \{Z_1, \ldots, Z_n\})$. Specially, $S_0 = \{\bar{\pi}\}$, with $\bar{\pi}$ is an arbitrary policy in Π — this is a valid choice since for any $\pi \in \Pi$,

$$\ell_2(\pi, \bar{\pi}; \{z_1, \ldots, z_n\}) = \sqrt{\frac{\sum_{i=1}^n \big(h(z_i, \pi(x_i)) - h(z_i, \bar{\pi}(x_i))\big)^2}{4 \sum_{i=1}^n c_i(z_i)^2}} \leq 1.$$

We shall let $\Lambda = 2\sqrt{\sum_{i=1}^n c_i(z_i)^2}$ to denote the normalization factor. The approximation operators are defined in a backward manner: for any $\pi \in \Pi$,

(1) define $A_J[\pi] = \operatorname*{argmin}_{\pi' \in S_J} \ell_2\big(\pi, \pi'; \{z_1, \ldots, z_n\}\big)$;

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

$$A_j[\pi] = \operatorname*{argmin}_{\pi' \in S_j} \ell_2\big(A_{j+1}[\pi], \pi'; \{z_1, \ldots, z_n\}\big).$$

Using the sequential approximation operators, we decompose the inner expectation term in (17) (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,

$$\begin{aligned}
&\sup_{\pi \in \Pi} \left| \frac{1}{n} \sum_{i \in [n]} \epsilon_i \big[h(z_i, \pi(x_i)) - h(z_i, A_J[\pi](x_i))\big] \right| \\
\leq &\frac{1}{n} \sqrt{n \sum_{i \in [n]} \Big(h\big(z_i, \pi(x_i)\big) - h(z_i, A_J[\pi](x_i))\Big)^2} \\
= &\frac{\Lambda}{\sqrt{n}} \cdot \ell_2(\pi, A_J(\pi); \{z_1, \ldots, z_n\}) \\
\leq &\frac{\Lambda}{\sqrt{n}} 2^{-J} \leq \frac{\Lambda}{n^{3/2}},
\end{aligned}$$

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, $\Xi_1 \leq \Lambda / n^{3/2}$.

Next, for any $j = 1, \ldots, J$ we use P_j to denote the projection of projecting a policy to S_j, i.e., $A_{j-1}[\pi] = P_{j-1}[A_j[\pi]]$. Once $A_j(\pi)$ is determined, $A_{j-1}(\pi)$ is also determined. For any $s > 0$,

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

we z is a shorthand for $\{z_1, \ldots, z_n\}$. For any $j = 1, \ldots, J$ and $m \in \mathbb{N}$, take

$$
s_{j,m} = \frac{\Lambda}{n2^{j-1/2}}\sqrt{\log\big(N_2(2^{-j},\Pi;Z)\cdot 2^{m+1}j^2\big)}.
$$

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

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

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

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

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

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

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

$$\begin{aligned}
\Xi_2 &= \mathbb{E}_\epsilon\left[\sup_{\pi\in\Pi}\left|\sum_{j=1}^{J}\frac{1}{n}\sum_{i\in[n]}\epsilon_i\Big[h(z_i, A_j[\pi](x_i)) - h(z_i, A_{j-1}[\pi](x_i))\Big]\right|\right] \\
&= \int_0^\infty \mathbb{P}_\epsilon\left(\sup_{\pi\in\Pi}\left|\sum_{j=1}^{J}\frac{1}{n}\sum_{i\in[n]}\epsilon_i\Big[h(z_i, A_j[\pi](x_i)) - h(z_i, A_{j-1}[\pi](x_i))\Big]\right| > s\right)\mathrm{d}s \\
&\le u_1 + \sum_{k=1}^{\infty}(u_{k+1}-u_k)\cdot 2^{-k+1} \\
&= \frac{4\Lambda}{n}\cdot\Big(\kappa(\Pi)+\sqrt{2}+1+\sum_{k=1}^{\infty}(\sqrt{k+2}-\sqrt{k+1})\cdot 2^{-k+1}\Big) \le \frac{4\Lambda}{n}\cdot\big(\kappa(\Pi)+7\big).
\end{aligned}$$

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

$$\begin{aligned}
\Xi_3 = \mathbb{E}_\epsilon\left[\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i h(z_i, \bar{\pi}(x_i))\right|\right] &= \int_0^\infty \mathbb{P}_\epsilon\left(\left|\frac{1}{n}\sum_{i\in[n]}\epsilon_i h(z_i, \bar{\pi}(x_i))\right| > s\right)\mathrm{d}s \\
&\le \int_0^\infty 2\exp\Big(-\frac{n^2 s^2}{\Lambda^2}\Big)\mathrm{d}s = \frac{3\Lambda}{n}.
\end{aligned}$$

Putting everything together,

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\big(x_i, a_i, y_i, \pi(x_i)\big)\Big|\right] &\le \frac{\Lambda}{n}\cdot(4\kappa(\Pi)+32) \\
&= \frac{2\sqrt{\sum_{i=1}^n c_i(z_i)^2}}{n}(4\kappa(\Pi)+32).
\end{aligned}$$

F.3. Proof of Lemma F.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 $z_1, \ldots, z_n$, define

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

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

$$n_i = \Big\lceil \frac{m}{\Lambda^2 n}\big\{h(z_i, \pi_{i,1}^*(x_i)) - h(z_i, \pi_{i,2}^*(x_i))\big\}^2\Big\rceil,$$

where we recall that $\Lambda^2 = 4\sum_{i=1}^n c_i(z_i)^2$. We then construct a new set of data

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

where z_i appears n_i times and

$$N = \sum_{i=1}^{n} n_i = \sum_{i=1}^{n}\Big\lceil \frac{m}{\Lambda^2}\big\{h(z_i, \pi_{i,1}^*(x_i)) - h(z_i, \pi_{i,2}^*(x_i))\big\}^2\Big\rceil \le m+n.$$

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 $H(\pi, \pi'; \tilde{x}) \le \gamma^2$. On the other hand,

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

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 $N \le m + n$, we further have

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

Since m is arbitrary, we take m to infinity and have $\ell_2(\pi, \pi'; z) \le \gamma$. By definition, S_0 is a γ-cover of Π under ℓ_2 with respect to $z_1, \ldots, z_n$, and therefore $N_2(\gamma, \Pi; z_1, \ldots, z_n) \le N_H(\gamma^2, \Pi)$.

F.4. Proof of Lemma D.3

By Yang et al. (2022, Lemma B12), $g_\delta(q)$ is differentiable in q, and

$$g_\delta'(q) = -\frac{\partial_q D_{\mathrm{KL}}(g(q) \,\|\, q)}{\partial_p D_{\mathrm{KL}}(g(q) \,\|\, q)} = \frac{g(q)/q - (1-g(q))/(1-q)}{\log\big(g(q)/(1-g(q))\big) - \log\big(q/(1-q)\big)}.$$

Also by Yang et al. (2022, Lemma B12), $g_\delta(q)$ is convex in q, so $g_\delta'(q)$ is increasing in q. Since $q \in [0.4, 0.6]$, $g_\delta'(q) \ge g_\delta'(0.4)$. From the dual form, we can check that $g(0.4) \ge 0.1$. Plugging in $q = 0.4$, we have

$$g_\delta'(0.4) = \frac{\frac{g(0.4)}{0.4} + \frac{g(0.4)}{0.6} - 5/3}{\log\big(g(0.4)/(1-g(0.4))\big) - \log(2/3)} = \frac{g(0.4)/0.24 - 5/3}{\log(g(0.4)/(1-g(0.4))) - \log(2/3)}.$$

Since the function $f(x) = \frac{x/0.24 - 5/3}{\log(x/(1-x)) - \log(2/3)}$ is increasing in x for $x \in (0, 0.4)$, we conclude that

$$g_\delta'(0.4) \ge \frac{1/2.4 - 5/3}{\log(1/9) - \log(2/3)} \ge 1/2,$$

completing the proof.