Faster Rates for Policy Learning

Alexander R Luedtke, Antoine Chambaz

Faster Rates for Policy Learning

Alexander R. Luedtke[1,2] and Antoine Chambaz[3,4]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, USA

[2]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, USA

[3]Modal'X, UPL, Univ Paris Nanterre, F92000 Nanterre France

[4]Division of Biostatistics, School of Public Health, UC Berkeley, USA

Abstract

This article improves the existing proven rates of 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, with regret decaying at a faster rate than the standard error of an efficient estimator of the value of an optimal policy. We also give a result from the classification literature that shows that faster regret decay is possible via plug-in estimation provided a margin condition holds. Four examples are considered. In these examples, the regret is expressed in terms of either the mean value or the median value; the number of possible actions is either two or finitely many; and the sampling scheme is either independent and identically distributed or sequential, where the latter represents a contextual bandit sampling scheme.

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. 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 a cumulated sum of rewards 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 [2, 3, 4], ensemble techniques [5], 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 slower than $n^{-1/2}$. Notable exceptions to this $n^{-1/2}$ restriction occur for outcome weighted learning (OWL) methods under a hard margin [2] and rates attainable by plug-in estimation strategies [6, 7, 8].

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 [9]. The second 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 [10]. 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.

All of our results are for the regret under a fixed data generating distribution. The fast 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. 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 [12]. 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 [5], so that one need 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.

We do not concern ourselves with measurability issues, with the understanding that some work may be needed to make these arguments precise [13].

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 $\mathbb{E}$ the expectation under 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

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

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. 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 functional

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

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

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

By the bound on Y and the strong positivity assumption, there exists a universal constant $M > 0$ such that $P(\sup_{\pi \in \Pi} |f_\pi(O)| \leq M) = 1$.

We refer the interested reader to [14, Section 25.3] and [7] for definitions and proofs of these facts. Although the class

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

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.

2 Main Results

2.1 Preliminary

Suppose that we observe mutually independent data-structures $O_1 \equiv (X_1, A_1, Y_1), \ldots, O_n \equiv (X_n, A_n, Y_n)$ drawn from P. Let $P_n \equiv n^{-1} \sum_{i=1}^n \text{Dirac}(O_i)$ denote the empirical measure. 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:

$$\text{Rem}_n \equiv \sup_{\pi \in \Pi} \left| \widehat{\mathcal{V}}(\pi) - \mathcal{V}(\pi) - (P_n - P) f_\pi \right|. \quad (5)$$

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 [15] or targeted minimum loss-based estimation (TMLE) [16, 17]. Without imposing empirical process conditions, one could use cross-validated estimating equations [18, 19], now also called double machine learning [20], or a cross-validated TMLE [21]. 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 [22]. For any of these listed approaches, obtaining the above uniformity over π is straightforward if Π is a Donsker class.

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

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

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 $o_P(\text{Rem}_n)$ term allows us to obtain an approximate solution to the ERM problem, which is useful because the optimization on the right is non-concave.

We now present a result that is similar to the groundbreaking results for ERMs based on general losses given in [11]. In Section 3.1, we discuss how our results relate to those in [11].

The upcoming theorem uses $\overline{\Pi}$ to denote 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 RHS set in the above display. Finally, we define

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

which roughly corresponds to the best approximation in Π of the closest optimal policy $\pi^\star \in \overline{\Pi}^\star$ to the estimated policy $\widehat{\pi}$. We will show that $\mathrm{EP}_n = o_P(n^{-1/2})$ under mild assumptions.

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

$$0 \leq \mathcal{R}(\widehat{\pi}) \leq \mathrm{EP}_n + [1 + o_P(1)]\,\mathrm{Rem}_n\,.$$

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

The proof of Theorem 1 is given in Section A. 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 $L^2(P)$ while that of the second one uses the stronger assumption that Π is Donsker. Three detailed remarks on this result 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 23].

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 [5].

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 [10] 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, $\widehat{\pi}$ is given by

$$\widehat{\pi}(x) \equiv \mathbb{1}\{\widehat{\gamma}(x) > 0\} - \mathbb{1}\{\widehat{\gamma}(x) \leq 0\}.$$

Extensions of the result of [10] have previously been presented in both the reinforcement learning literature [6] and in the optimal individualized treatment literature [7, 8]. We therefore omit any proof, and refer the reader to [10, Lemma 5.2] for further details.

The upcoming theorem uses $\|\cdot\|$ to denote the $L^2(P)$ norm and $\|\cdot\|_\infty$ 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 2. *Suppose MA holds. If* $\|\widehat{\gamma} - \gamma\| = o_P(1)$, *then* $0 \leq \mathcal{R}(\widehat{\pi}) \lesssim \|\widehat{\gamma} - \gamma\|^{2(1+\alpha)/(2+\alpha)}$. *If* $\|\widehat{\gamma} - \gamma\|_\infty = o_P(1)$, *then* $0 \leq \mathcal{R}(\widehat{\pi}) \lesssim \|\widehat{\gamma} - \gamma\|_\infty^{1+\alpha}$.

We now briefly describe this result, though we note that it has been discussed thoroughly elsewhere [10, 7, 8]. 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 $\gamma(X)$ is 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 3 (Estimating γ)**.** Doubly robust unbiased transformations [24] provide one way to estimate γ (Section 3.1 of [25]). In particular, one can regress (via any desired algorithm) the pseudo-outcome

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

against X, both for double robustness and for efficiency gains. Above $\widehat{P}(A|X)$ and $\widehat{\mathbb{E}}[Y|A,X]$ are estimators of $P(A|X)$ and $\mathbb{E}[Y|A,X]$. One can use cross-validation to avoid dependence on empirical process conditions for the estimators of $P(A|X)$ and $\mathbb{E}[Y|A,X]$. □

3 Three Remarks on Theorem 1

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 belongs 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.5 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 simlar 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 (only uses a Donsker condition) was given for general result for ERMs in [26, 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 shown: 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 EP_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 and reward regression. It is thus sufficient to assume that this product is $o_P(n^{-1/2})$ to guarantee that $\text{Rem}_n = o_P(n^{-1/2})$. It is worth noting that the product is $O_P(n^{-1})$ in a well-specified parametric model. On the other hand, the magnitude of EP_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 EP_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 $\widehat{\pi}$, even when $\text{Rem}_n = o_P(n^{-1/2})$. A notable exception to this lack of strict ordering between Rem_n and EP_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 the 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. High-probability regret upper bounds were derived, with leading constants that scale with the standard error of a semiparametric efficient estimator for policy evaluation. The authors argue that this leading constant demonstrates the importance of using semiparametric efficient value estimators to define the empirical risk used by their estimator. As we discussed in Section 3.2, we fully agree with the importance of using efficient estimators to estimate the empirical risk. Nonetheless, the present work shows that the regret of ERMs decays faster than the standard error of an efficient estimator of the value and so, the regret of ERMs that use this empirical risk will not necessarily scale with the standard error of this estimator.

Like in [1], our results are given under a fixed data generating distribution P. We implicitly leverage the behavior of γ near the decision boundary (zero) under this distribution P. Crucially, there is a problem-dependent constant that can be made arbitrarily large if one chooses a P for which $\gamma(X)$ concentrates a large amount of mass near the decision boundary. Hence, the minimax rate is not faster than $n^{-1/2}$ unless one constrains the class of distributions to which P can belong via a margin condition (see [23, 11]). The implication of this observation is encouraging: it would not be surprising to find that, without a margin condition, the minimax regret does in fact decay at the rate of the standard error of an efficient estimator of an optimal policy. The leading constant may also critically depend on

the efficient variance of such an estimator. It is worth studying whether the high-probability, finite sample results in [1] can be extended in this direction, thereby demonstrating a criterion for which efficient value estimation is mathematically indispensable for obtaining the optimal regret decay.

4 Extension of Main ERM Result and Three Examples

4.1 Higher Level Result

Suppose we observe $(O_1, \ldots, O_n)$ 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 independent and identically distributed (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, 4.3 and 4.4. We let $\|\cdot\|$ denote the $L^2(P)$ norm.

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

$$\Pi \text{ is totally bounded with respect to } \|\cdot\|. \tag{8}$$

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

$$\mathcal{V}(\cdot) \text{ is uniformly continuous on } \overline{\Pi} \text{ with respect to } \|\cdot\|. \tag{9}$$

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) and an estimator $\{\widehat{\mathcal{V}}(\pi) : \pi \in \Pi\}$ of $\{\mathcal{V}(\pi) : \pi \in \Pi\}$ such that, for a (possibly stochastic) rate $r_n \to +\infty$,

$$\pi \mapsto f_\pi \text{ is uniformly continuous from } \overline{\Pi} \text{ to } L^2(P), \quad \text{and} \tag{10}$$

$$\sup_{\pi \in \Pi} \left| r_n \left(\widehat{\mathcal{V}}(\pi) - \mathcal{V}(\pi) \right) - \widetilde{\mathbb{G}}_n f_\pi \right| = o_P(1). \tag{11}$$

In (10), both spaces are equipped with $\|\cdot\|$. In (11), $\widetilde{\mathbb{G}}_n \in \ell^\infty(\mathcal{F})$ 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

$$\widetilde{\mathbb{G}}_n \rightsquigarrow \widetilde{\mathbb{G}}_P \text{ in } \ell^\infty(\mathcal{F}), \tag{12}$$

where almost all sample paths of $\widetilde{\mathbb{G}}_P$ are uniformly continuous with respect to $\|\cdot\|$.

Finally we also assume that, for the same rate r_n as in (11), $\widehat{\pi} \in \Pi$ satisfies the ERM property

$$\widehat{\mathcal{V}}(\widehat{\pi}) \leq \inf_{\pi \in \Pi} \widehat{\mathcal{V}}(\pi) + o_P(r_n^{-1}). \tag{13}$$

Like (6), (13) 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 4 (More General ERM Result)**.** *If (8) through (13) 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 5. If $\widetilde{\mathbb{G}}_P$ is equal to the zero-mean Gaussian process with covariance given by

$\mathbb{E}[\mathbb{G}_P f \mathbb{G}_P g] = Pfg - PfPg$, then $\widetilde{\mathbb{G}}_P$ has almost surely uniformly continuous sample paths. In particular, [13, Example 1.5.10] shows that $\widetilde{\mathbb{G}}_P$ has almost surely uniformly continuous sample paths with respect to 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 this result via three 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 27]. The third example focuses again on the mean reward, but considers a non-i.i.d., contextual-bandit-type setting in which context-specific actions may be informed by earlier observations [8].

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 parallel between the LHS of (11) with $r_n \equiv n^{1/2}$ and $n^{1/2} \operatorname{Rem}_n$ from (5).

We prove the next lemma in Section B.2:

Lemma 6. *If* Π *is Donsker, then* (8), (9), (10) *and* (12) *are met.*

Therefore, Theorem 4 yields the following corollary:

Corollary 7. *In the context of Section 4.2, suppose that* $\widehat{\pi}$ *satisfies* (13) *with* $r_n \equiv n^{1/2}$ *and that*

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

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 $F_\pi(m) \equiv \mathbb{E}\left[P(Y \leq m | A = \pi(X), X)\right]$. Under some causal assumptions, F_π is the cumulative distribution function of the reward in the counterfactual 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

$$\mathcal{V}(\pi) \equiv \inf \left\{ m \in \mathbb{R} : 1/2 \leq F_\pi(m) \right\}. \tag{14}$$

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 (9). 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

$$0 < \inf_{\pi \in \Pi} \inf_{m \in [\mathcal{V}(\pi) \pm c]} \dot{F}_\pi(m) \leq \sup_{\pi \in \Pi} \sup_{m \in [\mathcal{V}(\pi) \pm c]} \dot{F}_\pi(m) < \infty. \tag{15}$$

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

$$\sup_{\pi \in \overline{\Pi}} \sup_{|k| \leq c} \left[\left| \dot{F}_\pi(\mathcal{V}(\pi) + k) - \dot{F}_\pi(\mathcal{V}(\pi)) \right| - \omega(|k|) \right] \leq 0. \tag{16}$$

Remark 8. 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

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

□

The next lemma gives conditions under which (9) holds. Its proof is given in Appendix B.3.

Lemma 9. *If (15) and (16) are met, then (9) holds.*

For every $\pi \in \mathcal{P}$, $\mathcal{V}(\pi)$ defined in (14) can be viewed as the evaluation at P of the functional

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

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

$$f_\pi(o) \equiv \frac{\mathbb{1}\{a = \pi(x)\}}{P(A = a | X = x) \dot{F}_\pi(\mathcal{V}(\pi))} \left[\mathbb{1}\{y \leq \mathcal{V}(\pi)\} - P\{Y \leq \mathcal{V}(\pi) | A = a, X = x\} \right] + \frac{P\{Y \leq \mathcal{V}(\pi) | A = \pi(x), X = x\} - 1/2}{\dot{F}_\pi(\mathcal{V}(\pi))}. \tag{17}$$

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 (11) 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 11 at the end of the present section.

With this choice of $\widetilde{\mathbb{G}}_n$, (12) is met and Π Donsker yields (8), 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 4 yields the following corollary:

Corollary 10. *In the context of Section 4.3, suppose that (15) and (16) are met. Let $\widehat{\pi}$ satisfy (13) with $r_n \equiv n^{1/2}$ and*

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

If Π is Donsker and (10) holds, then $\mathcal{R}(\widehat{\pi}) = o_P(n^{-1/2})$.

Since Π is Donsker, (10) 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.7 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.9 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 11. One can, for example, establish conditions under which the estimating equation-based estimator defined as a solution in v to

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

satisfies (11). 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 $\widehat{P}$. □

4.4 Example 3: Sequential Decisions to Maximize the Mean Reward of a Binary Action

The sampling design. This section builds upon [8]. Let $\{(X_n, Z_n(-1), Z_n(1)\}_{n\geq 1}$ be a sequence of random variables drawn independently from a distribution $\mathbb{P}$ such that, if $\{\tilde{A}_n\}_{n\geq 1}$ is a sequence of independent actions drawn from the conditional distribution of A (an action with support $\{-1, 1\}$) given X under the same P as in Section 1.3, then the resulting sequence $\{(X_n, \tilde{A}_n, Z_n(\tilde{A}_n))\}_{n\geq 1}$ is an i.i.d. sample from P. For notational simplicity, we assume that all the random variables $Z_n(-1)$ and $Z_n(1)$ take their values in $[0, 1]$.

In this example, however, the distribution ν_n of $(O_1, \ldots, O_n)$ does not write as a product because the actions $A_1, \ldots, A_n$ are not i.i.d. On the contrary, once sufficiently many observations have been accrued to carry out inference then, sequentially, each new randomized action A_n is drawn conditionally on X_n and an estimate derived from the previous observations $O_1 \equiv (X_1, A_1, Y_1), \ldots, O_{n-1} \equiv (X_{n-1}, A_{n-1}, Y_{n-1})$ yielded by the previous actions, where $Y_i \equiv Z_i(A_i)$ for $i = 1, \ldots, n-1$. Let us describe how the data-adaptive design unfolds.

Let $\{t_n\}_{n\geq 1}$ and $\{\xi_n\}_{n\geq 1}$ be two nonincreasing sequences with $t_1 \leq 1/2$, $\lim_n t_n > 0$ and $\lim_n \xi_n > 0$. For each $n \geq 1$, let G_n be a nondecreasing κ_n-Lipschitz functions approximating $u \mapsto \mathbb{1}\{u \geq 0\}$ in the sense that $G_n(u) = t_n$ for $u \leq -\xi_n$, $G_n(0) = 1/2$, $G_n(u) = 1 - t_n$ for $u \geq \xi_n$. We also suppose that $\limsup_n \kappa_n < \infty$. For each $n \geq 1$, let $\mathcal{Q}_n \equiv \{Q_\beta : \beta \in B_n\}$ be a working model consisting of functions mapping $\{-1, 1\} \times \mathcal{X}$ to $[0, 1]$. Each $Q_\beta \in \mathcal{Q}_n$ can be viewed as a candidate approximation to the function $(A, X) \mapsto \mathbb{1}\{A = 1\}\, \mathbb{E}_{\mathbb{P}}[Z(1)|X] + \mathbb{1}\{A = -1\}\, \mathbb{E}_{\mathbb{P}}[Z(-1)|X]$.

For some $n_0 \geq 1$, $A_1, \ldots, A_{n_0}$ are independent draws from the conditional distribution of A given X under P. For every $i = 1, \ldots, n_0$, let $g_i : \{-1, 1\} \times \mathcal{X} \to [0, 1]$ be defined pointwise by $g_i(a, x) \equiv P(A = a|X = x)$. Set $n > n_0$, suppose that we have fully specified how $O_1, \ldots, O_{n-1}$ have been sampled through the description of how $A_1, \ldots, A_{n-1}$ have

been randomly drawn from Rademacher laws[†] with parameters $g_1(1, X_1), \ldots, g_{n-1}(1, X_{n-1})$, respectively. Now, let us describe how O_n is sampled by specifying how action A_n is randomized, therefore completing the description of the sampling design. Based on $O_1, \ldots, O_{n-1}$, we define

$$\beta_n \equiv \arg\min_{\beta \in B_n} \sum_{i=1}^{n-1} \frac{\mathcal{L}(Q_\beta)(O_i)}{g_i(A_i|X_i)}$$

where the least-square loss function $\mathcal{L}$ is given by $\mathcal{L}(Q_\beta)(O) \equiv (Y - Q_\beta(A, X))^2$. This yields the mapping $g_n : \{-1, 1\} \times \mathcal{X} \to [t_n, 1 - t_n] \subset [0, 1]$ given by

$$1 - g_n(-1, x) \equiv g_n(1, x) \equiv G_n\big(Q_{\beta_n}(1, x) - Q_{\beta_n}(-1, x)\big).$$

Once X_n is observed, we sample A_n conditionally on $O_1, \ldots, O_{n-1}$ and X_n from the Rademacher law with parameter $g_n(1, X_n)$, we then carry out action A_n, observe $Y_n \equiv Z_n(A_n)$, and form $O_n \equiv (X_n, A_n, Y_n)$.

Remark 12. The sampling design looks for a trade-off between exploration and exploitation. Here, exploitation consists in using the current best estimates $Q_{\beta_n}(1, X_n)$ and $Q_{\beta_n}(-1, X_n)$ of $\mathbb{E}_{\mathbb{P}}[Z(1)|X = X_n]$ and $\mathbb{E}_{\mathbb{P}}[Z(-1)|X = X_n]$ to favor the action which seems to have the larger mean reward in context X_n. If the absolute value of the difference between the two estimates is larger than ξ_n, then the supposedly superior action is carried out with probability $(1 - t_n)$. Exploration consists in giving the supposedly inferior action a probability t_n to be undertaken nonetheless. This allows, for instance, to correct a possibly poor estimation of $\mathbb{E}_{\mathbb{P}}[Z(1)|X]$ and $\mathbb{E}_{\mathbb{P}}[Z(-1)|X]$ in some strata of $\mathcal{X}$. □

Remark 13. For any $g : \{-1, 1\} \times \mathcal{X} \to [0, 1]$, let $\mathbb{P}^g$ be the distribution of O defined by *(i)* sampling $(X, Z(-1), Z(1))$ from $\mathbb{P}$, *(ii)* sampling A from the Rademacher law with parameter $g(1, X)$, *(iii)* setting $O \equiv (X, A, Z(A))$. Note that $W \equiv g(A, X)$ is a deterministic function of g and O. Therefore, we can augment O with W, *i.e.*, substitute (O, W) for

[†]The Rademacher law with parameter $p \in [0, 1]$ is the law of $A \in \{-1, 1\}$ such that $A = 1$ with probability p.

O, while still denoting $(O, W) \sim \mathbb{P}^g$. A balanced design corresponds to $\mathbb{P}^b$, where $b : \{-1, 1\} \times \mathcal{X} \to [0, 1]$ only takes the value $1/2$. □

Optimal policy estimation. We now define set of candidate policies Π. In particular Π is the set of policies such that, for each $\pi \in \Pi$, there exists $\beta \in \cup_{n \geq 1} B_n$ for which

$$\begin{aligned} \pi(X) &\equiv \arg\max_{a \in \{-1,1\}} Q_\beta(a, X) \\ &= \mathbb{1}\{Q_\beta(1, X) \geq Q_\beta(-1, X)\} - \mathbb{1}\{Q_\beta(1, X) < Q_\beta(-1, X)\}, \end{aligned}$$

where the choice to let $A = 1$ when $Q_\beta(1, X) = Q_\beta(-1, X)$ in the second equality is a convention. The value $\mathcal{V}(\pi)$ of $\pi \in \Pi$ is still given by (1) and its (within class Π) regret $\mathcal{R}(\pi)$ by (2).

Recall the definition of $\mathcal{F}$ characterized by (3) and (4). Let $\widetilde{\mathbb{G}}_n \in \ell^\infty(\mathcal{F})$ be such that, for each $f \in \mathcal{F}$,

$$n^{1/2} \widetilde{\mathbb{G}}_n f \equiv \sum_{i=1}^{n} \left(f(O_i, W_i) - \mathbb{E}_{\mathbb{P}^{g_i}}[f(O_i, W_i) | O_1, \ldots, O_{i-1}] \right),$$

where we use the fact that g_i is deterministic when one conditions on $O_1, \ldots, O_{i-1}$ (*i.e.*, on nothing when $i = 1$). By exploiting the martingale structure of each $n^{1/2} \widetilde{\mathbb{G}}_n f$, it is possible to show that $\widetilde{\mathbb{G}}_n$ satisfies (12) under mild conditions (expressed in terms of the uniform entropy integral) on $\{\mathcal{Q}_n\}_{n \geq 1}$ [28, Lemma B.3]. Theorem 4 yields a corollary for our sequential design setting. We discuss the conditions of the corollary following its statement.

Corollary 14. *In the context of Section 4.4, suppose that $\{\mathcal{Q}_n\}_{n \geq 1}$ is chosen in such a way that Π is separable and $J_1(\eta_n, \Pi) = o(1)$ for any $\eta_n = o(1)$, where $J_1(\eta, \Pi)$ is the uniform entropy integral evaluated at η of Π wrt the envelope function constantly equal to one. Moreover, suppose that $\widehat{\pi}$ satisfies (13) with $r_n \equiv n^{1/2}$ and that*

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

Then $\mathcal{R}(\widehat{\pi}) = o_P(n^{-1/2})$.

Lemma 6 shows that (9) and (10) are met. Conditions (8) and (12) follow from Lemma A.7 in Appendix A and from the maximal inequality stated in [28, Lemma B.3], whose assumptions drive the conditions on $\{\mathcal{Q}_n\}_{n\geq 1}$ stated in Corollary 14. A concrete example of such a sequence can be found in [8, Section 4.4] (see also Section 4.1 therein for a brief reminder about the uniform integral entropy). Additional mild assumptions guaranteeing that an estimator $\widehat{\mathcal{V}}$ can be defined as requested by Corollary 14 can be adapted from [8]. Finally, we note for the ERM property (13) that each $\pi \in \Pi$ is known once one knows the $\beta \in \cup_{n\geq 1} B_n$ that indexes this π. Hence, it suffices to study the estimated value $\widehat{\mathcal{V}}(\pi)$ for each such π. The appropriate algorithm for carrying out this optimization will depend on the particular choice of classes B_n and the working model Q_β.

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. Despite this surprising phenomenon, we continue to advocate the use of first-order efficient value estimators.

Acknowledgements

Alex Luedtke gratefully acknowledges the support of the New Development Fund of the Fred Hutchinson Cancer Research Center. 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).

Appendix

A Proof of Main ERM Result

We begin this section with five lemmas and a corollary, whose proofs only require results from functional analysis. We then prove Lemma A.7 and Theorem 1, using results from empirical process theory.

In Lemmas A.1 through A.4 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 with respect to $\|\cdot\|$. 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.5 and the proof of Theorem 1 use several additional definitions, which are given immediately before Lemma A.5.

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, [29, 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 $\pi : \mathcal{X} \to \{-1, 1\}$,

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

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

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

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

Recall the definition (7) of $\overline{\Pi}^\star$.

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

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 $\overline{\Pi}^\star \neq \emptyset$.

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*

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

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 $\delta > 0$ such that $\mathcal{R}(\pi_m) \to 0$ and, for all $m \geq 1$,

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

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

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

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. □

Recall the definition (4) of $\mathcal{F}$. The next lemma uses the following definitions. We let $\ell^\infty(\mathcal{F})$ denote the metric space of all bounded functions $z : \mathcal{F} \to \mathbb{R}$, equipped with the supremum norm and, 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.5. *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.6. *Recall the definition of* g *from Lemma A.5. Let* $h : \ell^\infty(\mathcal{F}) \times [0, \infty) \to \mathbb{R}$ *be such that, for all* $z \in \ell^\infty(\mathcal{F})$,

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

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.5 and Corollary A.6. Fix a sequence $\{(z_m, r_m)\}_{m\geq 1}$ satisfying the conditions of the Lemma A.5. Observe that, for every $m \geq 1$,

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

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

$$\begin{aligned}
g(z_m, r_m) &= g(z_m, r_m) - g(z, r_m) + g(z, r_m) \\
&= g(z_m, r_m) - g(z, r_m) + \left[\sup_{\pi\in\Pi_{r_m}} z(f_\pi) - \limsup_{s\downarrow 0} \sup_{\pi^\star\in\overline{\Pi}^\star} \inf_{\pi\in B_\Pi(\pi^\star, s)} z(f_\pi) \right]. \\
&\leq \left[\sup_{\pi\in\Pi_{r_m}} z(f_\pi) - \limsup_{s\downarrow 0} \sup_{\pi^\star\in\overline{\Pi}^\star} \inf_{\pi\in B_\Pi(\pi^\star, s)} z(f_\pi) \right] + o(1). \qquad \text{(A.4)}
\end{aligned}$$

Let us show now that the RHS expression in (A.4) is $o(1)$.

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

$$\sup_{\pi\in\Pi_r} z(f_\pi) \leq z(f_{\pi_r}) + r. \qquad \text{(A.5)}$$

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

$$\|\pi_r^\star - \pi_r\| \leq \inf_{\pi^\star\in\overline{\Pi}^\star} \|\pi^\star - \pi_r\| + r \leq \sup_{\pi\in\Pi_r} \inf_{\pi^\star\in\overline{\Pi}^\star} \|\pi^\star - \pi\| + r. \qquad \text{(A.6)}$$

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

$$\begin{aligned} z(f_{\tilde{\pi}_r}) &\leq \inf_{\pi \in B_\Pi(\pi_r^\star, r)} z(f_\pi) + r \leq \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, r)} z(f_\pi) + r \\ &= \left(\limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi) + o_r(1) \right) + r, \end{aligned} \tag{A.7}$$

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.5) and (A.7), we thus have

$$\sup_{\pi \in \Pi_r} z(f_\pi) - \limsup_{s \downarrow 0} \sup_{\pi^\star \in \overline{\Pi}^\star} \inf_{\pi \in B_\Pi(\pi^\star, s)} z(f_\pi) \leq z(f_{\pi_r}) - z(f_{\tilde{\pi}_r}) + 2r + o_r(1). \tag{A.8}$$

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.6) and (A.11) imply that

$$\begin{aligned} \|f_{\pi_r} - f_{\tilde{\pi}_r}\| &\leq \left\|f_{\pi_r} - f_{\pi_r^\star}\right\| + \left\|f_{\pi_r^\star} - f_{\tilde{\pi}_r}\right\| \\ &\lesssim \|\pi_r - \pi_r^\star\| + \|\pi_r^\star - \tilde{\pi}_r\| \leq \sup_{\pi \in \Pi_r} \inf_{\pi^\star \in \overline{\Pi}^\star} \|\pi^\star - \pi\| + 2r. \end{aligned} \tag{A.9}$$

Because $\Pi_r \subset \{\pi \in \overline{\Pi} : \mathcal{R}(\pi) \leq r\}$, 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.9) 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.8),

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

Taking $r = r_m$ and using the above result reveals that the RHS expression in (A.4) is $o(1)$ indeed. This completes the proof of Lemma A.5.

In the context of Corollary A.6, let $\{(z_m, r_m)\}_{m\geq 1}$ 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.5 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.6 is complete. □

Lemma A.7. *If Π is Donsker, then $\mathcal{F}$ is also Donsker.*

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

$$\begin{aligned} 2f_\pi(O) &= |A + \pi(X)| \frac{Y - Q(A, X)}{P(A|X)} \\ &\quad + (1 + \pi(X))Q(1, X) + (1 - \pi(X))Q(-1, X) - 2\mathcal{V}(\pi). \end{aligned} \tag{A.10}$$

For future use, we first note that (A.1) and (A.2) imply, for any $\pi_1, \pi_2 \in \Pi$,

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

hence

$$\|f_{\pi_1} - f_{\pi_2}\| \lesssim \|\pi_1 - \pi_2\|. \tag{A.11}$$

Introduce $\phi : \mathbb{R}^5 \to \mathbb{R}$ given by $2\phi(u) = u_1|u_2 + u_3| + (1 + u_3)u_4 + (1 + u_3)u_5$ and f_1, f_2, f_4, f_5 be the function 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 $\phi \circ \mathbf{f} = \tilde{f}_\pi$ if $\mathbf{f} \in \tilde{\mathcal{F}}$ writes as $\mathbf{f} = (f_1, f_2, \pi, f_4, f_5)$. In light of (A.10), 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

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

(the bound on Y implies that $\|\gamma\|_\infty$ is finite). By [13, Theorem 2.10.6], whose conditions are obviously met, $\phi \circ \tilde{\mathcal{F}} = \{\tilde{f}_\pi : \pi \in \Pi\}$ 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, [13, Example 1.10.7] yields that $\{\tilde{f}_\pi - \lambda : \pi \in \Pi, \lambda \in \Lambda\}$ is Donsker, and so is its subset $\mathcal{F}$. □

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.7, 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:

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

Since $z \mapsto \sup_{f \in \mathcal{F}} |z(f)|$ is continuous and $\mathcal{F}$ is Donsker, the continuous mapping theorem [13, 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

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

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

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

Let g be defined as in Lemma A.5. Note that the second term in (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

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

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.6. Clearly the previous display yields the bounds

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

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 $\mathcal{F}$ with respect to $\|\cdot\|$ [13, Section 2.1], so Corollary A.6 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. □

B Additional Proofs

B.1 Sketch of Proof of Theorem 4

The proof of Theorem 4 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 4. Obviously (9) implies that $\mathcal{R}(\cdot)$ is uniformly continuous on $\overline{\Pi}$ with respect to $\|\cdot\|$. Assumption (8) 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 continous and the implication of Lemma A.3 holds, (CA) from Lemma A.4 also holds. Furthermore, (10) yields Lemma A.5, for which Corollary A.6 remains a valid corollary. We will not make use of Lemma A.7: we will instead use directly (12).

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 (13) and (5) by (11), we see that (A.13) from the proof of Theorem 1 can be replaced by

$$\begin{aligned} 0 \leq \mathcal{R}(\widehat{\pi}_n) &\lesssim r_n^{-1}\left[\widetilde{\mathbb{G}}_n f_{\widehat{\pi}_n} - \widetilde{\mathbb{G}}_n f_{\pi_s^\star}\right] + [\mathcal{V}(\pi^\star) - \mathcal{V}(\pi_s^\star)] + o_P(r_n^{-1}) \\ &\leq 2r_n^{-1}\sup_{f\in\mathcal{F}}|\widetilde{\mathbb{G}}_n f| + [\mathcal{V}(\pi^\star) - \mathcal{V}(\pi_s^\star)] + o_P(r_n^{-1}). \end{aligned} \tag{B.16}$$

By (12) and the continuous mapping theorem [13, 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 (9), 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

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

The proof concludes by noting that (12) includes the condition that almost all sample paths of $\widetilde{\mathbb{G}}_P$ are uniformly continuous on $\mathcal{F}$ with respect to $\|\cdot\|$, 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 6. If the bounded class Π is Donsker, then it is totally bounded in $L^2(P)$, so (8) 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

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

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

$$|\mathcal{V}(\pi_1) - \mathcal{V}(\pi_2)| \lesssim \|\pi_1 - \pi_2\| \tag{B.17}$$

for every $\pi_1, \pi_2 \in \overline{\Pi}$, hence (9). We also generalize (A.10), which becomes

$$f_\pi(O) = \sum_{a \in \mathcal{A}} \theta(\pi(X) - a) \left(\theta(A - a) \frac{Y - Q(A, X)}{P(A|X)} + Q(a, X) \right) - \mathcal{V}(\pi) \tag{B.18}$$

for every $\pi \in \overline{\Pi}$. The above equality and (B.17) imply $\|f_{\pi_1} - f_{\pi_2}\| \lesssim \|\pi_1 - \pi_2\|$ for every $\pi_1, \pi_2 \in \overline{\Pi}$, hence (10). The second part of the proof of Lemma A.7 can easily be generalized to derive that $\mathcal{F}$ is Donsker from (B.18) and the fact that Π is itself Donsker. Therefore, (12) is met and the proof is complete. □

B.3 Proofs for Section 4.3

We start by proving Lemma 9.

Proof of Lemma 9. 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 (15) 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

$$\begin{aligned} |F_{\pi_1}(m) - F_{\pi_2}(m)| &= \left| \mathbb{E}\left[\frac{\mathbb{1}\{A = \pi_1(X)\} - \mathbb{1}\{A = \pi_2(X)\}}{P(A|X)} \mathbb{1}\{Y \leq m\} \right] \right| \\ &\leq \frac{1}{2} \mathbb{E}\left[\frac{|\pi_1(X) - \pi_2(X)|}{P(A|X)} \mathbb{1}\{Y \leq m\} \right] \leq k_1 \|\pi_1 - \pi_2\| \end{aligned} \tag{B.19}$$

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} and (15) imply the existence of $\tilde{m} \in [m, \mathcal{V}(\pi_2)]$ such that

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

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

$$|\mathcal{V}(\pi_2) - m| \leq k_2 \left| F_{\pi_2}(\mathcal{V}(\pi_2)) - F_{\pi_2}(m) \right|. \tag{B.20}$$

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 $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq k_1 k_2 \|\pi_1 - \pi_2\|$, and *(ii)* $\|\pi_1 - \pi_2\| < c/k_1 k_2$ yields $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq c$. By combining these two results, we will thus prove that $\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) \leq k_1 k_2 \|\pi_1 - \pi_2\|$ for all $\|\pi_1 - \pi_2\|$ sufficiently small, and this will complete the proof ($k_1 k_2$ 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.19)

and (B.20) at $m = \mathcal{V}(\pi_1)$ establishes *(i)*:

$$\begin{aligned}\mathcal{V}(\pi_2) - \mathcal{V}(\pi_1) &\leq k_2 \left[F_{\pi_2}(\mathcal{V}(\pi_2)) - F_{\pi_2}(\mathcal{V}(\pi_1))\right] \\ &= k_2 \left[F_{\pi_1}(\mathcal{V}(\pi_1)) - F_{\pi_2}(\mathcal{V}(\pi_1))\right] \leq k_1 k_2 \left\|\pi_1 - \pi_2\right\|.\end{aligned}$$

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.20) at $m = \mathcal{V}(\pi_2) - c$,

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

By (B.19) 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 $k_1 \left\|\pi_1 - \pi_2\right\|$. Therefore, $\left\|\pi_1 - \pi_2\right\| \geq c/k_1 k_2$. □

Lemma B.8. *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.8. 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

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

Second note that, for every $\epsilon > 0$ small enough,

$$\left|F_{\pi_2}(m+\epsilon) - F_{\pi_2}(m) - F_{\pi_1}(m+\epsilon) + F_{\pi_1}(m)\right|$$

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

Returning to the first display completes the proof. □

Corollary B.9. *Under the conditions of Lemmas 9 and B.8, and the additional assumption (16), the map $\pi \mapsto \dot{F}_\pi(\mathcal{V}(\pi))$ is uniformly continuous on $\overline{\Pi}$ with respect to $\|\cdot\|$.*

Proof of Corollary B.9. Set $\pi_1, \pi_2 \in \overline{\Pi}$. By Lemma 9, 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 (15). For all such choices of π_1, π_2,

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

where the constant on the right does not depend on π_1, π_2. The final bound used Lemma B.8 and (16). 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 $\overline{\Pi}$ with respect to $\|\cdot\|$. □

References

[1] S Athey and S Wager. Efficient Policy Learning. *arXiv preprint arXiv:1702.02896*, 2017.

[2] Y Zhao, D Zeng, A Rush, and M Kosorok. Estimating individual treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.*, 107:1106–1118, 2012.

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

[4] D B Rubin and M J van der Laan. Statistical issues and limitations in personalized medicine research with clinical trials. *The International Journal of Biostatistics*, 8:Issue 1, Article 18, 2012.

[5] Alexander R Luedtke and Mark J van der Laan. Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, 12(1):305–332, 2016.

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

[7] A R Luedtke and M J van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2): 713–742, 2016.

[8] 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. *The Annals of Statistics (to appear)*, 2017.

[9] A R Luedtke and M J van der Laan. Comment. *Journal of the American Statistical Association*, 111(516):1526–1530, 2016.

[10] J Y Audibert and A B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[11] V Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[12] A Sheehy and J A Wellner. Uniform Donsker classes of functions. *The Annals of Probability*, pages 1983–2030, 1992.

[13] A W van der Vaart and J A Wellner. *Weak convergence and empirical processes*. Springer, Berlin Heidelberg New York, 1996.

[14] A W van der Vaart. *Asymptotic statistics*. Cambridge University Press, New York, 1998.

[15] M J van der Laan and J M Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York Berlin Heidelberg, 2003.

[16] M J van der Laan and D B Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006.

[17] M J van der Laan and S Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, New York, 2011.

[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):267–271, 2016.

[19] 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):61–95, 2014. doi: 10.1515/jci-2013-0022.

[20] V Chernozhukov, D Chetverikov, M Demirer, E Duflo, and C Hansen. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

[21] W Zheng and M J van der Laan. Targeted maximum likelihood estimation of natural direct effects. *Int. J. Biostat.*, 8(1):Art. 3, 42, 2012. ISSN 1557-4679.

[22] 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.

[23] A B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[24] D Rubin and M J van der Laan. A doubly robust censoring unbiased transformation. *Int. J. Biostat.*, 3:Art. 4, 21, 2007. ISSN 1557-4679.

[25] 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.

[26] V Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. ISBN 978-3-642-22146-0. 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].

[27] K A Linn, E B Laber, and L A Stefanski. Interactive Q-learning for Quantiles. *Journal of the American Statistical Association*, just-accepted:1–37, 2016.

[28] 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, supplementary material. *The Annals of Statistics (to appear)*, 2017.

[29] A Browder. *Mathematical analysis: an introduction*. Springer Science & Business Media, 2012.