Adaptive Welfare Maximization*

Kirill Ponomarev
University of Chicago

Liquiang Shi
Amazon Japan

January 29, 2026

Abstract

We consider the problem of learning optimal treatment policies from observational data. We propose an algorithm that combines doubly robust welfare estimation, to accommodate rich covariates and unknown propensity scores, and sample splitting, to adaptively select policy complexity. We show that the resulting treatment rule achieves the minimax-optimal rate of convergence in expected regret while selecting a suitable policy complexity with nearly oracle performance. Our analysis avoids unnecessarily restrictive assumptions commonly imposed on the data-generating process or on first-stage nonparametric estimators and yields a sharp characterization of the relevant universal constants. The practical performance of the proposed method is demonstrated in a simulation study.

*This is a revised version of a coauthored chapter in our Ph.D. dissertations at UCLA; see Ponomarev (2022) and Shi (2022). We thank Jinyong Hahn, Rosa Matzkin, Andres Santos, Denis Chetverikov, and seminar participants at UCLA for valuable feedback.

1 Introduction

Problems of treatment choice are ubiquitous in economics, arising in settings such as the provision of subsidies to disadvantaged households, bail decisions in pre-trial hearings, loan approval by banks, scholarship allocation by colleges, and personalized pricing by online retailers. In such environments, a decision-maker (DM) seeks to design a treatment rule that assigns each individual to one of several treatment options based on observable characteristics in order to maximize welfare (Manski, 2004). Designing an effective treatment rule is challenging for two main reasons. First, the DM often relies on observational data, which requires controlling for a rich set of covariates to identify the relevant welfare function. Second, the choice of policy complexity is constrained by institutional requirements such as transparency or non-discrimination and by a fundamental bias-variance trade-off: while more flexible, personalized rules can potentially achieve higher welfare, they are harder to estimate reliably from the data.

In this paper, we propose a policy learning algorithm that tackles the aforementioned practical challenges and has strong theoretical guarantees. Building on the Empirical Welfare Maximization (EWM) framework of Kitagawa and Tetenov (2018), the proposed algorithm combines two key components: doubly-robust welfare estimation (Athey and Wager, 2021) and model selection (Mbakop and Tabord-Meehan, 2021). Double robustness enables the use of flexible nonparametric estimators of the propensity score and outcome regression functions — obtained, for example, using modern machine learning methods — and ensures that the welfare function is estimated at the parametric rate under mild consistency and rate conditions (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018). Moreover, when the doubly-robust estimator is semiparametrically efficient (Chernozhukov, Escanciano, Ichimura, Newey, and Robins, 2022), we show that this efficiency translates into sharper performance guarantees even in finite samples. To choose an appropriate policy complexity, we consider a finite number of candidate policy classes with varying functional forms and complexities, and evaluate the best-in-class policies out-of-sample. The resulting procedure adapts to the optimal policy complexity for the underlying data-generating process (DGP) and achieves nearly oracle performance. For this reason, we call the proposed algorithm "Adaptive Welfare Maximization" (AWM).

Following the bulk of the existing literature, we focus on utilitarian (linear) welfare and evaluate policy performance in terms of expected regret, defined as the expected welfare loss relative to the optimal treatment rule in the population. Our contribution consists of two main results. First, we derive a finite-sample upper bound on the expected regret of the AWM rule, showing that it attains the parametric convergence rate and adaptively selects

the optimal policy complexity by balancing estimation error against potential welfare loss. We derive the bound under weaker assumptions than those typically imposed in the literature and precisely pin down the universal constant. We also show that using a semiparametrically (asymptotically) efficient welfare estimator leads to a tighter regret bound even in finite samples. Second, we establish a new finite-sample lower bound on the worst-case expected regret, which characterizes the fundamental performance limit of any data-dependent treatment rule. The lower bound matches the upper bound up to constants, implying that the AWM rule is minimax-rate optimal in expected regret. Together, these results provide strong theoretical support for the proposed method.

To assess the practical performance of the proposed rule, we conduct a simulation study in two practically relevant scenarios. In the first, several candidate policy classes contain the true optimal policy, ranging from relatively simple to unnecessarily complex. Consistent with our theoretical results, the AWM rule systematically selects the simplest relevant class and performs comparably to an oracle rule that knows the correct policy complexity. In the second scenario, none of the candidate classes is sufficiently rich to represent the optimal policy exactly. In this case, the AWM rule adaptively selects increasingly complex policy classes as the sample size grows.

This paper contributes to a large and growing literature on optimal treatment choice. Early work focused on unconstrained policy classes and proposed treatment rulesbased on the estimate of the conditional average treatment effect function. Manski (2004), Stoye (2009, 2012), and Tetenov (2012) studied minimax-regret rules; Dehejia (2005) and Chamberlain (2011) considered a Bayesian approach; Bhattacharya and Dupas (2012) introduced a budget constraint; and Hirano and Porter (2009) considered a limiting experiment. Related work in statistics includes the so-called Q-learning and A-learning approaches (e.g., Murphy (2003); Robins (2004); Qian and Murphy (2011); Shi, Fan, Song, and Lu (2018)).

This paper builds upon more recent literature focusing on policy classes with explicit constraints on complexity, with either binary or multivalued treatments. This line of work started with Kitagawa and Tetenov (2018), who introduced the EWM framework with binary treatments and showed that the EWM rule attains minimax-optimal rate of convergence for expected regret if the propensity scores are known. To accommodate observational settings with unknown propensity scores, Athey and Wager (2021) introduced a doubly-robust welfare estimator and established similar, although asymptotic, regret guarantees. Mbakop and Tabord-Meehan (2021) proposed regularizing the EWM objective to choose among several available policy classes and showed that the resulting treatment rules have oracle properties. Zhou, Athey, and Wager (2023) further extended the analysis of Athey and Wager (2021) to multivalued treatments, and Fang, Xi, and Xie (2025) combined it with model selection.

Our work builds upon the same ideas but imposes weaker assumptions and provides sharper theoretical guarantees. A more detailed comparison with existing results requires a formal setup, so we defer it to Section 3.3.

The rest of the paper is organized as follows. Section 2 gives the general setup; Section 3 describes the AWM procedure and presents the main theoretical results; Section 4 presents a simulation study; Section 5 concludes. All proofs are collected in the Appendix.

2 Setup

We adopt the standard potential outcomes framework of Neyman (1923) and Rubin (1974). Let $d \in \{0,1\}$ denote a binary treatment status, $Y(0), Y(1) \in \mathcal{Y}$ potential outcomes, and $X \in \mathcal{X}$ a vector of covariates. Let $m(x,d) = \mathbb{E}[Y(d) \,|\, X = x]$, for $d \in \{0,1\}$, and $\tau(x) = m(x,1) - m(x,0)$, denote the conditional mean and conditional average treatment effect (CATE) functions. Consider the problem of a utilitarian decision maker, who chooses a treatment rule $\pi : \mathcal{X} \to \{0,1\}$, based on covariates $X \in \mathcal{X}$, to maximize the *average welfare*, defined as

$$V_0(\pi) = \mathbb{E}[Y(\pi(X))].$$

The welfare function can be equivalently expressed as $V_0(\pi) = \mathbb{E}[Y(0)] + \mathbb{E}[\pi(X)\tau(X)]$. Since $\mathbb{E}[Y(0)]$ does not affect the optimal policy $\pi(\cdot)$, we will work with the *welfare gain*,[1]

$$V(\pi) = \mathbb{E}[\pi(X)\tau(X)]. \tag{1}$$

The first-best policy, $\pi^{FB}(x) = \mathbf{1}(\tau(x) \geqslant 0)$, is to treat individuals for whom the CATE is non-negative. However, without further restrictions, such policy may be overly complicated, hard to reliably estimate and implement (e.g., with multiple continuous covariates), or simply infeasible to the decision maker for institutional reasons (e.g., non-discriminatory laws). To discipline the problem, we restrict attention to a pre-specified class of feasible treatment rules Π (policy class), and focus on the constrained problem,

$$\pi^* \in \operatorname*{argmax}_{\pi \in \Pi} V(\pi). \tag{2}$$

Choosing a suitable policy class Π is essential in applications, as we discuss below.

We assume that the welfare function can be identified from the observable data. To accommodate endogenous treatment selection, we assume that instrumental variables $Z \in \mathcal{Z}$

[1] All results below are formulated in terms of regret, $V(\pi^*) - V(\pi)$, where π^* represents the optimal policy and π the implemented one. Since the term $\mathbb{E}[Y(0)]$ cancels out, the results are valid for $V_0(\pi)$ as stated.

are available such that $Z \perp \{D(z)\}_{z\in\mathcal{Z}}, Y(0), Y(1) \,|\, X$, where $\{D(z)\}_{z\in\mathcal{Z}}$ denote the potential treatments. When D is exogenous, i.e., $D \perp (Y(1), Y(0)) \,|\, X$, we set $Z = D$ in the notation. We denote the observed data vector by $W = (Y, D, X, Z)$ and assume that $W \sim P \in \mathbf{P}$, for a class of distributions $\mathbf{P}$ specified below.

Assumption 2.1 (Identification)**.**

1. *There is a weighting function $g(x, z) \in \mathcal{G}$ that identifies the treatment effect function $\tau_m(x, d) \in \mathcal{T}$ via*

$$\mathbb{E}_P[\tau_{\tilde{m}}(X, D) - g(X, Z)\tilde{m}(X, D) \,|\, X] = 0,$$

   for each $\tilde{m}(x, d) \in \mathcal{M}$.

2. *The welfare gain can be expressed as*

$$V(\pi) = \mathbb{E}_P[\pi(X)\tau(X)],$$

   where $\tau(X) = \mathbb{E}_P[\tau_m(X, D) \,|\, X]$.

The examples below, borrowed from Athey and Wager (2021), illustrate the scope of Assumption 2.1. The first example deals with a randomized control trial with binary treatment.

Example 1 (Exogenous Binary Treatment)**.** Suppose that the observed treatment is binary, $D \in \{0, 1\}$, and exogenous, $D \perp (Y(1), Y(0)) \,|\, X$. Then, we may take $Z = D$, and Assumption 2.1 holds with

$$\tau_m(x) = m(x,1) - m(x,0);$$

$$g(x, d) = \frac{d - p(x)}{p(x)(1 - p(x))},$$

where $p(x) = P(D = 1 \,|\, X = x)$ denotes the propensity score. Multivalued exogenous treatments can be accommodated by modifying the moment function and complexity measure for the policy class Π as in Fang, Xi, and Xie (2025). ■

The second example discusses settings in which the observed treatment is endogenous, e.g., due to non-compliance. As an example, consider a clinical trial in which patients are randomly assigned to two different treatment protocols. Since the patients are at will to choose any treatment they want after discussing the options with their doctors, some may end up crossing over. In this case, the original randomly assigned protocol serves as an instrumental variable for the actual treatment; see, e.g., Angrist, Gao, Hull, and Yeh (2025).

Example 2 (Endogenous Binary Treatments with Binary Instruments)**.** If a binary treatment $D \in \{0,1\}$ fails to satisfy the conditional independence restriction in Example 1, the CATE function cannot be point identified without further restrictions. To this end, suppose there is an instrumental variable $Z \in \{0,1\}$ satisfying $Z \perp (D(0), D(1), Y(0), Y(1)) \,|\, X$, and treatment selection is monotone, in the sense that $D(1) \geqslant D(0)$, almost surely (Imbens and Angrist, 1994). Then, one can point identify the conditional Local Average Treatment Effect (LATE) for a subpopulation of individuals with $D(1) > D(0)$ via

$$LATE(x) = \mathbb{E}[Y(1) - Y(0) \,|\, D(1) > D(0), X = x] = \frac{\mathbb{C}ov(Y, Z \,|\, X = x)}{\mathbb{C}ov(D, Z \,|\, X = x)}.$$

This causal parameter may not be relevant to the decision-maker who aims to maximize the average welfare across all individuals. In some settings, it may be reasonable to assume that that $CATE(x) = LATE(x)$ (i.e, if individual treatment effects are suitably homogeneous). Then, Assumption 2.1 holds with

$$\begin{aligned}
\tau_m(x) &= m(x,1) - m(x,0); \\
g(x,z) &= \frac{1}{\Delta(x)} \frac{z - s(x)}{s(x)(1 - s(x))}; \\
s(x) &= P(Z = 1 \,|\, X = x); \\
\Delta(x) &= P(D = 1 \,|\, Z = 1, X = x) - P(D = 0 \,|\, Z = 1, X = x).
\end{aligned}$$

If individuals select into treatment based on its perceived effectiveness, one might reasonably expect that $LATE(x) \geqslant CATE(x)$. Then, implementing an optimal treatment policy based on the assumption $LATE(x) = CATE(x)$ would lead to treating excessively. If the treatment is, for example, a medical test that is costly but potentially life-saving, this approach may be justified. Under stronger type-independence restrictions, settings with multi-valued instruments, such as "judge designs," can also be accommodated. ■

The final example illustrates that Assumption 2.1 accommodates settings in which the observed treatment is non-binary. As an example, consider an online retailer that has experimented with various price levels and is deciding whether to offer a small discount to a subset of its customers.

Example 3 (Continuous Treatments)**.** Let $D \in \mathcal{D}$ be a continuous treatment variable, and $\{Y(d)\}_{d \in \mathcal{D}}$ denote the corresponding potential outcomes. Suppose the decision maker aims to maximize

$$V(\pi) = \frac{d}{dv} \mathbb{E}[Y(D + v\pi(X))] \Big|_{v=0},$$

which is the average effect of an infinitesimal nudge following policy $\pi(x) \in \{0,1\}$. Suppose that D is exogenous in the sense that $\{Y(d)\}_{d\in\mathcal{D}} \perp D \,|\, X$. Denote

$$\tau_m(x,d) = \frac{\partial}{\partial v} m(x, d+v)\bigg|_{v=0}.$$

Then, under regularity conditions, using integration by parts (Powell, Stock, and Stoker, 1989), Assumption 2.1 can be shown to hold with

$$g(x,d) = -\frac{\partial}{\partial d} \log f(d \,|\, x),$$

where $f(d \,|\, x)$ denotes the conditional density of D given X. ■

In practice, the decision-maker observes a random sample $W_1, \ldots, W_n$ drawn i.i.d. from a distribution $P \in \mathbf{P}$, forms an estimator $\hat{V}_n(\pi)$ of $V(\pi)$, and solves

$$\hat{\pi}_n^{EWM} \in \operatorname*{argmax}_{\pi\in\Pi} \hat{V}_n(\pi). \tag{3}$$

This is precisely the EWM rule of Kitagawa and Tetenov (2018). In our framework, estimating $V(\pi)$ requires estimating the functions $m(x,d)$, $\tau_m(x,d)$, and $g(x,z)$ (to which we will refer as the *nuisance functions*) non-parametrically. Athey and Wager (2021) and Fang, Xi, and Xie (2025) showed that using a doubly-robust estimator $\hat{V}_n(\pi)$ leads to better performance guarantees for $\hat{\pi}_n^{EWM}$. We adopt the same approach below. Under Assumption 2.1, the welfare gain can be expressed as

$$V(\pi) = \mathbb{E}_P[\pi(X)\Gamma(W)], \tag{4}$$

where

$$\Gamma(W) = \tau_m(X,D) + g(X,Z)(Y - m(X,D)).$$

The moment condition in (4) is Neyman-orthogonal with respect to $m(x,d)$ and $g(x,z)$, and the corresponding doubly-robust estimator $\hat{V}_n(\pi)$ can be constructed using cross-fitting (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018). To this end, we impose the following assumptions.

Assumption 2.2 (DGP)**.** *All distributions $P \in \mathbf{P}$ satisfy the following conditions.*

1. *(Bounded moments)* $\mathbb{E}_P[Y^2] \leqslant B^2 < \infty$, $\mathbb{E}_P[\tau_m(X,D)^2] \leqslant B_\tau^2 < \infty$.
2. *(Overlap)* $\sup_{x\in\mathcal{X}, z\in\mathcal{Z}} |g(x,z)| \leqslant \eta^{-1}$ *for some* $\eta \in (0, 1/2)$.

Assumption 2.3 (First-stage Estimators)**.** *The first-stage estimators $\hat{m}(x,d)$, $\tau_{\hat{m}}(x,d)$, and $\hat{g}(x,z)$ satisfy the following conditions. For $m \in \mathcal{M}$, $\tau_m \in \mathcal{T}$, and $g \in \mathcal{G}$, for some $0 < \zeta_m, \zeta_g < 1$, with $\zeta_m + \zeta_g \geqslant 1$, and a positive sequence $a(n) \to 0$ as $n \to 0$,*

$$\mathbb{E}_P[(\hat{m}(X,D) - m(X,D))^2] \vee \mathbb{E}_P[(\tau_{\hat{m}}(X,D) - \tau_m(X,D))^2] \leqslant \frac{a(n)}{n^{\zeta_m}},$$

$$\mathbb{E}_P[(\hat{g}(X,Z) - g(X,Z))^2] \leqslant \frac{a(n)}{n^{\zeta_g}},$$

where (X,D,Z) is an independent sample drawn from P, for all $P \in \mathbf{P}$.

Assumption 2.2 imposes relatively weak conditions on the underlying DGP. Unlike the bulk of existing literature, we do not require boundedness of Y or sub-Gaussianity of the residuals $Y - m(X,D)$, conditional on X and D. The stated assumptions only require that $\mathbb{E}[\Gamma(W)^2] \leqslant B_\tau^2 + B^2/\eta^2 < \infty$. Although such assumptions would make the proofs more straightforward, they are not essential for the results, and thus we dispose of them. In the familiar setting of Example 1, the above assumptions are equivalent to $\mathbb{E}_P[Y^2] \leqslant B^2$ and $P(D = 1 \,|\, X = x) \in [\eta, 1-\eta]$, for all $x \in \mathcal{X}$.

Assumption 2.3 takes an agnostic view on how the estimates of the nuisance functions are obtained. It can be satisfied, for example, by machine learning estimators such as Lasso, random forest, or neural networks, under further assumptions on the latent low-dimensional structure of the underlying DGP. Notably, unlike Athey and Wager (2021), we do not require uniform consistency of the first-stage estimators, which implicitly requires further restrictions on the DGP, such as compact support or extra smoothness of regression functions and propensity scores, and may be overly restrictive.

When solving for the EWM policy in practice, the decision-maker faces a familiar "bias-variance" trade-off: A more complex rule may yield higher welfare but it is harder to estimate from the data. To this end, Mbakop and Tabord-Meehan (2021) proposed treating the policy class Π as a tuning parameter and using the existing approaches to model selection to obtain a treatment rule with oracle performance guarantees. We also take this approach and propose a data-dependent criterion that chooses a suitable complexity adaptively.

We assume that the decision-maker can choose between a finite number K of policy classes $\Pi_1, \ldots, \Pi_K$ with different complexity. These classes may be nested, overlapping, or non-overlapping. Since solving the EWM problem is typically computationally hard, we do not pursue settings with infinite number of policy classes, as in Mbakop and Tabord-Meehan (2021) and Fang, Xi, and Xie (2025), and focus only on the practical case with finite K.[2] A

[2]Dealing with an infinite number of classes typically requires stronger restrictions on the DGP and additional penalization, as explained in Mbakop and Tabord-Meehan (2021).

convenient way to measure complexity is via the VC dimension.

Definition 2.1 (VC-Dimension of the Policy Class)**.** *Let $\mathcal{G}$ be a collection of subsets of $\mathcal{X}$. Say that a set of points $\{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ is shattered by $\mathcal{G}$ if, for every subset $S \subseteq \{x_1, \ldots, x_d\}$, there exists a set $G \in \mathcal{G}$ such that $S = G \cap \{x_1, \ldots, x_d\}$. The VC-dimension of $\mathcal{G}$ is the cardinality of the largest set $\{x_1, \ldots, x_d\}$ that can be shattered by $\mathcal{G}$. Any policy $\pi \in \Pi_k$ takes the form $\pi(x) = \mathbf{1}(x \in G)$ for some $G \subseteq \mathcal{X}$, so we identify Π_k with a collection of sets $\mathcal{G}_k$ and define $VC(\Pi_k) = VC(\mathcal{G}_k)$.*

To state the main results, we assume that all relevant policy classes have finite VC dimension. This assumption can be relaxed as we discuss in Remark 1.

Assumption 2.4 (Policy Complexity)**.** *$VC(\Pi_k) < \infty$ for all $k \in \{1, \ldots, K\}$.*

The following examples illustrate.

Aggregation-based rules. Let $S_k : \mathcal{X} \to \mathcal{S}_k$, with $|\mathcal{S}_k| = N_k < \infty$, be a function that turns a covariate vector X into a summary statistic S_k that can take N_k different values. Consider the class of rules that depend on X only through S_k:

$$\Pi_k = \{\pi : \mathcal{X} \to \{0, 1\} : S_k(x) = S_k(x') \implies \pi(x) = \pi(x')\}.$$

Such class is finite and has VC dimension N_k. In the absence of further constraints on Π_k, the solution to (2) can be obtained analytically as

$$\pi_k(s) = \mathbf{1}(\tau_k(s) \geqslant 0),$$

where $\tau_k(s) = \mathbb{E}[Y(1) - Y(0) \,|\, S_k = s]$. By the law of iterated expectations, the function $\tau_k(x)$ is identified as $\tau_k(s) = \mathbb{E}\left[\Gamma(W) \,|\, S_k = s\right]$, and its empirical analog can be computed as

$$\hat{\tau}_k(s) = \frac{\sum_{i=1}^n \hat{\Gamma}^{-j(i)} \mathbf{1}(S_k(X_i) = s)}{\sum_{i=1}^n \mathbf{1}(S_k(X_i) = s)},$$

which gives a closed-form solution to (3).

Linear threshold rules. Let $x_k \in \mathbb{R}^{d_k}$ be a subvector of x and consider the classes of rules

$$\Pi_{m,k} = \{\pi_{m,k}(x) = \mathbf{1}(\beta_{m,k}' x_k \geqslant c_{m,k}) : (c_{m,k}, \beta_{m,k}')' \in \mathbb{R}^{d_k+1}\},$$

for $m = 1, \ldots, M$. Then, set

$$\Pi_k = \{\pi(x) = \prod_{m=1}^M \pi_{m,k}(x) : \pi_{m,k}(x) \in \Pi_{m,k}\}$$

The VC dimension of each $\Pi_{m,k}$ is at most $d_k + 2$, while the VC dimension of Π_k is finite although harder to precisely quantify (see Lemmas 2.6.15 and 2.6.17 in van der Vaart and Wellner, 1996). Kitagawa and Tetenov (2018) show that for such classes, the optimization problem in (3) can be re-formulated as a Mixed-Integer Linear Program (MILP), which can be converted into a sequence of linear programs via a branch-and-bound algorithm. Compared with the preceding class, the regions in the partition of $\mathcal{X}$ here are estimated from the data rather than set exogenously.

Decision trees. Trees represent decision rules recursively. A depth-zero decision tree, $T_0(x)$, is a constant decision rule $T_0(x) = 0$ or $T_0(x) = 1$ for all $x \in \mathcal{X}$. For any $k \geqslant 1$, a depth-k decision tree T_k is obtained by specifying a splitting variable $j \in \{1, \ldots, d_X\}$, a threshold $t \in \mathbb{R}$, and two depth-$(k-1)$ decision trees T_{k-1}^L, T_{k-1}^R, so that

$$T_k(x) = \mathbf{1}(x_j \leqslant t) T_{k-1}^L(x) + \mathbf{1}(x_j > t) T_{k-1}^R(x).$$

Letting Π_k denote the class of all depth-k decision trees over $\mathcal{X} \subseteq \mathbb{R}^{d_X}$, the VC-dimension of Π_k is of order $2^k \log(d_X)$, see Zhou, Athey, and Wager (2023). The aforementioned paper also proposes two methods for solving (3) in practice closely related to the branch-and-bound algorithm used for solving MILP. Similar to linear threshold rules, decision trees infer the appropriate partition of the covariate space $\mathcal{X}$ from the data, rather than setting it exogenously.

3 Adaptive Welfare Maximization

3.1 Implementation

We start by introducing the proposed policy learning algorithm. The first step is to obtain a doubly-robust estimator for the welfare function.

Algorithm 1 (Doubly-Robust Welfare Estimation)**.**
Input: *A data sample* $(W_i)_{i \in E}$ *of size* n_E*.*
Output: *A welfare estimate* $\hat{V}^{(E)}(\pi)$*, for any fixed policy* $\pi \in \bigcup_{k=1}^K \Pi_k$*.*

1. *Randomly split the sample into* J *evenly sized folds* $I_1, \ldots, I_J$ *of size* $\lfloor n_E/J \rfloor$*, distributing the remaining* $n_E - J\lfloor n_E/J \rfloor$ *observations arbitrarily.*

2. *For each* j*, compute the non-parametric estimators* $\hat{m}^{(-j)}(x, d)$*,* $\hat{\tau}^{(-j)}(x, d)$*, and* $\hat{g}^{(-j)}(x, z)$ *using the* $\frac{J-1}{J} n_I$ *observations in all folds except for* j*.*

3. *For each $i \in I_j$, for each $j \in \{1, \ldots, J\}$, compute*

$$\hat{\Gamma}^{(-j)}(W_i) = \hat{\tau}^{(-j)}(X_i, D_i) + \hat{g}^{(-j)}(X_i, Z_i)(Y_i - \hat{m}^{(-j)}(X_i, D_i)).$$

Note: *If the propensity score $p(X_i)$ is known, it may be plugged into $\hat{g}^{(-j)}(X_i, Z_i)$.*

4. *Compute the final stimator*

$$\hat{V}^{(E)}(\pi) = \frac{1}{n_E} \sum_{j=1}^{J} \sum_{i \in I_j} \pi(X_i) \hat{\Gamma}^{(-j)}(W_i).$$

■

Next, we describe the main algorithm. Following Mbakop and Tabord-Meehan (2021), the idea is to choose the optimal policy complexity using sample-splitting and compute the corresponding EWM rule.

Algorithm 2 (Adaptive Welfare Maximization)**.**

Input: *Data sample $(W_i)_{i=1}^n$; Policy classes Π_k for $k \in \{1, \ldots, K\}$.*

Output: *A policy $\hat{\pi}_n^{AWM} : \mathcal{X} \to \{0, 1\}$.*

1. *Randomly split the sample $(W_i)_{i=1}^n$ into the estimating $(W_i)_{i \in E}$ and hold-out $(W_i)_{i \in H}$ samples of sizes n_E and n_H. Let the superscripts (E) and (H) indicate that the corresponding object was computed using each of the two samples correspondingly.*

2. *For each policy class Π_k:*

   (i) Compute the EWM policy using $(W_i)_{i \in E}$,

$$\hat{\pi}_k^{(E)} = \underset{\pi \in \Pi_k}{\operatorname{argmax}} \hat{V}^{(E)}(\pi),$$

   where $\hat{V}^{(E)}(\pi)$ is computed as in Algorithm 1.

   (ii) Evaluate its performance in the holdout sample:

$$\hat{Q}_k = \frac{1}{n_H} \sum_{i \in H} \hat{\pi}_k^{(E)}(X_i) \hat{\Gamma}^{(E)}(W_i).$$

Note: *Computing $\hat{\Gamma}^{(E)}$ here does not require cross-fitting: the first-stage estimators $\hat{m}^{(E)}(x, d)$, $\tau_{\hat{m}}^{(E)}(x, d)$, and $\hat{g}^{(E)}(x, z)$ are computed using the full (E) sample.*

3. Select $\hat{k} = \operatorname{argmax}_{k\in\{1,\dots,K\}} \hat{Q}_k$ *and define*

$$\hat{\pi}_n^{AWM}(x) = \hat{\pi}_{\hat{k}}^{(E)}(x).$$

■

The downside of using a hold-out sample is that a share $n_H/(n_E + n_H)$ of observations is used only for out-of-sample evaluation. In Appendix A.1, we provide a cross-validation procedure that alleviates this concern but is more computationally intensive. The resulting treatment rules have exactly the same theoretical guarantees and perform similarly in simulations, with the CV-rule being consistently slightly better.

3.2 Theoretical Guarantees

To evaluate the performance of the AWM rule, we compare its welfare with the maximum welfare attainable within the class of policies $\Pi \equiv \bigcup_{k=1}^{K} \Pi_k$. Denote $\pi_P^* \in \operatorname{argmax}_{\pi\in\Pi} V(\pi)$ and $\pi_{k,P}^* \in \operatorname{argmax}_{\pi\in\Pi_k} V(\pi)$, where the subscript highlights the dependence of the policy rules on the underlying data-generating process. Note that by construction, $\pi_P^* = \pi_{k,P}^*$, for some k. Since the welfare under $\hat{\pi}_n^{AWM}$ is a random quantity, we focus on the *expected regret*:

$$R_P(\hat{\pi}_n^{AWM}) = \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n^{AWM})], \tag{5}$$

where $V(\hat{\pi}_n^{AWM}) = \mathbb{E}_P[\hat{\pi}_n^{AWM}(X)\Gamma(W) \mid \hat{\pi}_n^{AWM}]$. This criterion can also be viewed as risk under a specific data-dependent choice of the loss function $\ell_P(\pi, \pi') = |V_P(\pi) - V_P(\pi')|$, which measures the misclassification loss in welfare units. Other loss functions can be accommodated under suitable modifications of the moment conditions.

We prove two key results regarding the performance of AWM policy rule. Our first result is an upper bound on expected regret.

Theorem 1 (Regret Upper Bound)**.** *Let Assumptions 2.1–2.4 hold, and* $\hat{\pi}^{AWM}$ *be computed as in Algorithm 2. Then, for any* $P \in \mathbf{P}$*, for all* n *large enough,*

$$R_P(\hat{\pi}_n^{AWM}) \leqslant \min_{k\leqslant K} \left(\overline{C} \sqrt{\frac{\mathbb{E}_P[\Gamma(W)^2]}{n_E}} \sqrt{VC(\Pi_k)} + V(\pi_P^*) - V(\pi_{k,P}^*) \right) + \sqrt{\frac{K\mathbb{E}_P[\Gamma(W)^2]}{n_H}} + R_n, \tag{6}$$

where $\overline{C} \leqslant 58$ *is a universal constant, and* $R_n = o(n)$ *is the remainder term explicitly computed in Equation (A.10).*

The term $\sqrt{\mathbb{E}_P[\Gamma(W)^2]/n_E}$ in (6) reflects two desirable properties of the welfare estimator: semiparametric efficiency and double-robustness. To see a connection with efficiency, note

that $\mathbb{E}_P[\Gamma(W)^2] = \mathbb{V}ar_P(\Gamma(W)) + ATE_P^2$, where the second summand does not depend on the chosen estimator. Although any score function $\tilde{\Gamma}(W)$ satisfying $\mathbb{E}_P[\tilde{\Gamma}(W) \,|\, X] = \tau(X)$ can be used to form an estimate of $V(\pi)$, the efficient score $\Gamma(W)$ has the lowest variance, leading to the tightest bound. Despite the fact that semiparametric efficiency is an asymptotic concept, the reduced variance of the welfare estimator leads to better theoretical guarantees even in finite samples. We further discuss semiparametric efficiency in the context of policy learning in Remark 2. In turn, the double-robustness of $\hat{V}(\pi)$ ensures that it is first-order asymptotically equivalent to the oracle welfare estimate, $\tilde{V}(\pi) = n^{-1} \sum_{i=1}^n \pi(X_i)\Gamma(W_i)$ uniformly over $\pi \in \Pi$. As a result, expected regret of the AWM rule approaches zero at the parametric rate $n^{-1/2}$ under relatively weak consistency requirements on the first-stage estimators. This fact allows to accommodate practical settings with rich covariates and unknown propensity scores.

The minimum over k in (6) shows the adaptivity of the AWM rule: It optimally balances the policy complexity, $\sqrt{VC(\Pi_k)}$, and welfare loss, $V(\pi_P^*) - V(\pi_{k,P}^*)$. On the one hand, (6) implies that

$$R_P(\hat{\pi}_n^{AWM}) \leqslant \overline{C} \sqrt{\frac{\mathbb{E}_P[\Gamma(W)^2]}{n_E}} \sqrt{\min(VC(\Pi_k) : \pi_P^* \in \Pi_k)} + \sqrt{\frac{K\mathbb{E}_P[\Gamma(W)^2]}{n_H}} + R_n,$$

yielding the same regret bound as if the decision-maker knew the complexity of the simplest policy class Π_k containing the optimal policy π_P^*. On the other hand, is possible that the minimum in (6) is attained by the class k with $0 < V(\pi_P^*) - V(\pi_{k,P}^*) = O(n^{-1/2})$, i.e., a simple policy from a low complexity class is nearly optimal. In either case, the AWM rule optimally resolves the bias-variance trade-off, relative to the class Π. Our Monte Carlo experiments, presented in Section 4, suggest that this trade-off is very pronounced in practice, and the AWM successfully adapts to the underlying optimal complexity.

The final leading term in (6) reflects the "price" of model selection and increases with the number of classes. The seemingly restrictive dependence on K appears since our Assumption 2.2(1) only requires $\mathbb{E}_P[\Gamma(W)^2] < \infty$, and can be drastically improved under further restrictions. In particular, the relevant rate would be $K^{1/m}$ if $\mathbb{E}_P[\Gamma(W)^m] < \infty$, $\log(K)$ if $\Gamma(W)$ is sub-exponential, and $\sqrt{\log(K)}$ if $\Gamma(W)$ is sub-Gaussian.

Our second main result concerns rate-optimality. Let $\mathbf{P}_k = \{P \in \mathbf{P} : \pi_P^* \in \Pi_k\}$ denote a class of distributions such that the optimal treatment rule within Π belongs to Π_k. Note that $(\mathbf{P}_k)_{k \leqslant K}$ form a partition of $\mathbf{P}$. By Assumption 2.2 and the law of iterated expectations, we can bound $\sqrt{\mathbb{E}_P[\Gamma(W)^2]} \leqslant B/\eta \cdot L$, where $L = \sqrt{1 + B_\tau^2 \eta^2 / B^2}$. The bound in (6) provides

a guarantee on the worst-case performance of $\hat{\pi}_n^{AWM}$ within $\mathbf{P}_k$,

$$\sup_{P \in \mathbf{P}_k} R_P(\hat{\pi}_n^{AWM}) \leqslant \overline{C} L \frac{B}{\eta} \sqrt{\frac{VC(\Pi_k)}{n_E}} + L \frac{B}{\eta} \sqrt{\frac{K}{n_H}} + R_n. \tag{7}$$

We show that no policy rule can do substantially better.

Theorem 2 (Regret Lower Bound)**.** *Let Assumptions 2.2 and 2.4 hold. Then, for any k,*

$$\inf_{\hat{\pi}_n : W_1^n \to \{0,1\}} \sup_{P \in \mathbf{P}_k} R_P(\hat{\pi}_n) \geqslant \underline{C} \frac{B}{\eta} \sqrt{\frac{VC(\Pi_k) - 1}{n}} - \frac{0.6B}{n}, \tag{8}$$

for all $n \geqslant \max(5, \eta^{-1}(VC(\Pi_k) - 1))$, where $\underline{C} \geqslant 0.16$ is a universal constant.

The "worst-case" DGP-s, which attain the supremum in (8), are such that the individual treatment effects $Y(1) - Y(0)$ are highly variable, the covariate space $\mathcal{X}$ is rich, the distribution of X has high entropy, and yet the CATE function $\tau(X)$ is of the magnitude $n^{-1/2}$. In such settings, it is statistically hard to distinguish between individuals that should and should not be treated, so any policy learning rule is bound to make mistakes. Similar worst-case DGP's appear in the proofs of related results in Hirano and Porter (2009), Kitagawa and Tetenov (2018) and Athey and Wager (2021).

3.3 Discussion

Taken together, Theorems 1 and 2 provide a strong theoretical justification for using the AWM policy rule in practice, and refine the existing results in the literature. Specifically, we show that, in contrast to the EWM rule of Kitagawa and Tetenov (2018) and the PWM rule of Mbakop and Tabord-Meehan (2021), the AWM rule attains the optimal (parametric) rate of convergence in settings with rich covariates and unknown propensity scores. Our analysis does not require bounded outcomes and yields a sharper bound by leveraging a semiparametrically efficient welfare estimator. While related rate results have been obtained in Athey and Wager (2021) and Zhou, Athey, and Wager (2023), our approach additionally establishes adaptivity, provides finite-sample (rather than asymptotic) guarantees, and avoids restrictive tail assumptions. Moreover, we do not require uniform consistency of first-stage estimators, a condition that may be overly restrictive in practice

Combining doubly robust welfare estimation with model selection requires technical arguments that differ substantially from those in the existing literature. The paper closest to ours is Fang, Xi, and Xie (2025). Relative to that work, we restrict attention to binary policies but accommodate a richer set of environments, including endogenous treatment selection (as in Example 2) and continuous treatments (as in Example 3). Our analysis does

not require bounded outcomes or first-stage estimators and yields a sharp characterization of the relevant universal constants. In particular, the constant $\overline{C}$ in Theorem 1 is smaller than its counterparts in Theorem 2.1 of Kitagawa and Tetenov (2018) and Theorem 1 of Athey and Wager (2021). We further provide a new finite-sample lower bound on expected regret, establishing the minimax-rate optimality of the AWM rule. Unlike the lower bound of Kitagawa and Tetenov (2018), our result allows for weak overlap — formally, sequences of models $\mathbf{P}_n$ with $\eta_n \to 0$ and distributions $P_n \in \mathbf{P}_n$ such that $P_n(D = 1 \mid X = x) \to 0$ for some $x \in \mathcal{X}$ — and, unlike Athey and Wager (2021), is non-asymptotic.

We conclude this section with two technical remarks. The first one discusses alternative measures of policy complexity, and the second one points out another connection with semiparametric efficiency theory.

Remark 1 (Infinite VC Dimension and Multivalued Policy Rules)**.** A version of Theorem 1 holds for many policy classes of infinite VC dimension. To elaborate, let $N(\varepsilon, \mathcal{F}, ||\cdot||)$ denote the covering number, i.e., minimum number of $||\cdot||$-balls of radius ε required to cover a set $\mathcal{F}$. Define the classes of functions $\mathcal{F}_k = \{f(w) = \pi(x)\Gamma(w) : \pi \in \Pi_k\}$, for each $k = 1, \ldots, K$, and suppose that

$$\mathcal{E}(\Pi_k) = \int_0^\infty \sup_Q \sqrt{\log N(u\,||\Gamma||_{2,Q}\,, \mathcal{F}_k, ||\cdot||_{2,Q})}du < \infty, \tag{9}$$

where the supremum is taken over all finitely supported measures Q, and $||f||_{2,Q} = (\int f^2 dQ)^{1/2}$. This quantity $\mathcal{E}(\Pi_k)$ is known as the entropy integral and plays an important role in empirical process theory (see Chapter 2 in van der Vaart and Wellner, 1996). For a version of Theorem 1 to hold, it suffices to require that $\mathcal{E}(\Pi_k) < \infty$, for all Π_k. Specifically, as a simple corollary of our proofs (starting from Equation (A.6) in the proof of Lemma A.7), under Assumptions 2.1–2.3 and the above condition, we obtain the same bound on expected regret as in (6) with a smaller universal constant $\overline{C} = 4\sqrt{12}$, and $\mathcal{E}(\Pi_k)$ replacing $VC(\Pi_k)$. The entropy formulation also allows to accommodate multivalued policy rules, as in Fang, Xi, and Xie (2025), for which the VC dimension is not appropriate. ■

Remark 2 (On Semiparametric Efficiency in Policy Learning)**.** Intuitively, using efficient estimator of the welfare function should be beneficial: if $\hat{V}(\pi)$ is "close" to $V(\pi)$, its maximizer $\hat{\pi}$ should come "close" to maximizing $V(\pi)$. Although it is hard to study efficiency of $\hat{\pi}$ itself (it estimates an infinite dimensional parameter and, in many cases, has non-standard rates of convergence), some general results can be obtained for the maximum welfare, $\hat{V}(\hat{\pi})$. If the policy class Π is unrestricted, the optimal treatment rule is $\pi^{FB}(x) = \mathbf{1}(\tau(x) \geqslant 0)$, and the maximum welfare is $\max_{\pi\in\Pi} V(\pi) = \mathbb{E}[\max(\tau(X), 0)]$. Luedtke and Van Der Laan (2016)

showed that if the optimal treatment rule is unique, i.e., $P(\tau(X) = 0) = 0$, then $\max_{\pi \in \Pi} V(\pi)$ has a well-defined efficiency bound, and provided an estimator that attains it. Below, we argue that similar results hold for restricted policy classes.

We only give a sketch of the argument below and defer the details to Lemma A.13 in the Appendix. Suppose that the covariate space $\mathcal{X}$ is bounded, the model $\mathbf{P}$ satisfies Assumption 2.2 and all $P \in \mathbf{P}$ are dominated by a sigma-finite measure μ with bounded densities $dP/d\mu \leqslant C_\mu < \infty$. Suppose that $\psi_\pi(W) = \pi(X)\Gamma(W) - \mathbb{E}[\pi(X)\Gamma(W)]$ is the efficient influence function for $V(\pi)$, and $\sup_{\pi \in \Pi} |\hat{V}(\pi) - \tilde{V}(\pi)| = o_P(n^{-1/2})$, where $\tilde{V}(\pi) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\Gamma(W_i)$. Note that the welfare function $V : \Pi \to \mathbb{R}$ satisfies

$$
\begin{aligned}
&|V(\pi)| \leqslant ||\Gamma||_{2,P} \leqslant C_\Gamma < \infty, \\
&|V(\pi_1) - V(\pi_2)| \leqslant ||\Gamma||_{2,P} \, ||\pi_1 - \pi_2||_{2,P} \leqslant C_\Gamma C_\mu \, ||\pi_1 - \pi_2||_{2,\mu} \,,
\end{aligned}
$$

where $C_\Gamma = \sqrt{B_\tau^2 + B^2/\eta^2}$ implied by Assumption 2.2–(1). Thus, $V(\cdot)$ can be viewed as an element of the Banach space $C_b(\Pi)$ of continuous bounded functions on Π endowed with a sup-norm, $||V||_\infty = \sup_{\pi \in \Pi} |V(\pi)|$.

The stated assumptions imply that $\hat{V}_n(\pi)$ is semiparametrically efficient for $V(\pi)$ for each fixed $\pi \in \Pi$. Under further regularity conditions, using efficiency theory in Banach spaces (e.g., Chapter 5 in Bickel, Klaassen, Ritov, and Wellner, 1993), one can show that $\hat{V}(\cdot)$ is semiparametrically efficient for $V(\cdot)$ as an element of $C_b(\Pi)$. In particular,

$$
\sqrt{n}(\hat{V}_n(\cdot) - V(\cdot)) \quad \to_d \quad \mathbb{G}(\cdot), \quad \text{in } C_b(\Pi),
$$

where $\mathbb{G}(\cdot)$ is a centered Gaussian process with covariance kernel $\mathbb{C}ov(\mathbb{G}(\pi_1), \mathbb{G}(\pi_2)) = \mathbb{E}[\psi_{\pi_1}(W)\psi_{\pi_2}(W)]$, defining the efficiency bound for $V(\cdot)$. Now, consider a map $\psi : C(\Pi) \to \mathbb{R}$ defined as $\psi(V) = \max_{\pi \in \Pi} V(\pi)$. By, Proposition 4.12 in Bonnans and Shapiro (2013), $\psi(\cdot)$ is Hadamard directionally differentiable at V in direction H with derivative $\psi'_V(H) = \max_{\pi \in S^*(V)} H(\pi)$, where $S^*(V) = \text{argmax}_{\pi \in \Pi} V(\pi)$. Thus, $\psi(\cdot)$ is fully Hadamard differentiable if and only if $S^*(V)$ is a singleton. In the latter case, it follows from the Delta-method that $\hat{V}(\hat{\pi}) = \max_{\pi \in \Pi} \hat{V}_n(\pi)$ is semiparametrically efficient for $\max_{\pi \in \Pi} V(\pi)$. ■

4 A Simulation Study

4.1 Design

In this section, we present a simulation study designed to illustrate the performance of the proposed AWM procedure in practice. We consider a stylized data-generating process

to illustrate the main insights of our results and allow for straightforward visualization. Covariates $X = (X_1, X_2, X_3, X_4)$ are drawn independently from Uniform $[0, 1]$. The potential outcomes are defined as

$$Y(0) = 0.7(X_3 + X_4 + \varepsilon_0);$$
$$Y(1) = CATE(X_1, X_2) + 0.7(X_3 + X_4 + \varepsilon_1),$$

where $\varepsilon_0, \varepsilon_1 \sim \mathcal{N}(0, 1)$ are independent Normal errors. We consider two CATE functions: the first exhibits a positive effect when (X_1, X_2) lies within a rectangle ("Rectangle DGP"), and the second when they fall within an ellipse ("Ellipse DGP"), highlighted in black in Figure 1a. In each case, the first-best policy assigns treatment to all units within the corresponding region. The covariates X_3 and X_4 are irrelevant, but this fact is unknown to the algorithm. The propensity score is given by

$$P(D = 1 \mid X) = \Lambda\left(\log(0.5) + \frac{(X_1 + X_2 + X_3 + X_4)(\log(2) - \log(0.5))}{4}\right),$$

where $\Lambda(\cdot)$ denotes the logistic function. This specification ensures the propensity score lies in the interval $[1/3, 2/3]$.

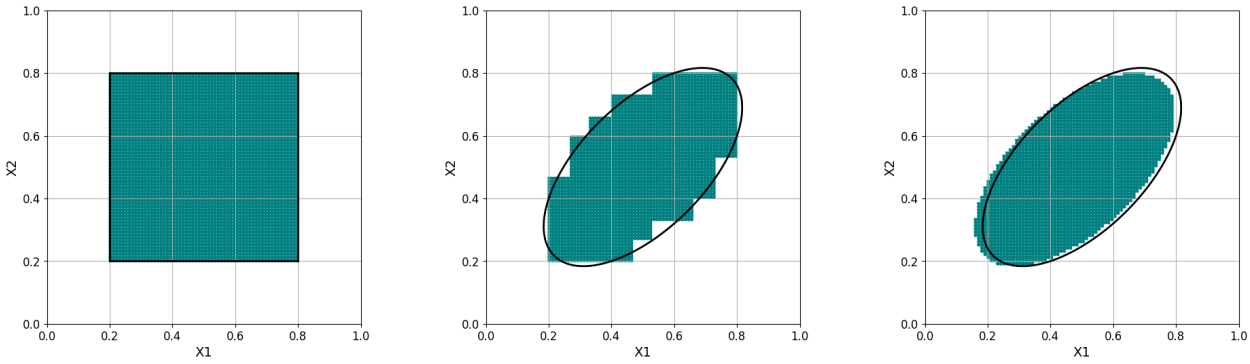
4.2 The Importance of Choosing Policy Class

In practice, the DM does not know what policy class is most suitable for the underlying DGP. The effects of choosing an incorrect policy class can be dramatic. To illustrate, we consider two families of policies more or less suitable for each of the above DGPs.

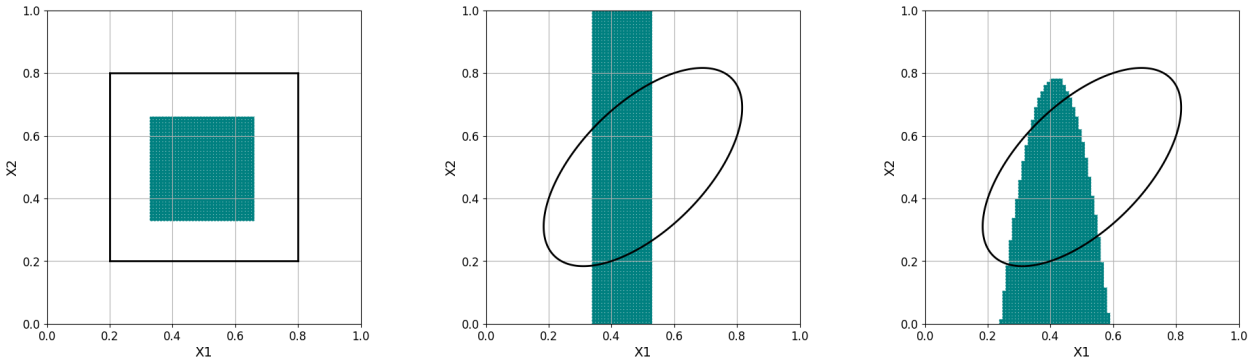
The first family consists of *discretized rules*, where X_1 and X_2 are discretized into bins to form a grid over the unit square. Each cell in the grid can be assigned to either treat or not treat, with model complexity governed by the number of bins along each axis. The rectangular region can be exactly recovered using discretized rules when the number of bins along each axis is a multiple of five (since the decision boundaries lie at $1/5$ and $4/5$), while approximating the elliptical region well requires a very large number of cells.

The second family of policies consists of *linear threshold rules*, in which covariates enter polynomially and complexity is determined by the number of included terms. The elliptical region can be exactly recovered by such a rule when second-order polynomial terms of X_1 and X_2 are included, while approximating the rectangular region well requires using very high-degree polynomials.

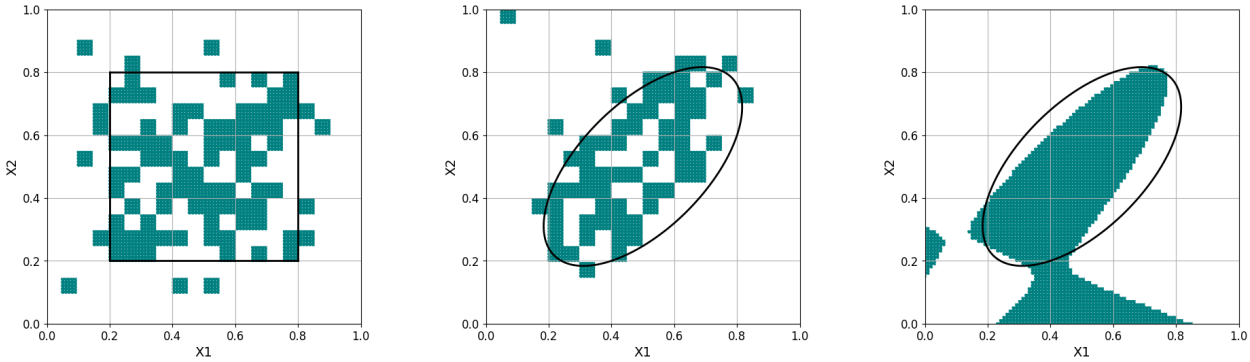
For each DGP, we generate a dataset with a relatively small sample size of 200. We estimate the outcome regressions using OLS and propensity score using logistic regression.



(a) EWM policies closely approximate the first-best rule with appropriate complexity.



(b) EWM policies underfit due to insufficient complexity.



(c) EWM policies overfit due to excessive complexity.

Figure 1: The role of policy complexity

For each family of policies, we fit EWM rules with different fixed complexities. Figure 1 presents the results. The first two columns feature discretized rules and the third one — linear threshold rules. Panel (a) shows the optimal choice of complexity, Panel (b) — underfitting, and Panel (c) — overfitting. It is visually clear that both under- and over-fitting may lead to treatment rules that are drastically different from the optimal rule.

4.3 Adaptive Welfare Maximization

Next, we show that the AWM policy successfully adapts to the underlying unknown complexity in each of the above DGPs. We generate 200 datasets for each sample size $n \in \{200, \ldots, 1600\}$ and compute the average regret for different procedures across these datasets. We compare the AWM rule against several EWM rules with fixed complexity levels. We implement AWM as described in Algorithm 3 in the Appendix, using 4-fold cross-validation and 5-fold cross-fitting. We estimate the nuisance parameters using OLS for the outcome regression models and logistic regression for the propensity score. To reduce the computational complexity, here we focus on discretized rules.[3]

4.3.1 Rectangle DGP

The optimal policy complexity for the Rectangle DGP is five bins per axis: it recovers the optimal treatment region while avoiding overfitting. Our theory suggests that as the sample size increases, AWM should increasingly favor this level of complexity over the alternatives. We let AWM rule select the number of bins per axis adaptively via cross-validation from the range $\{3, 4, \ldots, 10\}$. Figure 2 depicts the regret of AWM alongside EWM policies using 3, 5, and 10 bins per axis. Figure 3 additionally shows the proportion of times each complexity level is selected by AWM at each sample size.

The EWM rule with five bins per axis achieves the lowest regret, as it corresponds to the correctly specified policy class. The AWM rule achieves regret that is very close, despite not knowing the true complexity in advance. EWM with 3 or 10 bins performs worse, with 10 eventually outperforming 3 as the sample size increases and approximation error begins to dominate estimation error. Consistent with our theoretical results, the AWM rule selects mostly 5 and 4 bins at smaller sample sizes and increasingly favors 5 as the sample size grows.

[3]Computing the EWM discretized rule amounts to computing the CATE functions cell by cell, which is computationally straightforward. Computing the optimal linear threshold rules requires solving a Mixed Integer Linear Program; see Mbakop and Tabord-Meehan (2021).

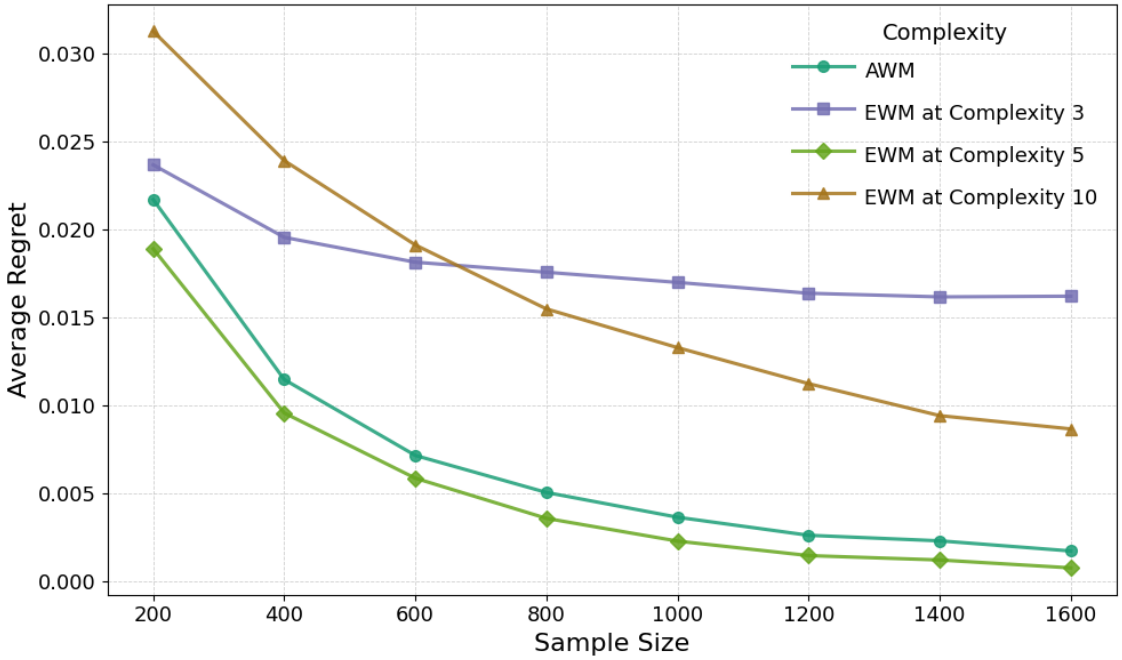


Figure 2: Average regret for the AWM and fixed-complexity EWM rules for the Rectangle DGP.

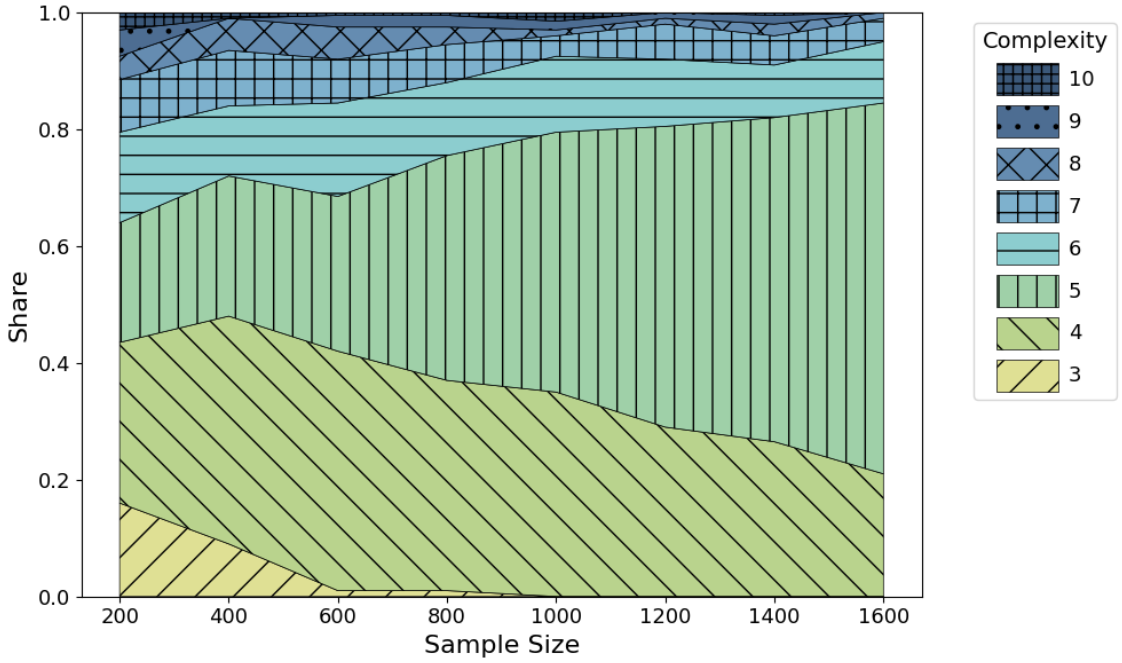


Figure 3: Share of complexity levels (number of bins per axis) selected by the AWM rule at each sample size for the Rectangle DGP.

4.3.2 Ellipse DGP

For the Ellipse DGP, the first-best policy cannot be exactly recovered by any discretized rule. As the sample size increases, we expect higher-complexity classes to perform better, since they can better approximate the curved decision boundary. Ideally, the AWM rule should also shift toward selecting more complex policies as the sample size grows. In this simulation, we let AWM select the number of bins per axis from 5 to 15. We plot the regret of EWM policies with fixed discretizations at 5, 10, and 15 bins, along with the regret of AWM, in Figure 4. Additionally, Figure 5 shows the share of complexity levels chosen by AWM across different sample sizes.

From Figure 5, we clearly see that as the sample size increases, AWM begins to favor higher-complexity classes. As a result, AWM maintains relatively low regret across all sample sizes, as shown in Figure 4. Among the fixed-complexity EWM policies, using 5 bins performs best at smaller sample sizes, while 10 bins becomes optimal as the sample size grows. The 15-bin model consistently overfits and performs worse. AWM initially favors 5 and 6 bins, and gradually shifts toward selecting 10 and 11 bins as more data become available.

In conclusion, these simulation results illustrate the ideal behavior of adaptively selecting policy complexity using cross-validation. When the optimal policy lies within the policy class, AWM eventually identifies the correct level of complexity and achieves regret close to the first-best. In settings where the optimal policy cannot be exactly represented, AWM adapts to the sample size and favors increasingly rich models as sample size increases, effectively balancing estimation and approximation error.

5 Conclusion

This paper proposed a policy learning algorithm called Adaptive Welfare Maximization. It is based a doubly-robust, semiparametrically efficient estimate of the welfare function, which allows to accommodate settings with rich covariates, where estimating the nuisance functions reliably requires using machine learning methods. Moreover, it automatically adapts to the unknown optimal policy complexity for a given DGP. Our proof strategy can be readily adopted and extended to settings with multivalued treatments or non-linear regret functions. We leave such extensions for future work. From a practical perspective, an important direction for future work is developing computational tools to scale the approach proposed here and in related work to large datasets.

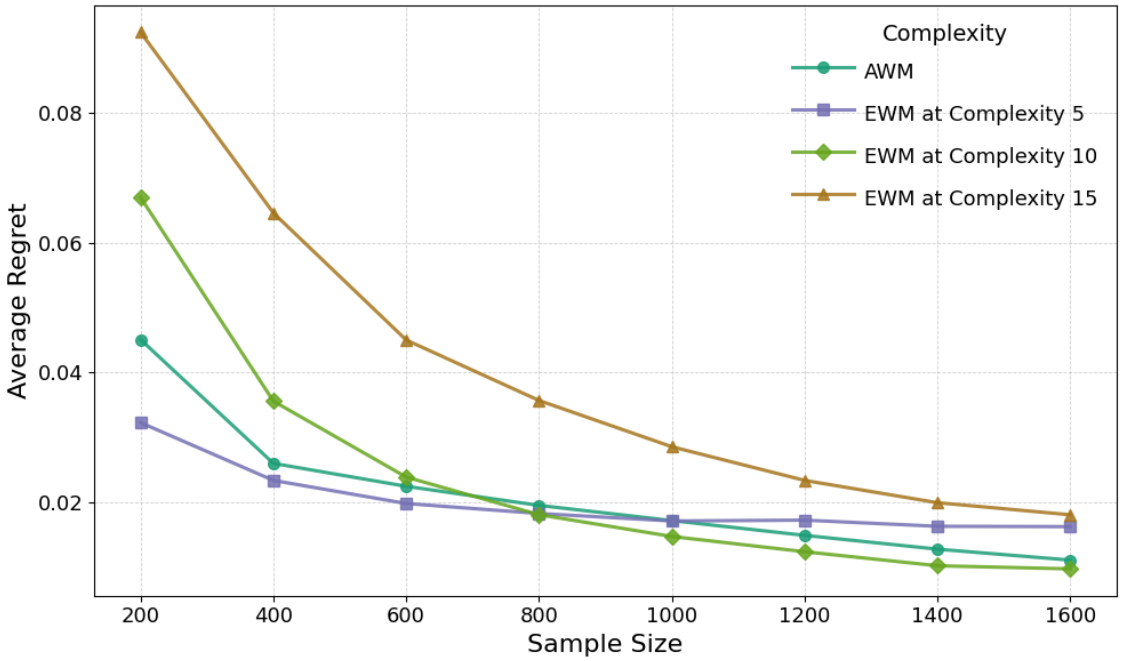


Figure 4: Average regret for AWM and fixed-complexity EWM rules for the Ellipse DGP.

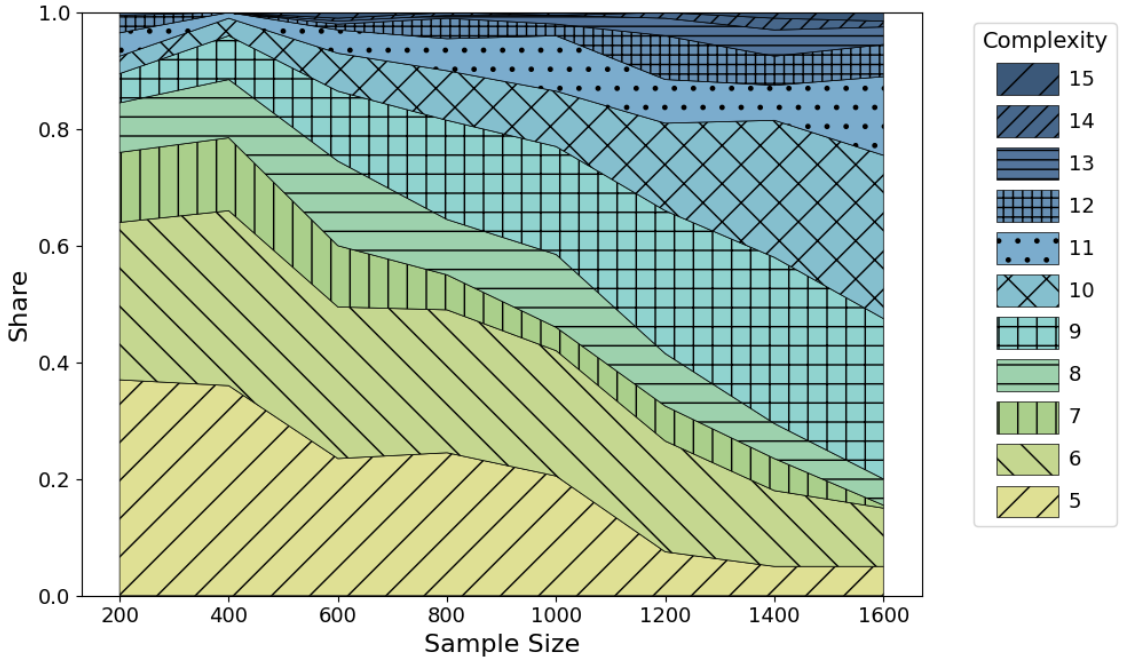


Figure 5: Share of complexity levels (number of bins per axis) selected by AWM at each sample size for the Ellipse DGP.

References

Angrist, J. D., C. Gao, P. Hull, and R. W. Yeh (2025): "Instrumental variables in randomized trials," *NEJM evidence*, 4(4), EVIDctw2400204.

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

Bhattacharya, D., and P. Dupas (2012): "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 167(1), 168–196.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and adaptive estimation for semiparametric models*, vol. 4. Johns Hopkins University Press Baltimore.

Bonnans, J. F., and A. Shapiro (2013): *Perturbation analysis of optimization problems*. Springer Science & Business Media.

Chamberlain, G. (2011): "Bayesian aspects of treatment choice," .

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

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022): "Locally robust semiparametric estimation," *Econometrica*, 90(4), 1501–1535.

Dehejia, R. H. (2005): "Program evaluation as a decision problem," *Journal of Econometrics*, 125(1-2), 141–173.

Fang, Y., J. Xi, and H. Xie (2025): "Model selection for multivalued-treatment policy learning in observational studies," *Journal of Business & Economic Statistics*, pp. 1–13.

Hirano, K., and J. R. Porter (2009): "Asymptotics for statistical treatment rules," *Econometrica*, 77(5), 1683–1701.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

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

Luedtke, A. R., and M. J. Van Der Laan (2016): "Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy," *Annals of statistics*, 44(2), 713.

Manski, C. F. (2004): "Statistical treatment rules for heterogeneous populations," *Econometrica*, 72(4), 1221–1246.

Mbakop, E., and M. Tabord-Meehan (2021): "Model selection for treatment choice: Penalized welfare maximization," *Econometrica*, 89(2), 825–848.

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

Neyman, J. S. (1923): "On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480)," *Annals of Agricultural Sciences*, 10, 1–51.

Ponomarev, K. (2022): "Essays in Econometrics," Ph.d. dissertation, University of California, Los Angeles.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.

Qian, M., and S. A. Murphy (2011): "Performance guarantees for individualized treatment rules," *Annals of statistics*, 39(2), 1180.

Robins, J. M. (2004): "Optimal structural nested models for optimal sequential decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data*, pp. 189–326. Springer.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, 66(5), 688.

Shevtsova, I. G. (2013): "On the absolute constants in the Berry–Esseen inequality and its structural and nonuniform improvements," *Informatika i Ee Primeneniya [Informatics and its Applications]*, 7(1), 124–125.

Shi, C., A. Fan, R. Song, and W. Lu (2018): "High-dimensional A-learning for optimal dynamic treatment regimes," *Annals of statistics*, 46(3), 925.

Shi, L. (2022): "Essays on Treatment Effect Estimation and Treatment Choice Learning.," Ph.d. dissertation, University of California, Los Angeles.

Stoye, J. (2009): "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 151(1), 70–81.

——— (2012): "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics*, 166(1), 138–156.

Tetenov, A. (2012): "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 166(1), 157–165.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak convergence*. Springer.

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

A Appendix

A.1 CV algorithm

Algorithm 3 (Cross-Validation for Policy Complexity)**.**
Input: *Data sample* $(W_i)_{i=1}^n$*; Policy classes* Π_k *for* $k \in \{1, \dots, K\}$*.*
Output: *Data-driven complexity choice* $\hat{k}$*.*

1. *Randomly split the sample* $(W_i)_{i=1}^n$ *into* L *samples, denoted* $S_1, \dots, S_L$*. For each* $l \in \{1, \dots, L\}$*, let the superscripts* (l) *and* $(-l)$ *indicate that the corresponding object was computed using only* S_l*, or only* $S_{-l} = \cup_{j \neq l} S_j$*, correspondingly.*

2. *For each policy class* Π_k*:*

   (i) For each $l \in \{1, \dots, L\}$*, compute the EWM policy:*

   $$\hat{\pi}_k^{(-l)} = \operatorname*{argmax}_{\pi \in \Pi_k} \hat{V}^{(-l)}(\pi),$$

   where $\hat{V}^{(-l)}(\pi)$ *is constructed as in Algorithm 1, and evaluate its performance out of sample:*

   $$\hat{Q}(\hat{\pi}_k^{(-l)}) = \frac{1}{|S_l|} \sum_{i \in S_l} \hat{\pi}_k^{(-l)}(X_i) \hat{\Gamma}^{(-l)}(W_i).$$

   Note: *Computing* $\hat{\Gamma}^{(-l)}$ *does not require cross-fitting, i.e., the first-stage estimators* $\hat{m}^{(-l)}(x, d)$*,* $\tau_{\hat{m}}^{(-l)}(x, d)$*, and* $\hat{g}^{(-l)}(x, z)$ *may be computed using the entire* $(-l)$ *sample.*

   (ii) Compute the cross-validation criterion

   $$\hat{Q}_k = \frac{1}{L} \sum_{l=1}^{L} \hat{Q}(\hat{\pi}_k^{(-l)}).$$

3. *Choose* $\hat{k} = \operatorname{argmax}_{k \in \{1, \dots, K\}} \hat{Q}_k$*.* ■

A.2 Known Results and Some Refinements

To keep the notation simple, we state all results for regular rather than outer expectations, but take into account the potential difference between the two throughout the proofs. For the symmetrization lemma below, the structure of the underlying probability space is important. For the detailed discussion, see Sections 2.1–2.3 in van der Vaart and Wellner (1996).

First, we state a well-known symmetrization inequality; see, e.g., Lemma 2.3.1. in van der Vaart and Wellner (1996), for reference.

Lemma A.1 (Symmetrization). *Let $W_1, \ldots, W_n$ be an i.i.d. sample and $\mathcal{F}$ a class of measurable functions $f : \mathcal{W} \mapsto \mathbb{R}$ such that $\mathbb{E}[f(W_i)] < \infty$ for all $f \in \mathcal{F}$. Then,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \mathbb{E}[f(W_i)]\right|\right] \leqslant 2 \cdot \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i f(W_i)\right|\right],$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables independent of $W_1, \ldots, W_n$.

Next, we state a useful maximal inequality for Orlicz norms. Let ψ be a strictly increasing, convex function satisfying $\psi(0) = 0$, and X be a random variable. The Orlisz norm $||X||_\psi$ is defined as

$$||X||_\psi = \inf\left\{C > 0 : \mathbb{E}\left(\psi\left(\frac{|X|}{C}\right)\right) \leqslant 1\right\}.$$

The following result is Exercise 2.2.8 in van der Vaart and Wellner (1996).

Lemma A.2 (Maximal Inequality with Orlisz Norms). *For any random variables $X_1, \ldots, X_m$ and any strictly increasing, convex function ψ,*

$$\mathbb{E}\left[\max_{j\leqslant m}|X_j|\right] \leqslant \psi^{-1}(m)\max_{j\leqslant m}||X_j||_\psi$$

Proof. For any $C > 0$,

$$\begin{aligned}\psi\left(\mathbb{E}\left[\max_{j\leqslant m}\frac{|X_j|}{C}\right]\right) &\leqslant \mathbb{E}\left[\max_{j\leqslant m}\psi\left(\frac{|X_j|}{C}\right)\right] \\ &\leqslant m\max_{j\leqslant m}\mathbb{E}\left[\psi\left(\frac{|X_j|}{C}\right)\right],\end{aligned}$$

where the first inequality holds because ψ is convex and non-decreasing. Therefore, for any C such that $\max_{j\leqslant m}\mathbb{E}\left[\psi\left(|X_j|/C\right)\right] \leqslant 1$, we have

$$\mathbb{E}\left[\max_{j\leqslant m}|X_j|\right] \leqslant C\psi^{-1}(m).$$

Choosing $C = \max_{j\leqslant m}||X_j||_\psi$ concludes the proof. ■

Next, we pin down the universal constant in Theorem 2.6.4. from van der Vaart and Wellner (1996).

Lemma A.3 (Covering Numbers for VC classes)**.** *For any VC-class $\mathcal{C}$ of sets, any probability measure Q, any $r \geqslant 1$, and $0 < \varepsilon < 1$,*

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{C})-1)}.$$

Proof. We closely follow the proof of Theorem 2.6.4. in van der Vaart and Wellner (1996). We start by referencing the main steps and introducing the necessary notation. First, note that $||\mathbf{1}_C - \mathbf{1}_D||_{Q,r} = Q^{1/r}(C \triangle D)$, so an ε^r-cover under $L_1(Q)$ produces an ε-cover under $L_r(Q)$. Therefore, the result for $r > 1$ follows immediately from the result for $r = 1$. Second, one can argue that it suffices to consider empirical type measures Q supported on a large enough finite set of distinct points $\{x_1, \ldots, x_n\}$. Third, it is more convenient to bound the packing number $D(\varepsilon, \mathcal{C}, L_1(Q))$ first and use the fact that $N(\varepsilon, \mathcal{C}, L_1)(Q)) \leqslant D(\varepsilon/2, \mathcal{C}, L_1(Q))$.

Each set $C \in \mathcal{C}$ can be identified with a binary vector $\mathbf{1}_C = (\mathbf{1}(x_i \in C))_{i=1}^n$, and the collection $\mathcal{C}$ can be identified with a binary matrix $\mathcal{Z}$ of size $n \times \#\mathcal{Z}$. Define $d(\mathbf{1}_{C_1}, \mathbf{1}_{C_2}) = n^{-1} \sum_{i=1}^n |\mathbf{1}_{C_1} - \mathbf{1}_{C_2}|$. Then, recalling that Q places probability $1/n$ on each x_i, $Q(C_1 \triangle C_2) = d(\mathbf{1}_{C_1}, \mathbf{1}_{C_2})$, so that $D(\varepsilon, \mathcal{C}, L_1(Q)) = D(\varepsilon, \mathcal{Z}, d)$. For simplicity of notation, assume that $\mathcal{Z}$ is ε-separated with respect to d, so the goal is to bound its size $\#\mathcal{Z}$ in terms of the VC dimension $V(\mathcal{C})$.

Denote $S = V(\mathcal{C}) - 1$ and fix an integer m such that $S \leqslant m < n$. For a subset $J \subset \{1, \ldots, n\}$ of size $\#J = m$, let $\mathcal{Z}_J$ denote the projection of $\mathcal{Z}$ onto $\{0,1\}^J$, and $\overline{\#\mathcal{Z}_J}$ denote the average size of $\mathcal{Z}_J$ over all subsets J or size m. Then, following the proof on Page 138 of van der Vaart and Wellner (1996), we arrive to the bound

$$\#\mathcal{Z} \leqslant \frac{\overline{\#\mathcal{Z}_J} n\varepsilon(m+1)}{\varepsilon n(m+1) - 2(n-m)S} \leqslant \frac{\varepsilon(m+1)\overline{\#\mathcal{Z}_J}}{\varepsilon(m+1) - 2S} \leqslant \frac{\varepsilon m \overline{\#\mathcal{Z}_J}}{\varepsilon m - 2S},$$

which holds without any extra constants. The number of points in any $\mathcal{Z}_J$ is equal to the number of subsets picked out by $\mathcal{C}$ from the points $\{x_i : i \in J\}$. By the Sauer-Shelah Lemma, this is bounded by $\sum_{j=0}^S \binom{m}{j}$, which is smaller than $(em/S)^S$ for $m \geqslant S$.[4] Therefore,

$$\#\mathcal{Z} \leqslant \left(\frac{e}{S}\right)^S \frac{m^{S+1}\varepsilon}{m\varepsilon - 2S}$$

holds for all integers m such that $S \leqslant m < n$. Denote the right-hand side of the preceding display by $f(m)$. This function is strictly decreasing until $m^* = 2(S+1)/\varepsilon$ and strictly increasing afterwards. Therefore, the optimal unconstrained choice is $m = m^*$, for which $f(m^*) = (2e/\varepsilon)^S (S+1)(1 + S^{-1})^S$. However, the argument leading to the upper bound on

[4] Indeed, for $t \in (0,1)$, $\sum_{j=0}^S \binom{m}{j} \leqslant \sum_{j=0}^S \binom{m}{j} \frac{t^j}{t^S} \leqslant \frac{(1+t)^m}{t^S}$. Set $t = \frac{S}{m}$ and use $(1 + S/m)^m \leqslant e^S$.

$\#\mathcal{Z}$ only applies to integer m such that $S \leqslant m < n$, while m^* may not be integer. To ensure that a similar bound holds for an integer value of m, we can simply use $f(m^* - 1)$ since somewhere between $m^* - 1$ and m^* there must be an integer, and $f(m)$ is decreasing on this interval. We have

$$\begin{aligned} f(m^* - 1) &= \left(\tfrac{e}{S}\right)^S \frac{(2(S+1)/\varepsilon - 1)^{S+1}\varepsilon}{(2(S+1)/\varepsilon - 1)\varepsilon - 2S} \\ &= \left(\tfrac{2e}{\varepsilon}\right)^S \tfrac{1}{1-\varepsilon/2}(S + 1 - \varepsilon/2)\left(1 + \tfrac{1-\varepsilon/2}{S}\right)^S \\ &\leqslant \left(\tfrac{2e}{\varepsilon}\right)^S (S+1)^{\frac{1}{1-\varepsilon/2}} \exp(1 - \varepsilon/2) \\ &\leqslant \left(\tfrac{2e}{\varepsilon}\right)^S (S+1) \cdot 2\sqrt{e}, \end{aligned}$$

for all $\varepsilon \in (0, 1)$ since the function $g(\varepsilon) = (1-\varepsilon/2)^{-1}\exp(1-\varepsilon/2)$ is monotonically increasing. Therefore, we obtain the bound

$$\#\mathcal{Z} \leqslant \left(\frac{2e}{\varepsilon}\right)^S (S+1) \cdot 2\sqrt{e},$$

and it remains to check that this bound still holds when $m^* - 1 < S$ or $m^* \geqslant n$. Note that $m^* - 1 \geqslant S$ for all $\varepsilon \in (0, 1)$. If $m^* \geqslant n$, by the Sauer-Shelah Lemma

$$\#\mathcal{Z} \leqslant \sum_{j=0}^{S} \binom{n}{j} \leqslant \left(\frac{en}{S}\right)^S \leqslant \left(\frac{em^*}{S}\right)^S \leqslant e\left(\frac{2e}{\varepsilon}\right)^S,$$

which certainly implies the bound in the previous display. Therefore, recalling that $\#\mathcal{Z} = D(\varepsilon, \mathcal{C}, L_1(Q))$,

$$\begin{aligned} N(\varepsilon, \mathcal{C}, L_1(Q)) &\leqslant D(\varepsilon/2, \mathcal{C}, L_1(Q)) \\ &\leqslant \left(\tfrac{4e}{\varepsilon}\right)^S (S+1) \cdot 2\sqrt{e} \\ &= \left(\tfrac{4e}{\varepsilon}\right)^{V(\mathcal{C})-1} V(\mathcal{C}) \cdot 2\sqrt{e} \\ &= \tfrac{1}{2\sqrt{e}} V(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\tfrac{1}{\varepsilon}\right)^{(V(\mathcal{C})-1)}, \end{aligned}$$

and the desired result follows. ■

Next, we state three lemmas about specific *VC-subgraph* classes of functions. A subgraph of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$C_f = \{(t, x) \in \mathbb{R} \times \mathcal{X} : t < f(x)\}.$$

A class of functions $\mathcal{F}$ is *VC-subgraph* if the class of all subgraphs

$$\mathcal{C}_{\mathcal{F}} = \{C_f : f \in \mathcal{F}\}$$

has a finite VC dimension. In this case we denote $V(\mathcal{F}) = V(\mathcal{C}_{\mathcal{F}})$.

The first result is Theorem 2.6.7. from van der Vaart and Wellner (1996). It is a direct corollary of the result for sets (our Lemma A.3) and holds with the same universal constant.

Lemma A.4 (Covering Number for VC-subgraph Classes)**.** *For a VC-class of functions with a measurable envelope function F and $r \geqslant 1$, for any probability measure Q with $||F||_{Q,r} > 0$,*

$$N(\varepsilon ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)},$$

for $0 < \varepsilon < 1$.

For a particular VC-subgraph class of functions, the bound in Lemma A.4 can be improved.

Lemma A.5 (A Simple VC-Subgraph Class)**.** *Let $\mathcal{G}$ denote a class of subsets of $\mathcal{X}$ with a finite VC dimension $V(\mathcal{G})$, and $F : \mathcal{X} \to \mathbb{R}$ be an arbitrary function. Define a class of functions:*

$$\mathcal{F} = \{\mathbf{1}(x \in G)F(x) : G \in \mathcal{G}\}.$$

Then, $\mathcal{F}$ is VC-subgraph with $V(\mathcal{F}) \leqslant V(\mathcal{G})$.

Proof. Let $VC(\mathcal{G}) = d$ and $D = \{(t_1, x_1), \ldots, (t_d, x_{d+1})\} \subset \mathbb{R} \times \mathcal{X}$ be an arbitrary set of points. By definition, D is shattered by $\mathcal{F}$ if for every subset $\{(t_j, x_j) : j \in J\}$ there is a function f with subgraph C_f such that $C_f \cap D = \{(t_j, x_j) : j \in J\}$. Equivalently, D is shattered by $\mathcal{F}$ if for every subset $J \subset \{1, \ldots, d+1\}$ there is a set $G \in \mathcal{G}$ satisfying

$$\begin{aligned} t_j &< \mathbf{1}(x_j \in G)F(x_j) \text{ for } j \in J \\ t_k &\geqslant \mathbf{1}(x_k \in G)F(x_k) \text{ for } k \notin J \end{aligned} \tag{A.1}$$

We will argue that D cannot be shattered by $\mathcal{F}$.

First, if there is (t_j, x_j) such that $t_j < 0$ and $t_j < F(x_j)$, then $t_j < \mathbf{1}(x_j \in G)F(x_j)$ holds for all $G \in \mathcal{G}$. In this case, any subset of D that does not include t_j, x_j cannot be picked out, so D cannot be shattered by $\mathcal{F}$. Similarly, if there is (t_k, x_k) such that $t_k \geqslant 0$ and $t_k \geqslant F(x_k)$, then $t_k \geqslant \mathbf{1}(x_k \in G)F(x_k)$ holds for all $G \in \mathcal{G}$. So, any subset of D that includes this point cannot be picked out, and D cannot be shattered by $\mathcal{F}$. Therefore, we will assume that each (t_j, x_j) satisfies either $t_j < 0, F(x_j) \geqslant 0$ or $t_j \geqslant 0, F(x_j) < 0$ for $j = 1, \ldots, d+1$.

By assumption, $\mathcal{G}$ does not shatter $\{x_1, \ldots, x_{d+1}\}$, meaning that there exist a subset $\{x_j\}_{j \in J}$ that $\mathcal{G}$ cannot pick out. Then, for every $G \in \mathcal{G}$ we have either $x_j \notin G$ for some $j \in J$ or $x_k \in G$ for some $k \notin J$. If the inequalities in (A.1) do not hold for this J for any G, then $\{(t_j, x_j)\}_{j \in J}$ cannot be picked out and D cannot be shattered by $\mathcal{F}$. Suppose the inequalities in (A.1) hold for some $G \in \mathcal{G}$. If $x_j \notin G$ for some $j \in J$, it must be that $t_j < 0$ and, according to the previous discussion, $F(x_j) \geqslant 0$. Then the set $J' = J \backslash (t_j, x_j)$ cannot be picked out. If $x_k \in G$ for some $k \notin J$, it must be that $t_k \geqslant 0$ and $F(x_k) < 0$, so the set $J'' = J \cup k$ cannot be picked out. Therefore, D cannot be shattered by $\mathcal{F}$, so $VC(\mathcal{F}) \leqslant VC(\mathcal{G})$. ■

Lemma A.6 (Covering Numbers for Special VC-Subgraph Classes)**.** *Let $\mathcal{F}$ be the class of functions defined in Lemma A.5. For any $r \geqslant 1$, probability measure Q with $||F||_{Q,r} > 0$, and $0 < \varepsilon < 1$,*

$$N(\varepsilon \, ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Proof. By Lemma A.5, $\mathcal{F}$ is VC-subgraph. For $r = 1$, note that

$$||f_1 - f_2||_{Q,1} = \mathbb{E}_Q[|\mathbf{1}_{G_1} - \mathbf{1}_{G_2}||F|] = P(C_{f_1} \triangle C_{f_2}) \, ||F||_{Q,1},$$

where $P = \lambda \times Q / ||F||_{Q,1}$ is a probability measure on $\mathbb{R} \times \mathcal{X}$ and λ is a Lebesgue measure on $\mathbb{R}$. Then, by Lemma A.3,

$$N(\varepsilon \, ||F||_{Q,1}, \mathcal{F}, L_1(Q)) = N(\varepsilon, \mathcal{C}_{\mathcal{F}}, L_1(P)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{(V(\mathcal{F})-1)}.$$

For $r > 1$, note that

$$||f_1 - f_2||_{Q,r}^r = \mathbb{E}_Q(|\mathbf{1}_{G_1} F - \mathbf{1}_{G_2} F||F|^{r-1}) = \frac{||f_1 - f_2||_{R,1}}{||F||_{R,1}} \mathbb{E}_Q(|F|^r),$$

for the probability measure R with density $|F|^{r-1}/\mathbb{E}_Q(|F|^{r-1})$ with respect to Q. Therefore,

$$||f_1 - f_2||_{Q,r} = \left(\frac{||f_1 - f_2||_{R,1}}{||F||_{R,1}}\right)^{1/r} ||F||_{Q,r},$$

so that by the previous argument applied to R instead of Q

$$N(\varepsilon \, ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant N(\varepsilon^r \, ||F||_{R,1}, \mathcal{F}, L_1(R)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)},$$

which completes the proof. ■

The last lemma is a version of the classical maximal inequality for Rademacher complexity of functional classes with finite VC-dimension with a pinned down universal constant.

Lemma A.7 (Finite-Sample Bound on Rademacher Complexity)**.** *Let $W_1, \ldots, W_n$ be an i.i.d. sample and $\xi_1, \ldots, \xi_n$ be i.i.d. Rademacher random variables independent of $W_1, \ldots, W_n$.*

1. *Let $\mathcal{F}$ be a VC-subgraph of functions with $f_0(w) = 0 \in \mathcal{F}$, a finite VC dimension $VC(\mathcal{F})$, and a measurable envelope F such that $S = \mathbb{E}[F^2] < \infty$. Then:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i f(W_i)\right|\right] \leqslant C\sqrt{\frac{VC(\mathcal{F})S}{n}},$$

   where $C = 4\sqrt{12}\int_0^1 \sqrt{1/(2e^{3/2}) + \log(16e) + 2\log(1/u)}du \leqslant 34$.

2. *In the special case when $\mathcal{F} = \{f(x) = \mathbf{1}(x \in G)F(x) : G \in \mathcal{G}\}$, for a VC-class of sets $\mathcal{G}$ and an arbitrary measurable function F with $S = \mathbb{E}[F^2] < \infty$, the above holds with $C = 4\sqrt{12}\int_0^1 \sqrt{1/(2e^{3/2}) + \log(4e) + 2\log(1/u)}du \leqslant 29$.*

Proof. Denote $\mathbb{G}_n^0(f) = n^{-1/2}\sum_{i=1}^n \xi_i f(W_i)$. By the Law of Iterated Expectations,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\mathbb{G}^0(f)\right|\right] = \frac{1}{\sqrt{n}}\mathbb{E}_{W_1^n}\left[\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{G}^0(f)\right|\right]\right] \tag{A.2}$$

We will use a chaining argument to bound the right hand side of (A.2). Let $\eta = 2\,||F||_{2,n}$, and define $\mathcal{F}_0 = \{f_0\}$ and $\mathcal{F}_j$ contain centers of the balls in the minimal $\eta 2^{-j}$-cover of $\mathcal{F}$ under $||\cdot||_{2,n}$, so that $|\mathcal{F}_j| = N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n})$. Let $\phi_j : \mathcal{F} \to \mathcal{F}_j$ be a map that for a given f finds the closest element of $\mathcal{F}_j$. For any $f_k \in \mathcal{F}_k$ define a chain $f_{k-l} = \phi_{k-l}(f_{k-l+1})$ for $l = 1, \ldots, k$. Then,

$$\mathbb{G}_n^0(f_k) = \sum_{j=1}^{k}(\mathbb{G}_n^0(f_j) - \mathbb{G}_n^0(f_{j-1})) \leqslant \sum_{j=1}^{k}\max_{g\in\mathcal{F}_j}|\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))|, \tag{A.3}$$

Let $\psi_2(x) = e^{x^2} - 1$ and $||\cdot||_{\psi_2}$ denote the corresponding Orlisz norm. By Lemma 2.2.7. in van der Vaart and Wellner (1996), the process $\mathbb{G}_n^0(f)$ is sub-Gaussian for the metric $d_n(f_1, f_2) = ||f_1 - f_2||_{2,n}$, and satisfies $||\mathbb{G}_n^0(f) - \mathbb{G}_n^0(g)||_{\psi_2} \leqslant \sqrt{6}\,||f - g||_{2,n}$, conditional on W_1^n. By Lemma A.2, the preceding sentence, and the fact that $||g - \phi_{j-1}(g)||_{2,n} \leqslant \eta 2^{-(j-1)}$,

$$\begin{aligned}\mathbb{E}_{\xi_1^n}\left[\max_{g\in\mathcal{F}_j}|\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))|\right] &\leqslant \psi_2^{-1}(|\mathcal{F}_j|)\max_{g\in\mathcal{F}_j}||\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))||_{\psi_2} \\ &\leqslant \sqrt{6}\cdot\psi_2^{-1}(N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n}))\cdot\eta 2^{-(j-1)}.\end{aligned} \tag{A.4}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}_k}|\mathbb{G}_n^0(f)|\right] &\leqslant \sqrt{6}\sum_{j=1}^{k}\psi_2^{-1}(N(\eta 2^{-j},\mathcal{F},||\cdot||_{2,n}))\eta 2^{-(j-1)}\\ &\overset{(a)}{\leqslant} 4\sqrt{6}\int_0^{\eta/2}\psi_2^{-1}(N(\varepsilon,\mathcal{F},||\cdot||_{2,n}))d\varepsilon\\ &= 4\sqrt{6}\int_0^{||F||_{2,n}}\sqrt{\log(N(\varepsilon,\mathcal{F},||\cdot||_{2,n})+1)}d\varepsilon\\ &\overset{(b)}{\leqslant} 4\sqrt{12}\int_0^{||F||_{2,n}}\sqrt{\log N(\varepsilon,\mathcal{F},||\cdot||_{2,n})}d\varepsilon,\end{aligned}$$

where (a) follows from counting the rectangles under the curve $\varepsilon\mapsto\psi_2^{-1}(N(\varepsilon,\mathcal{F},||\cdot||_{2,n}))$, and (b) follows from $\log(x+1)\leqslant 2\log(x)$ for $x\geqslant 2$. Conditional on W_1^n, the process $\mathbb{G}_n^0$ is separable, so by letting $k\to\infty$ in the previous display we conclude that

$$\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n^0(f)|\right]\leqslant 4\sqrt{12}\int_0^{||F||_{2,n}}\sqrt{\log N(\varepsilon,\mathcal{F},||\cdot||_{2,n})}d\varepsilon. \tag{A.5}$$

Denote $V\equiv VC(\mathcal{F})$ and $L=(2\sqrt{e})^{-1}$. With a change of variables $u=\varepsilon/||F||_{2,n}$,

$$\int_0^{||F||_{2,n}}\sqrt{\log N(\varepsilon,\mathcal{F},||\cdot||_{2,n})}d\varepsilon = ||F||_{2,n}\int_0^1\sqrt{\log N(u||F||_{2,n},\mathcal{F},||\cdot||_{2,n})}du. \tag{A.6}$$

Applying Lemma A.4 (or Lemma A.6 for the special case) with $r=2$ and $Q=P_n$,

$$\begin{aligned}\log N(u||F||_{2,n},\mathcal{F},||\cdot||_{2,n}) &\leqslant \log(LV)+V\log(16e)+2(V-1)\log\left(\tfrac{1}{u}\right)\\ &= V\left(L\tfrac{\log(LV)}{LV}+\log(16e)+2\tfrac{V-1}{V}\log\left(\tfrac{1}{u}\right)\right)\\ &\leqslant V\left(L/e+\log(16e)+2\log\left(\tfrac{1}{u}\right)\right),\end{aligned}$$

where the last line uses the fact that $\log(t)/t\leqslant 1/e$ for all $t>0$. Therefore,

$$\int_0^{||F||_{2,n}}\sqrt{\log N(\varepsilon,\mathcal{F},||\cdot||_{2,n})}d\varepsilon\leqslant\int_0^1\sqrt{L/e+\log(16e)+2\log(1/u)}du\cdot\sqrt{V||F||_{2,n}^2} \tag{A.7}$$

Combining (A.5) and (A.7), we obtain

$$\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n^0(f)|\right] \leqslant C\sqrt{V\,||F||_{2,n}^2}$$

where $C = 4\sqrt{12}\int_0^1 \sqrt{\sqrt{e}/2 + \log(16e) + 2\log(1/u)}du$ (or the same expression with $4e$ instead of $16e$ in the special case). By (A.2) and Jensen's inequality,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(W_i)\right|\right] \leqslant C\sqrt{\frac{VC(\mathcal{F})S}{n}},$$

which concludes the proof.

■

A.3 Proofs of Theorems 1 and 2

We break the proof of Theorem 1 into a sequence of lemmas. The first lemma establishes an upper bound on expected regret for the oracle EWM rule, assuming full knowledge of the nuisance functions in $\Gamma(W)$. It is a version of Theorem 1 in Kitagawa and Tetenov (2018) with a precisely pinned down constant.

Lemma A.8 (Regret Bound for Oracle EWM)**.** *Let Assumptions 2.1 and 2.2 hold,* Π *have a finite VC-dimension, and* $\tilde{V}_n(\pi) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\Gamma(W_i)$*. Consider the oracle EWM policy rule*

$$\tilde{\pi}_n^{EWM} \in \operatorname*{argmax}_{\pi\in\Pi} \tilde{V}_n(\pi).$$

Then, for any $P \in \mathbf{P}$*,*

$$R_P(\tilde{\pi}_n^{EWM}) \leqslant \overline{C}\cdot\sqrt{\mathbb{E}_P[\Gamma(W)^2]}\cdot\sqrt{\frac{VC(\Pi)}{n}},$$

for a universal constant $\overline{C} \leqslant 58$*.*

Proof. Since $\tilde{\pi}_n^{EWM}$ maximizes $\tilde{V}_n(\pi)$,

$$\begin{aligned} V(\pi^*) - V(\tilde{\pi}_n^{EWM}) &= V(\pi^*) - \tilde{V}_n(\tilde{\pi}_n^{EWM}) + \tilde{V}_n(\tilde{\pi}_n^{EWM}) - V(\tilde{\pi}_n^{EWM}) \\ &\leqslant V(\pi^*) - \tilde{V}_n(\pi^*) + \sup_{\pi\in\Pi}|\tilde{V}_n(\pi) - V(\pi)|. \end{aligned}$$

Thus, since $\mathbb{E}[\tilde{V}_n(\pi^*)] = V(\pi^*)$,

$$R_P(\tilde{\pi}_n^{EWM}) \leqslant \mathbb{E}_P\left[\sup_{\pi\in\Pi}|\tilde{V}_n(\pi) - V(\pi)\right].$$

Applying Lemma A.1 to the class $\mathcal{F} = \{\pi(x)\Gamma(w) : \pi \in \Pi\}$, which has an integrable envelope $|\Gamma(w)|$, and then applying part 2 of Lemma A.7 implies the stated result. ■

The second Lemma establishes that the doubly-robust and oracle welfare estimates are close to each other uniformly in $\pi \in \Pi$. It is a non-asymptotic version of Lemma 4 from Athey and Wager (2021), which we prove under weaker assumptions.

Lemma A.9 (Uniform Coupling). *Let assumptions 2.1, 2.2, and 2.3 hold, and* Π *have finite VC-dimension. Let* $\hat{V}_n(\pi)$ *be computed as in Algorithm 1 (denoting the sample size by* n*, for simplicity) and* $\tilde{V}_n(\pi) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\Gamma(W_i)$*. Then,*

$$\mathbb{E}\left[\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)|\right] \leqslant R_{1,n} + R_{2,n} + R_{3,n},$$

where

$$R_{1,n} = \overline{C}\sqrt{J \cdot B^2 \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_g}}},$$

$$R_{2,n} = \overline{C}\sqrt{J \cdot \frac{2(\eta^2+1)}{\eta^2} \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}},$$

$$R_{3,n} = \sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}},$$

and $\overline{C} \leqslant 58$ *is a universal constant.*

Proof. Denote the indices of the observations included in j-th fold by I_j, and recall that $\hat{m}^{(-j)}$, $\tau_{\hat{m}^{(-j)}}$ and $\hat{g}^{(-j)}$ denote the first-stage estimators computed using all observations excluding the j-th fold. For $i \in I_j$, denote $\hat{\Gamma}_i \equiv \hat{\Gamma}^{(-j)}(W_i)$, $\Gamma_i = \Gamma(W_i)$, and write the difference $\hat{\Gamma}_i - \Gamma_i$ as a sum of three terms

$$\begin{aligned}
\hat{\Gamma}_i - \Gamma_i &= (Y_i - m(X_i, D_i))(\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i)) \\
&\quad + \tau_{\hat{m}^{(-j)}}(X_i, D_i) - \tau_m(X_i, D_i) - g(X_i, Z_i)(\hat{m}^{(-j)}(X_i, D_i) - m(X_i, D_i)) \\
&\quad - (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))(\hat{m}^{(-j)}(X_i, D_i) - m(X_i, D_i)).
\end{aligned}$$

Denote the corresponding summands in $\hat{V}_n(\pi) - \tilde{V}_n(\pi)$ by $S_1(\pi), S_2(\pi)$, and $S_3(\pi)$. We will bound each term separately.

First Term. Write $S_1(\pi) = \sum_{j=1}^J S_1^{(j)}(\pi)$, where $\frac{n}{n_j} S_1^{(j)}(\pi)$ is equal to

$$\frac{1}{n_j}\sum_{i\in I_j} \pi(X_i)(Y_i - m(X_i, D_i))(\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i)).$$

By the law of iterated expectations, (recalling that $Z \perp (Y(0), Y(1), D(0), D(1)) \,|\, X$)

$$\mathbb{E}[\pi(X_i)(Y_i - m(X_i, D_i))(\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i)) \mid \hat{g}^{(-j)}] = 0.$$

Denote

$$V_{1,n}(j) = \mathbb{E}\left[\pi(X_i) \cdot \mathbb{E}[(Y_i - m(X_i, D_i))^2 | X_i, D_i] \cdot (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))^2 \mid \hat{g}^{(-j)}\right].$$

Applying, conditional on $\hat{g}^{(-j)}$, Lemma A.8 with $(Y_i - m(X_i, D_i)) \cdot (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))$ in place of Γ_i, we obtain

$$\frac{n}{n_j}\mathbb{E}_P\left[\sup_{\pi\in\Pi}|S_1^{(j)}(\pi)| \;\middle|\; \hat{g}^{(-j)}\right] \leqslant \overline{C}\sqrt{\frac{VC(\Pi)V_{1,n}(j)}{n_j}}.$$

By Assumption 2.3, $\pi(X_i) \leqslant 1$, and the law of total variance,

$$\mathbb{E}_P[V_{1,n}(j)] \leqslant B^2\frac{a((1-J^{-1})n)}{n^{\zeta_g}}.$$

Using the last two displays, the law of iterated expectations, and Jensen's inequality, for each $j \in \{1, \ldots, J\}$, we obtain

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|S_1^{(j)}(\pi)|\right] \leqslant \overline{C}\sqrt{\frac{n_j}{n}}\sqrt{B^2\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_g}}}.$$

Since supremum is sub-additive, using the inequality $\sum_{j=1}^J \sqrt{n_j/n} \leqslant \sqrt{J}$, and summing over j, we obtain

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|S_1(\pi)|\right] \leqslant \overline{C}\sqrt{J \cdot B^2 \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_g}}}.$$

Second Term. As above, write $S_2(\pi) = \sum_{j=1}^J S_2^{(j)}(\pi)$, where $\frac{n}{n_j}S_2^{(j)}(\pi)$ is equal to

$$\frac{1}{n_j}\sum_{i\in I_j}\pi(X_i)(\tau_{\hat{m}^{(-j)}}(X_i, D_i) - \tau_m(X_i, D_i) - g(X_i, Z_i)(\hat{m}^{(-j)}(X_i, D_i) - m(X_i, D_i)))$$

Denote the individual summands in the previous display by $f(W_i; \pi)$. By Assumption 2.1 and the law of iterated expectations,

$$\mathbb{E}_P[f(W_i; \pi) \,|\, \hat{m}^{(-j)}, \tau_{\hat{m}^{(-j)}}] = 0.$$

Denote $V_{2,n}(j) = \mathbb{E}_P[f(W_i;\pi)^2 \,|\, \hat{m}^{(-j)}, \tau_{\hat{m}^{(-j)}}]$. Applying, conditional on $\hat{m}^{(-j)}$ and $\tau_{\hat{m}^{(-j)}}$, Lemma A.8 with $(\tau_{\hat{m}^{(-j)}}(X_i,D_i) - \tau_m(X_i,D_i) - g(X_i,Z_i)(\hat{m}^{(-j)}(X_i,D_i) - m(X_i,D_i)))$ in place of Γ_i, we obtain

$$\frac{n}{n_j}\mathbb{E}\left[\sup_{\pi\in\Pi}|S_2^{(j)}(\pi)| \;\middle|\; \hat{g}^{(-j)}\right] \leqslant \overline{C}\sqrt{\frac{VC(\Pi)V_{2,n}(j)}{n_j}}$$

Using $\pi(X_i) \leqslant 1$, $(a+b)^2 \leqslant 2(a^2+b^2)$, and Assumptions 2.1, 2.2, and 2.3, we obtain

$$\mathbb{E}_P[V_{2,n}(j)] \leqslant 2\left(\frac{a((1-J^{-1})n)}{n^{\zeta_m}} + \frac{1}{\eta^2}\frac{a((1-J^{-1})n)}{n^{\zeta_m}}\right) = \frac{2(\eta^2+1)}{\eta^2}\frac{a((1-J^{-1})n)}{n^{\zeta_m}}.$$

By the last two displays, the law of iterated expectation, and Jensen's inequality,

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|S_2^{(j)}(\pi)|\right] \leqslant \overline{C}\sqrt{\frac{n_j}{n}}\sqrt{\frac{2(\eta^2+1)}{\eta^2}\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}}$$

Since supremum is sub-additive, summing up across j, using the fact that $\sum_{j=1}^{J}\sqrt{n_j/n} \leqslant \sqrt{J}$,

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|S_2(\pi)|\right] \leqslant \overline{C}\sqrt{J\cdot\frac{2(\eta^2+1)}{\eta^2}\cdot\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}}$$

Third Term. Let $j(i)$ denote the fold in which observation i belongs. We have:

$$S_3(\pi) = -\frac{1}{n}\sum_{i=1}^{n}\pi(X_i)(\hat{g}^{(-j(i))}(X_i,Z_i) - g(X_i,Z_i))(\hat{m}^{(-j(i))}(X_i,D_i) - m(X_i,D_i))$$

By Cauchy-Schwartz inequality (in $\mathbb{R}^n$) and $\pi(X_i) \leqslant 1$,

$$\begin{aligned}\sup_{\pi\in\Pi}|S_3(\pi)| \;&\leqslant \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\hat{g}^{(-j(i))}(X_i,Z_i) - g(X_i,Z_i))^2} \\ &\quad\times\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\hat{m}^{(-j(i))}(X_i,D_i) - m(X_i,D_i))^2}.\end{aligned}$$

Taking expectations on both sides, using Cauchy-Schwartz inequality, and recalling Assumption 2.3, we obtain

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|S_3(\pi)|\right] \leqslant \sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}},$$

and the proof is complete. ■

The third lemma is a finite-sample version of Theorem 1 from Athey and Wager (2021), which we prove under weaker assumptions.

Lemma A.10 (EWM with Doubly-Robust Welfare Estimates)**.** *Let Assumptions 2.1, 2.2 and 2.3 hold, and* Π *have finite VC-dimension. Let* $\hat{V}_n(\pi)$ *be computed as in Algorithm 1, and*

$$\hat{\pi}_n \in \operatorname*{argmax}_{\pi \in \Pi} \hat{V}_n(\pi).$$

Then,

$$R_P(\hat{\pi}_n) \leqslant \overline{C} \cdot \sqrt{\mathbb{E}_P[\Gamma(W)^2]} \cdot \sqrt{\frac{VC(\Pi)}{n}} + 2R_n,$$

for a universal constant $\overline{C} \leqslant 58$ *and* $R_n = o(n^{-1/2})$ *as defined in Lemma A.9.*

Proof. Recall that $\tilde{V}_n(\pi) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\Gamma(W_i)$. Note that

$$\begin{aligned} V(\pi^*) - V(\hat{\pi}_n) &= V(\pi^*) - \hat{V}_n(\hat{\pi}_n) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) \\ &\leqslant V(\pi^*) - \hat{V}_n(\pi^*) + \tilde{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) + \sup_{\pi\in\Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)| \\ &\leqslant \{V(\pi^*) - \tilde{V}_n(\pi^*)\} + \sup_{\pi\in\Pi} |\tilde{V}_n(\pi) - V(\pi))| + 2\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)|. \end{aligned}$$

Taking expectations and applying Lemmas A.8 and A.9, we obtain

$$R_P(\hat{\pi}_n) \leqslant \overline{C}\sqrt{\mathbb{E}_P[\Gamma(W)^2]}\sqrt{\frac{VC(\Pi)}{n}} + 2R_n,$$

where $R_n = o(n^{-1/2})$ defined precisely in Lemma A.9. ■

The fourth lemma establishes an upper bound on expected regret for the "oracle AWM" rule computed assuming full knowledge of the nuisance functions in $\Gamma(W)$.

Lemma A.11 (Adaptation for Oracle AWM)**.** *Let Assumptions 2.1, 2.2, and 2.4 hold, and* $\tilde{\pi}_n^{AWM}$ *be defined as in Algorithm 2 but with* $\tilde{V}_n(\pi) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\Gamma(W_i)$ *instead of* $\hat{V}_n(\pi)$*. For* $P \in \mathbf{P}$*, let* $\pi_P^* \in \operatorname{argmax}_\Pi V(\pi)$ *and* $\mathbf{P}_k = \{P \in \mathbf{P} : \pi_P^* \in \Pi_k\}$*. Then, for any* $P \in \mathbf{P}_k$

$$R_P(\tilde{\pi}_n^{AWM}) \leqslant \min_{k\leqslant K}\left(\overline{C}\sqrt{\frac{\mathbb{E}_P[\Gamma(W)^2]}{n_E}}\sqrt{VC(\Pi_k)} + V(\pi_P^*) - V(\pi_{k,P}^*)\right) + \sqrt{\frac{K\mathbb{E}_P[\Gamma(W)^2]}{n_H}},$$

where $\overline{C} \leqslant 58$ *is a universal constant.*

Proof. To simplify notation, we suppress the dependence of population quantities on P. To distinguish between the oracle and feasible estimators, we use the notation $\tilde{A}$ instead of $\hat{A}$ for all in-sample quantities. Further, we denote $\pi_k^* \in \operatorname{argmax}_{\pi\in\Pi_k} V(\pi)$, and

$$\tilde{V}^{(H)}(\pi) = \frac{1}{n_H}\sum_{i\in H} \pi(X_i)\Gamma(W_i).$$

For any $k \leqslant K$, we can expand

$$\begin{aligned} V(\pi^*) - V(\tilde{\pi}_n^{AWM}) &= V(\pi^*) - V(\pi_k^*) \\ &\quad + V(\pi_k^*) - \tilde{Q}_{\tilde{k}} \\ &\quad + \tilde{Q}_{\tilde{k}} - V(\tilde{\pi}_n^{AWM}). \end{aligned}$$

Recall from Algorithm 2 that $\tilde{k} = \text{argmax}_{k \leqslant K} \tilde{Q}_k$. Thus:

$$\begin{aligned} V(\pi_k^*) - \tilde{Q}_{\tilde{k}} &\leqslant V(\pi_k^*) - \tilde{Q}_k \\ &= \{V(\pi_k^*) - V(\tilde{\pi}_k^{(E)})\} + \{V(\tilde{\pi}_k^{(E)}) - \tilde{V}^{(H)}(\tilde{\pi}_k^{(E)})\}. \end{aligned}$$

By the Law of Iterated Expectations, the second summand has mean zero. By Lemma A.8, the mean of the first summand is bounded by

$$\mathbb{E}_P[V(\pi_k^*) - V(\tilde{\pi}_k^{(E)})] \leqslant \overline{C}\sqrt{\mathbb{E}_P[\Gamma(W)^2]}\sqrt{\frac{VC(\Pi_k)}{n_E}}.$$

Next, recall from Algorithm 2 that $\tilde{\pi}_n^{AWM} = \tilde{\pi}_{\tilde{k}}^{(E)}$. Thus,

$$\tilde{Q}_{\tilde{k}} - V(\tilde{\pi}_n^{AWM}) = \tilde{V}^{(H)}(\tilde{\pi}_{\tilde{k}}^{(E)}) - V(\tilde{\pi}_{\tilde{k}}^{(E)}) \leqslant \max_{k \leqslant K}\{\tilde{V}^{(H)}(\tilde{\pi}_k^{(E)}) - V(\tilde{\pi}_k^E)\}.$$

Since

$$\mathbb{E}_P[(\tilde{V}^{(H)}(\tilde{\pi}_k^{(E)}) - V(\tilde{\pi}_k^E))^2 \mid \tilde{\pi}_k^E] \leqslant \frac{\mathbb{E}_P[\Gamma(W)^2]}{n_H},$$

using Lemma A.2 we obtain:

$$\mathbb{E}_P\left[\max_{k \leqslant K}\{\tilde{V}^{(H)}(\tilde{\pi}_k^{(E)}) - V(\tilde{\pi}_k^E)\}\right] \leqslant \sqrt{K}\max_{k \leqslant K}\mathbb{E}[(\tilde{V}^{(H)}(\tilde{\pi}_k^{(E)}) - V(\tilde{\pi}_k^E))^2]^{1/2} \leqslant \sqrt{\frac{K\mathbb{E}_P[\Gamma(W)^2]}{n_H}}.$$

Combining the above results, we obtain, for any $k \in \{1, \ldots, K\}$,

$$R_P(\tilde{\pi}_n^{AWM}) \leqslant V_P(\pi_P^*) - V_P(\pi_{k,P}^*) + \overline{C}\sqrt{\mathbb{E}_P[\Gamma(W)^2]}\sqrt{\frac{VC(\Pi_k)}{n_E}} + \sqrt{\frac{K\mathbb{E}_P[\Gamma(W)^2]}{n_H}}.$$

Taking a minimum over $k \leqslant K$ gives the stated result. ■

The last lemma addresses the remainder terms.

Lemma A.12 (Remainder Terms)**.** *Let $(W_i)_{i \in E}$ and $(W_i)_{i \in H}$ denote the estimation and hold-out samples. In the notation of Lemma A.9,*

1. *For every fixed* $\pi \in \Pi$,
$$\mathbb{E}_P[\hat{V}^{(E)}(\pi) - \tilde{V}^{(E)}(\pi)] \leqslant R_{3,n_E}.$$

2. *For any* $\hat{\pi}_k^{(E)}$ *computed using the estimation sample,*
$$\mathbb{E}_P[\hat{V}^{(H)}(\hat{\pi}_k^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_k^{(E)})] \leqslant R_{3,n_E}.$$

Proof. The first claim follows from the proof of Lemma A.9. Recall the terms $S_1(\pi)$, $S_2(\pi)$, and $S_3(\pi)$ introduced there. The expectations of the first two terms are equal to zero, and the expectation of the third term is shown to be less than R_{3,n_H}.

Now, we term to the second claim. To simplify the notation, we replace the arguments of the functions $\Gamma(W_i)$, $m(D_i, X_i)$, $g(X_i, Z_i)$, $\tau_m(D_i, X_i)$, and their estimated counterparts, with a subscript i reflecting the observation (from the hold-out sample) at which they are evaluated. Moreover, we drop the superscript (E) since all quantities are estimated on the same sample. With this in mind, we can expand $\hat{\Gamma}_i - \Gamma_i$ as a sum of three terms:

$$\hat{\Gamma}_i - \Gamma_i = (\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i)) + (Y_i - m_i)(\hat{g}_i - g_i) - (\hat{m}_i - m_i)(\hat{g}_i - g_i).$$

Let S_1, S_2 and S_3 denote the corresponding sums in $\hat{V}^{(H)}(\hat{\pi}_k^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_k^{(E)})$. Then, by Assumption 2.1-2 and the Law of Iterated Expectations,

$$\mathbb{E}[S_1 \,|\, (W_i)_{i\in E}] = \mathbb{E}\left[\, \hat{\pi}_k(X_i) \cdot \mathbb{E}[(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i)) \,|\, X_i, (W_i)_{i\in E}] \;\middle|\; (W_i)_{i\in E} \,\right] = 0.$$

Further, by the Law of Iterated Expectations and conditional exogeneity of Z_i,

$$\mathbb{E}[S_2 \,|\, (W_i)_{i\in E}] = \mathbb{E}\left[\, \hat{\pi}_k(X_i) \cdot \mathbb{E}[Y_i - m_i \,|\, X_i, D_i, (W_i)_{i\in E}] \cdot (\hat{g}_i - g_i) \;\middle|\; (W_i)_{i\in E} \,\right] = 0.$$

Finally, by Cauchy-Schwartz inequality (in $\mathbb{R}^{n_H}$) and $\hat{\pi}_k(X_i)^2 \leqslant 1$,

$$S_3 \leqslant \sqrt{\tfrac{1}{n_H}\textstyle\sum_{i\in H}(\hat{m}_i - m_i)^2} \cdot \sqrt{\tfrac{1}{n_H}\textstyle\sum_{i\in H}(\hat{g}_i - g_i)^2}.$$

Taking expectations on both sides, applying Cauchy-Schwartz inequality again, and using the Law of Iterated Expectations, we obtain

$$\mathbb{E}[S_3] \leqslant \sqrt{\mathbb{E}[(\hat{m}_i - m_i)^2] \cdot \mathbb{E}[(\hat{g}_i - g_i)^2]} \leqslant R_{3,n_E},$$

and the proof is complete. ■

A.3.1 Proof of Theorem 1

Recall that $\hat{\pi}_n^{AWM} = \hat{\pi}_{\hat{k}}^{(E)}$, $\pi^* \in \text{argmax}_{\pi\in\Pi} V(\pi)$, and $\pi_k^* \in \text{argmax}_{\pi\in\Pi_k} V(\pi)$. Consider an expansion

$$V(\pi^*) - V(\hat{\pi}_{n,\hat{k}}) = V(\pi^*) - V(\pi_k^*) + \underbrace{V(\pi_k^*) - \hat{Q}_{\hat{k}}}_{(I)} + \underbrace{\hat{Q}_{\hat{k}} - V(\hat{\pi}_{\hat{k}}^{(E)})}_{(II)}. \tag{A.8}$$

Consider term (I). Since $\hat{Q}_{\hat{k}} \geqslant \hat{Q}_k$, for any $k \in \{1, \ldots, K\}$, we can bound

$$\begin{aligned}(I) &\leqslant V(\pi_k^*) - \hat{Q}_k \\ &\leqslant \left\{V(\pi_k^*) - V(\hat{\pi}_k^{(E)})\right\} + \left\{\tilde{V}^{(H)}(\hat{\pi}_k^{(E)}) - V(\hat{\pi}_k^{(E)})\right\} + \left\{\tilde{V}^{(H)}(\hat{\pi}_k^{(E)}) - \hat{V}^{(H)}(\hat{\pi}_k^{(E)})\right\}.\end{aligned}$$

By Lemmas A.10 and A.12, and the Law of Iterated Expectations,

$$\mathbb{E}_P[(I)] \leqslant \overline{C}\sqrt{\mathbb{E}_P[\Gamma(W)^2]}\sqrt{\frac{VC(\Pi_k)}{n_E}} + 2R_{n_E} + R_{3,n_E},$$

where R_{n_E} and R_{3,n_E} are as defined in Lemma A.9.

Next, consider

$$(II) = \left\{\hat{V}^{(H)}(\hat{\pi}_{\hat{k}}^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_{\hat{k}}^{(E)})\right\} + \left\{\tilde{V}^{(H)}(\hat{\pi}_{\hat{k}}^{(E)}) - V(\hat{\pi}_{\hat{k}}^{(E)})\right\}. \tag{A.9}$$

For the first summand in (A.9), we can bound

$$\begin{aligned}\mathbb{E}_P\left[\hat{V}^{(H)}(\hat{\pi}_{\hat{k}}^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_{\hat{k}}^{(E)})\right] &\leqslant \mathbb{E}_P\left[\max_{k\leqslant K}|\hat{V}^{(H)}(\hat{\pi}_k^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_k^{(E)})|\right] \\ &\leqslant K\max_{k\leqslant K}\mathbb{E}_P\left[\left|\hat{V}^{(H)}(\hat{\pi}_k^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_k^{(E)})\right|\right].\end{aligned}$$

To bound the above expression, using the notation of Lemma A.12 and omitting the superscript (E) for the first-stage estimators, for simplicity, we expand

$$\hat{\Gamma}_i - \Gamma_i = (\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i)) + (Y_i - m_i)(\hat{g}_i - g_i) - (\hat{m}_i - m_i)(\hat{g}_i - g_i).$$

By the triangle inequality,

$$\begin{aligned}\mathbb{E}_P\left[\left|\hat{V}^{(H)}(\hat{\pi}_k^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_k^{(E)})\right|\right] &= \mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_k^{(E)}(X_i)(\hat{\Gamma}_i - \Gamma_i)\right|\right] \\ &\leqslant \mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_k^{(E)}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))\right|\right] \\ &+ \mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_k^{(E)}(X_i)(Y_i - m_i)(\hat{g}_i - g_i)\right|\right] \\ &+ \mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_k^{(E)}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)\right|\right]\end{aligned}$$

By Assumption 2.1-1, the Law of Iterated Expectations, $\hat{\pi}_k^{(E)}(X_i) \leqslant 1$, and Assumption 2.3, we obtain

$$\begin{aligned}&\mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_{l,k}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))\right|^2\right] \\ &\qquad\leqslant \mathbb{E}_P\left[\tfrac{1}{n_H^2}\textstyle\sum_{i\in H}(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))^2\right] \\ &\qquad\leqslant \tfrac{1}{n_H}\mathbb{E}_P[(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))^2] \\ &\qquad\leqslant \tfrac{2}{n_H}(\mathbb{E}_P[(\tau_{\hat{m},i} - \tau_{m,i})^2] + \mathbb{E}[g_i^2(\hat{m}_i - m_i)^2]) \\ &\qquad\leqslant \tfrac{2}{n_H}\cdot\tfrac{\eta^2+1}{\eta^2}\tfrac{a((1-J^{-1})n_E)}{(n_E)^{\zeta_m}}.\end{aligned}$$

As a result, using the inequality $\mathbb{E}[|A|] \leqslant \mathbb{E}[|A|^2]^{1/2}$,

$$\mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_k^{(E)}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))\right|\right] \leqslant \sqrt{2\cdot\tfrac{n_E}{n_H}\cdot\tfrac{\eta^2+1}{\eta^2}\tfrac{a((1-J^{-1})n_E)}{(n_E)^{1+\zeta_m}}}.$$

Using a similar argument, instrument exogeneity, and $\mathbb{E}[(Y_i - m_i)^2 \,|\, X_i, D_i] \leqslant B^2$, we obtain:

$$\mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_k^{(E)}(X_i)(Y_i - m_i)(\hat{g}_i - g_i)\right|^2\right] \leqslant \tfrac{1}{n_H}\mathbb{E}_P[(Y_i - m_i)^2(\hat{g}_i - g_i)^2] \leqslant \tfrac{B^2}{n_H}\cdot\tfrac{a((1-J^{-1})n_E)}{(n_E)^{\zeta_g}},$$

and thus

$$\mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_{i\in H}\hat{\pi}_k^{(E)}(X_i)(Y_i - m_i)(\hat{g}_i - g_i)\right|\right] \leqslant \sqrt{\tfrac{n_E}{n_H}\cdot B^2\cdot\tfrac{a((1-J^{-1})n_E)}{(n_E)^{1+\zeta_g}}}.$$

Finally, by Cauchy-Schwartz inequality (in $\mathbb{R}^{n_H}$) and $\hat{\pi}_k^{(E)}(X_i) \leqslant 1$,

$$\left|\tfrac{1}{n_H}\textstyle\sum_i\hat{\pi}_k^{(E)}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)\right| \leqslant \sqrt{\tfrac{1}{n_H}\textstyle\sum_{i\in H}(\hat{m}_i - m_i)^2}\cdot\sqrt{\tfrac{1}{n_H}\textstyle\sum_{i\in H}(\hat{g}_i - g_i)^2}$$

Taking expectations on both sides, applying Cauchy-Schwartz inequality and the Law of

Iterated Expectations,

$$\mathbb{E}_P\left[\left|\tfrac{1}{n_H}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)\right|\right] \leqslant \sqrt{\mathbb{E}_P[(\hat{m}_i - m_i)^2]\cdot\mathbb{E}_P[(\hat{g}_i - g_i)^2]} \leqslant \sqrt{\tfrac{a((1-J^{-1})n_E)^2}{n_E^{\zeta_m+\zeta_g}}}.$$

Combining the above results, we obtain

$$\mathbb{E}_P\left[\left|\hat{V}^{(H)}(\hat{\pi}_k^{(E)}) - \tilde{V}^{(H)}(\hat{\pi}_k^{(E)})\right|\right] \leqslant \sqrt{2\cdot\tfrac{n_E}{n_H}\cdot\tfrac{\eta^2+1}{\eta^2}\cdot\tfrac{a((1-J^{-1})n_E)}{(n_E)^{1+\zeta_m}}} + \sqrt{\tfrac{n_E}{n_H}\cdot B^2\cdot\tfrac{a((1-J^{-1})n_E)}{(n_E)^{1+\zeta_g}}} + R_{3,n_E},$$

where R_{3,n_E} is defined in Lemma A.9 (and the preceding display).

For the second summand in (A.9), arguing as in the proof of Lemma A.11,

$$\mathbb{E}_P\left[\tilde{V}^{(H)}(\hat{\pi}_{\hat{k}}^{(E)}) - V(\hat{\pi}_{\hat{k}}^{(E)})\right] \leqslant \sqrt{\tfrac{K\mathbb{V}ar_P(\Gamma(W))}{n_H}}.$$

Combining the bounds on (I) and (II), we obtain, for any $k \in \{1,\ldots,K\}$,

$$R_P(\hat{\pi}_n^{AWM}) \leqslant V(\pi_P^*) - V(\pi_k^*) + \overline{C}\sqrt{\mathbb{E}_P[\Gamma(W)^2]}\sqrt{\tfrac{VC(\Pi_k)}{n_E}} + \sqrt{\tfrac{K\mathbb{V}ar_P(\Gamma(W))}{n_H}} + \tilde{R}_n,$$

where $\tilde{R}_n = o(n^{-1/2})$ is given by

$$\tilde{R}_n = R_{n_E} + 3R_{3,n_E} + \sqrt{2\cdot\tfrac{n_E}{n_H}\cdot\tfrac{\eta^2+1}{\eta^2}\tfrac{a((1-J^{-1})n_E)}{(n_E)^{1+\zeta_m}}} + \sqrt{\tfrac{n_E}{n_H}\cdot B^2\cdot\tfrac{a((1-J^{-1})n_E)}{(n_E)^{1+\zeta_g}}}. \tag{A.10}$$

For any $P \in \mathbf{P}_k$, $V(\pi_P^*) - V(\pi_k^*) = 0$, so the above display implies the stated result. ■

A.3.2 Proof of Theorem 2

Let $\mathbf{P}$ denote the class of DGP's satisfying Assumption 2.2. Below, we construct a subclass of $\mathbf{P}$ for which the worst-case regret can be bounded from below by a term proportional to $B/\eta\sqrt{(VC(\Pi)-1)/n}$. Let $x_1,\ldots,x_d$, where $d = VC(\Pi)-1$, be a set shattered by Π with the largest possible cardinality. Let

$$X \in \{x_1,\ldots,x_d\},\ \ P(X = x_j) = \tfrac{1}{d},\ \text{for all } j;$$
$$T \in \{0,1\},\ \ P(T=1) = p,\ \ T \perp (X, Y_0, Y_1).$$

Further, let $Y_0 = 0$, and, given a parameter vector $c = (c_1,\ldots,c_d) \in \{-1,1\}^d$,

$$Y_1 \,|\, X = x_j = \begin{cases} A & \text{w.p. } \frac{1}{2}(1 + c_j\frac{\gamma}{A}) \\ -A & \text{w.p. } \frac{1}{2}(1 - c_j\frac{\gamma}{A}) \end{cases},$$

where $\gamma/A \leqslant 1$, and w.p. stands for "with probability." Then, for $Y = TY_1 + (1-T)Y_0$,

$$\begin{aligned} \mathbb{E}[Y^2] &= pA^2, \\ \tau(x_j) &= \mathbb{E}[Y_1 - Y_0 | X = x_j] = \gamma c_j. \end{aligned}$$

For any $c \in \{-1,1\}^d$, the joint distribution of $W = (Y, X, T)$ constructed above belongs to $\mathbf{P}$ as long as $p \in [\eta, 1-\eta]$ and $pA^2 \leqslant B^2$. We will specify suitable p and A below.

Let $\mathbf{P}_\mathcal{C} = \{P_{W|C=c} : c \in \{-1,1\}^d\} \subset \mathbf{P}$ denote the set of distributions of W constructed above. Let π_P^* denote the optimal treatment rule when the distribution of the data is P, and $\pi_c^* \equiv \pi^*_{P_{W|C=c}}$. By construction, $\pi_c^*(x_j) = \mathbf{1}(c_j = 1) \in \Pi$, since the class Π shatters $\{x_1, \ldots, x_d\}$. For any data-dependent policy $\hat{\pi}_n$,

$$V(\pi_c^*) - V(\hat{\pi}_n) = \frac{\gamma}{d} \sum_{j=1}^d c_j (\pi_c^*(x_j) - \hat{\pi}_n(x_j)) = \frac{\gamma}{d} \sum_{j=1}^d \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j)).$$

Then, for any distribution $\mu \in \Delta(\{-1,1\}^d)$,

$$\begin{aligned} \sup_{P \in \mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] &\geqslant \max_{P \in \mathcal{P}_\mathcal{C}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n))] \\ &\geqslant \int \mathbb{E}_{P_{W_1^n|C=c}}[V(\pi_c^*) - V(\hat{\pi}_n))] d\mu(c) \\ &= \frac{\gamma}{d} \sum_{j=1}^d \int \int \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j)) dP_{W_1^n|C=c} d\mu(c) \\ &= \frac{\gamma}{d} \sum_{j=1}^d P_{W_1^n, C_j}(\mathbf{1}(C_j = 1) \neq \hat{\pi}_n(x_j)) \\ &\geqslant \gamma \cdot \inf_\pi P_{W_1^n, C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n)). \end{aligned} \tag{A.11}$$

Here, $P_{W_1^n, C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n))$ is the probability of misclassification of $\mathbf{1}(C_j = 1)$ using W_1^n. By Theorem 2.1. in Devroye and Lugosi (1996), the infimum is attained by the Bayes Classifier, $\pi^*(W_1^n) = \mathbf{1}(P(C_j = 1|W_1^n) > 0.5)$, and is equal to

$$\begin{aligned} P(\mathbf{1}(C_j = 1) \neq \pi^*(W_1^n)) &= \tfrac{1}{2} P(\, P(C_j = 1 \,|\, W_1^n) \leqslant 0.5 \,|\, C_j = 1) \\ &\quad + \tfrac{1}{2} P(\, P(C_j = 1 \,|\, W_1^n) > 0.5 \,|\, C_j = -1). \end{aligned} \tag{A.12}$$

We bound this quantity from below for a specific distribution $\mu(\cdot)$ of C. Let $C_j \in \{-1, 1\}$ be i.i.d. with $P(C_j = 1) = 1/2$ and $C = (C_1, \ldots, C_d)$. The joint distribution of $W = (Y, X, T)$

given $C = c$ is

$$P(Y = y, X = x_j, T = t \,|\, C = c) = \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}(1 + c_j \frac{\gamma}{A})\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}(1 - c_j \frac{\gamma}{A})\frac{p}{d} & y = -A, t = 1 \end{cases} .$$

Moreover,

$$P(Y = y, X = x_k, T = t) = \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}\frac{p}{d} & y = -A, t = 1 \end{cases} ,$$

and

$$\begin{aligned} P(Y = y, X = x_k, T = t \,|\, C_j = 1) \;&= \mathbf{1}(k \neq j) P(Y = y, X = x_j, T = t) \\ &\quad + \mathbf{1}(k = j) \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}(1 + \frac{\gamma}{A})\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}(1 - \frac{\gamma}{A})\frac{p}{d} & y = -A, t = 1 \end{cases} , \end{aligned}$$

so that

$$\frac{P(Y = y, X = x_k, T = t \,|\, C_j = 1)}{P(Y = y, X = x_k, T = t)} = \mathbf{1}(k \neq j) + \mathbf{1}(k = j) \begin{cases} 1 & y = 0, t = 0 \\ 1 + \frac{\gamma}{A} & y = A, t = 1 \\ 1 - \frac{\gamma}{A} & y = -A, y = 1 \end{cases} .$$

Therefore,

$$P(C_j = 1 \,|\, W_1^n) = \frac{P(W_1^n \,|\, C_j = 1) P(C_j = 1)}{P(W_1^n)} = \frac{1}{2}\left(1 + \frac{\gamma}{A}\right)^{N_j^+} \left(1 - \frac{\gamma}{A}\right)^{N_j^-} ,$$

where $N_j^+ = \#\{i : X_i = x_j, Y_i = A, T_i = 1\}$ and $N_j^- = \#\{i : X_i = x_j, Y_i = -A, T_i = 1\}$. The tuple $(N_j^+, N_j^-, n - N_j^+ - N_j^-)$ has a multinomial distribution

$$\begin{aligned} &P(N_j^+ = k_1, N_j^- = k_2 \,|\, C_j = 1) \\ &\qquad = \binom{n}{k_1}\binom{n - k_1}{k_2} \left(\frac{1}{2}(1 + \frac{\gamma}{B})\frac{p}{d}\right)^{k_1} \left(\frac{1}{2}(1 - \frac{\gamma}{B})\frac{p}{d}\right)^{k_2} \left(1 - \frac{p}{d}\right)^{n - k_1 - k_2} . \quad \text{(A.13)} \end{aligned}$$

Consider the first summand in (A.12). Denote $a = \gamma/A \leqslant 1$, for brevity, and proceed

conditional on $C_j = 1$. Note that

$$\begin{aligned} P(P(C_j = 1|W_1^n) \leqslant 0.5) &= P((1+a)^{N_j^+}(1-a)^{N_j^-} \leqslant 1) \\ &\geqslant P((1-a^2)^{N_j^+} \leqslant 1 \,|\, N_j^+ \leqslant N_j^-) \cdot P(N_j^+ \leqslant N_j^-) \\ &= P(N_j^+ \leqslant N_j^-). \end{aligned}$$

Let $D_i^+ = \mathbf{1}(X_i = x_j, Y_i = A, T_i = 1)$, $D_i^- = \mathbf{1}(X_i = x_j, Y_i = -A, T_i = 1)$. Then, $\mathbb{E}[D_i^+ - D_i^-] = ap/d$, $\mathbb{V}ar[D_i^+ - D_i^-] = p/d - (ap/d)^2$, and, by direct computation $\mathbb{E}[|D_i^+ - D_i^-|^3] \leqslant p/d$. Letting Z_n denote the studentised version of $n^{-1}\sum_{i=1}^n (D_i^+ - D_i^-)$ and $\Phi(\cdot)$ denote the Standard Normal CDF, using Berry-Esseen inequality we obtain

$$\begin{aligned} P(N_j^+ \leqslant N_j^-) &= P(\tfrac{1}{n}\textstyle\sum_{i=1}^n (D_i^+ - D_i^-) \leqslant 0) \\ &= P\left(Z_n \leqslant \frac{-\sqrt{n}ap/d}{\sqrt{p/d-(ap/d)^2}}\right) \\ &\geqslant \Phi\left(\frac{-\sqrt{n}ap/d}{\sqrt{p/d-(ap/d)^2}}\right) - \frac{K}{\sqrt{n}}\frac{p/d}{(p/d)^{1/2}(1-a^2p/d)^{3/2}}, \end{aligned}$$

where $K < 0.469$ Shevtsova (2013). Choose $a = \frac{c}{\sqrt{n}}\sqrt{\frac{d}{p}}$ for some $c \in (0,1)$ and n large enough to ensure $a \leqslant 1$. Then,

$$P(N_j^+ \leqslant N_j^-) \geqslant \Phi\left(-\frac{c}{\sqrt{1-c^2/n}}\right) - \frac{K}{\sqrt{n}}\frac{\sqrt{p/d}}{(1-c^2/n)^{3/2}}.$$

It is easy to verify that the second summand in (A.12) can be bounded in exactly the same way. Thus, recalling that $\gamma = aA = c\frac{A}{\sqrt{p}}\sqrt{\frac{d}{n}}$,

$$\sup_{P\in\mathbf{P}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] \geqslant c\frac{A}{\sqrt{p}}\sqrt{\frac{d}{n}} \cdot \left\{\Phi\left(-\frac{c}{\sqrt{1-c^2/n}}\right) - \frac{K}{\sqrt{n}}\frac{\sqrt{p/d}}{(1-c^2/n)^{3/2}}\right\}.$$

Choosing $p = \eta$, $A = B/\sqrt{\eta}$, and simplifying,

$$\sup_{P\in\mathbf{P}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] \geqslant \frac{B}{\eta}\sqrt{\frac{d}{n}} \cdot c \cdot \Phi\left(-\frac{c}{\sqrt{1-c^2/n}}\right) - \frac{B}{n}\frac{Kc}{(1-c^2/n)^{3/2}}$$

For $n \geqslant 5$, the maximum value of the function $c \to c\Phi(-c/\sqrt{1-c^2/n})$ is at least 0.16 (attained at some $c \in [0.5, 1]$), and the function $c \mapsto c/(1-c^2/n)$ is monotonically increasing on $c \in [0,1]$ with the maximum value of at most 1.2. Plugging in these values and $K = 0.469$

gives the final bound

$$\sup_{P\in\mathbf{P}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] \geqslant 0.16\frac{B}{\eta}\sqrt{\frac{d}{n}} - \frac{0.6B}{n},$$

valid for $n \geqslant \max(5, \frac{d}{\eta})$. Since the lower bound is valid for any measurable map $\hat{\pi}_n$, we may take an infimum over all such $\hat{\pi}_n$. Repeating the argument with Π_k and $\mathbf{P}_k$ in place of Π and $\mathbf{P}$ gives the stated result.

A.3.3 Proof of Remark 2

Lemma A.13 (Semiparametric Efficiency of Welfare Function)**.** *Suppose that*

(i) The covariate space $\mathcal{X}$ is bounded, the model $\mathbf{P}$ satisfies Assumption 2.2 and all $P \in \mathbf{P}$ are dominated by a sigma-finite measure μ with a bounded density $dP/dQ \leqslant C_Q < \infty$.

(ii) The entropy integral $\mathcal{E}(\Pi)$, defined in (9), is finite. The class Π contains a countable subclass Π_0 such that for each $\pi \in \Pi$ there exists a sequence $\pi_{0,m}$ such that $\pi_{0,m}(x) \to \pi(x)$, for each x.

(iii) The tangent space $T(P)$ is a closed linear subspace of $L_0^2(P)$.

(iii) $V_P(\pi) = \mathbb{E}_P[\pi(X)\Gamma(W)]$, and $\psi(\pi)(W) = \pi(X)\Gamma(W) - \mathbb{E}_P[\pi(X)\Gamma(W)] \in T(P)$.

(v) Letting $\hat{V}(\pi)$ denote a feasible estimator and $\tilde{V}(\pi) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\Gamma(W_i)$ an oracle one, $\sup_{\pi\in\Pi} |\hat{V}(\pi) - \tilde{V}(\pi)| = o_P(n^{-1/2})$.

Then, $\hat{V}(\cdot)$ is semiparametrically efficient for $V_P(\cdot)$ in $C_b(\Pi)$, and

$$\sqrt{n}(\hat{V}(\cdot) - V(\cdot)) \quad \to_d \quad \mathbb{G}(\cdot),$$

where $\mathbb{G}(\cdot)$ is a tight centered Gaussian process with $\mathbb{C}ov(\mathbb{G}(\pi_1), \mathbb{G}(\pi_2)) = \mathbb{E}_P[\psi(\pi_1)\psi(\pi_2)]$.

Proof. We will verify the conditions of Theorem 5.2.1 in Bickel, Klaassen, Ritov, and Wellner (1993). We need to establish that $P \mapsto V_P(\cdot)$ is (weakly) path-wise differentiable, derive the form of the efficient influence function, and argue that $\hat{V}(\cdot)$ attains the efficiency bound.

Fix some $P \in \mathbf{P}$, and let $\{P_{t,h}\} \subseteq \mathbf{P}$ denote a regular parametric sub-model with a score function $h \in T(P)$ and density $p_{t,h} = dP_{t,h}/dQ$ satisfying

$$\int \left(\frac{\sqrt{p_{t,h}} - \sqrt{p}}{t} - \frac{1}{2}\sqrt{p}h\right)^2 dQ \to 0.$$

It is without loss of generality to assume that h is bounded, since the set of all bounded functions is dense in $L_2^0(P)$. Using linearity of $\pi \mapsto V_P(\pi)$, boundedness of $\pi(x)$ and h, and Cauchy-Schwartz inequality, it is straightforward to verify that

$$\sup_{\pi \in \Pi} \left| \frac{V_{P_{t,h}}(\pi) - V_P(\pi)}{t} - \mathbb{E}_P[\psi(\pi)h] \right| \to 0.$$

Thus, $P \mapsto V_P(\cdot)$ is path-wise differentiable with derivative $V_P'(h) : T(P) \to C(\Pi)$ given by

$$V_P'(h)(\pi) = \mathbb{E}[\psi(\pi)h].$$

By the Riesz-Markov theorem, every bounded linear functional $L : C(\Pi) \to \mathbb{R}$ takes the form

$$L(v) = \int_\Pi v(\pi) d\mu_L(\pi),$$

where μ_L is a finite-signed Borel measure on Π. Applying Fubini's theorem twice, we obtain

$$L(V_P'(h)) = \int_\Pi \mathbb{E}_P[\psi(\pi)h] d\mu_L(\pi) = \mathbb{E}_P[\psi(\bar{\pi}_{\mu_L})h],$$

where $\bar{\pi}_L(x) = \int_\Pi \pi(x) d\mu_L(\pi)$. Thus, $\psi(\bar{\pi}_L)$ is the canonical gradient of $P \mapsto V_P(\cdot)$ in the direction L. Define a mapping $\Psi : \mathcal{W} \to C(\Pi)$ as

$$\Psi(W)(\pi) = \pi(X)\Gamma(W) - \mathbb{E}_P[\pi(X)\Gamma(W)].$$

By Fubini's Theorem, $L(\Psi(W)) = \psi(\bar{\pi}_L)(W)$, so $\Psi(W)(\cdot)$ is the efficient influence function for $V(\pi)$. Now, consider the oracle estimator $\tilde{V}(\pi)$. Assumptions (i)–(ii) ensure that the conditions of Theorem 3.10.12 in van der Vaart and Wellner (1996) are met, and thus $\tilde{V}(\cdot)$ is a regular estimator. Since the influence function of $\tilde{V}(\cdot)$ is precisely $\Psi(W)(\cdot)$, by Theorem 5.2.1 in Bickel, Klaassen, Ritov, and Wellner (1993), it is semiparametrically efficient and converges weakly to $\mathbb{G}(\cdot)$. By Assumption (v), $\hat{V}(\pi)$ and $\tilde{V}(\pi)$ are first-order asymptotically equivalent uniformly over $\pi \in \Pi$, meaning that $\hat{V}(\pi)$ is semiparametrically efficient. ■