Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2020, Vol. 56, No. 3, 2162–2188
https://doi.org/10.1214/19-AIHP1034

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES
www.imstat.org/aihp

Performance guarantees for policy learning

Alex Luedtke[a,b] and Antoine Chambaz[c,d]

[a]*Department of Statistics, University of Washington, Seattle, USA*
[b]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, USA*
[c]*MAP5 (UMR CNRS 8145), Université de Paris, Paris, France. E-mail: antoine.chambaz@parisdescartes.fr*
[d]*Division of Biostatistics, School of Public Health, UC Berkeley, Berkeley, USA*

Received 20 April 2017; revised 18 September 2019; accepted 22 October 2019

Abstract. This article gives performance guarantees for the regret decay in optimal policy estimation. We give a margin-free result showing that the regret decay for estimating a within-class optimal policy is second-order for empirical risk minimizers over Donsker classes when the data are generated from a fixed data distribution that does not change with sample size, with regret decaying at a faster rate than the standard error of an efficient estimator of the value of an optimal policy. We also present a result giving guarantees on the regret decay of policy estimators for the case that the policy falls within a restricted class and the data are generated from local perturbations of a fixed distribution, where this guarantee is uniform in the direction of the local perturbation. Finally, we give a result from the classification literature that shows that faster regret decay is possible via plug-in estimation provided a margin condition holds. Three examples are considered. In these examples, the regret is expressed in terms of either the mean value or the median value, and the number of possible actions is either two or finitely many.

Résumé. Cet article présente des garanties de performance concernant la vitesse à laquelle le regret s'amenuise dans le cadre de l'estimation d'une politique d'action optimale. Si la politique optimale est définie comme optimale relativement à un ensemble de politiques formant une classe de Donsker, et si elle est estimée par minimisation sur cet ensemble d'une estimation du regret vu comme une fonction sur celui-ci, alors un premier résultat révèle que la vitesse est de second ordre dès lors que les observations sont générées sous une loi qui ne change pas à mesure que leur nombre augmente. Plus spécifiquement, le regret de l'estimateur de la politique optimale s'amenuise plus rapidement que l'écart type d'un estimateur efficace de la valeur d'une politique optimale. Ce résultat ne nécessite pas le recours à une hypothèse de marge. Un second résultat porte sur la vitesse à laquelle le regret de l'estimateur de la politique optimale s'amenuise lorsque les observations sont générées sous des lois définies comme des perturbations locales d'une loi de référence fixe, la garantie de performance étant alors uniforme relativement aux directions de perturbation. Finalement, un troisième résultat montre qu'il est possible d'atteindre des vitesses plus rapides en mettant en œuvre une procédure d'estimation par substitution à la condition qu'une hypothèse de marge soit satisfaite. Ce résultat s'inspire de la littérature consacrée à la classification. Trois exemples illustrent nos trouvailles. Dans ceux-ci, le regret s'exprime en termes de valeur moyenne ou de valeur médiane, et les actions envisageables sont au nombre de deux ou bien en nombre fini.

MSC: 62F12; 62G20; 62H30; 62P10

Keywords: Individualized treatment rules; Personalized medicine; Policy learning; Precision medicine

1. Introduction

1.1. *Objective*

We consider an experiment where a player repeatedly chooses and carries out one among two actions to receive a random reward (the case of finitely many actions is also considered) in a batch setting. Each time, the reward depends on both the action undertaken and a random context preceding it, which is given to the player before she makes her choice. The law of the context and conditional law of the reward given the context and action are fixed throughout the experiment. The player's objective is to obtain as large an expected reward as possible.

In this framework, a policy is a rule that maps any context to an action. The value of a policy is the expectation of the reward in the experiment where the action carried out is the action recommended by the policy. Given a class Π of candidate policies, the regret of a policy $\pi \in \Pi$ is the difference between the largest value achievable within Π and the value of π.

Learning the optimal policy within a class of candidate policies is meaningful whenever the goal is to make recommendations. This is, for instance, the case in personalized medicine, also known as precision medicine. There, the context would typically consist of the description of a patient, the actions would correspond to two strategies of treatment, and the policies are rather called individualized treatment rules.

The objective of this article is not to establish optimal regret bounds for optimal policy estimators. It is, rather, to show that rates faster than $n^{-1/2}$ can be demonstrated under much more general conditions than have previously been discussed in the policy learning literature.

1.2. *A brief literature review*

There has been a surge of interest in developing flexible methods for estimating optimal policies in recent years. Here we give a deeply abbreviated overview, and refer the reader to [1] for a recent overview of the literature. Exciting developments in policy learning over the last several years include outcome weighted learning (OWL) [22,33,34], ensemble techniques [15], and empirical risk minimizers (ERMs) [1], to name a few. Each of these works has established some form of regret bound for the estimator, with most of these regret rates no faster than $n^{-1/2}$. Notable exceptions to this $n^{-1/2}$ restriction occur for OWL methods under a hard margin [34] and rates attainable by plug-in estimation strategies [6,8,10,16,20]. Asymptotically minimax results that scale as $n^{-1/2}$ when the covariate distribution is discrete have also been established [19]. Stoye [24] furthered this line of work by establishing finite-sample minimax results for certain classes of distributions, and also noted that a data-free rule is finite-sample minimax when the policy class is not restricted [24] – we note here that these results will not impact the results in our work, given that we primarily focus on restricted policy classes (except in Section 2.3, where we consider unrestricted policy classes).

We give our results in Section 2. Our first result pertains to empirical risk minimization within a Donsker class, where the optimal policy is defined as the so-called value maximizer within this class. These estimators were recently studied in [1], and were also discussed for continuous treatments in [17]. The second result studies the performance of these estimators under local fluctuations of the data generating distribution. These sort of local results were studied for unconstrained policy classes in [9], and are also studied for restricted classes starting in version 2 of Athey and Wager's technical report (2017) [1]. The third result links the optimal policy problem to the classification literature, which has shown that faster regret decay rates are attainable by plug-in estimators under certain margin conditions [2]. Section 3 gives three remarks, one relating our results to the pioneering results of Koltchinskii [11], another studying which term in our regret bound dominates the rate, and the third relating our results to those of Athey and Wager [1]. Section A proves our results regarding empirical risk minimization.

Two of our results are for the regret under a fixed data generating distribution. The faster-than-$n^{-1/2}$ rate that we will establish for ERMs (independent of any margin assumption) will not transfer to the minimax setting unless some form of margin assumption is imposed. Indeed, the results that we establish under local asymptotics do not scale faster than $n^{-1/2}$. We do not give high probability results, though these are often a straightforward extension of convergence in probability results when working within Donsker classes. Throughout this work, when we refer to Donsker classes, we are referring to P-Donsker classes, where P is the data generating distribution, rather than universal/uniform Donsker classes [23]. While our ERM results are not useful for non-P-Donsker classes, we note that they also serve as an oracle guarantee for an ensemble procedure, such as [15], so that one needs not know in advance whether or not all of the classes over which they are optimizing are P-Donsker for the result to be useful.

The so-called empirical process plays a prominent role in our study. More specifically, several key events under scrutiny concern suprema of the empirical process over a variety of well-chosen classes of functions. In order to circumvent measurability issues that are potentially delicate, we assume that two pivotal classes of functions are separable (for instance, countable). Alternatively, we might have relied on a separable version of the empirical process or on convergence in outer probability [32].

1.3. *Formalization*

Let $X \in \mathcal{X}$ denote a vector of covariates describing the context preceding the action, $A \in \{-1, 1\}$ denote the action undertaken, and Y denote the corresponding reward, where here larger rewards are preferable. We denote by P the distribution of $O \equiv (X, A, Y)$ and by $\mathbb{E}$ and $\mathbb{V}$ the expectation and variance under P (the expectation under a generic distribution P' of O is denoted by $\mathbb{E}_{P'}$). For simplicity we assume throughout that, under P, Y is uniformly bounded and that there exists some $\delta \in (0, 1/2]$ so that $P(A = 1|X)$ falls in $[\delta, 1 - \delta]$ with probability one. The second assumption is known as the strong positivity assumption. While stronger than we need, these assumptions simplify our analysis.

Let $\mathcal{P}$ be the class of all policies, the subset of $L^2(P)$ consisting of functions mapping $\mathcal{X}$ to $\{-1, 1\}$. The value of a policy $\pi \in \mathcal{P}$ is given by

$$\mathcal{V}(\pi) \equiv \mathbb{E}\big[\mathbb{E}\big[Y|A=\pi(X), X\big]\big]. \tag{1}$$

We emphasize here that, though not indicated in the notation, $\mathcal{V}(\pi)$ is a functional of the distribution P. Under some causal assumptions that we will not explore here, $\mathcal{V}(\pi)$ can be identified with the mean reward if, possibly contrary to fact, action $\pi(X)$ is carried out in context X.

Now, let $\Pi \subset \mathcal{P}$ be the class of candidate policies. By choice, Π is separable, in that there exists a countable $\Pi_0 \subset \Pi$ such that every $\pi \in \Pi$ is the pointwise limit of a sequence of elements of Π_0 (for instance, Π itself could be countable). The regret (within class Π) of π is defined as the difference between $\mathcal{V}(\pi)$ and the optimal value $\mathcal{V}^\star \equiv \sup_{\pi \in \Pi} \mathcal{V}(\pi)$: for all $\pi \in \Pi$,

$$\mathcal{R}(\pi) \equiv \mathcal{V}^\star - \mathcal{V}(\pi). \tag{2}$$

We extend the definition of $\mathcal{R}$ to $\mathcal{P}$. Obviously, $\mathcal{R}(\pi) \geq 0$ for all $\pi \in \Pi$, but $\mathcal{R}$ is not necessarily nonnegative over $\mathcal{P}$.

For every $\pi \in \mathcal{P}$, $\mathcal{V}(\pi)$ can be viewed as the evaluation at P of the real-valued functional

$$P' \mapsto \mathbb{E}_{P'}\big[\mathbb{E}_{P'}\big[Y|A=\pi(X), X\big]\big]$$

from the nonparametric model of distributions P' satisfying the same constraints as P. This functional is pathwise differentiable at P relative to the maximal tangent space with an efficient influence function f_π given by

$$f_\pi(o) \equiv \frac{\mathbf{1}\{a=\pi(x)\}}{P(A=a|X=x)}\big(y - \mathbb{E}[Y|A=a, X=x]\big) + \mathbb{E}\big[Y|A=\pi(x), X=x\big] - \mathcal{V}(\pi). \tag{3}$$

By the bound on Y and the strong positivity assumption, there exists a universal constant $M > 0$ such that $P(\sup_{\pi \in \Pi} |f_\pi(O)| \leq M) = 1$. As stated earlier, the strong positivity assumption is stronger than required. In particular, if we make the weaker positivity assumption that $P(A=1|X) \in (0,1)$ with probability one and $\mathbb{E}[P(A=1|X)^{-2}] < \infty$ then, by boundedness of Y, $\{f_\pi : \pi \in \Pi\}$ will still have an envelope function in $L^2(P)$, which will suffice for the conclusion of our main result (Theorem 1) to remain valid, though the proof would need to be modified.

We refer the interested reader to [31, Section 25.3] and [16] for definitions and proofs of these facts. Although the class

$$\mathcal{F} \equiv \{f_\pi : \pi \in \Pi\} \subset L^2(P) \tag{4}$$

will play a central role in the rest of this article, it is not necessary to master the derivation or properties of efficient influence functions to appreciate the contributions of this work. Note that the separability of Π implies that the closure $\overline{\Pi}$ of Π under $L^2(P)$ norm, $\mathcal{F}$ and $\{f_\pi : \pi \in \overline{\Pi}\}$ are separable too. Therefore, all the forthcoming suprema of the empirical process are measurable.

2. Main results

2.1. *Preliminary*

Suppose that we observe mutually independent and identically distributed (i.i.d.) data-structures $O_1 \equiv (X_1, A_1, Y_1), \ldots, O_n \equiv (X_n, A_n, Y_n)$ drawn from P. We denote the corresponding empirical measure by $P_n \equiv n^{-1} \sum_{i=1}^n \mathrm{Dirac}(O_i)$. For every function f mapping an observation to the real line, we set $Pf \equiv \mathbb{E}[f(X, A, Y)]$ and $P_n f \equiv n^{-1} \sum_{i=1}^n f(O_i)$.

Our theorem will attain fast rates under the assumption that one has available an estimator $\{\widehat{\mathcal{V}}(\pi) : \pi \in \Pi\}$ of $\{\mathcal{V}(\pi) : \pi \in \Pi\}$ that makes the following term small:

$$\mathrm{Rem}_n \equiv \sup_{\pi \in \Pi} \big|\widehat{\mathcal{V}}(\pi) - \mathcal{V}(\pi) - (P_n - P) f_\pi\big|. \tag{5}$$

If empirical process conditions are imposed on the estimators for the reward regression, $\mathbb{E}[Y|A, X]$, and the action mechanism, $P(A|X)$, then $\widehat{\mathcal{V}}$ could be defined using estimating equations [28] or targeted minimum loss-based estimation (TMLE) [29,30]. Without imposing empirical process conditions, one could use cross-validated estimating equations [18, 26], now also called double machine learning [7], or a cross-validated TMLE [35]. Note that cross-validated estimating equations in this scenario represent a special case of a cross-validated TMLE, where one uses the squared cross-validated efficient influence function as loss [5]. For any of these listed approaches, obtaining the above uniformity over π is straightforward if Π is a Donsker class. It is plausible that Rem_n will be small for these listed approaches precisely because they make use of an estimate of f_π, a function which is known to be a gradient of the parameter taking as input a distribution P and returning the value of the rule π at P. If one either uses a cross-validated approach or enforces Donsker conditions on (consistent) estimators of the nuisance functions $(a,x) \mapsto \mathbb{E}[Y|A=a, X=x]$ and $(a,x) \mapsto P(A=a|X=x)$, then the fact that f_π is a gradient ensures that Rem_n is equal to an $o_P(n^{-1/2})$ term plus a second-order term that can be upper bounded by a constant times the product of the $L^2(P)$ distances between the estimated and true $(a,x) \mapsto \mathbb{E}[Y|A=a, X=x]$ and $(a,x) \mapsto P(A=a|X=x)$ functions [28], which is $O_P(n^{-1})$ in a correctly specified parametric model.

2.2. *Empirical risk minimizers*

Our first objective will be to establish a faster than $n^{-1/2}$ rate of regret decay for a value-based estimator $\widehat{\pi}$ of an optimal policy, given by any ERM $\widehat{\pi} \in \Pi$ satisfying

$$\widehat{\mathcal{V}}(\widehat{\pi}) \geq \sup_{\pi \in \Pi} \widehat{\mathcal{V}}(\pi) - o_P(\mathrm{Rem}_n). \tag{6}$$

Note that (6) is a requirement on the behavior of the optimization algorithm on the realized sample, rather than a statistical or probabilistic condition. The $o_P(\mathrm{Rem}_n)$ term allows us to obtain an approximate solution to the ERM problem, which is useful because the optimization on the right is non-concave. In practice, it is possible that the approximation error will converge to zero more slowly than Rem_n. We discuss this case briefly following Theorem 1.

2.2.1. *Main, fixed-P ERM result*

We now present a result providing regret decay guarantees under data generated from the fixed distribution P that is similar to the groundbreaking results for ERMs based on general losses given in [11]. In Section 3.1, we discuss how our result relates to those in [11].

Recall that $\overline{\Pi}$ is the closure of Π under $L^2(P)$ norm. The set

$$\begin{aligned}\overline{\Pi}^\star &\equiv \{\pi^\star \in \overline{\Pi} : \mathcal{V}(\pi^\star) = \mathcal{V}^\star\} \\ &= \{\pi^\star \in \overline{\Pi} : \mathcal{R}(\pi^\star) = 0\} \subset \{\pi^\star \in \overline{\Pi} : \mathcal{R}(\pi^\star) \leq 0\}\end{aligned} \tag{7}$$

also plays an important role. If Π is Donsker, then Lemma A.3 in Section A guarantees that $\overline{\Pi}^\star$ is nonempty and coincides with the right-hand side (RHS) set in the above display. It may be the case that $\overline{\Pi}^\star$ contain more than one policy.

To develop an intuition, let us derive an initial upper bound on the regret $\mathcal{R}(\widehat{\pi})$ of $\widehat{\pi}$. In light of (5) and (6),

$$\begin{aligned}0 \leq \mathcal{R}(\widehat{\pi}) &= \mathcal{V}(\pi^\star) - \mathcal{V}(\widehat{\pi}) \\ &= [\mathcal{V}(\pi^\star) - \widehat{\mathcal{V}}(\pi^\star)] + [\widehat{\mathcal{V}}(\widehat{\pi}) - \mathcal{V}(\widehat{\pi})] + [\widehat{\mathcal{V}}(\pi^\star) - \widehat{\mathcal{V}}(\widehat{\pi})] \\ &\leq [\mathcal{V}(\pi^\star) - \widehat{\mathcal{V}}(\pi^\star)] + [\widehat{\mathcal{V}}(\widehat{\pi}) - \mathcal{V}(\widehat{\pi})] + o_P(1)\mathrm{Rem}_n \\ &\leq (P_n - P)(f_{\widehat{\pi}} - f_{\pi^\star}) + (2 + o_P(1))\mathrm{Rem}_n.\end{aligned}$$

Much of the challenge of the proof of the upcoming theorem is to control the first RHS term. It turns out that the pivotal random expression of interest is

$$\mathcal{E}_n \equiv \liminf_{s \downarrow 0} \inf_{\pi^\star \in \overline{\Pi}^\star} \sup_{\pi \in \Pi : \|\pi - \pi^\star\| \leq s} (P_n - P)(f_{\widehat{\pi}} - f_\pi),$$

where $\|\cdot\|$ denotes the $L^2(P)$ norm and the dependence of $\mathcal{E}_n$ on $\widehat{\pi}$ is suppressed in the notation. The above roughly corresponds to the best approximation in Π of the closest optimal policy $\pi^\star \in \overline{\Pi}^\star$ to the estimated policy $\widehat{\pi}$.

Critically, $\mathcal{E}_n$ will allow us to transfer the study of the regret $\mathcal{R}(\widehat{\pi})$ of $\widehat{\pi}$ to the study of an empirical process, an object that we will be able to study using asymptotic equicontinuity arguments. This is similar to how classical semiparametric estimation schemes of pathwise differentiable, but possibly nonlinear, functionals transform the estimation problem to that of a linear functional [3], thereby enabling a proof of $n^{-1/2}$ convergence rates and the application of well-established central limit theorems. Indeed, in our case we will show via well-established asymptotic equicontinuity arguments that $\mathcal{E}_n = o_P(n^{-1/2})$ under mild assumptions. We note that the form of our $\mathcal{E}_n$ is slightly more complicated than that of empirical process terms in classical semiparametric estimation settings because $\overline{\Pi}^\star$ may not be a singleton.

Theorem 1 (Main Fixed-P ERM Result). *If $\widehat{\pi}$ satisfies (5), (6), and if Π is Donsker, then*

$$0 \le \mathcal{R}(\widehat{\pi}) \le \mathcal{E}_n + \big[2 + o_P(1)\big]\mathrm{Rem}_n. \tag{8}$$

If also $\mathrm{Rem}_n = o_P(n^{-1/2})$, then $\mathcal{E}_n = o_P(n^{-1/2})$, and therefore $\mathcal{R}(\widehat{\pi}) = o_P(n^{-1/2})$.

The proof of Theorem 1 is given in Section A. If the $o_P(\mathrm{Rem}_n)$ term in (6) is replaced by an arbitrary approximation error term AE_n, then the above result changes to $\mathcal{R}(\widehat{\pi}) \le \mathcal{E}_n + \mathrm{Rem}_n + \mathrm{AE}_n$, and the conclusion that $\mathcal{R}(\widehat{\pi}) = o_P(n^{-1/2})$ remains valid if Rem_n and AE_n are $o_P(n^{-1/2})$.

We note that the size of the class Π never shows up explicitly in the bound in Theorem 1. All we used about Π was that it was Donsker, which allowed us to invoke asymptotic equicontinuity to control the term $\mathcal{E}_n$, which appears as a part of the $o_P(n^{-1/2})$ second-order term in the bound on the right-hand side of (8). One could apply maximal inequalities (e.g., Chapter 2.14 in [32]) to get explicit bounds on how changing the size of the class Π will impact the rate of convergence of the term $\mathcal{E}_n$. Nonetheless, in some cases the $\mathcal{E}_n$ term will not be the dominant term on the right-hand side of (8). Indeed, if the reward regression, $(a,x) \mapsto \mathbb{E}[Y|A=a, X=x]$, and/or the action mechanism, $(a,x) \mapsto P(A=a|X=x)$, is difficult to estimate (e.g., very nonsmooth), then the term Rem_n may dominate the right-hand side of (8) unless the Donsker class Π is very large.

The above rate on $\mathcal{R}(\widehat{\pi})$ is faster than the rate of convergence of the standard error of an efficient estimator of the value $\mathcal{V}(\pi)$ of a policy $\pi \in \Pi$, which converges at rate $n^{-1/2}$ in all but degenerate cases. Note that our proof of the first result of the above theorem only uses that Π is totally bounded in $L^2(P)$ (a notion defined in Appendix A.1) whereas that of the second result uses the stronger assumption that Π is Donsker (see [32, Example 1.5.10] or [31, Lemma 18.15]).

2.2.2. *Main uniform local ERM result*

The next result concerns the local properties of the ERM estimator. Local properties of policy estimators within unrestricted classes Π were studied in [9].

Define $L_1^2(P)$ as the set of functions $x \mapsto \phi_1(x)$ in $L^2(P)$ satisfying $\mathbb{E}[\phi_1(X)] = 0$, $L_2^2(P)$ as the set of functions $(a,x) \mapsto \phi_2(a,x)$ in $L^2(P)$ satisfying $P(\mathbb{E}[\phi_2(A,X)|X] = 0) = 1$, and $L_3^2(P)$ as the set of functions $(y,a,x) \mapsto \phi_3(y,a,x)$ in $L^2(P)$ satisfying $P(\mathbb{E}[\phi_3(Y,A,X)|A,X] = 0) = 1$. In a slight abuse of notation, for $o = (y,a,x)$ we will sometimes write $\phi_1(o)$ to denote $\phi_1(x)$, and similarly for $\phi_2(o)$ and $\phi_3(o)$. For $j = 1,2,3$, let $\Phi_j \subset L_j^2(P)$ be a separable (for instance, countable) class of functions ϕ_j that satisfy a *uniformly sub-exponential* property in the sense that there exist parameters (b_1, σ_1^2), (b_2, σ_2^2) and (b_3, σ_3^2) in $(0,\infty)^2$ such that, P-almost surely,

$$\log \mathbb{E}\big[\exp\big\{\lambda\phi_1(X)\big\}\big] \le \frac{\lambda^2\sigma_1^2}{2} \quad \text{for all } |\lambda| \le 1/b_1, \tag{9}$$

$$\log \mathbb{E}\big[\exp\big\{\lambda\phi_2(A,X)\big\}|X\big] \le \frac{\lambda^2\sigma_2^2}{2} \quad \text{for all } |\lambda| \le 1/b_2, \tag{10}$$

$$\log \mathbb{E}\big[\exp\big\{\lambda\phi_3(Y,A,X)\big\}|A,X\big] \le \frac{\lambda^2\sigma_3^2}{2} \quad \text{for all } |\lambda| \le 1/b_3. \tag{11}$$

We impose that Φ_2 contain the zero function. Let $\Phi = \Phi_1 \times \Phi_2 \times \Phi_3$, and note that Φ is also separable.

For $\epsilon \in \mathbb{R}$, $|\epsilon| \le 1/\max\{b_1, b_2, b_3\}$, and $\phi = (\phi_1, \phi_2, \phi_3) \in \Phi$, define a fluctuation $P_{\phi,\epsilon}$ of P by its conditional densities:

$$\frac{dP_{\phi,\epsilon}}{dP}(x) = c_{1,\phi}(\epsilon)\exp\big\{\epsilon\phi_1(x)\big\}, \tag{12}$$

$$\frac{dP_{\phi,\epsilon}}{dP}(a|x) = c_{2,\phi}(\epsilon|x)\exp\big\{\epsilon\phi_2(a,x)\big\}, \tag{13}$$

$$\frac{dP_{\phi,\epsilon}}{dP}(y|a,x) = c_{3,\phi}(\epsilon|a,x)\exp\big\{\epsilon\phi_3(y,a,x)\big\}, \tag{14}$$

where above $c_{j,\phi}$ are normalizing constants that ensures that the (conditional) densities define probability measures. We think of $P_{\phi,\epsilon}$ as a fluctuation of P in the direction of $\sum_{j=1}^3 \phi_j$ (the one-dimensional parametric collection of distributions thus defined goes through P at $\epsilon = 0$ with a score equal to $\sum_{j=1}^3 \phi_j$). We will be particularly interested in $P_{\phi,\epsilon}$ for $\epsilon = n^{-1/2}$. For notational simplicity, we will proceed as if $\max_{j\in\{1,2,3\}} b_j \le 1$ instead of assuming that n is large enough to guarantee that $n^{-1/2} \le 1/\max_{j\in\{1,2,3\}} b_j$. In Lemma A.11 in the appendix, we show that the sequence of collections of distributions $\{P^n_{\phi,\epsilon=n^{-1/2}} : \phi \in \Phi\}_{n\geq 1}$ is what we call *uniformly contiguous* with respect to (w.r.t.) the sequence $\{P^n\}_{n\geq 1}$, in the sense that, for any sequence $\{B_n\}_{n\geq 1}$ of measurable sets with $\lim_{n\to\infty} P^n(B_n) = 0$,

$$\lim_{n\to\infty} \sup_{\phi\in\Phi} P^n_{\phi,\epsilon=n^{-1/2}}(B_n) = 0. \tag{15}$$

Note that uniform contiguity of the collection $\{P^n_{\phi,\epsilon=n^{-1/2}} : \phi \in \Phi\}_{n\geq 1}$ w.r.t. $\{P^n\}_{n\geq 1}$ implies that, for any $\phi \in \Phi$, $\{P^n_{\phi,\epsilon=n^{-1/2}}\}_{n\geq 1}$ is contiguous w.r.t. $\{P^n\}_{n\geq 1}$ as developed by Le Cam [13].

In what follows, for $\phi \in \Phi$ and a real-valued ϵ, we denote expectations over $P_{\phi,\epsilon}$ by $\mathbb{E}_{\phi,\epsilon}$, and we respectively define the value, the optimal value, and the regret of a rule $\pi \in \overline{\Pi}$ under $P_{\phi,\epsilon}$ as $\mathcal{V}_{\phi,\epsilon}(\pi) \equiv \mathbb{E}_{\phi,\epsilon}[\mathbb{E}_{\phi,\epsilon}[Y|A=\pi(X), X]]$, $\sup_{\pi^\star\in\Pi} \mathcal{V}^\star_{\phi,\epsilon} \equiv \sup_{\pi^\star\in\Pi} \mathcal{V}_{\phi,\epsilon}(\pi^\star)$, and $\mathcal{R}_{\phi,\epsilon}(\pi) \equiv \sup_{\pi^\star\in\Pi} \mathcal{V}^\star_{\phi,\epsilon} - \mathcal{V}_{\phi,\epsilon}(\pi)$. We continue to use $\|\cdot\|$ to denote the $L^2(P)$ norm and also use the notation $\langle\cdot,\cdot\rangle$ to denote the inner product that this norm induces.

For $\phi \in \Phi$, we define $\overline{\Pi}^\star_{\phi,\epsilon}$ as the set of rules in $\overline{\Pi}$ maximizing the value under the distribution $P_{\phi,\epsilon}$:

$$\overline{\Pi}^\star_{\phi,\epsilon} \equiv \{\pi^\star \in \overline{\Pi} : \mathcal{V}_{\phi,\epsilon}(\pi^\star) = \mathcal{V}^\star_{\phi,\epsilon}\}.$$

We note that Lemma A.3, which applies to any distribution in our model, not just to the distribution P, shows that $\overline{\Pi}^\star_{\phi,\epsilon}$ is nonempty and that the regret is zero at the minimizer.

The below theorem assumes that Π is totally bounded in $L^2(P)$. We note that $\overline{\Pi}$ is compact if Π is totally bounded (see Lemma A.1), and recall that Π is totally bounded in $L^2(P)$ if Π is P-Donsker. Furthermore, we note here that the below theorem applies to *any* policy estimator $\widehat{\pi}_1$ satisfying the given condition, including the ERM $\widehat{\pi}$ or the plug-in estimators that will be presented in Section 2.3 provided they achieve the required rate of convergence. Finally, for a given $\pi^\star \in \overline{\Pi}^\star$, introduce

$$V_*^{1/2}(\pi^\star) \equiv \|f_{\pi^\star} - f_{-\pi^\star}\|. \tag{16}$$

As was noted in version 1 of [1], $V_*^{1/2}(\pi^\star)$ does not depend on the choice of $\pi^\star \in \overline{\Pi}^\star$. Hence, we will write $V_*^{1/2} \equiv V_*^{1/2}(\pi^\star)$, where $\pi^\star$ is any arbitrary element of $\overline{\Pi}^\star$.

Theorem 2 (Main Uniform Local ERM Result). *Suppose that $n^{1/2}\mathcal{R}(\widehat{\pi}_1)$ converges to zero in probability under sampling from P and that Π is totally bounded. Let $\epsilon_n \equiv n^{-1/2}$ for all $n \geq 1$. For any $t > 0$, it holds that*

$$\begin{aligned} o(1) &= \sup_{\phi\in\Phi} P^n_{\phi,\epsilon_n}\Big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \sup_{\pi^\star\in\overline{\Pi}^\star} \inf_{\pi\in\overline{\Pi}^\star_{\phi,\epsilon_n}} \langle f_\pi - f_{\pi^\star}, \phi_1 + \phi_3\rangle + t\Big\} \\ &\geq \sup_{\phi\in\Phi} P^n_{\phi,\epsilon_n}\big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \|\phi_1 + \phi_3\| V_*^{1/2} + 2t\big\} - o(1). \end{aligned}$$

The uniform nature of the above result may help to make it more informative of finite-sample performance of $\widehat{\pi}_1$ than would be a non-uniform result. If $\widehat{\pi}_1$ is the empirical risk minimizer $\widehat{\pi}$, then, by Theorem 1 and the fact that Π Donsker implies that Π is totally bounded, the two conditions in the first sentence of the above theorem statement can be replaced by the condition that Π is Donsker. Also note that the inequality stated in the theorem is a simple consequence of the preceding equality, the Cauchy–Schwarz inequality and Lemma A.14 stated and proven in Appendix A.2. Moreover, noting that Lemma A.8 in the appendix shows that the $L^2(P)$-magnitude of the perturbations in Φ is uniformly bounded by some $C > 0$, the above shows that, for all $t > 0$,

$$\sup_{\phi\in\Phi} P^n_{\phi,\epsilon_n}\big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > C V_*^{1/2} + t\big\} = o(1).$$

As such, any policy estimator with fixed-P regret decaying at a rate faster than $n^{-1/2}$ has local regret that is uniformly bounded by a (C, P)-dependent constant in probability.

For certain distributions P, this result may be of limited interest. For example, [9] focused on perturbations of distributions P for which $\mathbb{E}[Y|A=1,X] - \mathbb{E}[Y|A=-1,X] = 0$ almost surely. In this setting, any policy estimator $\widehat{\pi}_1$ will have regret zero, thereby achieving $n^{1/2}\mathcal{R}(\widehat{\pi}_1) = o_P(1)$ with much to spare. The results of [9] are more informative in this setting, since they provide a $n^{-1/2}$ rate bound on the regret that does not require the $L^2(P)$-magnitude of perturbations in Φ to be bounded. However, our Theorem 2 provides guarantees under local perturbations of *any* distribution P, where we have shown that these conditions are always satisfied for ERMs over Donsker classes. Moreover, our asymptotics are uniform over perturbations in Φ, which may make our results more reflective of finite-sample performance.

The above theorem results in the following important corollary, which shows that the regret decays faster than $n^{-1/2}$ whenever $\overline{\Pi}^\star$ is a singleton, and furthermore that this convergence is uniform over local fluctuation directions $\phi \in \Phi$.

Corollary 3. *If the conditions of Theorem 2 hold and* $\overline{\Pi}^\star$ *is a singleton, then, for all* $t > 0$,

$$\sup_{\phi\in\Phi} P^n_{\phi,\epsilon_n}\left\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > t\right\} = o(1).$$

The above result follows immediately from Theorem 2 once it is noted that $\sup_{\pi_1^\star,\pi_2^\star\in\overline{\Pi}^\star}\langle f_{\pi_1^\star} - f_{\pi_2^\star}, \phi_1 + \phi_3\rangle = 0$ if $\overline{\Pi}^\star$ is a singleton. Hence, the proof is omitted.

Three detailed remarks on Theorems 1 and Theorem 2 are given in Section 3.

2.3. *Plug-in estimators*

Section 3.1 discusses why we do not expect a faster than $O_P(n^{-1})$ rate of regret decay for $\widehat{\pi}$, even within a parametric model. A notable exception to this rule of thumb occurs if Π is of finite cardinality and $P(A|X)$ is known, since in this case large deviation bounds suggest very fast rates of convergence for ERMs. The same phenomenon has been noted and extensively studied in the classification literature (see the review in [25]).

We now show that faster rates are attainable under some conditions if one uses a different estimation procedure. We point this out because, in our experience, this alternative estimation strategy can sometimes yield better estimates than a value-based strategy [15].

Let $\gamma(X) \equiv \mathbb{E}[Y|A = 1, X] - \mathbb{E}[Y|A = -1, X]$ denote the conditional average action effect. In this subsection, we assume that Π is an unrestricted class, *i.e.*, that Π contains all functions mapping $\mathcal{X}$ to $\{-1, 1\}$. Audibert and Tsybakov [2] presented the surprising fact that plug-in classifiers can attain much faster rates (faster than n^{-1}). In our setting, the plug-in policy that we will study first defines an estimator $\widehat{\gamma}$ of γ, and then determines the action to undertake based on the sign of the resulting estimate. Formally, $\widehat{\pi}_{\widehat{\gamma}}$ is given by

$$\widehat{\pi}_{\widehat{\gamma}}(x) \equiv \mathbf{1}\left\{\widehat{\gamma}(x) > 0\right\} - \mathbf{1}\left\{\widehat{\gamma}(x) \leq 0\right\}.$$

Extensions of the result of [2] have previously been presented in both the reinforcement learning literature [8] and in the optimal individualized treatment literature [6,10,16,20]. We therefore omit any proof, and refer the reader to [2, Lemma 5.2] for further details.

The upcoming theorem uses $\|\cdot\|_\infty$ to denote the $L^\infty(P)$ norm. Let $\lesssim$ denote "less than or equal to up to a universal positive multiplicative constant" and consider the following "margin assumption":

(MA) For some $\alpha > 0$, $P(0 < |\gamma(X)| \leq t) \lesssim t^\alpha$ for all $t > 0$.

Theorem 4. *Suppose MA holds. If* $\|\widehat{\gamma} - \gamma\| = o_P(1)$, *then* $0 \leq \mathcal{R}(\widehat{\pi}_{\widehat{\gamma}}) \lesssim \|\widehat{\gamma} - \gamma\|^{2(1+\alpha)/(2+\alpha)}$. *If* $\|\widehat{\gamma} - \gamma\|_\infty = o_P(1)$, *then* $0 \leq \mathcal{R}(\widehat{\pi}_{\widehat{\gamma}}) \lesssim \|\widehat{\gamma} - \gamma\|_\infty^{1+\alpha}$.

We now briefly describe this result, though we note that it has been discussed thoroughly elsewhere [2,6,16]. Note that MA does not place any restriction on the decision boundary $\{x \in \mathcal{X} : \gamma(x) = 0\}$ where no action is superior to the other, but rather only places a restriction on the probability that $\gamma(X)$ be near zero.

If $\alpha = 1$ and $\gamma(X)$ is absolutely continuous (under P), then MA corresponds to $\gamma(X)$ having bounded density near zero. The case that $\alpha = 1$ is of particular interest for the sup-norm result because then the regret bound is quadratic in the rate of convergence of $\widehat{\gamma}$ to γ. As $\alpha \to 0$, more mass is placed near the decision boundary (zero) and the above result yields the rate of convergence of $\widehat{\gamma}$ to γ. As $\alpha \to \infty$, a vanishingly small amount of mass is concentrated near zero and the above result recovers very fast rates of convergence when $\widehat{\gamma}$ converges to γ uniformly.

Remark 5 (Estimating γ). Doubly robust unbiased transformations [21] provide one way to estimate γ (Section 3.1 of [27]). In particular, one can regress (via any desired algorithm) the pseudo-outcome

$$\widehat{\Gamma}(O) \equiv \frac{A}{\widehat{P}(A|X)}\left(Y - \widehat{\mathbb{E}}[Y|A, X]\right) + \widehat{\mathbb{E}}[Y|A = 1, X] - \widehat{\mathbb{E}}[Y|A = -1, X]$$

against X, both for double robustness and for efficiency gains. Above $(a, x) \mapsto \widehat{P}(A = a|X = x)$ and $(a, x) \mapsto \widehat{\mathbb{E}}[Y|A = a, X = x]$ are estimators of $(a, x) \mapsto P(A = a|X = x)$ and $(a, x) \mapsto \mathbb{E}[Y|A = a, X = x]$. One can use cross-validation to avoid dependence on empirical process conditions for the estimators of $(a, x) \mapsto P(A = a|X = x)$ and $(a, x) \mapsto \mathbb{E}[Y|A = a, X = x]$. In practice, one can ensure that $\widehat{P}(A|X)$ is bounded away from zero and one by truncating this quantity in $[\tilde{\delta}, 1 - \tilde{\delta}]$ for $\tilde{\delta} > 0$. To ensure that the truncation does not destroy the good statistical properties of $\widehat{P}(A|X)$, one could select $\tilde{\delta}$ via cross-validation w.r.t. the cross-entropy loss, which heavily penalizes estimators for selecting $\widehat{P}(A = a|X)$ near zero when an individual with covariate level X has observed action $A = 1 - a$.

3. Three remarks on ERM results

3.1. *Relation to the rates of Koltchinskii* [11]

Our Theorem 1 is related to [11, Theorem 4] and to the discussion of classification problems in Section 6.1 of this landmark article, although we *(i)* make no assumptions on the behavior of γ near the decision boundary ("margin assumptions"), *(ii)* give a fixed-P rather than a minimax result, *(iii)* do not require that the regret minimizer belong to the set Π, and *(iv)* make a slightly weaker complexity requirement on the class Π (only requiring Π Donsker). We thus have weaker conditions (only constrain the complexity of Π using a general Donsker condition) and, consequently, a weaker implication. Our result also differs from that of [11] due to the need for us to control the extra remainder term arising from (5).

Our Lemmas A.4 and A.6 in Section A are key to giving this general "Donsker implies fast rate" implication, even when Π does not achieve the infimum on the regret. Once this crucial lemma is established and one has dealt with the existence of the Rem_n remainder term, one could use similar arguments to those used in [11, Section 4] to control the regret (though, the proofs in our work are self-contained). We were not able to find a similar result in the classification literature that shows that Π Donsker suffices to attain a fast rate on misclassification error (the classification analogue of our regret). A result that makes a similarly weak complexity requirement (it only uses a Donsker condition) was given for general result for ERMs in [12, Theorem 4.5]. In particular, that result shows that, if Π is Donsker and $\overline{\Pi}^{\star}$ is a singleton, then one attains a fast rate. In the policy learning context, this is a weaker result than what we have stated: Π Donsker implies a fast rate, without any assumption on the cardinality of $\overline{\Pi}^{\star}$. We note also that our result could readily be extended to standard classification problems: the lack of a remainder term Rem_n only makes the problem easier.

3.2. *Tightening Theorem* 1

Tightening Theorem 1 would require careful consideration of the rate of convergence of two $o_P(n^{-1/2})$ terms, namely Rem_n and $\mathcal{E}_n$. On the one hand, if one uses a cross-validated estimator, then the rate of Rem_n is typically dominated by the rate of a doubly robust term, which is in turn upper bounded by the product of the $L^2(P)$ norm rates of convergence of the estimated action mechanism $(a, x) \mapsto P(A = a|X = x)$ and reward regression $(a, x) \mapsto \mathbb{E}(Y|A = a, X = x)$. It is thus sufficient to assume that this product is $o_P(n^{-1/2})$ to guarantee that $\mathrm{Rem}_n = o_P(n^{-1/2})$. It is worth noting that the product is typically $O_P(n^{-1})$ in a well-specified parametric model. On the other hand, the magnitude of $\mathcal{E}_n$ is controlled by both the size of class Π and the behavior of γ near the decision boundary, hence the need for a margin assumption like MA.

We do not believe there is a general ordering between the rate of convergence of Rem_n and $\mathcal{E}_n$ that applies across all problems, and so it does not appear that the size of Π nor the use of an efficient choice of influence function fully determines the rate of regret decay of $\widehat{\pi}$ under fixed-P asymptotics, even when $\mathrm{Rem}_n = o_P(n^{-1/2})$. A notable exception to this lack of strict ordering between Rem_n and $\mathcal{E}_n$ occurs when the action mechanism is known: in this case, the size of Π and the behavior of γ on the decision boundary fully control the rate of regret decay whenever a cross-validated estimator is used. We also note that, as far as we can tell, there does not appear to be any cost to using an efficient value function estimator when the model is nonparametric. It is not clear if using the efficient influence function is always preferred in more restrictive, semiparametric models: there may need to be a careful tradeoff between the efficiency of the influence function and the corresponding Rem_n.

3.3. *Relation to results of Athey and Wager* [1]

In a recent technical report [1], Athey and Wager showed that policy learning regret rates on the order of $O_P(n^{-1/2})$ are attainable by ERMs. Here we focus on the latest version (version 4) of that report. In that work, Athey and Wager derive high-probability regret upper bounds, with leading constants that scale with the standard error $S_*^{1/2}$ of an estimator for policy evaluation. When a semiparametric efficient estimator is used, their leading constant scales with $V_*^{1/2}$ (16). The results were shown to remain valid under uniform asymptotics in which the magnitude of the conditional average action effect changes with the sample size n. Section 4 of that report studies local asymptotics under a fixed policy class Π when the conditional average action effect decays at rate $n^{-1/2}$. There, the authors argue that, under these local asymptotics, the appearance of $S_*^{1/2}$ in the leading constant of their regret guarantee demonstrates the importance of using semiparametric efficient value estimators when defining a policy ERM, as compared to using inefficient value estimators such as those based on inverse probability weighting in settings where the propensity score is known. In contrast to the results of Athey and Wager, our findings in Theorems 1 and 2 do not indicate this improvement in performance of ERMs based on efficient value estimators over those based on inefficient value estimators. Specifically, Theorem 1 or, more precisely, its generalization in Theorem 6, can be used to show that the regret of an ERM based on an inverse probability weighted value estimator is $o_P(n^{-1/2})$ under sampling from a fixed distribution P. Our Theorem 2 shows that this ERM based on an inefficient value estimator will achieve a regret guarantee that scales with the efficient standard error $V_*^{1/2}$ times the $L^2(P)$-magnitude of the score indexing the perturbation of the fixed distribution P.

This apparent discrepancy between our result and those of Athey and Wager appears to occur for several reasons. In contrast to their result, our Theorem 1 is able to indicate faster than $n^{-1/2}$ rates for ERMs because it focuses on the fixed-P setting, whereas their result is uniform over a large collection of distributions. Our Theorem 2 is able to indicate a uniform local regret bound that scales with $V_*^{1/2}$ for a (possibly inefficient) ERM because the regret bound scales with the $L^2(P)$-magnitude of the perturbation ϕ of P. Notably, this bound can be made arbitrarily loose by simply increasing the $L^2(P)$-magnitude of ϕ. Nonetheless, this bound provides a tighter result than does the result of Athey and Wager when $\|\phi\|$ is small. This is not too surprising, since decreasing the $L^2(P)$-magnitude of ϕ has the effect of making the estimation problem behave more like the fixed-P setting, where Theorem 1 shows that faster than $n^{-1/2}$ rates are possible.

4. Extension of main fixed-P ERM result and two more examples

4.1. *Higher level result*

Suppose we observe $(O_1, \dots, O_n)$ drawn from a distribution ν_n, where each $O_i \equiv (X_i, A_i, Y_i)$ takes its values in $\mathcal{O} \equiv \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ with $\mathcal{A} \subset [-1, 1]$. Like in Section 1.3, $X_i \in \mathcal{X}$ denotes a vector of covariates describing the context preceding the ith action, $A_i \in \mathcal{A}$ denotes the action undertaken in this context and $Y_i \in \mathcal{Y}$ is the corresponding reward. Unlike in Section 1.3, there may be more than two actions and the observations are not necessarily i.i.d. This extends the setting introduced in Section 1.3, which can be recovered with $\mathcal{A} = \{-1, 1\}$ and ν_n equal to a product measure. Throughout we also assume that there exists a distribution P with support on $\mathcal{O}$ that will have to do with our limit process. The requirements of this distribution P will become clear in what follows and the worked examples of Sections 4.2 and 4.3.

Let Π be a class of policies, *i.e.*, a set of mappings from $\mathcal{X}$ to $\mathcal{A}$, and let $\overline{\Pi}$ be its $L^2(P)$ closure. We request that Π is not too large, in the sense that

$$\Pi \text{ is totally bounded w.r.t. } \|\cdot\|. \tag{17}$$

The value of a policy $\pi \in \overline{\Pi}$ is quantified via $\mathcal{V}(\pi)$. As in our earlier result, the regret is defined as $\mathcal{R}(\pi) \equiv \mathcal{V}^\star - \mathcal{V}(\pi)$, where $\mathcal{V}^\star \equiv \sup_{\pi \in \Pi} \mathcal{V}(\pi)$.

We also introduce the condition

$$\mathcal{V}(\cdot) \text{ is uniformly continuous on } \overline{\Pi} \text{ w.r.t. to } \|\cdot\|. \tag{18}$$

Let $\ell^\infty(\mathcal{F})$ denote the metric space of all bounded functions $z : \mathcal{F} \to \mathbb{R}$, equipped with the supremum norm. We assume that there exists a class $\{f_\pi : \pi \in \overline{\Pi}\} \subset L^2(P)$ of mappings from $\mathcal{O}$ to $[-M, M]$ (M not relying on π), an estimator $\{\widehat{\mathcal{V}}(\pi) : \pi \in \Pi\}$ of $\{\mathcal{V}(\pi) : \pi \in \Pi\}$ and a process $\widetilde{\mathbb{G}}_n \in \ell^\infty(\mathcal{F})$ such that, for a (possibly stochastic) rate $r_n \to +\infty$,

$$\pi \mapsto f_\pi \text{ is uniformly continuous from } \overline{\Pi} \text{ to } L^2(P), \quad \text{and} \tag{19}$$

$$\sup_{\pi \in \Pi} \left| r_n\big(\widehat{\mathcal{V}}(\pi) - \mathcal{V}(\pi)\big) - \widetilde{\mathbb{G}}_n f_\pi \right| = o_P(1). \tag{20}$$

In (19), both spaces are equipped with $\|\cdot\|$. In (20), $\widetilde{\mathbb{G}}_n \in \ell^\infty(\mathcal{F})$ is a stochastic process on $\mathcal{F} \equiv \{f_\pi : \pi \in \Pi\}$ that may or may not be equal to the empirical process, but must satisfy

$$\widetilde{\mathbb{G}}_n \rightsquigarrow \widetilde{\mathbb{G}}_P \text{ in } \ell^\infty(\mathcal{F}), \tag{21}$$

where almost all sample paths of $\widetilde{\mathbb{G}}_P$ are uniformly continuous w.r.t. $\|\cdot\|$.

Finally we also assume that, for the same rate r_n as in (20), $\widehat{\pi} \in \Pi$ satisfies the ERM property

$$\widehat{\mathcal{V}}(\widehat{\pi}) \geq \sup_{\pi \in \Pi} \widehat{\mathcal{V}}(\pi) - o_P\big(r_n^{-1}\big). \tag{22}$$

Like (6), (22) is a requirement on the behavior of the optimization algorithm on the realized sample, rather than a statistical or probabilistic condition. The following result generalizes Theorem 1.

Theorem 6 (More General ERM Result). *If* (17) *through* (22) *hold, then* $\mathcal{R}(\widehat{\pi}) = o_P(r_n^{-1})$.

A sketch of the proof is given in Appendix B. We only outline where the proof would deviate from that of Theorem 1.

Remark 7. If $\widetilde{\mathbb{G}}_P$ is equal to the zero-mean Gaussian process with covariance given by $\mathbb{E}[\mathbb{G}_P f \mathbb{G}_P g] = Pfg - Pf Pg$, then $\widetilde{\mathbb{G}}_P$ has almost surely uniformly continuous sample paths. In particular, [32, Example 1.5.10] shows that $\widetilde{\mathbb{G}}_P$ has almost surely uniformly continuous sample paths w.r.t. the standard deviation semimetric, and Section 2.1 in that same reference shows that, for bounded $\mathcal{F}$, one can replace this semimetric by that on $L^2(P)$.

In the remainder of this section, we give a taste of the broad applicability of Theorem 6 via two more examples. The first example is a simple extension of the framework of Section 2 that allows for more than two possible actions. The second example substitutes the median reward for the mean reward (for a similar setting, see [14]).

4.2. *Example* 1: *Maximizing the mean reward of a discrete action*

In this example, there are $\text{card}(\mathcal{A}) \in [2, \infty)$ (finitely many) candidate actions to undertake (think of a discretized dose in the personalized medicine framework). Without loss of generality, we assume that $\mathcal{A} \subset [0, 1]$. We assume that under the distribution P of $O \equiv (X, A, Y)$, the reward Y is uniformly bounded and there exists some $\delta > 0$ so that $\min_{a \in \mathcal{A}} P(A = a|X) \geq \delta$ with probability one. Here too, the value of a policy $\pi \in \overline{\Pi}$ is given by (1), the (within class Π) optimal value is $\mathcal{V}^\star \equiv \sup_{\pi \in \Pi} \mathcal{V}(\pi)$ and the regret of $\pi \in \overline{\Pi}$ is defined as in (2). Finally, we observe $O_1 \equiv (X_1, A_1, Y_1), \ldots, O_n \equiv (X_n, A_n, Y_n)$ drawn i.i.d. from P.

For each $\pi \in \overline{\Pi}$ and $o \in \mathcal{O}$, let $f_\pi(o)$ be given by (3). Let $\widetilde{\mathbb{G}}_n \equiv n^{1/2}(P_n - P) \in \ell^\infty(\mathcal{F})$ be the empirical process on $\mathcal{F} = \{f_\pi : \pi \in \Pi\}$. Note the equality between the LHS of (20) with $r_n \equiv n^{1/2}$ and $n^{1/2}\text{Rem}_n$ from (5).

We prove the next lemma in Section B.2:

Lemma 8. *If* Π *is Donsker, then* (17), (18), (19) *and* (21) *are met.*

Therefore, Theorem 6 yields the following corollary:

Corollary 9. *In the context of Section* 4.2, *suppose that* $\widehat{\pi}$ *satisfies* (22) *with* $r_n \equiv n^{1/2}$ *and that*

$$\sup_{\pi \in \Pi} \big| n^{1/2}\big(\widehat{\mathcal{V}}(\pi) - \mathcal{V}(\pi)\big) - \widetilde{\mathbb{G}}_n f_\pi \big| = o_P(1).$$

If Π *is Donsker, then* $\mathcal{R}(\widehat{\pi}) = o_P(n^{-1/2})$.

4.3. *Example* 2: *Maximizing the median reward of a binary action*

In this example, we use the same i.i.d. (from P) observed data structure as in Section 1.3, including the strong positivity assumption and the bounds on A and Y. For every policy $\pi \in \mathcal{P}$, define $F_\pi : \mathbb{R} \to \mathbb{R}$ pointwise by $F_\pi(m) \equiv \mathbb{E}[P(Y \leq m|A = \pi(X), X)]$. Under some causal assumptions, F_π is the cumulative distribution function of the reward in the world where action $\pi(x)$ is taken in each context $x \in \mathcal{X}$.

We define the value of π as the median rather than the mean reward, *i.e.*, as

$$\mathcal{V}(\pi) \equiv \inf\big\{m \in \mathbb{R} : 1/2 \leq F_\pi(m)\big\}. \tag{23}$$

Let $\Pi \subset \mathcal{P}$ be the class of candidate policies. Recall that the regret (within class Π) of $\pi \in \Pi$ takes the form $\mathcal{R}(\pi) \equiv \sup_{\pi \in \Pi} \mathcal{V}(\pi) - \mathcal{V}(\pi)$.

Let us now turn to (18). Assume that there exists $c > 0$ such that, for each $\pi \in \Pi$, F_π is continuously differentiable in the neighborhood $[\mathcal{V}(\pi) \pm c]$ with derivative $\dot{F}_\pi(m)$ at m in this neighborhood, where

$$0 < \inf_{\pi \in \overline{\Pi}} \inf_{m \in [\mathcal{V}(\pi) \pm c]} \dot{F}_\pi(m) \leq \sup_{\pi \in \overline{\Pi}} \sup_{m \in [\mathcal{V}(\pi) \pm c]} \dot{F}_\pi(m) < \infty. \tag{24}$$

In addition, assume that there exists a function $\omega : [0, \infty) \to [0, \infty)$ with $\lim_{m \downarrow 0} \omega(m) = \omega(0) = 0$ such that

$$\sup_{\pi \in \overline{\Pi}} \sup_{|u| \leq c} \left[\left| \dot{F}_\pi\big(\mathcal{V}(\pi) + u\big) - \dot{F}_\pi\big(\mathcal{V}(\pi)\big) \right| - \omega\big(|u|\big) \right] \leq 0. \tag{25}$$

Remark 10. A sufficient, but not necessary, condition for such an ω to exist is that F_π is twice continuously differentiable with the absolute range of the second derivative $\ddot{F}_\pi$ on $[\mathcal{V}(\pi) \pm c]$ bounded away from infinity uniformly in $\pi \in \overline{\Pi}$. Indeed, Taylor's theorem then shows that one can take

$$\omega(m) \equiv m \times \sup_{\pi \in \overline{\Pi}} \sup_{\tilde{m} \in [\mathcal{V}(\pi) \pm c]} \left| \ddot{F}_\pi(\tilde{m}) \right|.$$

The next lemma gives conditions under which (18) holds. Its proof is given in Appendix B.3.

Lemma 11. *If (24) and (25) are met, then (18) holds.*

For every $\pi \in \mathcal{P}$, $\mathcal{V}(\pi)$ defined in (23) can be viewed as the evaluation at P of the real-valued functional

$$P' \mapsto \inf\big\{m \in \mathbb{R} : 1/2 \leq \mathbb{E}_{P'}\big[P'\big(Y \leq m | A = \pi(X), X\big)\big]\big\}$$

from the nonparametric model of distributions P' satisfying the same constraints as P. This functional is pathwise differentiable at P relative to the maximal tangent space with an efficient influence function f_π given by

$$\begin{aligned} f_\pi(o) \equiv{}& \frac{\mathbf{1}\{a = \pi(x)\}}{P(A = a | X = x)\dot{F}_\pi(\mathcal{V}(\pi))} \big[\mathbf{1}\big\{y \leq \mathcal{V}(\pi)\big\} - P\big\{Y \leq \mathcal{V}(\pi) | A = a, X = x\big\}\big] \\ &+ \frac{P\{Y \leq \mathcal{V}(\pi) | A = \pi(x), X = x\} - 1/2}{\dot{F}_\pi(\mathcal{V}(\pi))}. \end{aligned} \tag{26}$$

Observe that above $\dot{F}_\pi(\mathcal{V}(\pi))$ only enters f_π as a multiplicative constant, and therefore an estimating equation-based estimator or TMLE for $\mathcal{V}(\pi)$ can be asymptotically linear for $\mathcal{V}(\pi)$ without estimating $\dot{F}_\pi$. We, in particular, suppose that we have an estimator satisfying (20) with $r_n = n^{1/2}$ and $\widetilde{\mathbb{G}}_n \equiv n^{1/2}(P_n - P) \in \ell^\infty(\mathcal{F})$ with $\mathcal{F} \equiv \{f_\pi : \pi \in \Pi\}$, see Remark 13 at the end of the present section.

With this choice of $\widetilde{\mathbb{G}}_n$, (21) is met and Π Donsker yields (17), as seen in the proof of Theorem 1 (the same argument applies because the range of each $\pi \in \mathcal{P}$ is bounded in $[-1, 1]$). Theorem 6 yields the following corollary:

Corollary 12. *In the context of Section 4.3, suppose that (24) and (25) are met. Let $\widehat{\pi}$ satisfy (22) with $r_n \equiv n^{1/2}$ and*

$$\sup_{\pi \in \Pi} \left| n^{1/2}\big(\widehat{\mathcal{V}}(\pi) - \mathcal{V}(\pi)\big) - \widetilde{\mathbb{G}}_n f_\pi \right| = o_P(1).$$

If Π is Donsker and (19) holds, then $\mathcal{R}(\widehat{\pi}) = o_P(n^{-1/2})$.

Since Π is Donsker, (19) can be derived under regularity conditions by using essentially the same techniques as in the proof of Theorem 1. A slight modification to Lemma A.5 is needed. The main regularity condition consists in assuming that the real-valued function $\pi \mapsto \dot{F}_\pi(\mathcal{V}(\pi))$ over $\overline{\Pi}$ equipped with $\|\cdot\|$ is uniformly continuous, see Corollary B.2 in Appendix B. It is tedious, though not difficult, to complement this main regularity condition with secondary conditions. It would suffice to restrict the (uniform in π) behavior for all $P\{Y \leq v | A = \pi(x), X = x\}$ across all real v.

Remark 13. One can, for example, establish conditions under which the estimating equation-based estimator defined as a solution in v to

$$\frac{\mathbf{1}\{a = \pi(x)\}[\mathbf{1}\{Y \leq v\} - \widehat{P}\{Y \leq v | A = \pi(X), X\}]}{\widehat{P}(A = a | X = x)} + \widehat{P}\big\{Y \leq v | A = \pi(X), X\big\} = 1/2,$$

satisfies (20). In the above display, $\widehat{P}$ denotes an estimate of (certain conditional probabilities under) P. One could use cross-validated estimating equations to avoid the need for any empirical process conditions on $\widehat{P}$.

5. Discussion

We have presented fast rates of regret decay in optimal policy estimation, *i.e.*, rates of decay that are faster than the rate of decay of the standard error of an efficient estimator of the value of any given policy in the candidate class. Our method of proof for our primary result, Theorem 1, leverages the fact that the empirical process over a Donsker class converges in distribution to a Gaussian process, and that the sample paths of this limiting process are (almost surely) uniformly continuous. The downside of our analysis, namely passing to the limit and then studying the behavior of the limiting process, is that it does not appear to allow one to obtain a faster than $o_P(n^{-1/2})$ rate of convergence for the regret.

It would be of interest to replace our limiting argument by finite sample results that would allow one to exploit the finite sample equicontinuity of the empirical process to demonstrate faster rates. Nonetheless, we note that the problem-dependent margin condition will often have a major impact on the extent to which the rate can be improved. Furthermore, as we discussed in Section 3.2, the existence of the remainder term Rem_n necessitates a careful consideration of whether or not the empirical process term represents the dominant error term, or whether the second-order remainder that appears due to the non-linearity of the value parameter represents the dominant error term. In restricted, semiparametric models, we suspect that, depending on the underlying margin, an inefficient estimator of the value may yield a faster rate of regret decay than an efficient estimator in the fixed-P setting. Despite this surprising phenomenon, we continue to advocate the use of first-order efficient value estimators.

Appendix A: Proofs of main ERM results

A.1. *Proof of fast rate of convergence at fixed P (Theorem 1)*

We begin this section with six lemmas and a corollary. The proofs of Lemmas A.1, through A.4, Lemma A.6 and Corollary A.7 only require results from functional analysis. The proofs of Lemma A.5 and Theorem 1 use results from empirical process theory.

In Lemmas A.1 through A.4 and Lemma A.6 to follow, all topological results make use of the strong topology on $L^2(P)$. A set $S \subset L^2(P)$ is called totally bounded if, for every $\epsilon > 0$, there exists a finite collection of radius-ϵ open balls that covers S. When we refer to convergence in these lemmas, we refer to convergence w.r.t. $\|\cdot\|$. A set S is sequentially compact if every sequence of elements of S has a convergent subsequence. A set $S \subset L^2(P)$ is compact if every open cover has a finite subcover. Sequential compactness and compactness are equivalent for subsets of metric spaces. Lemma A.6 and the proof of Theorem 1 use several additional definitions, which are given immediately before Lemma A.6.

Lemma A.1. *If Π is totally bounded, then $\overline{\Pi}$ is compact.*

Proof. The Hilbert space $L^2(P)$ is a complete metric space. Therefore, [4, Corollary 6.65] implies that Π is relatively compact. In other words, $\overline{\Pi}$ is compact. □

For brevity, introduce $Q(A, X) \equiv \mathbb{E}(Y|A, X)$. Note that $\gamma(X) = Q(1, X) - Q(-1, X)$.

Lemma A.2. *Both $\mathcal{V}$ and $\mathcal{R}$ are continuous on $\overline{\Pi}$.*

Proof. The key to this proof is the following simple remark: for every policy $\pi : \mathcal{X} \to \{-1, 1\}$,

$$2Q\big(\pi(X), X\big) = \big(1 + \pi(X)\big)Q(1, X) + \big(1 - \pi(X)\big)Q(-1, X). \tag{A.1}$$

Choose arbitrarily $\pi_1, \pi_2 \in \overline{\Pi}$. By (A.1) and the Cauchy–Schwarz inequality, it holds that

$$2\big|\mathcal{V}(\pi_1) - \mathcal{V}(\pi_2)\big| = \big|\mathbb{E}\big[\gamma(X)(\pi_1 - \pi_2)(X)\big]\big| \leq \|\gamma\| \times \|\pi_1 - \pi_2\|. \tag{A.2}$$

This proves the continuity of $\mathcal{V}$. That of $\mathcal{R}$ follows immediately. □

Lemma A.3. *If Π is totally bounded, then* $\inf_{\pi\in\overline{\Pi}} \mathcal{R}(\pi) = 0$, $\overline{\Pi}^{\star} \equiv \{\pi^{\star} \in \overline{\Pi} : \mathcal{R}(\pi^{\star}) = 0\}$ *is equal to* $\{\pi^{\star} \in \overline{\Pi} : \mathcal{R}(\pi^{\star}) \leq 0\}$, *and* $\overline{\Pi}^{\star} \neq \varnothing$.

Proof. By Lemma A.1, $\overline{\Pi}$ is compact. By Lemma A.2, $\mathcal{R}$ is continuous. Thus, $\mathcal{R}$ admits and achieves a minimum $\mathcal{R}(\bar{\pi}) = \inf_{\pi\in\overline{\Pi}} \mathcal{R}(\pi)$ on $\overline{\Pi}$. Since $\inf_{\pi\in\Pi} \mathcal{R}(\pi) = 0$, we know that $\mathcal{R}(\bar{\pi}) \leq 0$. In fact, a contradiction argument shows that $\mathcal{R}(\bar{\pi}) = 0$. In other words, $\overline{\Pi}^\star = \{\pi^\star \in \overline{\Pi} : \mathcal{R}(\pi^\star) \leq 0\}$ and $\bar{\pi} \in \overline{\Pi}^\star$, hence $\overline{\Pi}^\star \neq \varnothing$.

Indeed, assume that $\mathcal{R}(\bar{\pi}) < 0$. Because $\bar{\pi} \in \overline{\Pi}$, there exists a sequence $\{\pi_m\}_{m\geq 1}$ of elements of Π such that $\|\pi_m - \bar{\pi}\| \to 0$. By continuity of $\mathcal{R}$ (see Lemma A.2), $\mathcal{R}(\pi_m) \to \mathcal{R}(\bar{\pi}) < 0$. Thus, $\inf_{\pi\in\Pi} \mathcal{R}(\pi) < 0$. Contradiction. □

Lemma A.4. *If* Π *is totally bounded, then*

$$\limsup_{r\downarrow 0} \sup_{\pi\in\overline{\Pi}:\mathcal{R}(\pi)\leq r} \inf_{\pi^\star\in\overline{\Pi}^\star} \left\|\pi^\star - \pi\right\| = 0. \tag{CA}$$

Proof. We argue by contraposition. Suppose CA does not hold. Then there exists a sequence $\{\pi_m\}_{m\geq 1}$ of elements of $\overline{\Pi}$ and $t > 0$ such that $\mathcal{R}(\pi_m) \to 0$ and, for all $m \geq 1$,

$$\inf_{\pi^\star\in\overline{\Pi}^\star} \left\|\pi_m - \pi^\star\right\| > t. \tag{A.3}$$

We now give a contradiction argument to show that $\{\pi_m\}_{m\geq 1}$ does not have a convergent subsequence. Suppose there exists a subsequence $\{\pi_{m_k}\}_{k\geq 1}$ such that $\|\pi_{m_k} - \pi_\infty\| \to 0$ for some $\pi_\infty \in L^2(P)$. Now, note that *(i)* $\pi_\infty \in \overline{\Pi}$, the closure of Π, *(ii)* $\mathcal{R}(\pi_{m_k}) \to 0$, and *(iii)* $\mathcal{R}(\pi_{m_k}) \to \mathcal{R}(\pi_\infty)$ by Lemma A.2. Consequently, $\mathcal{R}(\pi_\infty) = 0$. Since $\pi_\infty \in \overline{\Pi}$, this reveals that $\pi_\infty \in \overline{\Pi}^\star$ and

$$\inf_{\pi^\star\in\overline{\Pi}^\star} \left\|\pi_{m_k} - \pi^\star\right\| \leq \|\pi_{m_k} - \pi_\infty\| \to 0,$$

in contradiction with (A.3). Thus, there does not exist a convergent subsequence $\{\pi_{m_k}\}_{k\geq 1}$ of $\{\pi_m\}_{m\geq 1}$, completing the contradiction argument.

We now return to the contraposition argument. The existence of a sequence $\{\pi_m\}_{m\geq 1}$ of elements of $\overline{\Pi}$ not having a convergent subsequence implies that $\overline{\Pi}$ is not sequentially compact and, therefore, that it is not compact. By Lemma A.1, Π is not totally bounded. This completes the proof. □

Lemma A.5. *If* Π *is Donsker, then* $\mathcal{F}$ *is also Donsker.*

Proof. Similar to (A.1), the key is the following remark: for every policy $\pi : \mathcal{X} \to \{-1, 1\}$,

$$\begin{aligned} 2f_\pi(O) = {} & \left|A + \pi(X)\right| \frac{Y - Q(A,X)}{P(A|X)} \\ & + \big(1 + \pi(X)\big)Q(1,X) + \big(1 - \pi(X)\big)Q(-1,X) - 2\mathcal{V}(\pi). \end{aligned} \tag{A.4}$$

For future use, we first note that (A.2) implies, for any $\pi_1, \pi_2 \in \Pi$, that

$$\left|f_{\pi_1}(O) - f_{\pi_2}(O)\right| \lesssim \left|\pi_1(X) - \pi_2(X)\right| + \|\pi_1 - \pi_2\|$$

hence

$$\|f_{\pi_1} - f_{\pi_2}\| \lesssim \|\pi_1 - \pi_2\|. \tag{A.5}$$

Introduce $\psi : \mathbb{R}^5 \to \mathbb{R}$ given by $2\psi(u) = u_1|u_2 + u_3| + (1 + u_3)u_4 + (1 + u_3)u_5$ and let f_1, f_2, f_4, f_5 be the functions given by $f_1(o) \equiv (y - Q(a,x))/P(A = a|X = x)$, $f_2(o) \equiv x$, $f_4(o) \equiv Q(1,x)$ and $f_5(o) \equiv Q(-1,x)$. Let $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_4$, $\mathcal{F}_5$ be the singletons $\{f_1\}$, $\{f_2\}$, $\{f_4\}$ and $\{f_5\}$, each of them a Donsker class. Let $\tilde{\mathcal{F}} \equiv \mathcal{F}_1 \times \mathcal{F}_2 \times \Pi \times \mathcal{F}_4 \times \mathcal{F}_5$ and note that $\psi \circ \mathbf{f} = \tilde{f}_\pi$ if $\mathbf{f} \in \tilde{\mathcal{F}}$ writes as $\mathbf{f} = (f_1, f_2, \pi, f_4, f_5)$. In light of (A.4), observe now that, for every $\mathbf{f}_1 = (f_1, f_2, \pi_1, f_4, f_5)$, $\mathbf{f}_2 = (f_1, f_2, \pi_2, f_4, f_5) \in \tilde{\mathcal{F}}$, it holds that

$$\left|\psi \circ \mathbf{f}_1(o) - \psi \circ \mathbf{f}_2(o)\right| \lesssim \left|\pi_1(x) - \pi_2(x)\right|$$

(the bound on Y implies that $\|\gamma\|_\infty$ is finite). By [32, Theorem 2.10.6], of which the conditions are obviously met, $\psi \circ \tilde{\mathcal{F}} = \{\tilde{f}_\pi : \pi \in \Pi\}$ is Donsker. Because $\Lambda \equiv \{\mathcal{V}(\pi) : \pi \in \Pi\}$ (viewed as a set of constant functions with a uniformly bounded sup-norm) is also Donsker, [32, Example 1.10.7] it follows that $\{\tilde{f}_\pi - \lambda : \pi \in \Pi, \lambda \in \Lambda\}$ is Donsker, and so is its subset $\mathcal{F}$. □

Recall the definition (4) of $\mathcal{F}$. The next lemma uses the following definition: for any $r > 0$, $\Pi_r \equiv \{\pi \in \Pi : \mathcal{R}(\pi) \leq r\}$. If $\overline{\Pi}^\star$ is nonempty (for instance, if Π is totally bounded by Lemma A.3) we let, for any $\pi^\star \in \overline{\Pi}^\star$ and $s > 0$, $B_\Pi(\pi^\star, s) \equiv \{\pi \in \Pi : \|\pi^\star - \pi\| \leq s\}$ denote the intersection of the radius-s $L^2(P)$ ball centered at $\pi^\star$ and the collection Π. Because $\overline{\Pi}^\star \subset \overline{\Pi}$, $B_\Pi(\pi^\star, s)$ is nonempty.

Lemma A.6. *Define* $g : \ell^\infty(\mathcal{F}) \times (0, \infty) \to \mathbb{R}$ *as*

$$g(z, r) \equiv \sup_{\pi \in \Pi_r} z(f_\pi) - \limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi).$$

Let $z \in \ell^\infty(\mathcal{F})$ *be* $\|\cdot\|$*-uniformly continuous and let* $\{(z_m, r_m)\}_{m \geq 1}$ *be a sequence with values in* $\ell^\infty(\mathcal{F}) \times (0, \infty)$ *such that* $\sup_{f \in \mathcal{F}} |z_m(f) - z(f)| + |r_m| \to 0$. *If* Π *is totally bounded, then*

$$\limsup_{m \to \infty} g(z_m, r_m) \leq 0.$$

The following corollary will prove useful.

Corollary A.7. *Recall the definition of g from Lemma A.6. Let* $h : \ell^\infty(\mathcal{F}) \times [0, \infty) \to \mathbb{R}$ *be such that, for all* $z \in \ell^\infty(\mathcal{F})$,

$$h(z, r) \equiv \max\bigl(g(z, r), 0\bigr) \text{ if } r > 0 \text{ and } h(z, 0) \equiv 0.$$

Let $z \in \ell^\infty(\mathcal{F})$ *be* $\|\cdot\|$*-uniformly continuous. If* Π *is totally bounded, then h is continuous at* $(z, 0)$.

Proof of Lemma A.6 and Corollary A.7. Fix a sequence $\{(z_m, r_m)\}_{m \geq 1}$ satisfying the conditions of the Lemma A.6. Observe that, for every $m \geq 1$,

$$\begin{aligned}\bigl|g(z_m, r_m) - g(z, r_m)\bigr| &\leq \Bigl|\sup_{\pi \in \Pi_{r_m}} z_m(f_\pi) - \sup_{\pi \in \Pi_{r_m}} z(f_\pi)\Bigr| \\ &\quad + \Bigl|\limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z_m(f_\pi) - \limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi)\Bigr| \\ &\leq 2 \sup_{\pi \in \Pi} \bigl|z_m(f_\pi) - z(f_\pi)\bigr|\end{aligned}$$

where the above RHS expression is $o(1)$ because $z_m \to z$ in $\ell^\infty(\mathcal{F})$. Therefore,

$$\begin{aligned}g(z_m, r_m) &= g(z_m, r_m) - g(z, r_m) + g(z, r_m) \\ &= g(z_m, r_m) - g(z, r_m) + \Bigl[\sup_{\pi \in \Pi_{r_m}} z(f_\pi) - \limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi)\Bigr]. \\ &\leq \Bigl[\sup_{\pi \in \Pi_{r_m}} z(f_\pi) - \limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi)\Bigr] + o(1). \end{aligned} \tag{A.6}$$

Let us show now that the RHS expression in (A.6) is $o(1)$.

For any $r > 0$, there exists a $\pi_r \in \Pi_r$ such that

$$\sup_{\pi \in \Pi_r} z(f_\pi) \leq z(f_{\pi_r}) + r. \tag{A.7}$$

Furthermore, there exists a $\pi_r^\star \in \overline{\Pi}^\star$ such that

$$\bigl\|\pi_r^\star - \pi_r\bigr\| \leq \inf_{\pi^\star \in \overline{\Pi}^\star} \bigl\|\pi^\star - \pi_r\bigr\| + r \leq \sup_{\pi \in \Pi_r} \inf_{\pi^\star \in \overline{\Pi}^\star} \bigl\|\pi^\star - \pi\bigr\| + r. \tag{A.8}$$

Likewise, there exists $\tilde{\pi}_r \in B_\Pi(\pi_r^\star, r)$ such that

$$\begin{aligned}z(f_{\tilde{\pi}_r}) &\leq \inf_{\pi \in B_\Pi(\pi_r^\star, r)} z(f_\pi) + r \leq \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, r)} z(f_\pi) + r \\ &= \Bigl(\limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi) + o_r(1)\Bigr) + r,\end{aligned} \tag{A.9}$$

where the above equality holds by the definition of the limit superior (the $o_r(1)$ above represents the term's behavior as $r \to 0$). In light of (A.7) and (A.9), we thus have

$$\sup_{\pi \in \Pi_r} z(f_\pi) - \limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi) \le z(f_{\pi_r}) - z(f_{\tilde{\pi}_r}) + 2r + o_r(1). \tag{A.10}$$

By the $\|\cdot\|$-uniform continuity of z, $|z(f_{\pi_r}) - z(f_{\tilde{\pi}_r})| = o_r(1)$ if $\|f_{\pi_r} - f_{\tilde{\pi}_r}\| = o_r(1)$. Let us show that the latter condition is met. Because $\tilde{\pi}_r \in B_\Pi(\pi_r^\star, r)$, the triangle inequality, (A.5) and (A.8) imply that

$$\begin{aligned} \|f_{\pi_r} - f_{\tilde{\pi}_r}\| &\le \|f_{\pi_r} - f_{\pi_r^\star}\| + \|f_{\pi_r^\star} - f_{\tilde{\pi}_r}\| \\ &\lesssim \|\pi_r - \pi_r^\star\| + \|\pi_r^\star - \tilde{\pi}_r\| \le \sup_{\pi \in \Pi_r} \inf_{\pi^\star \in \overline{\Pi}^\star} \|\pi^\star - \pi\| + 2r. \end{aligned} \tag{A.11}$$

Because $\Pi_r \subset \{\pi \in \overline{\Pi} : \mathcal{R}(\pi) \le r\}$, Lemma A.4 (which applies because Π is totally bounded) implies that

$$\limsup_{r \downarrow 0} \sup_{\pi \in \Pi_r} \inf_{\pi^\star \in \overline{\Pi}^\star} \|\pi - \pi^\star\| = 0,$$

from which we deduce that the RHS of (A.11) is $o_r(1)$. In summary, $\|f_{\pi_r} - f_{\tilde{\pi}_r}\| = o_r(1)$, hence $|z(f_{\pi_r}) - z(f_{\tilde{\pi}_r})| = o_r(1)$ and consequently, by (A.10),

$$\sup_{\pi \in \Pi_r} z(f_\pi) - \limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi) = o_r(1).$$

Taking $r = r_m$ and using the above result reveals that the RHS expression in (A.6) is $o(1)$ indeed. This completes the proof of Lemma A.6.

In the context of Corollary A.7, let $\{(z_m, r_m)\}_{m \ge 1}$ be a sequence with values in $\ell^\infty(\mathcal{F}) \times [0, \infty)$ and such that $\sup_{f \in \mathcal{F}} |z_m(f) - z(f)| + |r_m| \to 0$. Lemma A.6 implies that $h(z_m, r_m) \to h(z, 0) \equiv 0$. Since the sequence was arbitrarily chosen, h is indeed continuous at $(z, 0)$ and the proof of Corollary A.7 is complete. □

Proof of Theorem 1. In this proof, we make the dependence of $\widehat{\pi}$ on n explicit by writing $\widehat{\pi}_n$. By Lemma A.5, Π Donsker implies $\mathcal{F}$ Donsker. Consider the empirical process $\mathbb{G}_n$ as the element of $\ell^\infty(\mathcal{F})$ characterized by $\mathbb{G}_n f \equiv n^{1/2}(P_n - P)f$ for every $f \in \mathcal{F}$. We let $\mathbb{G}_P \in \ell^\infty(\mathcal{F})$ denote the zero-mean Gaussian process with covariance given by $\mathbb{E}[\mathbb{G}_P f \mathbb{G}_P g] = Pfg - PfPg$.

Since Π Donsker implies Π totally bounded, Lemma A.3 guarantees the existence of a $\pi^\star \in \overline{\Pi}^\star$. For any $s > 0$ and $\pi_s^\star \in B_\Pi(\pi^\star, s) \subset \Pi$, (6) then (5) combined with (A.2) yield in turn the first and second inequalities below:

$$\begin{aligned} 0 \le \mathcal{R}(\widehat{\pi}_n) &= \mathcal{V}^\star - \mathcal{V}(\widehat{\pi}_n) = \big[\mathcal{V}(\pi_s^\star) - \mathcal{V}(\widehat{\pi}_n)\big] + \big[\mathcal{V}(\pi^\star) - \mathcal{V}(\pi_s^\star)\big] \\ &= (\widehat{\mathcal{V}} - \mathcal{V})(\widehat{\pi}_n) - (\widehat{\mathcal{V}} - \mathcal{V})(\pi_s^\star) + \big[\widehat{\mathcal{V}}(\pi_s^\star) - \widehat{\mathcal{V}}(\widehat{\pi}_n)\big] + \big[\mathcal{V}(\pi^\star) - \mathcal{V}(\pi_s^\star)\big] \\ &\le (\widehat{\mathcal{V}} - \mathcal{V})(\widehat{\pi}_n) - (\widehat{\mathcal{V}} - \mathcal{V})(\pi_s^\star) + \big[\mathcal{V}(\pi^\star) - \mathcal{V}(\pi_s^\star)\big] + o_P(\mathrm{Rem}_n) \\ &\le n^{-1/2}[\mathbb{G}_n f_{\widehat{\pi}_n} - \mathbb{G}_n f_{\pi_s^\star}] + s + \big[2 + o_P(1)\big]\mathrm{Rem}_n \qquad \text{(A.12)} \\ &\le 2n^{-1/2} \sup_{f \in \mathcal{F}} |\mathbb{G}_n f| + s + \big[2 + o_P(1)\big]\mathrm{Rem}_n \qquad \text{(A.13)} \end{aligned}$$

(we already derived the three first lines in Section 2.2 to develop an intuition about Theorem 1).

Since $z \mapsto \sup_{f \in \mathcal{F}} |z(f)|$ is continuous and $\mathcal{F}$ is Donsker, the continuous mapping theorem [32, Theorem 1.3.6] implies that the leftmost term in the above RHS sum is $O_P(n^{-1/2})$. Therefore, (A.13) and $\mathrm{Rem}_n = o_P(n^{-1/2})$ imply

$$0 \le \mathcal{R}(\widehat{\pi}_n) \lesssim s + O_P\big(n^{-1/2}\big) + \big[1 + o_P(1)\big]o_P\big(n^{-1/2}\big)$$

where the random terms do not depend on s. By letting s go to zero, we obtain $\mathcal{R}(\widehat{\pi}_n) = O_P(n^{-1/2})$. The remainder of the proof tightens this result to $\mathcal{R}(\widehat{\pi}_n) = o_P(n^{-1/2})$.

Let us go back to (A.12). Since $\mathrm{Rem}_n = o_P(n^{-1/2})$, it also yields the tighter bound

$$0 \le \mathcal{R}(\widehat{\pi}_n) \lesssim n^{-1/2} \liminf_{s \downarrow 0} \inf_{\pi^\star \in \overline{\Pi}^\star} \sup_{\pi_s^\star \in B_\Pi(\pi^\star, s)} [\mathbb{G}_n f_{\widehat{\pi}_n} - \mathbb{G}_n f_{\pi_s^\star}] + \big[2 + o_P(1)\big]o_P\big(n^{-1/2}\big). \tag{A.14}$$

Let g be defined as in Lemma A.6. Note that the second term in the RHS of (A.14) does not depend on $\widehat{\pi}_n$. Let $\{t_n\}_{n\geq 1}$ be a sequence with positive values such that $t_n \downarrow 0$. As $\widehat{\pi}_n$ trivially falls in $\Pi_{\mathcal{R}(\widehat{\pi}_n)+t_n}$, we can take a supremum over $\pi \in \Pi_{\mathcal{R}(\widehat{\pi}_n)+t_n}$. Multiplying both sides of (A.14) by $n^{1/2}$, we see that

$$\begin{aligned}0 \leq n^{1/2}\mathcal{R}(\widehat{\pi}_n) &\leq \sup_{\pi\in\Pi_{\mathcal{R}(\widehat{\pi}_n)+t_n}} \mathbb{G}_n f_\pi - \limsup_{s\downarrow 0}\sup_{\pi^\star\in\overline{\Pi}^\star}\inf_{\pi\in B_\Pi(\pi^\star,s)} \mathbb{G}_n(f_\pi) + o_P(1)\\ &= g\big(\mathbb{G}_n, \mathcal{R}(\widehat{\pi}_n)+t_n\big) + o_P(1).\end{aligned}$$

Above we used $\mathcal{R}(\widehat{\pi}_n)+t_n$ rather than $\mathcal{R}(\widehat{\pi}_n)$ to avoid separately handling the cases where $\overline{\Pi}^\star \cap \Pi$ is and is not empty.

The conclusion is at hand. Recall the definition of h from Corollary A.7. Clearly the previous display yields the bounds

$$0 \leq n^{1/2}\mathcal{R}(\widehat{\pi}_n) \leq h\big(\mathbb{G}_n, \mathcal{R}(\widehat{\pi}_n)+t_n\big) + o_P(1). \tag{A.15}$$

We have already established that $\mathcal{R}(\widehat{\pi}_n) = o_P(1)$, hence $0 < t_n \leq \mathcal{R}(\widehat{\pi}_n) + t_n = o_P(1)$ as well. Because $\mathcal{F}$ is Donsker, $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$ in distribution on $\ell^\infty(\mathcal{F})$. Therefore, $(\mathbb{G}_n, \mathcal{R}(\widehat{\pi}_n)+t_n) \rightsquigarrow (\mathbb{G}_P, 0)$ in distribution on $\ell^\infty(\mathcal{F}) \times [0,\infty)$. Almost all sample paths of $\mathbb{G}_P$ are uniformly continuous on $\mathcal{F}$ w.r.t. $\|\cdot\|$ [32, Section 2.1], so Corollary A.7 applies almost surely and the continuous mapping theorem yields $h(\mathbb{G}_n, \mathcal{R}(\widehat{\pi}_n)+t_n) \rightsquigarrow h(\mathbb{G}_P, 0)$ in distribution in $\ell^\infty(\mathcal{F})$. This convergence also occurs in probability since the limit is almost surely constant (zero). By (A.15), $0 \leq \mathcal{R}(\widehat{\pi}_n) = o_P(n^{-1/2})$. This completes the proof. □

A.2. *Proof of uniform local convergence results* (*Theorem* 2)

Eight lemmas precede the proof of Theorem 2. Except for one lemma, the statements are followed by the proofs. The long proof of Lemma A.12 is postponed at the end of the section.

Lemma A.8. *Let $b \equiv 3\max_{j\in\{1,2,3\}} b_j$ and $\sigma^2 \equiv \max_{j\in\{1,2,3\}}\sigma_j^2$. For all $\phi = (\phi_1,\phi_2,\phi_3) \in \Phi$ and $\bar{\phi}(O) \equiv \phi_1(X) + \phi_2(A,X) + \phi_3(O)$, the following sub-exponential property is satisfied*:

$$\log\mathbb{E}\big[\exp\{\lambda\bar{\phi}(O)\}\big] \leq \frac{\lambda^2\sigma^2}{2} \quad \textit{for all } |\lambda| \leq 1/b. \tag{A.16}$$

Consequently, $\sup_{\phi\in\Phi}\|\bar{\phi}\|$ *is finite and there exists a constant $c > 0$ that depends only on b and σ^2 such that*

$$\log\mathbb{E}\big[\big(\bar{\phi}(O)^2 + \bar{\phi}(O)^4\big)\exp\{\lambda\bar{\phi}(O)\}\big] \leq c + \lambda^2\sigma^2 \quad \textit{for all } |\lambda| \leq 1/2b. \tag{A.17}$$

Proof. For all λ with $|\lambda| < 1/b$, the convexity of the exponential function yields

$$\mathbb{E}\big[\exp\{\lambda\bar{\phi}(O)\}\big] \leq \frac{1}{3}\sum_{j=1}^{3}\mathbb{E}\big[\exp\{3\lambda\phi_j(O)\}\big] \leq \max_{j\in\{1,2,3\}}\mathbb{E}\big[\exp\{3\lambda\phi_j(O)\}\big]. \tag{A.18}$$

By the monotonicity of the logarithm, this shows that

$$\log\mathbb{E}\big[\exp\{\lambda\bar{\phi}(O)\}\big] \leq \max_{j\in\{1,2,3\}}\log\mathbb{E}\big[\exp\{3\lambda\phi_j(O)\}\big].$$

Using that $|\lambda| < 1/b$ implies that $3|\lambda| < 1/b_j$ for all $j \in \{1,2,3\}$, (9), (10), (11), the tower rule and (A.18) imply that

$$\log\mathbb{E}\big[\exp\{\lambda\bar{\phi}(O)\}\big] \leq \max_{j\in\{1,2,3\}}\frac{\lambda^2\sigma_j^2}{2} = \frac{\lambda^2\sigma^2}{2},$$

thus proving (A.16).

To prove that $\sup_{\phi\in\Phi}\|\bar{\phi}\|$ is finite and (A.17), introduce $U \equiv \bar{\phi}(O)/b$, note that

$$\frac{U^2}{2!} + \frac{U^4}{4!} + \frac{U^8}{8!} \leq \frac{e^U + e^{-U}}{2},$$

and recall that both $\mathbb{E}[\exp\{U\}]$ and $\mathbb{E}[\exp\{-U\}]$ are upper bounded by $\exp\{\sigma^2/2b^2\}$. Therefore, $\mathbb{E}[\bar{\phi}(O)^2] = \|\bar{\phi}\|^2$, $\mathbb{E}[\bar{\phi}(O)^4]$ and $\mathbb{E}[\bar{\phi}(O)^8]$ are upper bounded by $2b^2\exp\{\sigma^2/2b^2\}$, $(4!)b^4\exp\{\sigma^2/2b^2\}$ and $(8!)b^8\exp\{\sigma^2/2b^2\}$, respectively, all uniformly w.r.t. $\phi \in \Phi$. Then (A.17) follows from the Cauchy–Schwarz and triangle inequalities and (A.16). □

Lemma A.9. *For all $\phi \in \Phi$ and $\epsilon \in \mathbb{R}$, $|\epsilon| \leq 1/\max\{b_1, b_2, b_3\}$,*

$$0 < c_{1,\phi}(\epsilon), c_{2,\phi}(\epsilon|X), c_{3,\phi}(\epsilon|A, X) \leq 1$$

P-almost surely. Moreover, P-almost surely,

$$\lim_{\epsilon\to 0} \inf_{\phi\in\Phi} \min\bigl\{c_{1,\phi}(\epsilon), c_{2,\phi}(\epsilon|X), c_{3,\phi}(\epsilon|A, X)\bigr\} = 1.$$

Proof. We show that $0 < c_{2,\phi}(\epsilon|X) \leq 1$ and $\lim_{\epsilon\to 0} \inf_{\phi\in\Phi} c_{2,\phi}(\epsilon|X)$ P-almost surely. The proofs of the counterparts to that result for $c_{1,\phi}$ and $c_{3,\phi}$ are analogous.

Fix $\phi \in \Phi$ and $\epsilon \in \mathbb{R}$, $|\epsilon| \leq 1/\max\{b_1, b_2, b_3\}$. It is easy to see that $c_{2,\phi}(\epsilon|X)$ is well-defined and positive P-almost surely (indeed, $\mathbb{E}[\exp\{\epsilon\phi_2(A, X)\}|X]$ is finite by (10), and it is positive because $\exp\{\epsilon\phi_2(A, X)\}$ is positive — the three statements are meant P-almost surely), and therefore $c_{2,\phi}(\epsilon|X)^{-1}$ is well-defined P-almost surely. By Jensen's inequality, and because $\mathbb{E}[\phi_2(A, X)|X] = 0$ P-almost surely,

$$c_{2,\phi}(\epsilon|X)^{-1} = \mathbb{E}\bigl[\exp\bigl\{\epsilon\phi_2(A, X)\bigr\}|X\bigr] \geq \exp\bigl\{\epsilon\mathbb{E}\bigl[\phi_2(A, X)|X\bigr]\bigr\} = 1$$

with P-probability one. Moreover, in view of (10),

$$\sup_{\phi\in\Phi} c_{2,\phi}(\epsilon|X)^{-1} = \sup_{\phi\in\Phi} \mathbb{E}\bigl[\exp\bigl\{\epsilon\phi_2(A, X)\bigr\}|X\bigr] \leq \exp\bigl\{\epsilon^2\sigma_2^2/2\bigr\} \xrightarrow[\epsilon\to 0]{} 1$$

P-almost surely. This completes the proof. □

Lemma A.10. *For every $\phi = (\phi_1, \phi_2, \phi_3) \in \Phi$ with and $\epsilon \in (0, 1/b]$, where b is the same quantity as that appearing in Lemma A.8, the conditional expectation $Q_{\phi,\epsilon}(A, X)$ of Y given (A, X) under $P_{\epsilon,\phi}$ satisfies*

$$Q_{\phi,\epsilon}(A, X) = c_{3,\phi}(\epsilon|A, X)\mathbb{E}\bigl[e^{\epsilon\phi_3(O)}Y|A, X\bigr] \tag{A.19}$$

$P_{\epsilon,\phi}$-almost surely.

Proof. Let $\mathbb{E}_{\phi,\epsilon}$ denote the expectation under $P_{\phi,\epsilon}$ and $c_\phi(\epsilon) : \{-1, 1\} \times \mathcal{X} \to \mathbb{R}$ be the function given by $c_\phi(\epsilon)(A, X) \equiv c_{1,\phi}(\epsilon)c_{2,\phi}(\epsilon|X)c_{3,\phi}(\epsilon|A, X)$. Set arbitrarily $\eta : \{-1, 1\} \times \mathcal{X} \to \mathbb{R}_+$, a nonnegative, measurable function. By definition of $P_{\phi,\epsilon}$ (14) and the tower rule, the next equalities hold true:

$$\begin{aligned}
\mathbb{E}_{\phi,\epsilon}\bigl[Y\eta(A, X)\bigr] &= \mathbb{E}\bigl[c_\phi(\epsilon)(A, X)e^{\epsilon(\phi_1(X)+\phi_2(A,X)+\phi_3(O))}Y\eta(A, X)\bigr] \\
&= \mathbb{E}\bigl[c_\phi(\epsilon)(A, X)e^{\epsilon(\phi_1(X)+\phi_2(A,X))}\mathbb{E}\bigl[e^{\epsilon\phi_3(O)}Y|A, X\bigr]\eta(A, X)\bigr] \\
&= \mathbb{E}_{\phi,\epsilon}\bigl[c_{3,\phi}(\epsilon|A, X)\mathbb{E}\bigl[e^{\epsilon\phi_3(O)}Y|A, X\bigr]\eta(A, X)\bigr].
\end{aligned}$$

The result follows by definition of the conditional expectation. □

Lemma A.11. *The sequence of collections of distributions $\{P^n_{\phi,\epsilon=n^{-1/2}} : \phi \in \Phi\}_{n\geq 1}$ is uniformly contiguous w.r.t. the sequence $\{P^n\}_{n\geq 1}$,* i.e. (15) *holds.*

Proof. Let $\{B_n\}_{n\geq 1}$ be a sequence of measurable sets for which $\lim_{n\to\infty} P^n(B_n) = 0$. For notational convenience, for $\phi = (\phi_1, \phi_2, \phi_3) \in \Phi$, we let $\bar{\phi}$ denote $o \mapsto \sum_{j=1}^3 \phi_j(o)$. For any $n \in \mathbb{N}$, $\phi \in \Phi$,

$$\begin{aligned}
P^n_{\phi,\epsilon=n^{-1/2}}(B_n) &= \int \Biggl[\prod_{i=1}^n \bigl\{c_{1,\phi}(\epsilon)c_{2,\phi}(\epsilon|x_i)c_{3,\phi}(\epsilon|a_i, x_i)\bigr\}\Biggr] \\
&\qquad \times \exp\Biggl\{n^{-1/2}\sum_{i=1}^n \bar{\phi}(o_i)\Biggr\}\mathbf{1}_{B_n}\,dP^n(o_1, \ldots, o_n) \\
&\leq \int \exp\Biggl\{n^{-1/2}\sum_{i=1}^n \bar{\phi}(o_i)\Biggr\}\mathbf{1}_{B_n}\,dP^n(o_1, \ldots, o_n)
\end{aligned}$$

$$
\begin{aligned}
&\le P^n(B_n)^{1/2}\mathbb{E}\left[\exp\left\{2n^{-1/2}\sum_{i=1}^n\bar{\phi}(o_i)\right\}\right]^{1/2}\\
&= P^n(B_n)^{1/2}\mathbb{E}\left[\exp\left\{2n^{-1/2}\bar{\phi}(O)\right\}\right]^{n/2},
\end{aligned}
$$

where the first inequality holds by Lemma A.9, the second inequality holds by the Cauchy–Schwarz inequality, and the final equality holds because $O_1,\dots,O_n$ on the line above are i.i.d. Using the sub-exponential property of $\bar{\phi}$ given in Lemma A.8 at $\lambda = 2n^{-1/2}$, we have

$$
\mathbb{E}\left[\exp\left\{2n^{-1/2}\bar{\phi}(O)\right\}\right]^{n/2} \le \exp\left\{\sigma^2\right\}
$$

whenever $n \ge 4b^2$, thereby showing that $P^n_{\phi,\epsilon=n^{-1/2}}(B_n) \le P^n(B_n)^{1/2}\exp\{\sigma^2\}$ for n large enough. We take the supremum over $\phi \in \Phi$ then let n go to infinity to complete the proof. □

Lemma A.12. *It holds that*

$$
\sup_{\phi\in\Phi}\sup_{\pi_1,\pi_2\in\overline{\Pi}}\left\{\mathcal{V}_{\phi,\epsilon}(\pi_1) - \mathcal{V}_{\phi,\epsilon}(\pi_2) - \mathcal{V}(\pi_1) + \mathcal{V}(\pi_2) - \epsilon\langle\phi_1+\phi_3, f_{\pi_1} - f_{\pi_2}\rangle\right\} \le o(\epsilon). \tag{A.20}
$$

Moreover, there exists a constant C depending only on the bounds on the support of $Y \sim P$, the bounds from the strong positivity assumption, and on the parameters (b_1,σ_1^2), (b_2,σ_2^2), (b_3,σ_3^2) of (9), (10) *and* (11) *such that, for all $\epsilon > 0$ small enough,*

$$
\sup_{\phi\in\Phi}\sup_{\pi_1,\pi_2\in\overline{\Pi}}\left\{\mathcal{V}_{\phi,\epsilon}(\pi_1) - \mathcal{V}_{\phi,\epsilon}(\pi_2) - \mathcal{V}(\pi_1) + \mathcal{V}(\pi_2) - C\epsilon\|\pi_1 - \pi_2\|\right\} \le 0. \tag{A.21}
$$

The proof of Lemma A.12 is postponed at the end of the section.

Lemma A.13. *If Π is totally bounded and $\widehat{\pi}_1$ is such that $\mathcal{R}(\widehat{\pi}_1) = o_P(1)$ under sampling from P, then $\inf_{\pi^\star\in\overline{\Pi}^\star}\|\pi^\star - \widehat{\pi}_1\|$ converges to zero in probability under sampling from P.*

Proof. Fix $t > 0$. For any $r > 0$, we have that

$$
\begin{aligned}
&P\left\{\inf_{\pi^\star\in\overline{\Pi}^\star}\|\pi^\star - \widehat{\pi}_1\| > t\right\}\\
&\quad\le P\left\{\mathcal{R}(\widehat{\pi}_1) > r\right\} + P\left\{\mathcal{R}(\widehat{\pi}_1) \le r, \inf_{\pi^\star\in\overline{\Pi}^\star}\|\pi^\star - \widehat{\pi}_1\| > t\right\}\\
&\quad\le P\left\{\mathcal{R}(\widehat{\pi}_1) > r\right\} + P\left\{\mathcal{R}(\widehat{\pi}_1) \le r, \sup_{\pi\in\overline{\Pi}:\mathcal{R}(\pi)\le r}\inf_{\pi^\star\in\overline{\Pi}^\star}\|\pi^\star - \pi\| > t\right\}\\
&\quad\le P\left\{\mathcal{R}(\widehat{\pi}_1) > r\right\} + \mathbf{1}\left\{\sup_{\pi\in\overline{\Pi}:\mathcal{R}(\pi)\le r}\inf_{\pi^\star\in\overline{\Pi}^\star}\|\pi^\star - \pi\| > t\right\}.
\end{aligned}
$$

By Lemma A.4, there exists a small enough $r_0 > 0$ such that $\sup_{\pi\in\overline{\Pi}:\mathcal{R}(\pi)\le r}\inf_{\pi^\star\in\overline{\Pi}^\star}\|\pi^\star - \pi\| \le t$ for all $r < r_0$. Consequently, the indicator function above is equal to zero if r is sufficiently small. By assumption, $\mathcal{R}(\widehat{\pi}_1)$ converges to zero in probability under sampling from P, and so $P\{\mathcal{R}(\widehat{\pi}_1) > r\} = o(1)$. □

Lemma A.14. *It holds that* $\limsup_{\epsilon\to 0}\sup_{\phi\in\Phi,\pi^\star\in\overline{\Pi}^\star,\pi^\star_{\phi,\epsilon}\in\overline{\Pi}^\star_{\phi,\epsilon}}\|f_{\pi^\star_{\phi,\epsilon}} - f_{\pi^\star}\|^2 \le V_*$.

Proof of Lemma A.14. We show that $\|f_{\pi_1} - f_{\pi_2}\|^2 \le V_*$ for all $\pi_1,\pi_2 \in \overline{\Pi}^\star$, hence the result. Fix $\pi_1,\pi_2 \in \overline{\Pi}^\star$, hence $\mathcal{V}(\pi_1) = \mathcal{V}(\pi_2) = \mathcal{V}^\star$, and recall that $\tilde{f}_\pi : o \mapsto f_\pi(o) + \mathcal{V}(\pi)$ for all $\pi \in \overline{\Pi}$. By using that π_1 and π_2 take their values in $\{-1,1\}$, we first derive the following equalities:

$$
\begin{aligned}
P\left[\mathbf{1}\{\pi_1 \ne \pi_2\}(\tilde{f}_{\pi_1} - \tilde{f}_{-\pi_1})^2\right] &= P\left[(\tilde{f}_{\pi_1} - \tilde{f}_{\pi_2})^2\right]\\
&= P\left[\left(f_{\pi_1} - f_{\pi_2} + \mathcal{V}(\pi_1) - \mathcal{V}(\pi_2)\right)^2\right]\\
&= \|f_{\pi_1} - f_{\pi_2}\|^2
\end{aligned} \tag{A.22}
$$

$$\text{and}\quad P\big[\mathbf{1}\{\pi_1 \neq \pi_2\}(\tilde{f}_{\pi_1} - \tilde{f}_{-\pi_1})\big] = P[\tilde{f}_{\pi_1} - \tilde{f}_{\pi_2}] = P\big[f_{\pi_1} - f_{\pi_2} + \mathcal{V}(\pi_1) - \mathcal{V}(\pi_2)\big] = 0. \tag{A.23}$$

Combining (A.22), (A.23) and Lemma A.15 below then yields

$$\begin{aligned}\|f_{\pi_1} - f_{\pi_2}\|^2 &= \mathbb{V}\big[\big(\mathbf{1}\{\pi_1 \neq \pi_2\}(\tilde{f}_{\pi_1} - \tilde{f}_{-\pi_1})\big)(O)\big] \\ &\le \mathbb{V}\big[(\tilde{f}_{\pi_1} - \tilde{f}_{-\pi_1})(O)\big] \\ &= \mathbb{V}\big[(f_{\pi_1} - f_{-\pi_1})(O)\big] = V_*,\end{aligned}$$

hence the result. □

Lemma A.15. *If the random variables $A \in \{0,1\}$ and $B \in \mathbb{R}$ are such that $\mathbb{E}(AB) = 0$, then $\mathbb{V}(AB) \le \mathbb{V}(B)$.*

Proof. If $\mathbb{E}(B^2)$ is infinite, then the inequality obviously holds. Otherwise, we note that

$$\big[B - \mathbb{E}(B)\big]^2 \ge A\big[B - \mathbb{E}(B)\big]^2 = AB^2 - 2\mathbb{E}(B)AB + \big[\mathbb{E}(B)\big]^2 A$$

hence

$$\mathbb{V}(B) \ge \mathbb{E}\big[AB^2\big] + \mathbb{E}(A)\big[\mathbb{E}(B)\big]^2 \ge \mathbb{E}\big[AB^2\big] = \mathbb{V}(AB).$$

This completes the proof. □

We are now in a position to prove the main locally uniform ERM result from the main text.

Proof of Theorem 2. Fix $\phi \in \Phi$, $\epsilon \in \mathbb{R}$, and $t > 0$. Let $\bar{\phi}$ be given by $\bar{\phi}(O) \equiv \phi_1(X) + \phi_3(O)$. Fix $\pi^\star \in \overline{\Pi}^\star$ and $\pi^\star_{\phi,\epsilon} \in \overline{\Pi}^\star_{\phi,\epsilon}$. First, we note that (A.21) in Lemma A.12 implies the first inequality below:

$$\begin{aligned}\mathcal{R}_{\phi,\epsilon}(\widehat{\pi}_1) &= \mathcal{V}^\star_{\phi,\epsilon} - \mathcal{V}_{\phi,\epsilon}(\widehat{\pi}_1) \\ &= \big[\mathcal{V}^\star_{\phi,\epsilon} - \mathcal{V}_{\phi,\epsilon}(\pi^\star)\big] + \big[\mathcal{V}_{\phi,\epsilon}(\pi^\star) - \mathcal{V}_{\phi,\epsilon}(\widehat{\pi}_1) - \mathcal{V}(\pi^\star) + \mathcal{V}(\widehat{\pi}_1)\big] + \mathcal{R}(\widehat{\pi}_1) \\ &\le \big[\mathcal{V}^\star_{\phi,\epsilon} - \mathcal{V}_{\phi,\epsilon}(\pi^\star)\big] + C\epsilon\|\widehat{\pi}_1 - \pi^\star\| + \mathcal{R}(\widehat{\pi}_1) \\ &\le \big[\mathcal{V}_{\phi,\epsilon}(\pi^\star_{\phi,\epsilon}) - \mathcal{V}_{\phi,\epsilon}(\pi^\star) - \mathcal{V}(\pi^\star_{\phi,\epsilon}) + \mathcal{V}(\pi^\star)\big] + C\epsilon\|\widehat{\pi}_1 - \pi^\star\| + \mathcal{R}(\widehat{\pi}_1),\end{aligned}$$

the second inequality being due to $\mathcal{V}(\pi^\star) - \mathcal{V}(\pi^\star_{\phi,\epsilon}) \ge 0$. Second, we use (A.20) in Lemma A.12 to derive from the previous display the following key-inequality:

$$\mathcal{R}_{\phi,\epsilon}(\widehat{\pi}_1) \le \epsilon\langle\bar{\phi}, f_{\pi^\star_{\phi,\epsilon}} - f_{\pi^\star}\rangle + C\epsilon\|\widehat{\pi}_1 - \pi^\star\| + \mathcal{R}(\widehat{\pi}_1) + o(\epsilon),$$

where the term $o(\epsilon)$ is deterministic and does not depend on $(\phi, \widehat{\pi}_1, \pi^\star)$. The above holds for any $\pi^\star_{\phi,\epsilon} \in \overline{\Pi}^\star_{\phi,\epsilon}$, hence

$$\mathcal{R}_{\phi,\epsilon}(\widehat{\pi}_1) \le \epsilon \inf_{\pi^\star_{\phi,\epsilon} \in \overline{\Pi}^\star_{\phi,\epsilon}} \langle\bar{\phi}, f_{\pi^\star_{\phi,\epsilon}} - f_{\pi^\star}\rangle + C\epsilon\|\widehat{\pi}_1 - \pi^\star\| + \mathcal{R}(\widehat{\pi}_1) + o(\epsilon).$$

Taking suprema over unspecified quantities gives the bound

$$\mathcal{R}_{\phi,\epsilon}(\widehat{\pi}_1) \le \epsilon \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon} \in \overline{\Pi}^\star_{\phi,\epsilon}} \langle\bar{\phi}, f_{\pi^\star_{\phi,\epsilon}} - f_{\tilde{\pi}^\star}\rangle + C\epsilon\|\widehat{\pi}_1 - \pi^\star\| + \mathcal{R}(\widehat{\pi}_1) + o(\epsilon).$$

Noting that $\pi^\star \in \overline{\Pi}^\star$ was arbitrary, we see that

$$\mathcal{R}_{\phi,\epsilon}(\widehat{\pi}_1) \le \epsilon \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon} \in \overline{\Pi}^\star_{\phi,\epsilon}} \langle\bar{\phi}, f_{\pi^\star_{\phi,\epsilon}} - f_{\tilde{\pi}^\star}\rangle + C\epsilon \inf_{\pi^\star \in \overline{\Pi}^\star}\|\widehat{\pi}_1 - \pi^\star\| + \mathcal{R}(\widehat{\pi}_1) + o(\epsilon).$$

Thus, for $\epsilon_n = n^{-1/2}$,

$$
\begin{aligned}
&P^n_{\phi,\epsilon_n}\Big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon_n} \in \overline{\Pi}^\star_{\phi,\epsilon_n}} \langle \bar{\phi}, f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\rangle + 3t\Big\} \\
&\quad \le \mathbf{1}\big\{o(1) > t\big\} + P^n_{\phi,\epsilon_n}\Big\{C \inf_{\pi^\star \in \overline{\Pi}^\star} \|\widehat{\pi}_1 - \pi^\star\| > t\Big\} + P^n_{\phi,\epsilon_n}\big\{n^{1/2}\mathcal{R}(\widehat{\pi}_1) > t\big\},
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\sup_{\phi\in\Phi} P^n_{\phi,\epsilon_n}\Big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon_n} \in \overline{\Pi}^\star_{\phi,\epsilon_n}} \langle \bar{\phi}, f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\rangle + 3t\Big\} \\
&\quad \le o(1) + \sup_{\phi\in\Phi} P^n_{\phi,\epsilon_n}\Big\{C \inf_{\pi^\star \in \overline{\Pi}^\star} \|\widehat{\pi}_1 - \pi^\star\| > t\Big\} + \sup_{\phi\in\Phi} P^n_{\phi,\epsilon_n}\big\{n^{1/2}\mathcal{R}(\widehat{\pi}_1) > t\big\}.
\end{aligned}
$$

We know on the one hand that $P^n\{C \inf_{\pi^\star\in\overline{\Pi}^\star} \|\widehat{\pi}_1 - \pi^\star\| > t\} = o(1)$ by Lemma A.13 (which can be applied because Π is totally bounded and $\mathcal{R}(\widehat{\pi}_1) = o_P(n^{-1/2})$ by assumption) and on the other hand that $P^n\{n^{1/2}\mathcal{R}(\widehat{\pi}_1) > t\} = o(1)$ by assumption. Hence, by the uniform contiguity of $\{P^n_{\phi,\epsilon_n} : \phi \in \Phi\}_{n\ge 1}$ w.r.t. $\{P^n\}_{n\ge1}$, both the second and third terms in the above display are $o(1)$. This gives the first inequality in the theorem statement.

We now obtain the second inequality. Let $\{t_n\}_{n\ge1}$ be a real-valued sequence such that $t_n = o(1)$. Noting that, for all $\tilde{\pi}^\star \in \overline{\Pi}^\star$ and $\pi^\star_{\phi,\epsilon_n} \in \overline{\Pi}^\star_{\phi,\epsilon_n}$ it holds that

$$
\langle \bar{\phi}, f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\rangle \le \|\bar{\phi}\| \times \|f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\|,
$$

we see that

$$
\begin{aligned}
&P^n_{\phi,\epsilon_n}\Big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon_n} \in \overline{\Pi}^\star_{\phi,\epsilon_n}} \langle \bar{\phi}, f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\rangle + t\Big\} \\
&\quad \ge P^n_{\phi,\epsilon_n}\Big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \|\bar{\phi}\| \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon_n} \in \overline{\Pi}^\star_{\phi,\epsilon_n}} \|f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\| + t\Big\} \\
&\quad \ge P^n_{\phi,\epsilon_n}\big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \|\bar{\phi}\|\big(V_*^{1/2} + t_n\big) + t\big\} \\
&\qquad \times \mathbf{1}\Big\{V_*^{1/2} > \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon_n} \in \overline{\Pi}^\star_{\phi,\epsilon_n}} \|f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\| - t_n\Big\} \\
&\quad \ge P^n_{\phi,\epsilon_n}\big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \|\bar{\phi}\|\big(V_*^{1/2} + t_n\big) + t\big\} \\
&\qquad \times \mathbf{1}\Big\{V_*^{1/2} > \sup_{\tilde{\phi}\in\Phi, \tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\tilde{\phi},\epsilon_n} \in \overline{\Pi}^\star_{\tilde{\phi},\epsilon_n}} \|f_{\pi^\star_{\tilde{\phi},\epsilon_n}} - f_{\tilde{\pi}^\star}\| - t_n\Big\} \\
&\quad \ge P^n_{\phi,\epsilon_n}\Big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \|\bar{\phi}\| V_*^{1/2} + \sup_{\tilde{\phi}\in\Phi} \|\tilde{\phi}_1 + \tilde{\phi}_3\| t_n + t\Big\} \\
&\qquad \times \mathbf{1}\Big\{V_*^{1/2} > \sup_{\tilde{\phi}\in\Phi, \tilde{\pi}^\star \in \overline{\Pi}^\star, \pi^\star_{\tilde{\phi},\epsilon_n} \in \overline{\Pi}^\star_{\tilde{\phi},\epsilon_n}} \|f_{\pi^\star_{\tilde{\phi},\epsilon_n}} - f_{\tilde{\pi}^\star}\| - t_n\Big\}.
\end{aligned}
$$

By Lemma A.14, it is possible to choose $\{t_n\}_{n\ge1}$ in such a way that the indicator on the right is one for all n large enough. Moreover, for n large enough, it is also the case that $\sup_{\tilde{\phi}\in\Phi} \|\tilde{\phi}_1 + \tilde{\phi}_3\| t_n \le t$ hence, in conclusion,

$$
\begin{aligned}
&P^n_{\phi,\epsilon_n}\Big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \sup_{\tilde{\pi}^\star \in \overline{\Pi}^\star} \inf_{\pi^\star_{\phi,\epsilon_n} \in \overline{\Pi}^\star_{\phi,\epsilon_n}} \langle \bar{\phi}, f_{\pi^\star_{\phi,\epsilon_n}} - f_{\tilde{\pi}^\star}\rangle + t\Big\} \\
&\quad \ge P^n_{\phi,\epsilon_n}\big\{n^{1/2}\mathcal{R}_{\phi,\epsilon_n}(\widehat{\pi}_1) > \|\bar{\phi}\| V_*^{1/2} + 2t\big\}
\end{aligned}
$$

for n large enough. This completes the proof. □

Proof of Lemma A.12. Set arbitrarily $\phi = (\phi_1, \phi_2, \phi_3) \in \Phi$, $\epsilon \in (0, 1/b]$ and $\pi_1, \pi_2 \in \overline{\Pi}$. We can assume without loss of generality that ϕ_2 is the zero function because, for any $\pi \in \overline{\Pi}$, $\mathcal{V}_{\phi,\epsilon}(\pi) = \mathcal{V}_{(\phi_1,0,\phi_3),\epsilon}(\pi)$. For notational simplicity, let $Q \circ \pi_j(X)$ and $Q_{\phi,\epsilon} \circ \pi_j(X)$ stand for $Q(\pi_j(X), X)$ and $Q_{\phi,\epsilon}(\pi_j(X), X)$, respectively, for $j = 1, 2$. Moreover, let $Q_a(X)$ and $Q_{\phi,\epsilon,a}(X)$ stand for $Q(a, X)$ and $Q_{\phi,\epsilon}(a, X)$, respectively, for $a = -1, 1$. Recall the definition of $c_\phi(\epsilon)$ given in the proof of Lemma A.10 and let $c_{3,\phi}(\epsilon)$ be the function given by $c_{3,\phi}(\epsilon)(A, X) \equiv c_{3,\phi}(\epsilon|A, X)$. Finally, let η, $\bar{\phi}$ and sgn denote the functions given by $\eta(O) \equiv AY/P(A|X)$, $\bar{\phi}(O) \equiv \phi_1(X) + \phi_3(O)$, and $\operatorname{sgn}\{u\} \equiv \mathbf{1}\{u > 0\} - \mathbf{1}\{u < 0\}$.

For simplicity, we show (A.21) then build upon its proof to derive (A.20). First, we note that

$$\begin{aligned}
&\mathcal{V}_{\phi,\epsilon}(\pi_1) - \mathcal{V}_{\phi,\epsilon}(\pi_2) - \mathcal{V}(\pi_1) + \mathcal{V}(\pi_2)\\
&\quad = P\big[c_\phi(\epsilon)e^{\epsilon\bar{\phi}}(Q_{\phi,\epsilon} \circ \pi_1 - Q_{\phi,\epsilon} \circ \pi_2) - (Q \circ \pi_1 - Q \circ \pi_2)\big]\\
&\quad = P\big(c_\phi(\epsilon)e^{\epsilon\bar{\phi}} - 1\big)(Q_{\phi,\epsilon} \circ \pi_1 - Q_{\phi,\epsilon} \circ \pi_2)\\
&\qquad + P\big[(Q_{\phi,\epsilon} - Q) \circ \pi_1 - (Q_{\phi,\epsilon} - Q) \circ \pi_2\big]\\
&\quad = P\big(c_\phi(\epsilon)e^{\epsilon\bar{\phi}} - 1\big)(Q_{\phi,\epsilon,1} - Q_{\phi,\epsilon,-1}) \operatorname{sgn}\{\pi_1 - \pi_2\}\\
&\qquad + P\big[(Q_{\phi,\epsilon,1} - Q_1) - (Q_{\phi,\epsilon,-1} - Q_{-1})\big] \operatorname{sgn}\{\pi_1 - \pi_2\}\\
&\quad = P\big(c_\phi(\epsilon)e^{\epsilon\bar{\phi}} - 1 - \epsilon c_\phi(\epsilon)\bar{\phi}\big)(Q_{\phi,\epsilon,1} - Q_{\phi,\epsilon,-1}) \operatorname{sgn}\{\pi_1 - \pi_2\} && \text{(A.24)}\\
&\qquad + \epsilon P c_\phi(\epsilon)\bar{\phi}(Q_{\phi,\epsilon,1} - Q_{\phi,\epsilon,-1}) \operatorname{sgn}\{\pi_1 - \pi_2\} && \text{(A.25)}\\
&\qquad + \epsilon P c_{3,\phi}(\epsilon)\phi_3\eta \operatorname{sgn}\{\pi_1 - \pi_2\} && \text{(A.26)}\\
&\qquad + P\big[(Q_{\phi,\epsilon,1} - Q_1) - (Q_{\phi,\epsilon,-1} - Q_{-1}) - \epsilon c_{3,\phi}(\epsilon)\phi_3\eta\big] \operatorname{sgn}\{\pi_1 - \pi_2\}. && \text{(A.27)}
\end{aligned}$$

Below, we study in turn the four terms in the RHS of the above display in order to prove (A.21). Afterwards, the study of (A.25) and (A.26) will be resumed and refined to derive (A.20).

Studying term (A.24). Because $|Y|$ is bounded under P by, say, B the Cauchy-Schwartz inequality implies that the absolute value of (A.24) is upper bounded by

$$\begin{aligned}
&2B\big\|c_\phi(\epsilon)e^{\epsilon\bar{\phi}} - 1 - \epsilon c_\phi(\epsilon)\bar{\phi}\big\| \times P\mathbf{1}\{\pi \neq \pi_2\}\\
&\quad = 2B\|\pi_1 - \pi_2\| \times \big\|c_\phi(\epsilon)e^{\epsilon\bar{\phi}} - 1 - \epsilon c_\phi(\epsilon)\bar{\phi}\big\|.
\end{aligned}$$

By the triangle inequality and Lemma A.9, we first note that

$$\begin{aligned}
\big\|c_\phi(\epsilon)e^{\epsilon\bar{\phi}} - 1 - \epsilon c_\phi(\epsilon)\bar{\phi}\big\| &\leq \big\|c_\phi(\epsilon) - 1\big\| + \big\|c_\phi(\epsilon)\big(e^{\epsilon\bar{\phi}} - 1 - \epsilon\bar{\phi}\big)\big\|\\
&\leq \big\|c_\phi(\epsilon) - 1\big\| + \big\|e^{\epsilon\bar{\phi}} - 1 - \epsilon\bar{\phi}\big\|.
\end{aligned}$$

Second, invoking Lemma A.9 and the triangle inequality again,

$$\begin{aligned}
\big|c_\phi(\epsilon)(O) - 1\big| &= \big|c_{1,\phi}(\epsilon)\big[c_{3,\phi}(\epsilon|A, X) - c_{1,\phi}(\epsilon)^{-1}\big]\big|\\
&\leq \big|c_{3,\phi}(\epsilon|A, X) - c_{1,\phi}(\epsilon)^{-1}\big|\\
&\leq \big|c_{3,\phi}(\epsilon|A, X)\big[1 - c_{3,\phi}(\epsilon|A, X)^{-1}\big]\big| + \big|c_{1,\phi}(\epsilon)^{-1} - 1\big|\\
&\leq \big|c_{3,\phi}(\epsilon|A, X)^{-1} - 1\big| + \big|c_{1,\phi}(\epsilon)^{-1} - 1\big|.
\end{aligned}$$

In light of Lemma A.9 and (9), if $0 < \epsilon \leq 1/b_1$ then

$$0 \leq \frac{c_{1,\phi}(\epsilon)^{-1} - 1}{\epsilon} = \frac{Pe^{\epsilon\phi_1} - 1}{\epsilon} \leq \frac{\exp\{\epsilon^2\sigma_1^2/2\} - 1}{\epsilon} \xrightarrow[\epsilon \to 0]{} 0.$$

Likewise, in light of (11), if $0 < \epsilon \leq 1/b_3$ then

$$0 \leq \frac{c_{3,\phi}(\epsilon|A, X)^{-1} - 1}{\epsilon} = \frac{\mathbb{E}[e^{\epsilon\phi_3(O)}|A, X] - 1}{\epsilon} \leq \frac{\exp\{\epsilon^2\sigma_3^2/2\} - 1}{\epsilon} \xrightarrow[\epsilon \to 0]{} 0, \tag{A.28}$$

P-almost surely. Therefore, $\sup_{\tilde{\phi}\in\Phi,,\tilde{\phi}_2=0} \epsilon^{-1}\|c_{\tilde{\phi}}(\epsilon)-1\| = o(1)$. Furthermore, the Taylor–Lagrange theorem guarantees the existence, for every $o \in \mathcal{O}$, of $\tilde{\epsilon}(o) \in (0,\epsilon)$ such that

$$\begin{aligned}4\big(e^{\epsilon\bar{\phi}(o)} - 1 - \epsilon\bar{\phi}(o)\big)^2 &= \epsilon^4\bar{\phi}(o)^4 e^{2\tilde{\epsilon}(o)\bar{\phi}(o)} \\ &\le \epsilon^4\bar{\phi}(o)^4\big(e^{2\tilde{\epsilon}(o)\bar{\phi}(o)} + e^{-2\tilde{\epsilon}(o)\bar{\phi}(o)}\big) \\ &\le \epsilon^4\bar{\phi}(o)^4\big(e^{2\epsilon\bar{\phi}(o)} + e^{-2\epsilon\bar{\phi}(o)}\big).\end{aligned}$$

If $0 < \epsilon \le 1/4b$ it then follows from (A.17) in Lemma A.8 that

$$4\epsilon^{-2}\big\|e^{\epsilon\bar{\phi}} - 1 - \epsilon\bar{\phi}\big\|^2 \le 2\epsilon^2\exp\big\{c + 4\epsilon^2\sigma^2\big\} \underset{\epsilon\to 0}{\longrightarrow} 0,$$

hence $\sup_{\phi\in\Phi,\phi_2=0}\epsilon^{-1}\|e^{\epsilon\bar{\phi}} - 1 - \epsilon\bar{\phi}\| = o(1)$. In conclusion,

$$\sup_{\phi\in\Phi}\big|P\big(c_\phi(\epsilon)e^{\epsilon\bar{\phi}} - 1 - \epsilon c_\phi(\epsilon)\bar{\phi}\big)(Q_{\phi,\epsilon,1} - Q_{\phi,\epsilon,-1})\operatorname{sgn}\{\pi_1 - \pi_2\}\big| \le 2B\|\pi_1 - \pi_2\|o(\epsilon). \tag{A.29}$$

Studying term (A.25) *(first pass).* Because $|Y|$ is upper bounded by B under P, the Cauchy–Schwarz inequality implies that the absolute value of (A.25) is upper bounded by

$$\begin{aligned}2B\epsilon\|\pi_1 - \pi_2\| \times \big\|c_\phi(\epsilon)\bar{\phi}\big\| &\le 2B\epsilon\|\pi_1 - \pi_2\| \times \|\bar{\phi}\| \\ &\le 2B\epsilon\|\pi_1 - \pi_2\| \times \sup_{\phi\in\Phi}\|\bar{\phi}\|,\end{aligned}$$

where $\|c_\phi(\epsilon)\bar{\phi}\| \le \|\bar{\phi}\|$ is a consequence of Lemma A.9 and Lemma A.8 guarantees that $\sup_{\phi\in\Phi}\|\bar{\phi}\|$ is finite. In conclusion,

$$\sup_{\phi\in\Phi}\big|\epsilon P c_\phi(\epsilon)\bar{\phi}(Q_{\phi,\epsilon,1} - Q_{\phi,\epsilon,-1})\operatorname{sgn}\{\pi_1 - \pi_2\}\big| \le 2B\|\pi_1 - \pi_2\|\epsilon. \tag{A.30}$$

Studying term (A.26) *(first pass).* The same argument as in the study of term (A.25) applies, because $|\eta|$ is upper bounded by $B\delta^{-1}$ under P. It yields

$$\sup_{\phi\in\Phi}\big|\epsilon P c_{3,\phi}(\epsilon)\eta\operatorname{sgn}\{\pi_1 - \pi_2\}\big| \le 2B\delta^{-1}\|\pi_1 - \pi_2\|\epsilon. \tag{A.31}$$

Studying term (A.27). For both $a \in \{-1, 1\}$, introduce

$$T_a \equiv P\big[Q_{\phi,\epsilon,a} - Q_a - \epsilon c_{3,\phi}(\epsilon)\phi_3\eta a\mathbf{1}\{A = a\}\big]\operatorname{sgn}\{\pi_1 - \pi_2\}$$

and note that term (A.27) equals $T_1 - T_{-1}$. In light of Lemma A.10, using repeatedly the tower rule yields that, for each $a \in \{-1, 1\}$, T_a equals

$$\begin{aligned}&E\bigg[\bigg((Q_{\phi,\epsilon,a} - Q_a)(X) - \frac{a\mathbf{1}\{A = a\}A}{P(A = a|X)}\epsilon c_{3,\phi}(\epsilon|A,X)E\big[\phi_3(O)Y \mid A, X\big]\bigg)\operatorname{sgn}\big\{(\pi_1 - \pi_2)(X)\big\}\bigg] \\ &\quad= E\big[\big((Q_{\phi,\epsilon,a} - Q_a)(X) - \epsilon c_{3,\phi}(\epsilon|A = a, X)E\big[\phi_3(O)Y \mid A = a, X\big]\big)\operatorname{sgn}\big\{(\pi_1 - \pi_2)(X)\big\}\big] \\ &\quad= E\big(E\big[\big(c_{3,\phi}(\epsilon|A,X)e^{\epsilon\phi_3(O)} - 1 - \epsilon c_{3,\phi}(\epsilon|A,X)\phi_3(O)\big)Y \mid A = a, X\big]\operatorname{sgn}\big\{(\pi_1 - \pi_2)(X)\big\}\big) \\ &\quad= E\bigg[\frac{\mathbf{1}\{A = a\}Y}{P(A = a|X)}\big(c_{3,\phi}(\epsilon|A,X)e^{\epsilon\phi_3(O)} - 1 - \epsilon c_{3,\phi}(\epsilon|A,X)\phi_3(O)\big)\operatorname{sgn}\big\{(\pi_1 - \pi_2)(X)\big\}\bigg].\end{aligned}$$

Therefore, applying the Cauchy–Schwarz inequality implies that the absolute value of term (A.27) is upper bounded by

$$\begin{aligned}&2B\delta^{-1}\big|P\big(c_{3,\phi}(\epsilon)e^{\epsilon\phi_3} - 1 - \epsilon c_{3,\phi}(\epsilon)\phi_3\big)\operatorname{sgn}\{\pi_1 - \pi_2\}\big| \\ &\quad\le 2B\delta^{-1}\big\|c_{3,\phi}(\epsilon)e^{\epsilon\phi_3} - 1 - \epsilon c_{3,\phi}(\epsilon)\phi_3\big\| \times P\mathbf{1}\{\pi \ne \pi_2\} \\ &\quad= 2B\delta^{-1}\|\pi_1 - \pi_2\| \times \big\|c_{3,\phi}(\epsilon)e^{\epsilon\phi_3} - 1 - \epsilon c_{3,\phi}(\epsilon)\phi_3\big\|.\end{aligned}$$

Looking back at the study of term (A.24), we finally conclude that

$$\sup_{\phi\in\Phi}\bigl|P\bigl[(Q_{\phi,\epsilon,1}-Q_1)-(Q_{\phi,\epsilon,-1}-Q_{-1})-\epsilon c_{3,\phi}(\epsilon)\phi_3\eta\bigr]\operatorname{sgn}\{\pi_1-\pi_2\}\bigr|$$
$$\le 2B\delta^{-1}\|\pi_1-\pi_2\|o(\epsilon). \tag{A.32}$$

We are in a position to conclude. In view of (A.29), (A.30), (A.31), (A.32) and because $\sup_{\pi_1,\pi_2}\|\pi_1-\pi_2\|\le 2$, (A.21) holds with $C=8B\delta^{-1}+1$, for instance. This completes the first part of the proof, and we now move on to the second part of the proof.

Studying terms (A.25) *and* (A.26) *(second pass).* We first note that

$$\begin{aligned}\langle\bar\phi, f_{\pi_1}-f_{\pi_2}\rangle &= \bigl\langle c_\phi(\epsilon)\phi_1, f_{\pi_1}-f_{\pi_2}\bigr\rangle+\bigl\langle c_{3\phi}(\epsilon)\phi_3, f_{\pi_1}-f_{\pi_2}\bigr\rangle\\ &\quad-\bigl\langle\bigl(c_\phi(\epsilon)-1\bigr)\phi_1+\bigl(c_{3\phi}(\epsilon)-1\bigr)\phi_3, f_{\pi_1}-f_{\pi_2}\bigr\rangle.\end{aligned}$$

Applying the Cauchy–Schwarz and the triangle inequalities then reveals that the absolute value of the rightmost term in the above display is upper bounded by

$$\bigl(\bigl\|\bigl(c_\phi(\epsilon)-1\bigr)\phi_1\bigr\|+\bigl\|\bigl(c_{3\phi}(\epsilon)-1\bigr)\phi_3\bigr\|\bigr)\times\|f_{\pi_1}-f_{\pi_2}\|,$$

where it is easy to check that the first factor converges to 0 uniformly in $\phi\in\Phi$ as ϵ goes to 0. Indeed, for both $j=1,3$,

$$\bigl\|\bigl(c_{j\phi}(\epsilon)-1\bigr)\phi_j\bigr\|\le\bigl\|\phi_j^2\bigr\|\times\bigl\|\bigl(c_{j\phi}(\epsilon)-1\bigr)^2\bigr\|\le e^c\Bigl\|\sup_{\phi\in\Phi}\bigl(c_{j\phi}(\epsilon)-1\bigr)^2\Bigr\|$$

by the Cauchy–Schwarz inequality and Lemma A.8, and the above RHS expression converges to 0 as ϵ goes to 0 by Lemma A.9 and the dominated convergence theorem. In summary,

$$\sup_{\phi\in\Phi}\sup_{\pi_1,\pi_2\in\overline{\Pi}}\bigl|\langle\bar\phi, f_{\pi_1}-f_{\pi_2}\rangle-\bigl\langle c_\phi(\epsilon)\phi_1, f_{\pi_1}-f_{\pi_2}\bigr\rangle-\bigl\langle c_{3\phi}(\epsilon)\phi_3, f_{\pi_1}-f_{\pi_2}\bigr\rangle\bigr|=o(1). \tag{A.33}$$

Introduce $\tilde f_\pi: o\mapsto f_\pi(o)+\mathcal{V}(\pi)$ for all $\pi\in\overline{\Pi}$ and note that

$$(\tilde f_{\pi_1}-\tilde f_{\pi_2})(O)=\biggl[\frac{A(Y-Q(A,X))}{P(A|X)}+(Q_1-Q_{-1})(X)\biggr]\operatorname{sgn}\bigl\{(\pi_1-\pi_2)(X)\bigr\}. \tag{A.34}$$

Recall that $\mathbb{E}[\phi_3(O)|A,X]=0$ P-almost surely, and the fact that $\eta(O)\equiv AY/P(A|X)$. It thus holds that

$$\bigl\langle c_{3\phi}(\epsilon)\phi_3, f_{\pi_1}-f_{\pi_2}\bigr\rangle=\bigl\langle c_{3\phi}(\epsilon)\phi_3, \tilde f_{\pi_1}-\tilde f_{\pi_2}\bigr\rangle=Pc_{3\phi}(\epsilon)\phi_3\eta\operatorname{sgn}\{\pi_1-\pi_2\}. \tag{A.35}$$

Furthermore, recalling that ϕ_1 is a function of X such that $P\phi_1=0$, it also holds that

$$\begin{aligned}\bigl\langle c_\phi(\epsilon)\phi_1, f_{\pi_1}-f_{\pi_2}\bigr\rangle &= \bigl\langle c_\phi(\epsilon)\phi_1, \tilde f_{\pi_1}-\tilde f_{\pi_2}\bigr\rangle\\ &= Pc_\phi(\epsilon)\phi_1(Q_1-Q_{-1})\operatorname{sgn}\{\pi_1-\pi_2\}\\ &= Pc_\phi(\epsilon)\bar\phi(Q_{\phi,\epsilon,1}-Q_{\phi,\epsilon,-1})\operatorname{sgn}\{\pi_1-\pi_2\}\\ &\quad -Pc_\phi(\epsilon)\phi_1\bigl[(Q_{\phi,\epsilon,1}-Q_1)-(Q_{\phi,\epsilon,-1}-Q_{-1})\bigr]\operatorname{sgn}\{\pi_1-\pi_2\}.\end{aligned} \tag{A.36}$$

For both $a\in\{-1,1\}$, introduce

$$\tilde T_a\equiv Pc_\phi(\epsilon)\phi_1(Q_{\phi,\epsilon,a}-Q_a)\operatorname{sgn}\{\pi_1-\pi_2\},$$

so that the above rightmost term equals $\tilde T_1-\tilde T_{-1}$. Let $C_\phi(X)\equiv\mathbb{E}[c_\phi(O)|X]$, and note that $P(|C_{\phi(X)}|\le 1)=1$ by Lemma A.9. In light of Lemma A.10 and the strong positivity assumption, applying the tower rule and the Cauchy–Schwarz and triangle inequalities yields that

$$\begin{aligned}|\tilde T_a| &= \biggl|\mathbb{E}\biggl[\frac{\mathbf{1}\{A=a\}}{P(A|X)}C_\phi(X)\phi_1(X)\operatorname{sgn}\bigl\{(\pi_1-\pi_2)(X)\bigr\}\bigl(c_{3\phi}(\epsilon|A,X)e^{\epsilon\phi_3(O)}-1\bigr)Y\biggr]\biggr|\\ &\le B\delta^{-1}P\bigl|\phi_1\bigl[c_{3\phi}(\epsilon)e^{\epsilon\phi_3}-1\bigr]\bigr|\end{aligned}$$

$$\begin{aligned}
&\le B\delta^{-1}\|\phi_1\| \times \big\|c_{3\phi}(\epsilon)e^{\epsilon\phi_3} - 1\big\| \\
&\le B\delta^{-1}\|\phi_1\| \times \big\|e^{\epsilon\phi_3} - c_{3\phi}(\epsilon)^{-1}\big\| \\
&\le B\delta^{-1}\sup_{\phi\in\Phi}\|\bar{\phi}\| \times \Big(\sup_{\phi\in\Phi}\big\|c_{3\phi}(\epsilon)^{-1} - 1\big\| + \sup_{\phi\in\Phi}\big\|e^{\epsilon\phi_3} - 1\big\|\Big).
\end{aligned}$$

Now, Lemma A.8 guarantees that the first above supremum is finite and the bound (A.28) implies that the second one is $o(\epsilon)$. In addition, the mean value theorem guarantees the existence, for every $o \in \mathcal{O}$, of $\tilde{\epsilon}(o) \in (0, \epsilon)$ such that

$$\begin{aligned}
\big(e^{\epsilon\phi_3(o)} - 1\big)^2 &= \epsilon^2\phi_3(o)^2 e^{2\tilde{\epsilon}(o)\phi_3(o)} \\
&\le \epsilon^2\phi_3(o)^2\big(e^{2\tilde{\epsilon}(o)\phi_3(o)} + e^{-2\tilde{\epsilon}(o)\phi_3(o)}\big) \\
&\le \epsilon^2\phi_3(o)^2\big(e^{2\epsilon\phi_3(o)} + e^{-2\epsilon\phi_3(o)}\big).
\end{aligned}$$

If $0 < \epsilon \le 1/4b$ it then follows from (A.17) in Lemma A.8 that

$$\big\|e^{\epsilon\phi_3} - 1\big\|^2 \le 2\epsilon^2 \exp\big\{c + 4\epsilon^2\sigma^2\big\} \xrightarrow[\epsilon\to 0]{} 0,$$

hence $\sup_{\phi\in\Phi}\|e^{\epsilon\phi_3} - 1\| = o(1)$. Going back to (A.36), we thus obtain that

$$\sup_{\phi\in\Phi}\sup_{\pi_1,\pi_2\in\overline{\Pi}}\big|\langle c_\phi(\epsilon)\phi_1, f_{\pi_1} - f_{\pi_2}\rangle - Pc_\phi(\epsilon)\bar{\phi}(Q_{\phi,\epsilon,1} - Q_{\phi,\epsilon,-1})\,\mathrm{sgn}\{\pi_1 - \pi_2\}\big| = o(1). \tag{A.37}$$

In conclusion, by combining our initial decomposition of $\mathcal{V}_{\phi,\epsilon}(\pi_1) - \mathcal{V}_{\phi,\epsilon}(\pi_2) - \mathcal{V}(\pi_1) + \mathcal{V}(\pi_2)$ as the sum of (A.24), (A.25), (A.26) and (A.27) with (A.29) and (A.32) on the one hand and (A.33) (A.35) and (A.37) on the other hand, we show that (A.20) holds true, and thus complete the proof. □

Appendix B: Additional proofs

B.1. *Sketch of proof of Theorem 6*

The proof of Theorem 6 is very similar to that of Theorem 1. We only sketch it and point out the places where they differ.

Sketch of Proof of Theorem 6. Obviously (18) implies that $\mathcal{R}(\cdot)$ is uniformly continuous on $\overline{\Pi}$ w.r.t. $\|\cdot\|$. Assumption (17) then shows that the implication of Lemma A.3 also holds in our context, *i.e.*, that $\inf_{\pi\in\overline{\Pi}}\mathcal{R}(\pi) = 0$ and that $\overline{\Pi}^\star$ is not empty. As $\mathcal{R}(\cdot)$ is uniformly continuous and the implication of Lemma A.3 holds, (CA) from Lemma A.4 also holds. Furthermore, (19) yields Lemma A.6, for which Corollary A.7 remains a valid corollary. We will not make use of Lemma A.5: we will instead use directly (21).

We now have the tools needed to modify the proof of Theorem 1. Using the results we have obtained thus far, and replacing (6) by (22) and (5) by (20), we see that (A.13) from the proof of Theorem 1 can be replaced by

$$\begin{aligned}
0 \le \mathcal{R}(\widehat{\pi}_n) &\lesssim r_n^{-1}[\widetilde{\mathbb{G}}_n f_{\widehat{\pi}_n} - \widetilde{\mathbb{G}}_n f_{\pi_s^\star}] + \big[\mathcal{V}(\pi^\star) - \mathcal{V}(\pi_s^\star)\big] + o_P\big(r_n^{-1}\big) \\
&\le 2r_n^{-1}\sup_{f\in\mathcal{F}}|\widetilde{\mathbb{G}}_n f| + \big[\mathcal{V}(\pi^\star) - \mathcal{V}(\pi_s^\star)\big] + o_P\big(r_n^{-1}\big).
\end{aligned} \tag{B.1}$$

By (21) and the continuous mapping theorem [32, Theorem 1.3.6], $\sup_{f\in\mathcal{F}}|\widetilde{\mathbb{G}}_n f| = O_P(1)$. Hence, the leading term in the final inequality is $O_P(r_n^{-1})$, where this term does not depend on s. The final term also does not depend on s. By (18), the middle term above goes to zero as $s \downarrow 0$, where this convergence is uniform in both $\pi^\star \in \overline{\Pi}^\star$ and $\pi_s^\star \in B_\Pi(\pi^\star, s)$. Thus, $\mathcal{R}(\widehat{\pi}_n) = O_P(r_n^{-1})$.

We tighten this result to $\mathcal{R}(\widehat{\pi}_n) = o_P(r_n^{-1})$ as in the proof of Theorem 1. In particular, nearly identical arguments to those used in that proof show that

$$0 \le r_n\mathcal{R}(\widehat{\pi}_n) \le g\big(\widetilde{\mathbb{G}}_n, \mathcal{R}(\widehat{\pi}_n) + t_n\big) + o_P(1) \le h\big(\widetilde{\mathbb{G}}_n, \mathcal{R}(\widehat{\pi}_n) + t_n\big) + o_P(1).$$

The proof concludes by noting that (21) includes the condition that almost all sample paths of $\widetilde{\mathbb{G}}_P$ are uniformly continuous on $\mathcal{F}$ w.r.t. $\|\cdot\|$, and thus the right-hand side above is $o_P(1)$. In conclusion $\mathcal{R}(\widehat{\pi}_n) = o_P(r_n^{-1})$. □

B.2. *Proof for Section* 4.2

Proof of Lemma 8. If the bounded class Π is Donsker, then it is totally bounded in $L^2(P)$, so (17) is met. By Lemma A.1, $\overline{\Pi}$ is then compact. Let θ be a Lipschitz function from $[-2, 2]$ to $[0, 1]$ such that $\theta(0) = 1$ and $\theta(u) = 0$ if $|u| \geq \min_{a \neq a' \in \mathcal{A}} |a - a'|$. Introducing θ is merely a trick to generalize Lemma A.2. In particular, (A.1) becomes

$$Q\big(\pi(X), X\big) = \sum_{a \in \mathcal{A}} \theta\big(\pi(X) - a\big) Q(a, X)$$

for every policy $\pi : \mathcal{X} \to \mathcal{A}$, yielding

$$\big|\mathcal{V}(\pi_1) - \mathcal{V}(\pi_2)\big| \lesssim \|\pi_1 - \pi_2\| \tag{B.2}$$

for every $\pi_1, \pi_2 \in \overline{\Pi}$, hence (18). We also generalize (A.4), which becomes

$$f_\pi(O) = \sum_{a \in \mathcal{A}} \theta\big(\pi(X) - a\big)\left(\theta(A - a)\frac{Y - Q(A, X)}{P(A|X)} + Q(a, X)\right) - \mathcal{V}(\pi) \tag{B.3}$$

for every $\pi \in \overline{\Pi}$. The above equality and (B.2) imply $\|f_{\pi_1} - f_{\pi_2}\| \lesssim \|\pi_1 - \pi_2\|$ for every $\pi_1, \pi_2 \in \overline{\Pi}$, hence (19). The second part of the proof of Lemma A.5 can easily be generalized to derive that $\mathcal{F}$ is Donsker from (B.3) and the fact that Π is itself Donsker. Therefore, (21) is met and the proof is complete. □

B.3. *Proofs for Section* 4.3

We start by proving Lemma 11.

Proof of Lemma 11. Set arbitrarily $\pi_1, \pi_2 \in \overline{\Pi}$ and $m \in [\mathcal{V}(\pi_2) \pm c]$. We suppose without loss of generality that $\mathcal{V}(\pi_1) \leq \mathcal{V}(\pi_2)$. We will use the fact that (24) implies that $F_{\pi_1}(\mathcal{V}(\pi_1)) = F_{\pi_2}(\mathcal{V}(\pi_2)) = 1/2$ several times in this proof.

Firstly, note that

$$\begin{aligned}\big|F_{\pi_1}(m) - F_{\pi_2}(m)\big| &= \left|\mathbb{E}\left[\frac{\mathbf{1}\{A = \pi_1(X)\} - \mathbf{1}\{A = \pi_2(X)\}}{P(A|X)}\mathbf{1}\{Y \leq m\}\right]\right| \\ &\leq \frac{1}{2}\mathbb{E}\left[\frac{|\pi_1(X) - \pi_2(X)|}{P(A|X)}\mathbf{1}\{Y \leq m\}\right] \leq k_1\|\pi_1 - \pi_2\|\end{aligned} \tag{B.4}$$

where k_1 is a finite, positive constant that only depends on the lower bound on $P(A|X)$ from the strong positivity assumption.

Secondly, the continuous differentiability of F_{π_2} in $[\mathcal{V}(\pi_2) \pm c]$ and (24) imply the existence of $\tilde{m} \in [m, \mathcal{V}(\pi_2)]$ such that

$$\begin{aligned}\big|F_{\pi_2}\big(\mathcal{V}(\pi_2)\big) - F_{\pi_2}(m)\big| &= \big|\mathcal{V}(\pi_2) - m\big| \times \dot{F}_{\pi_2}(\tilde{m}) \\ &\geq \big|\mathcal{V}(\pi_2) - m\big| \times \inf_{\pi \in \overline{\Pi}} \inf_{m \in [\mathcal{V}(\pi) \pm c]} \dot{F}_\pi(m).\end{aligned}$$

Therefore, there exists a finite, positive constant k_2 such that

$$\big|\mathcal{V}(\pi_2) - m\big| \leq k_2\big|F_{\pi_2}\big(\mathcal{V}(\pi_2)\big) - F_{\pi_2}(m)\big|. \tag{B.5}$$

The remainder of this proof is broken into two parts: we will show that *(i)* $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq c$ implies $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq k_1k_2\|\pi_1 - \pi_2\|$, and *(ii)* $\|\pi_1 - \pi_2\| < c/k_1k_2$ yields $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq c$. By combining these two results, we will thus prove that $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq k_1k_2\|\pi_1 - \pi_2\|$ for all $\|\pi_1 - \pi_2\|$ sufficiently small, and this will complete the proof (k_1k_2 does not depend on π_1 or π_2).

Recall that $F_{\pi_1}(\mathcal{V}(\pi_1)) = F_{\pi_2}(\mathcal{V}(\pi_2)) = 1/2$. If $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq c$, then combining (B.4) and (B.5) at $m = \mathcal{V}(\pi_1) \in [\mathcal{V}(\pi_2) \pm c]$ establishes *(i)*:

$$\begin{aligned}\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) &\leq k_2\big[F_{\pi_2}\big(\mathcal{V}(\pi_2)\big) - F_{\pi_2}\big(\mathcal{V}(\pi_1)\big)\big] \\ &= k_2\big[F_{\pi_1}\big(\mathcal{V}(\pi_1)\big) - F_{\pi_2}\big(\mathcal{V}(\pi_1)\big)\big] \leq k_1k_2\|\pi_1 - \pi_2\|.\end{aligned}$$

We argue *(ii)* by contraposition. Suppose that $\mathcal{V}(\pi_1) < \mathcal{V}(\pi_2) - c$. By the monotonicity of cumulative distribution functions, $F_{\pi_2}(\mathcal{V}(\pi_1)) \leq F_{\pi_2}(\mathcal{V}(\pi_2) - c)$. Combining this with (B.5) at $m = \mathcal{V}(\pi_2) - c \in [\mathcal{V}(\pi_2) \pm c]$,

$$F_{\pi_2}\big(\mathcal{V}(\pi_2)\big) - F_{\pi_2}\big(\mathcal{V}(\pi_1)\big) \geq F_{\pi_2}\big(\mathcal{V}(\pi_2)\big) - F_{\pi_2}\big(\mathcal{V}(\pi_2) - c\big) \geq k_2^{-1} c.$$

By (B.4) and the fact that $F_{\pi_2}(\mathcal{V}(\pi_2)) = F_{\pi_1}(\mathcal{V}(\pi_1)) = 1/2$, the LHS expression is smaller than $k_1\|\pi_1 - \pi_2\|$. Therefore, $\|\pi_1 - \pi_2\| \geq c/k_1 k_2$. □

Lemma B.1. *Suppose the existence of a deterministic* $L > 0$ *such that, for all* $m \in \mathbb{R}$ *and sufficiently small* $\epsilon > 0$, *with* P*-probability one,*

$$P(m < Y \leq m + \epsilon | X) \leq L\epsilon.$$

Then, for all $m \in \mathbb{R}$ *and* $\pi_1, \pi_2 \in \overline{\Pi}$, $|\dot{F}_{\pi_2}(m) - \dot{F}_{\pi_1}(m)| \lesssim \|\pi_1 - \pi_2\|$.

Proof of Lemma B.1. Set arbitrarily $\pi_1, \pi_2 \in \overline{\Pi}$ and $m \in \mathbb{R}$. In this proof, the universal positive multiplicative constants attached to the $\lesssim$-inequalities do not depend on π_1, π_2, m. First, observe that

$$\big|\dot{F}_{\pi_2}(m) - \dot{F}_{\pi_1}(m)\big| = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon}\big|F_{\pi_2}(m+\epsilon) - F_{\pi_2}(m) - F_{\pi_1}(m+\epsilon) + F_{\pi_1}(m)\big|.$$

Second, mimicking the argument that previously lead to (B.4) we note that, for every $\epsilon > 0$ small enough,

$$\begin{aligned}
&\big|F_{\pi_2}(m+\epsilon) - F_{\pi_2}(m) - F_{\pi_1}(m+\epsilon) + F_{\pi_1}(m)\big| \\
&\quad \leq \frac{1}{2}\mathbb{E}\left[\frac{|\pi_1(X) - \pi_2(X)|}{P(A|X)}\mathbf{1}\{m < Y \leq m + \epsilon\}\right] \\
&\quad \lesssim \mathbb{E}\big[\big|\pi_1(X) - \pi_2(X)\big|\mathbf{1}\{m < Y \leq m + \epsilon\}\big] \\
&\quad = \mathbb{E}\big[\big|\pi_1(X) - \pi_2(X)\big|P(m < Y \leq m + \epsilon|X)\big] \\
&\quad \leq L\epsilon\mathbb{E}\big[\big|\pi_1(X) - \pi_2(X)\big|\big] \lesssim \epsilon\|\pi_1 - \pi_2\|.
\end{aligned}$$

Returning to the first display completes the proof. □

Corollary B.2. *Under the conditions of Lemmas 11 and B.1, and the additional assumption (25), the map* $\pi \mapsto \dot{F}_\pi(\mathcal{V}(\pi))$ *is uniformly continuous on* $\overline{\Pi}$ *w.r.t.* $\|\cdot\|$.

Proof of Corollary B.2. Set $\pi_1, \pi_2 \in \overline{\Pi}$. By Lemma 11, we can choose π_2 sufficiently close to π_1 in $L^2(P)$ (where "sufficiently close" does not depend on the choice of π_1) so that $|\mathcal{V}(\pi_1) - \mathcal{V}(\pi_2)| \leq c$, where c is the constant from (24). For all such choices of π_1, π_2,

$$\begin{aligned}
\big|\dot{F}_{\pi_2}\big(\mathcal{V}(\pi_2)\big) - \dot{F}_{\pi_1}\big(\mathcal{V}(\pi_1)\big)\big| &\leq \big|\dot{F}_{\pi_2}\big(\mathcal{V}(\pi_2)\big) - \dot{F}_{\pi_1}\big(\mathcal{V}(\pi_2)\big)\big| + \big|\dot{F}_{\pi_1}\big(\mathcal{V}(\pi_2)\big) - \dot{F}_{\pi_1}\big(\mathcal{V}(\pi_1)\big)\big| \\
&\lesssim \|\pi_1 - \pi_2\| + \omega\big(\|\pi_1 - \pi_2\|\big),
\end{aligned}$$

where the constant on the right does not depend on π_1, π_2. The final bound used Lemma B.1 and (25), hence the appearance of $\omega(\cdot)$. As $\omega(0) = 0$ and ω is continuous at zero, one can choose $\|\pi_1 - \pi_2\|$ sufficiently small so that the right-hand side is less than any $\epsilon > 0$. As "sufficiently small" does not depend on the choice of π_1, $\pi \mapsto \dot{F}_\pi(\mathcal{V}(\pi))$ is uniformly continuous on $\overline{\Pi}$ w.r.t. $\|\cdot\|$. □

Acknowledgements

The authors wish to thank warmly the two reviewers, associate editor and editor for their careful reviews. They led to a significantly enhanced version of the original manuscript. Alex Luedtke gratefully acknowledges the support of the New Development Fund of the Fred Hutchinson Cancer Research Center and the NIH under award number DP2-LM013340. Antoine Chambaz acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-13-BS01-0005 (project SPADRO) and that this research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsoring agencies.

References

[1] S. Athey and S. Wager. Efficient policy learning. arXiv preprint arXiv:1702.02896v4 (2017).

[2] J. Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** (2) (2007) 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

[3] P. J. Bickel, C. A. J. Klaassen, Y. Ritov and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993. MR1245941

[4] A. Browder. *Mathematical Analysis: An Introduction*. Springer-Verlag, New York, 2012. MR1411675 https://doi.org/10.1007/978-1-4612-0715-3

[5] P. Chaffee and M. J. van der Laan. Targeted minimum loss based estimation based on directly solving the efficient influence curve equation. Technical report, UC Berkeley Division of Biostatistics Working Paper Series, 2011.

[6] A. Chambaz, W. Zheng and M. J. van der Laan. Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward. *Ann. Statist.* **45** (6) (2017) 2537–2564. MR3737901 https://doi.org/10.1214/16-AOS1534

[7] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins. Double machine learning for treatment and causal parameters. *Econometrica* **21** (2018) C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[8] A.-M. Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems* 172–180, 2011.

[9] K. Hirano and J. R. Porter. Asymptotics for statistical treatment rules. *Econometrica* **77** (5) (2009) 1683–1701.

[10] T. Kitagawa and A. Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* **86** (2) (2018) 591–616. MR3783340 https://doi.org/10.3982/ECTA13288

[11] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** (6) (2006) 2593–2656. MR2329442 https://doi.org/10.1214/009053606000001019

[12] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Mathematics* **2033**. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2829871 https://doi.org/10.1007/978-3-642-22147-7

[13] L. Le Cam. Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3** (1960) 37–98. MR0126903 https://doi.org/10.1111/ectj.12097

[14] K. A. Linn, E. B. Laber and L. A. Stefanski. Interactive Q-learning for quantiles. *J. Amer. Statist. Assoc.* **just-accepted** (2016) 1–37.

[15] A. R. Luedtke and M. J. van der Laan. Super-learning of an optimal dynamic treatment rule. *Int. J. Biostat.* **12** (1) (2016) 305–332. MR3505699 https://doi.org/10.1515/ijb-2015-0052

[16] A. R. Luedtke and M. J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** (2) (2016) 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

[17] A. R. Luedtke and M. J. van der Laan. Comment: "Personalized dose finding using outcome weighted learning". *J. Amer. Statist. Assoc.* **111** (516) (2016) 1526–1530. MR3601708 https://doi.org/10.1080/01621459.2016.1242427

[18] A. R. Luedtke and M. J. van der Laan. Corrigendum to: Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of Causal Inference* **3** (2) (2016) 267–271.

[19] C. F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica* **72** (4) (2004) 1221–1246. MR2064712 https://doi.org/10.1111/j.1468-0262.2004.00530.x

[20] M. Qian and S. Murphy. Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** (2011) 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864

[21] D. Rubin and M. J. van der Laan. A doubly robust censoring unbiased transformation. *Int. J. Biostat.* **3** (2007), Art. 4, 21. MR2306842 https://doi.org/10.2202/1557-4679.1052

[22] D. B. Rubin and M. J. van der Laan. Statistical issues and limitations in personalized medicine research with clinical trials. *Int. J. Biostat.* **8** (1) (2012), Article 18.

[23] A. Sheehy and J. A. Wellner. Uniform Donsker classes of functions. *Ann. Probab.* **20** (1992) 1983–2030. MR1188051 https://doi.org/10.1111/ectj.12097

[24] J. Stoye. Minimax regret treatment choice with finite samples. *J. Econometrics* **151** (1) (2009) 70–81. MR2538273 https://doi.org/10.1016/j.jeconom.2009.02.013

[25] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** (1) (2004) 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

[26] M. J. van der Laan and A. R. Luedtke. Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of Causal Inference* **3** (1) (2014) 61–95. MR1188051 https://doi.org/10.1515/jci-2013-0022

[27] M. J. van der Laan and A. R. Luedtke. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. Technical report 329, Division of Biostatistics, University of California, Berkeley, 2014. MR3505699 https://doi.org/10.1515/ijb-2015-0052

[28] M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York, 2003. MR1958123 https://doi.org/10.1007/978-0-387-21700-0

[29] M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

[30] M. J. van der Laan and D. B. Rubin. Targeted maximum likelihood learning. *Int. J. Biostat.* **2** (1) (2006), Article 11. MR2306500 https://doi.org/10.2202/1557-4679.1043

[31] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, New York, 1998. MR1652247 https://doi.org/10.1017/CBO9780511802256

[32] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[33] B. Zhang, A. A. Tsiatis, M. Davidian, M. Zhang and E. Laber. Estimating optimal treatment regimes from a classification perspective. *Stat* **68** (1) (2012) 103–114.

[34] Y. Zhao, D. Zeng, A. Rush and M. Kosorok. Estimating individual treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** (2012) 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

[35] W. Zheng and M. J. van der Laan. Targeted maximum likelihood estimation of natural direct effects. *Int. J. Biostat.* **8** (1) (2012), Art. 3, 42. MR2923277 https://doi.org/10.2202/1557-4679.1361