Policy Learning with Abstention

Ayush Sawarni **Jikai Jin** **Justin Whitehouse** **Vasilis Syrgkanis**
Stanford University

Abstract

Policy learning algorithms are regularly leveraged in domains such as personalized medicine and advertising to develop individualized treatment regimes. However, a critical deficit of existing algorithms is that they force a decision even when predictions are uncertain, a risky approach in high-stakes settings. The ability to abstain, that is, to defer to a safe default or an expert, is crucial but largely unexplored in this context. To remedy this, we introduce a framework for policy learning with abstention, in which policies that choose not to assign a treatment to some customers/patients receive a small, additive reward on top of the value of a random guess. We propose a two-stage learner that first identifies a set of near-optimal policies and then constructs an abstention class based on disagreements between the policies. We establish fast $O(1/n)$-type regret guarantees for the abstaining policy when propensities are known, and show how to extend these guarantees to the unknown-propensity case via a doubly robust (DR) objective. Furthermore, we demonstrate that our abstention framework is a versatile tool with direct applications to several other core problems in policy learning. We use our algorithm as a black box to obtain improved guarantees under margin conditions without the common realizability assumption. We also show that abstention provides a natural connection to both distributionally robust policy learning, where it acts as a hedge against small data shifts, and safe policy improvement, where the goal is to improve upon a baseline policy with high probability.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Policy learning algorithms guide high-stakes decisions in domains like personalized medicine and public policy by constructing individualized treatment rules from observational data. Yet, a critical deficit of existing methods is their failure to abstain when faced with high uncertainty. In safety-critical applications, forcing a decision when evidence is weak can be harmful; the responsible action may be to defer to a human expert or a trusted baseline policy. One way to facilitate this is through allowing learned policies to *abstain* — that is, instead of just assigning a unit to treatment or control (denoted "1" or "0" respectively), they can abstain by outputting a special symbol "$*$".

While abstaining algorithms have been developed in the classification literature (Bousquet and Zhivotovskiy, 2021), generalizing these algorithms to the setting of policy learning from observational data is a non-trivial task. The primary difficulty is the counterfactual nature of the problem: the learner only ever observes an outcome under treatment or control for any given unit, never both. Because of this, the learner must transform samples (either via inverse propensity weighting or through a doubly-robust correction) into "pseudo-outcomes" — quantities representing the unit-specific contrasts in outcomes between treatment and control.

In this work, we present a framework formalizing policy learning with abstention. The goal of the learner is to use observational data $Z = (X, D, Y)$ consisting of covariates $X \in \mathcal{X}$, treatment $D \in \{0, 1\}$, and outcome $Y \in \mathbb{R}$ to learn a high reward *abstaining treatment policy* $\pi : \mathcal{X} \to \{0, 1, *\}$. Here, π is assumed to exist in some pre-defined policy class Π that implicitly encodes constraints of the problem (e.g. Π is the class of all depth 3 decision trees, thus enforcing explainability). Our per-section contributions are as follows:

1. In Section 3, we formalize the problem of policy learning with abstention. We introduce a form of "safe" or abstaining regret, which incentives abstention in regions of uncertainty by offering a

small, additive reward p on top of the value of a random guess.

2. Also in Section 3, we introduce our primary algorithm for policy learning with abstention (Algorithm 1). Our algorithm first de-biases observations via inverse propensity weighting (IPW) when propensities are known (Section 3.1), or by performing a *doubly-robust* correction when propensities are unknown (Section 3.2). We prove the algorithm obtains fast $O(1/n)$ high probability regret rates under the safe/abstaining regret, and in the worst-case obtains $O(1/\sqrt{n})$ standard regret, matching rates due to Athey and Wager (2021).

3. In Section 4.2, we discuss applications of Algorithm 1 to a variety of practically-relevant policy learning problems. This includes the problem of policy learning without standard margin assumptions (Theorems 4.1 and 4.2), safe-policy improvement (Algorithm 3), and learning under distribution shift (Proposition 4.3).

In addition to the above, we also conduct simulations to compare our safe policy improvement algorithm (Algorithm 3) to relevant benchmarks (Thomas et al., 2015). In sum, our work illustrates the practicality and importance of studying policy learning with abstention, and further provides insight into how abstention can impact other areas of policy learning as well.

Related work on abstention. Classification with abstention/a reject option traces back to Chow (1970), who characterize the optimal error–reject trade-off, prescribing abstention whenever the posterior risk of mis-classification exceeds the reject cost. Herbei and Wegkamp (2006) consider classification with a reject option in a statistical learning setting, and Bartlett and Wegkamp (2008) introduce a convex "reject-hinge" loss to make abstention trainable in large-scale settings. Other researchers have developed abstaining classification algorithms for SVMs (Grandvalet et al., 2009), multiclass classification (Ramaswamy et al., 2015), boosting (Cortes et al., 2016), online learning (Cortes et al., 2018; Pasteris et al., 2024; Neu and Zhivotovskiy, 2020; van der Hoeven et al., 2022), and even deep learning settings (Geifman and El-Yaniv, 2017; Geifman and El-Yaniv, 2019). Madras et al. (2018) study a setting where the prediction of an abstaining algorithm is replaced by a downstream expert, but don't provide regret/excess risk guarantees. Most closely related to this paper is the work of Bousquet and Zhivotovskiy (2021), who develop an empirical risk minimization-based algorithm for abstention that enables fast $O(1/n)$ risk minimization rates without standard margin/realizability assumptions.

Policy Learning. Our work builds off of the literature on policy learning, particularly works on welfare maximization/regret minimization (Athey and Wager, 2021; Hirano and Porter, 2009; Kitagawa and Tetenov, 2018; Manski, 2004). Also related to our work is the literature of safe policy improvement, which aims to produce a policy that, with high-probability, improves over some baseline "safe" policy. Existing approaches include using hypothesis testing to select between the baseline and learned policy (Thomas et al., 2015; Cho et al., 2025), bootstrap the baseline policy with a learned policy in regions of low confidence (Laroche et al., 2019; Simão et al., 2019), and directly minimizing "negative regret" (Ghavamzadeh et al., 2016). Our work provides new perspective onto safe policy improvement, showing how abstaining algorithms can naturally be used to construct improved policies. Other complementary areas that are not directly considered in the work are policy evaluation (Dudík et al., 2011; Karampatziakis et al., 2021), offline reinforcement learning (Moodie et al., 2007, 2012; Murphy, 2003), policy learning under partial identification (Kallus and Zhou, 2018), and inference on values of optimal treatment policies (Luedtke and Van Der Laan, 2016; Chen et al., 2023; Whitehouse et al., 2025).

2 NOTATION AND PRELIMINARIES

We assume the learner observes i.i.d. draws $Z = (X, D, Y)$ with covariates $X \in \mathcal{X}$, realized action $D \in \{0, 1\}$, and outcome $Y \in [0, 1]$. Let $(Y(1), Y(0))$ denote potential outcomes, and write

$$g_o(d, x) := \mathbb{E}\big[Y(d) \mid X = x\big],$$
$$\tau_o(x) := g_o(1, x) - g_o(0, x),$$

as respectively the expected outcome mapping and conditional average treatment effect (CATE). We denote the propensity score by $p_o(x) := \mathbb{P}(D{=}1 \mid X{=}x)$. We assume the following of the data generating process.

Assumption 2.1 (Data-generating process.)**.** We assume the following conditions hold throughout:

(i) *Unconfoundedness:* $(Y(1), Y(0)) \perp D \mid X$.

(ii) *Strict overlap:* $p_o(x) \in [\kappa, 1 - \kappa]$ for some known $\kappa \in (0, \frac{1}{2}]$.

We let $\hat{g}, \hat{p}, \hat{\tau}$ denote generic estimates of g_o, p_o, τ_o, respectively. For $q \geq 1$ and a distribution P_X on $\mathcal{X}$, define $\|f\|_{P_X, q} := \big(\mathbb{E}_{P_X}|f(X)|^q\big)^{1/q}$.

Policies. A policy $\pi : \mathcal{X} \to \{0,1\}$ prescribes a binary treatment to a unit with covariates X. Let Π denote a class of feasible treatment policies. We assume Π has bounded complexity, controlled through the VC dimension (Vapnik, 2013).

Assumption 2.2 (Policy class complexity)**.** The policy class Π has finite VC dimension $d < \infty$.

The in-class optimal policy is $\pi^* \in \arg\max_{\pi \in \Pi} V(\pi)$, where $V(\pi) := \mathbb{E}_{P_X}[Y(\pi(X))]$ denotes the expected welfare/reward under policy π. We also consider the *Bayes policy*, which maximizes value over all measurable binary policies: $\pi^B(x) := \arg\max_{d\in\{0,1\}} \mathbb{E}\big[Y(d) \mid X = x\big] = \mathbb{1}\{\tau_o(x) \geq 0\}$.

Policy value. For binary policies, we can rewrite the policy value $\mathbb{E}_{P_X}[Y(\pi(X))]$ as

$$V(\pi) := \mathbb{E}\Big[\pi(X)\,Y(1) + \big(1-\pi(X)\big)\,Y(0)\Big] = \mathbb{E}\Big[\pi(X)\tfrac{YD}{p_o(X)} + (1-\pi(X))\tfrac{Y(1-D)}{1-p_o(X)}\Big].$$

Likewise, we define the conditional value for policy π at realized covariates x as

$$v(\pi,x) := \mathbb{E}\Big[\pi(X)\,Y(1) + \big(1-\pi(X)\big)\,Y(0) \,\Big|\, X = x\Big] = \pi(x)\,g_o(1,x) + \big(1-\pi(x)\big)\,g_o(0,x). \quad (1)$$

Let $\mathbb{E}_n$ denote the expectation with respect to the empirical sample distribution. We define the inverse propensity weighted (IPW) analogue of Equation (1) as

$$V_n(\pi) \;:=\; \mathbb{E}_n\Big[\pi(X)\,\tfrac{YD}{p_o(X)} + \big(1-\pi(X)\big)\,\tfrac{Y(1-D)}{1-p_o(X)}\Big].$$

Additionally, we define the normalized IPW contribution for a policy π at an observation $z = (x,d,y)$:

$$f_\pi(x,d,y) := \kappa\bigg(\pi(x)\,\frac{yd}{p_o(x)} + \big(1-\pi(x)\big)\frac{y(1-d)}{1-p_o(x)}\bigg).$$

This quantity is the per-sample term whose average recovers the policy value up to the factor κ; the extra κ factor ensures the function is uniformly bounded. We use $f(z)$ and $f(x,d,y)$ interchangeably. By $Y \in [0,1]$ and $p_o(X) \in [\kappa, 1-\kappa]$, we have $f_\pi(Z) \in [0,1]$ a.s., and $\mathbb{E}\big[f_\pi(Z)\big] = \kappa\,V(\pi)$. As a notational shorthand, write $\mathbb{E}[f_\pi]$ and $\mathbb{E}_n[f_\pi]$ for $\mathbb{E}[f_\pi(Z)]$ and $\mathbb{E}_n[f_\pi(Z)]$, respectively. We use this notation throughout. In particular, when we write $\mathbb{E}_n\,|f_{\hat\pi} - f_\pi|$, we mean the empirical average of pointwise differences between the normalized IPW scores: $\frac{1}{n}\sum_{i=1}^n |\, f_{\hat\pi}(x_i,d_i,y_i) - f_\pi(x_i,d_i,y_i)\,|$.

3 POLICY LEARNING WITH ABSTENTION

We now present our framework for *policy learning with abstention.* In this framework, the learner can choose to defer on recommending a treatment to any unit based on their observed covariates X. For instance, the learner may not want to prescribe treatment when the estimated CATE for a unit is small. When the learner abstains, they receive a small, additive reward over the value of a random guess.

More formally, an *abstaining policy* is a mapping $\pi : \mathcal{X} \to \{0,1,*\}$, where $\pi(X) = *$ denotes the deferral of treatment decision. When the learner abstains, they receive reward $\frac{Y(1)+Y(0)}{2} + p$, where $p \geq 0$ is some fixed bonus. We define the expected reward mapping under abstention as

$$g_o(*,x) := \mathbb{E}\Big[\tfrac{Y(1)+Y(0)}{2} + p \,\Big|\, X = x\Big].$$

Likewise, we define the expected reward/welfare $V^{(p)}(\pi)$ of an abstaining policy with bonus p (and its empirical analogue $V_n^{(p)}(\pi)$) as

$$V^{(p)}(\pi) := \mathbb{E}\Big[\mathbb{1}\{\pi(X) \neq *\}\; v(\pi,X) \;+\; \mathbb{1}\{\pi(X) = *\}\; g_o(*,X)\Big],$$

$$V_n^{(p)}(\pi) := \mathbb{E}_n\Big[\pi(X)\,\tfrac{YD}{p_o(X)} + \big(1-\pi(X)\big)\,\tfrac{Y(1-D)}{1-p_o(X)} + \mathbb{1}\{\pi(X) = *\}\,\Big(\tfrac{YD}{2p_o(X)} + \tfrac{Y(1-D)}{2(1-p_o(X))} + p\Big)\Big].$$

Note that we have $V^{(p)}(\pi) = V(\pi)$ and $V_n^{(p)}(\pi) = V_n(\pi)$ for any policy π that does not abstain.

The goal of the learner is to use either experimental (Section 3.1) or observational (Section 3.2) data to learn an abstaining policy with small *abstaining regret*, which is defined with respect to a binary policy class Π as

$$\mathrm{Reg}_n^{(p)}(\pi) := V(\pi^*) - V^{(p)}(\pi), \quad (2)$$

where again π^* is the in-class optimal policy and π is some potentially abstaining policy. We define the *classical regret* just as $\mathrm{Reg}_n(\pi) := \mathrm{Reg}_n^{(0)}(\pi)$. The learner is constrained to returning a policy $\widetilde\pi$ that aligns with some $\pi \in \Pi$ when it does not abstain, i.e. $\widetilde\pi(x) = \pi(x)$ when $\widetilde\pi(x) \neq *$.

3.1 Known Propensities

We first describe an algorithm for learning an abstaining policy when the treatment assignment mechanism (i.e. propensity) for the data is known. Our algorithm (Algorithm 1) works by first computing an empirical welfare maximizer $\widehat\pi$ from Π using half of the data (where welfare is determined with respect to the IPW empirical value, V_n). Then, it computes a set of "near-optimal" policies whose empirical welfare is close to that of the empirical risk minimizer, modifying these policies to abstain precisely when they disagree $\widehat{\pi}$. Finally, the algorithm returns the empirical abstaining welfare maximizer in this restricted set of policies on the second half of the data. The following theorem shows that Algorithm 1 obtains fast abstaining regret rates, per the formulation of regret in Equation (2).

Theorem 3.1. *Fix $p > 0$ and $\delta \in (0,1)$. Under Assumption 2.1 and Assumption 2.2 and the construction in Algorithm 1, with probability at least $1-\delta$,*

$$\mathrm{Reg}_n^{(p)}(\widetilde{\pi}) := V(\pi^*) \;-\; V^{(p)}(\,\widetilde{\pi}\,) \;\lesssim\; \frac{d\,\log\frac{n}{d} \;+\; \log\frac{1}{\delta}}{p\,n\,\kappa^2}.$$

More details and the full proof are provided in Appendix Appendix A.

Remark 3.2. Theorem 3.1 analyzes policy learning with an abstention bonus and establishes an $O(1/n)$ fast rate when performance is measured against the best *binary* policy in the class. This benchmark is standard in the abstention literature (see, *e.g.*, (Bousquet and Zhivotovskiy, 2021, Section 1.2)). In Section 4.3, we will discuss a natural connection between this regret bound and distributional robustness.

We now provide some intuition for the proof. Typically, to obtain fast regret rates for policy learning, there must exist a *margin* on the CATE, i.e. the existence of some value $h > 0$ such that

$$\mathbb{P}(|\tau_o(X)| \geq h) = 1. \tag{3}$$

The abstention bonus p can be viewed as a "synthetic" margin — if the learner abstains when $|\tau_o(X)| < p$, they automatically accumulate higher reward than any binary treatment assignment policy. Likewise, when $|\tau_o(X)| \geq p$, the learner will be more certain in their decisions, and hence unlikely to abstain. While this intuition is just heuristic, we provide a fully rigorous proof in Appendix A. One important thing to note is that, in Algorithm 1, we never actually need to estimate the CATE. This is of particular importance when the CATE may be a highly complicated function and is thus difficult to capture using statistical learning methods.

When the additive bonus is $p = 0$, our algorithm can no longer be expected to obtain fast $O(1/n)$ rates. The following proposition shows that, in this undiscounted setting, Algorithm 1 still obtains $O(1/\sqrt{n})$ regret rates outlined, which are known to be generally unimprovable without margin or realizability assumption (Athey and Wager, 2021). In this setting, we can convert an abstaining policy into a binary one by having it assign a treatment uniformly at random in the set $\{0,1\}$. The following proposition provides this slower regret rate, showing there is no added risk

Algorithm 1 Policy Learning with Abstention

Input: Samples $\{(X_i, D_i, Y_i)\}_{i=1}^n$, policy class Π, overlap κ, confidence $\delta \in (0,1)$, bonus p, VC dimension d
Set $\alpha \leftarrow \sqrt{\frac{d\log\frac{n}{d}+\log\frac{1}{\delta}}{n}}$.
Split: Partition the samples into two sets of size $n/2$: $\mathcal{D}_1, \mathcal{D}_2$.
EWM: $\widehat{\pi} = \arg\max_{\pi\in\Pi}\; V_n(\pi)$ (computed on $\mathcal{D}_1$).
Select Near-optimal policies:
$\widehat{\Pi} \leftarrow \{\pi \in \Pi :$
$\quad V_n(\widehat{\pi}) - V_n(\pi) \leq \frac{c}{\kappa}\Big(\alpha^2 + \alpha\sqrt{\mathbb{E}_n\,|f_{\widehat{\pi}} - f_\pi|}\Big)\Big\}$.
Abstention projection: For each $\pi \in \widehat{\Pi}$, define
$\pi'(X) = \begin{cases}\pi(X), & \text{if } \pi(X) = \widehat{\pi}(X),\\ *, & \text{otherwise,}\end{cases}$
$\widetilde{\Pi} \leftarrow \{\,\pi' : \pi \in \widehat{\Pi}\,\}$.
EWM with abstention: $\widetilde{\pi} \in \arg\max_{\pi\in\widetilde{\Pi}}\; V_n^{(p)}(\pi)$ (evaluate $V_n^{(p)}$ on $\mathcal{D}_2$).
Return $\widetilde{\pi}$.

of using our algorithm over just empirical welfare maximization even in setting without the additive bonus. We prove this result in Appendix A.

Proposition 3.3 (ERM benchmark at $p = 0$)**.** *Let π_p be the output of Algorithm 1 for any fixed $p \in (0,1)$. Consider its value under p=0, denoted $V^0(\cdot)$. Then, with probability at least $1-\delta$,*

$$\mathrm{Reg}_n(\pi_p) := V(\pi^*) - V^0(\pi_p) \lesssim \frac{1}{\kappa}\sqrt{\frac{d\,\log\frac{n}{d} \;+\; \log\frac{1}{\delta}}{n}}.$$

The proof is provided in Appendix Appendix A.

Remark 3.4. Proposition 3.3 holds for the output of Algorithm 1 run with *any* fixed bonus $p > 0$. This is because the first stage of Algorithm 1 selects a set of near-optimal policies, each with regret $\widetilde{O}(1/\sqrt{n})$. For any two policies π_1, π_2 in this near-optimal set, the abstaining policy π' that follows them where they agree and abstains on their disagreement set has, at bonus $p = 0$, value $V^0(\pi') = \frac{1}{2}(V(\pi_1) + V(\pi_2))$ on the abstention region. Thus π' also remains within $\widetilde{O}(1/\sqrt{n})$ of $V(\pi^*)$, independently of the choice of p used during the second stage. In particular, a larger p may produce policies that abstain more but this does *not* degrade the worst-case $O(1/\sqrt{n})$ classical regret guarantee.

3.2 Unknown Propensities: Doubly Robust Learner

In observational settings, where either the treatment policy is unknown or choice of treatment is endogenous, we can no longer directly run Algorithm 1. One might think to salvage the algorithm by replacing the propensity $p_o(X)$ in the definition of V_n with an ML estimate $\widehat{p}(X)$. However, unless $\widehat{p}$ converges to p_o in probability at fast, parametric rates[1], Algorithm 1 will exhibit sub-optimal regret. Instead, one must leverage more sophisticated methods to still enable low-regret learning.

Taking inspiration from the literature on semiparametric estimation (Chernozhukov et al., 2018; Bang and Robins, 2005; Foster and Syrgkanis, 2023; Athey and Wager, 2021), we introduce a *doubly-robust* analogue of Algorithm 1 that furnishes fast regret rates even when the propensity is unknown. Our algorithm uses ML estimates $\widehat{g}, \widehat{p}$ of the regression $g_o(d, x)$ and propensity $p_o(x)$ to "de-bias" observed outcomes Y into more robust pseudo-outcomes. These nuisance estimates are assumed to be independent of the sample. In more detail, we define

$$\widehat{\varphi}(x, d, y) := \widehat{g}(d, x) + \left(\frac{d \cdot D}{\widehat{p}(x)} + \frac{(1-d)(1-D)}{1-\widehat{p}(x)}\right)(y - \widehat{g}(d, x))$$

This quantity is known as pseudo-outcomes because when $\widehat{g} = g_o$ and $\widehat{p} = p_o$, one can check that $\mathbb{E}[\widehat{\varphi}(X, d, Y) \mid X] = \mathbb{E}[Y(d) \mid X]$. With pseudo-outcomes, we can then define the doubly-robust abstaining welfare/value on the sample, $V^{(p)}_{n,\mathrm{DR}}$, via

$$V^{(p)}_{n,\mathrm{DR}} := \mathbb{E}_n\Big[\mathbb{1}_{\pi(X)\neq *}\Big\{\pi(X)\widehat{\varphi}(X,1,Y) + (1-\pi(X))\widehat{\varphi}(X,0,Y)\Big\} + \mathbb{1}_{\pi(X)=*}\Big(\tfrac{\hat{\varphi}(X,1,Y)+\hat{\varphi}(X,0,Y)}{2} + p\Big)\Big]$$

When $p = 0$, we arrive at the non-abstaining sample welfare $V_{n,\mathrm{DR}} := V^{(p)}_{n,\mathrm{DR}}$. To obtain a doubly-robust analogue of Algorithm 1, use an expanded α [2] and simply replace every occurrence of V_n, $V^{(p)}_n$ by $V_{n,\mathrm{DR}}$ and $V^{(p)}_{n,\mathrm{DR}}$, respectively. See Algorithm 4 in Appendix B.

We now state the main theorem of this subsection.

Theorem 3.5. *For any $p > 0$, suppose Algorithm 1 is run with $V^{(p)}_{n,\mathrm{DR}}, V_{n,\mathrm{DR}}$ in place of $V^{(p)}_n, V_n$ and expanded α. Further, suppose Assumptions 2.1 and 2.2 hold and that the learner is given nuisance estimates $\widehat{g}, \widehat{p}$ that are independent of the sample. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{Reg}^{(p)}_n(\widetilde{\pi}) \lesssim \frac{d \log \frac{n}{d} + \log \frac{1}{\delta}}{p\, n\, \kappa^2} + \frac{\mathrm{Err}^2_{\mathrm{DR}}}{p\, \kappa^2},$$

where $\mathrm{Err}_{\mathrm{DR}}$ is a known upper bound on the product error given by:

$$\mathbb{E}\left[(\hat{p}(X) - p_o(X))^2 \sum_{d=0}^{1} (\hat{g}(d,X) - g_o(d,X))^2\right]^{1/2}.$$

See Appendix Appendix B for the full proof and supporting lemmas.

The above theorem yields a bound on regret that has two terms — one quickly decaying term followed by another, random term whose decay is non-obvious. We now discuss conditions under which this second term is negligible. Note that one can obtain (via Cauchy-Schwarz) the following upper bound on the product error in terms of individual error rates in nuisance functions,

$$\|\hat{p} - p\|_{P_X,4} \sum_{d\in\{0,1\}} \|\, \hat{g}(d,\cdot) - g_o(d,\cdot)\|_{P_X,4}$$

Thus, for the second term in Theorem 3.5 to be negligible, we need $\mathrm{Err}_{\mathrm{DR}} \le c_\delta/\sqrt{n}$ with high probability, which will occur if

$$\max\Big\{\|\hat{p} - p\|_{P_X,4},\ \|\hat{g}(0,\cdot) - g_o(0,\cdot)\|_{P_X,4},\ \|\hat{g}(1,\cdot) - g_o(1,\cdot)\|_{P_X,4}\Big\} \lesssim n^{-1/4}$$

with probability at least $1-\delta$. In particular, this can be accomplished under a variety of learnability assumptions (finite VC-dimension, learnability by trees, etc.)

Remark 3.6. In the statement of Theorem 3.5, we assume that the nuisance estimates $\widehat{g}$ and $\widehat{p}$ are independent of the sample. In practice, this could be accomplished by actually performing a three-fold split of the data, reserving the third fold for nuisance estimation and using the first two folds as outlined in Algorithm 1. Another more sophisticated approach would be to use $\mathcal{D}_2$ to produce nuisance estimates $\widehat{g}_1, \widehat{p}_1$ to use in defining $V_{n,\mathrm{DR}}$ (which is defined in terms of samples in $\mathcal{D}_1$), and analogously use $\mathcal{D}_1$ to build estimates $\widehat{g}_2, \widehat{p}_2$ used in definition $V^{(p)}_{n,\mathrm{DR}}$.

4 APPLICATIONS

In the previous section, we established a formal framework for performing policy learning with abstention. We now explore connections between our results in Section 3 (in particular, Algorithm 1) and other aspects of policy learning. This includes developing (non-abstaining) algorithms for policy learning without standard margin assumptions, safely improving policies relative some baseline treatment strategy, and relating abstaining value/welfare $V^{(p)}$ to policy learning in the presence of distribution shift.

[1] In particular, one would need $\|p_o - \widehat{p}\|_{P_X,2} = O_{\mathbb{P}}(n^{-1/2})$

[2] We may simply increase the constant in Step 5 of Algorithm 4; when the nuisance product error is $o_p(n^{-1/2})$, this yields the same rate as Theorem 3.1.

4.1 Fast Learning Rates Without Standard Margin Assumptions

In policy learning, fast regret rates are sometimes achievable when the CATE is deterministically bounded away from zero, as outlined in the margin condition of Equation (3). In particular, prior works have shown that $O(n^{-1})$ rates are possible in the *realizable* setting, where the Bayes-optimal policy π^B lies in the class Π (Kitagawa and Tetenov, 2018; Luedtke and Chambaz, 2020). However, it is generally impossible to know if a policy class Π (say, all depth 3 decision trees) contains the optimal treatment policy in advance. Thus, one should aim to obtain fast regret rates in *agnostic* settings, where Π may not contain the optimal treatment strategy.

In the classification literature, recent works achieve fast learning rates in agnostic settings (Ben-David and Urner, 2014; Bousquet and Zhivotovskiy, 2021). These works eschew the realizability assumption, instead imposing a *finite combinatorial diameter* on the policy class. In words, a class of policies Π which bounds the maximal number of points of disagreement between any two policies in the class. More formally, the combinatorial diameter is defined by

$$D := \max_{\pi_1,\pi_2\in\Pi} \sum_{x\in\mathcal{X}} \mathbb{1}\{\pi_1(x) \neq \pi_2(x)\}\,.$$

Note that D may be infinite even for classes with finite VC dimension.

We prove an analogous result for policy learning (Algorithm 4.1) under the assumption of a finite combinatorial diameter. Further, we consider settings where the policy class does not have a finite combinatorial diameter but instead we have access to a CATE estimation oracle (Algorithm 4.2). The proofs are deferred to Appendix C. We start by stating our main theorem for agnostic policy learning under an assumption of finite combinatorial dimension.

Theorem 4.1. *Assume the margin condition $\mathbb{P}(|\tau_o(X)| \geq h) = 1$ and that Assumptions 2.1 and 2.2 hold. Further, assume Π has finite combinatorial diameter D. Then the output π_{final} of Algorithm 2 satisfies*

$$\mathrm{Reg}_n(\pi_{final}) := V(\pi^*) - V(\pi_{final}) \lesssim \frac{D \;+\; d\log\frac{n}{d} \;+\; \log\frac{1}{\delta}}{\kappa^2\, h\, n}$$

with probability at least $1-\delta$.

Proof details are provided in Appendix Appendix C.

Next, suppose we have an oracle for estimator $\widehat{\tau}_o$ that satisfies for any $(X, D, Y) \sim \mathbb{P}$ the following convergence rate with probability greater than $1-\delta$,

$$\|\widehat{\tau} - \tau\|_{\mathbb{P},2} \leq c_\delta n^{-\beta}.$$

Algorithm 2 Policy Learning under Margin Assumption

Input: Samples $\{(X_i, D_i, Y_i)\}_{i=1}^n$, policy class Π, overlap κ, margin $h > 0$, confidence δ, Mode$\in$ $\{\mathsf{FiniteD}, \mathsf{CATE\text{-}Oracle}\}$.
Split: Partition the sample into three equal parts: $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ (sizes $n/3$).
Abstention stage: On $\mathcal{D}_1 \cup \mathcal{D}_2$, run Algorithm 1 with bonus $p = h/2$ to obtain a policy $\widetilde{\pi}$ that may abstain. Let $\mathcal{X}_{rem} \subseteq \mathcal{X}$ be the covariate set where $\widetilde{\pi}$ abstains.
if Mode=FiniteD **then**
  Refine on $\mathcal{X}_{rem}$ (finite D): On $\mathcal{D}_3$, run EWM over all the policies on $\mathcal{X}_{rem}$.
  $\phi \leftarrow \arg\max \mathbb{E}_n[v(\pi, X) | X \in \mathcal{X}_{rem}]$.
else if Mode=CATE-Oracle **then**
  Refine on $\mathcal{X}_{rem}$ (CATE oracle): On $\mathcal{D}_3$, estimate $\widehat{\tau}$ on $\mathcal{X}_{rem}$ and set $\phi(x) \leftarrow \mathbb{1}\{\widehat{\tau}(x) > 0\}$ for $x \in \mathcal{X}_{rem}$.
end if
Define: $\pi_{\text{final}}(x) = \begin{cases} \widetilde{\pi}(x), & x \notin \mathcal{X}_{rem}, \\ \phi(x), & x \in \mathcal{X}_{rem} \end{cases}$
Output: π_{final}.

Theorem 4.2. *Assume the margin condition $\mathbb{P}(|\tau_o(X)| \geq h) = 1$ and that Assumptions 2.1 and 2.2 hold. Further, suppose $\widehat{\tau}$ satisfies the regression-oracle condition with exponent $\beta > 0$ and let π^B be the Bayes optimal policy. Then the output π_{final} of Algorithm 2 satisfies*

$$V(\pi^*) - V(\pi_{final}) \lesssim \mathrm{Reg}_1 + \mathrm{Reg}_2 + \mathrm{Reg}_3$$

where,

$$\mathrm{Reg}_1 \lesssim n^{-1}\,\frac{d\log\frac{n}{d} \;+\; \log\frac{1}{\delta}}{\kappa^2\, h},$$

$$\mathrm{Reg}_2 \lesssim n^{-2\beta}\frac{c_\delta\left(V(\pi^B(X)) - V(\pi^*(X)\right)^{1-2\beta}}{h^{2-2\beta}},$$

$$\mathrm{Reg}_3 \lesssim n^{-\frac{1}{2}-\beta}\,\frac{\left(d\log\frac{n}{d} + \log\frac{1}{\delta}\right)^{\frac{1}{2}-\beta}}{\kappa^{1-2\beta}h^{2-2\beta}}.$$

with probability at least $1-\delta$.

Proof details are provided in Appendix Appendix C.

Theorem 4.2 unifies the regression-rate guarantee (Luedtke and Chambaz, 2020) and fast rate in the realizable case (Kitagawa and Tetenov, 2018). In particular, when $\beta = 1/2$, one recovers n^{-1} regret. More generally, if $\beta < 1/2$ but the policy class Π is expressive enough that $V(\pi^B(X)) - V(\pi^*(X))$ is small (e.g., $\leq c/n^{-\alpha}$ for some $\alpha > 0$), the resulting bound improves upon a direct plug-in classifier

Algorithm 3 Safe Policy Learning with Abstention

Select a sequence of abstention bonuses $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$, with $p_1 < \ldots < p_k$.
Split the dataset into D_{train} and D_{test} of size n_{train} and n_{test} respectively.
for $p \in \mathcal{P}$ **do**
  Run Algorithm 1 for $V^{(p)}(\pi)$ on D_{train} to obtain an abstaining policy $\widetilde{\pi}$
  Replace abstention with the baseline policy

$$\hat{\pi}(x) \;=\; \begin{cases} \omega(x), & \tilde{\pi}(x) = *, \\ \tilde{\pi}(x), & \text{otherwise.} \end{cases} \qquad (4)$$

  `Hypothesis test:`
  Estimate $V_n(\widehat{\pi}) - V_n(\omega)$ on D_{test}.
  Compute the one-sided $(1-\delta)$ lower confidence bound: $\text{LCB} = V_n(\widehat{\pi}) - V_n(\omega) - z_{1-\frac{\delta}{k}} \frac{\hat{\sigma}_k}{\sqrt{n_{\text{test}}}}$.
  if LCB > 0 **then return** $\hat{\pi}$ **else** continue.
end for

($\pi(X) = \mathbb{1}\{\hat{\tau}(X) > 0\}$). Thus, either a strong CATE oracle or an "almost" realizable policy class is sufficient to surpass the $1/\sqrt{n}$ regret.

4.2 Safe Policy Improvement

Safe policy improvement (SPI) addresses the deployment problem under uncertainty: one should adopt a new policy only if it can be certified to outperform a baseline policy ω with high probability. Following recent work on safe policy improvement Thomas et al. (2015); Cho et al. (2025), we use sample splitting (into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$) to separate the tasks of *candidate policy selection* and *safety testing*. This splitting enables one-sided lower confidence bounds (LCBs) on the improvement over the baseline: $V(\hat{\pi}) - V(\omega)$. We treat Algorithm 1 as a black box to propose a $\{0, 1, *\}$-valued policy $\widetilde{\pi}$ learned on $\mathcal{D}_{train}$. To obtain a deployable binary policy, we impute abstentions with the baseline policy to obtain a candidate policy Equation (4).

On D_{test}, we estimate $V(\hat{\pi}) - V(\omega)$ via IPW/DR and compute a one-sided LCB at level $1-\delta$; we generate a sequence of candidate policies by running the abstention learner over a grid of bonuses, $\mathcal{P} = \{p_1 < \cdots < p_k\}$, which trades off overlap (more abstention) against improvement (more overrides), and apply a Bonferroni adjustment with $z_{1-\delta/k} = \Phi^{-1}(1-\delta/k)$. Algorithm 3 then: (i) selects the grid, (ii) learns abstaining policies on D_{train}, (iii) replaces abstentions with ω, and (iv) tests each on D_{test}, returning the first with LCB > 0.

Safe Policy Improvement Experiments We compare Algorithm 3 with an Empirical Welfare Maximizer (EWM) and prior work on synthetic data. In particular, we consider the two variants of High-Confidence Policy Improvement (HCPI) described in Thomas et al. (2015): one based on finite-sample confidence intervals and one based on a t-test statistic. For EWM, we choose the empirical welfare maximizer as the candidate policy and test it against the baseline in exactly the same way as in Algorithm 3 (with $k = 1$).

The design comprises two variants: (1) we use a baseline policy at a fixed baseline–optimal gap and vary the noise variance (0.01–1.0) (see Figure 1); and (2) we vary the baseline–optimal value gap at a fixed noise level (see Figure 2). Each parameter setting is replicated 100 times; policy value is computed by IPW (with known propensities). For Algo 3, we use a grid of abstention bonuses $\mathcal{P} = \{0, 0.01, 0.05, 0.10, 0.20\}$. We report value gain, mistake rate (probability of returning a policy worse than the baseline), and improvement rate (probability of returning a policy strictly better than the baseline). Full configuration and implementation details are deferred to Appendix D.

We evaluate all methods at significance level $\delta = 0.05$. Across more than 2,000 repetitions, our Safe Policy Learner controls the Type-I error at or below 0.05. In the very small–sample regime ($n \leq 500$), EWM performs better than the alternatives, but once $n \geq 1000$ our method dominates: *(i)* it achieves the highest improvement rate and the largest mean value gain among accepted policies across baseline–optimal gaps (Figure 2), and *(ii)* it maintains both higher power and competitive mistake rates as noise increases (Figure 1). Overall, for moderate-to-large samples, Algorithm 3 delivers the best safety–power tradeoff, improving more often and by more, while meeting the δ-level error control.

4.3 Robustness to Distribution Shifts

The abstention value $V^{(p)}$ has a natural interpretation as protection against outcome distribution shift. Specifically, consider a setting where the true data distribution of $(Y(0), Y(1)) \mid X$ is different from our training data. This would happen when our observational data is outdated and cannot reflect the true effect of the current treatment of interest.

While deterministic policies are optimal without such an outcome distribution shift, they can be problematic otherwise. To mathematically formulate this intuition, we assume the true potential outcome distribution lies in some W_1-ball of the distribution that generates our observations, *i.e.*

$$\mathbb{P}_{\texttt{test}} \in \mathcal{P}_\alpha(\mathbb{P}_{\texttt{train}}) \;:=\; \Big\{\mathbb{P} :\; W_1(\mathbb{P}, \mathbb{P}_{\texttt{train}}) \leq \alpha\Big\}.$$

The robust objective is then to maximize a policy's

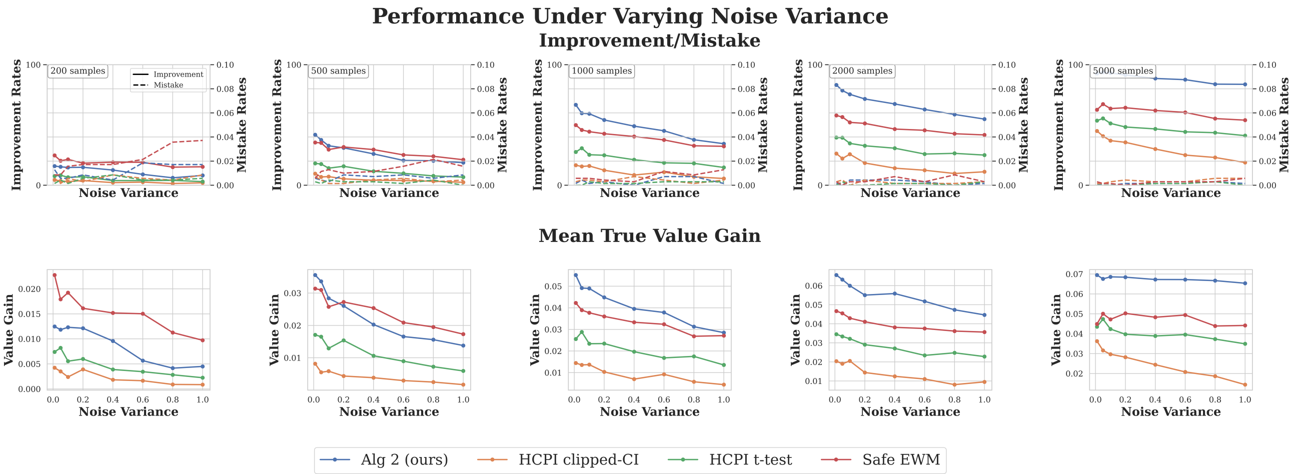

Figure 1: Safe Policy Improvement — performance across varying noise variance. We compare Algorithm 2 (ours), HCPI (two variants), and Safe EWM.

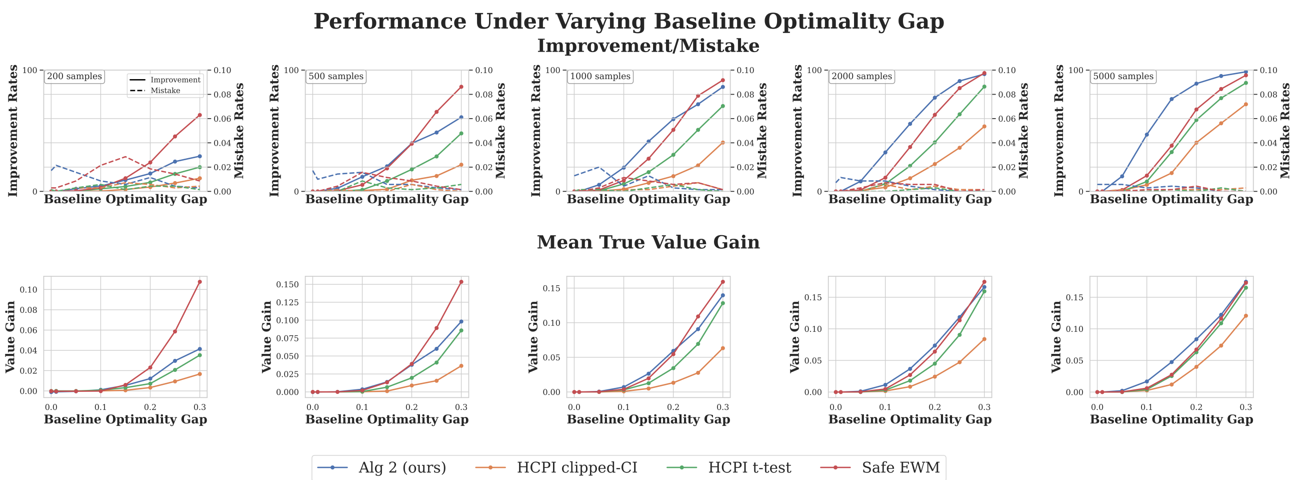

Figure 2: Safe Policy Improvement — performance across varying gap between baseline and optimal policy value. We compare Algorithm 3 (ours), HCPI (two variants), and Safe EWM.

worst–case value over $\mathcal{P}_\alpha(\mathbb{P}_{\texttt{train}})$. Specifically, consider policies π taking values in $\{0, 1, \frac{1}{2}\}$, where $\frac{1}{2}$ denotes a uniform random choice between the two treatments. In this model, the robust objective coincides with maximizing the abstention value $V^{(p)}$ for a specific bonus level, as formalized below.

Proposition 4.3. *(i). For any (possibly random) policy $\pi : \mathcal{X} \mapsto [0, 1]$, there exists $\hat{\pi} : \mathcal{X} \mapsto \{0, \frac{1}{2}, 1\}$ such that $\pi \in \{0, 1\} \Rightarrow \pi = \hat{\pi}$ and $\min_{\mathbb{P}\in\mathcal{P}_\alpha} V(\hat{\pi}) > \min_{\mathbb{P}\in\mathcal{P}_\alpha} V(\pi)$.*
(ii). Let $\pi : \mathcal{X} \mapsto \{0, \frac{1}{2}, 1\}$ and $\tilde{\pi} : \mathcal{X} \mapsto \{0, 1, *\}$ be obtained by replacing all $\pi(X) = \frac{1}{2}$ with $*$, then we have

$$\min_{\mathbb{P}\in\mathcal{P}_\alpha} V(\pi) = V^{(\alpha/2)}(\pi) - \alpha. \tag{5}$$

Distributional uncertainty acts like an outcome–level penalty of size α when the policy commits to an arm, and a half–penalty under randomization; hence, maximizing the robust objective is equivalent to maximizing $V^{(p)}$ with $p = \alpha/2$.

5 CONCLUSION

In this paper, we introduced a framework for policy learning with abstention. Building off of the classification with abstention literature, we formulated a value/welfare that incentivizes abstention by offering a small, additive reward on top of the value of a random guess in regions of uncertainty. We described algorithms that obtain fast abstaining regret rates in both known and unknown propensity settings, and showed how to apply our algorithm to various disparate policy learning problems. There are still many other interesting open directions related to policy learning with abstention. First, our framework only handles binary treatments. There are many settings (such as in medicine) where designing policies to optimize a continuous treatment (or dosage) may be more appropriate. Second, implementing Algorithm 1 efficiently for popular policy classes such as decision trees is also an interesting open direction. Likewise, we evaluate the performance of algorithms through regret, but one could also consider loss-based modes of policy evaluation as well. Finally, a complementary line of work applies pessimism and generalization bounds to ensure safe deployment of learned policies (Sakhi et al., 2023; Gabbianelli et al., 2024; Sakhi et al., 2024). These pessimistic/penalized objectives can, at a high level, replace the EWM step in our Algorithm 1 and may alleviate the need for the uniform overlap assumption by requiring only coverage of the optimal action. However, the main technical contribution of our work is not the EWM step itself but the construction of the near-optimal set and the abstention projection (Steps 6–9), which relies on localized Rademacher analysis for bounded IPW/DR scores. Combining pessimistic objectives with the abstention projection is a promising direction for future work.

Acknowledgements

Vasilis Syrgkanis is supported by NSF Award IIS-2337916. Ayush Sawarni is partially supported by NSF Award IIS-2337916.

References

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

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability theory and related fields*, 135(3):311–334.

Bartlett, P. L. and Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840.

Ben-David, S. and Urner, R. (2014). The sample complexity of agnostic learning under deterministic labels. In *Conference on Learning Theory*, pages 527–542. PMLR.

Bousquet, O. and Zhivotovskiy, N. (2021). Fast classification rates without standard margin assumptions. *Information and Inference: A Journal of the IMA*, 10(4):1389–1421.

Chen, Q., Austern, M., and Syrgkanis, V. (2023). Inference on optimal dynamic policies via softmax approximation. *arXiv preprint arXiv:2303.04416*.

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

Cho, B., Pop, A.-R., Gan, K., Corbett-Davies, S., Nir, I., Evnine, A., and Kallus, N. (2025). Cspi-mt: Calibrated safe policy improvement with multiple testing for threshold policies. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 165–176.

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

Cortes, C., DeSalvo, G., Gentile, C., Mohri, M., and Yang, S. (2018). Online learning with abstention. In *International conference on machine learning*, pages 1059–1067. PMLR.

Cortes, C., DeSalvo, G., and Mohri, M. (2016). Boosting with abstention. *Advances in neural information processing systems*, 29.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921.

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

Foster, D. J. and Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908.

Gabbianelli, G., Neu, G., and Papini, M. (2024). Importance-weighted offline learning done right. In *Proceedings of the 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 614–634. PMLR.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*.

Geifman, Y. and El-Yaniv, R. (2019). Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR.

Ghavamzadeh, M., Petrik, M., and Chow, Y. (2016). Safe policy improvement by minimizing robust baseline regret. *Advances in Neural Information Processing Systems*, 29.

Grandvalet, Y., Rakotomamonjy, A., Keshet, J., and Canu, S. (2009). Support vector machines with a reject option. In *Advances in Neural Information Processing Systems 21 (NeurIPS 2008)*, pages 537–544. Conference held in 2008; published in NeurIPS 21.

Guntuboyina, A. and Sen, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1):379–411.

Herbei, R. and Wegkamp, M. H. (2006). Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4):709–721.

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

Kallus, N. and Zhou, A. (2018). Confounding-robust policy improvement. *Advances in neural information processing systems*, 31.

Karampatziakis, N., Mineiro, P., and Ramdas, A. (2021). Off-policy confidence sequences. In *International Conference on Machine Learning*, pages 5301–5310. PMLR.

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

Laroche, R., Trichelair, P., and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International conference on machine learning*, pages 3652–3661. PMLR.

Lepskii, O. (1992). Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697.

Luedtke, A. and Chambaz, A. (2020). Performance guarantees for policy learning. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(3):2162–2188.

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

Madras, D., Pitassi, T., and Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 6147–6157.

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

Moodie, E. E., Chakraborty, B., and Kramer, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics*, 40(4):629–645.

Moodie, E. E., Richardson, T. S., and Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455.

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

Neu, G. and Zhivotovskiy, N. (2020). Fast rates for online prediction with abstention. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 3030–3048. PMLR.

Pasteris, S., Rumi, A., Thiessen, M., Saito, S., Miyauchi, A., Vitale, F., and Herbster, M. (2024). Bandits with abstention under expert advice. *Advances in Neural Information Processing Systems*, 37:126059–126087.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2015). Consistent algorithms for multiclass classification with a reject option. *arXiv preprint arXiv:1505.04137*.

Sakhi, O., Aouali, I., Alquier, P., and Chopin, N. (2024). Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 80706–80755. Curran Associates, Inc.

Sakhi, O., Chopin, N., and Alquier, P. (2023). PAC-Bayesian offline contextual bandits with guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29777–29799. PMLR.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Simão, T. D., Laroche, R., and Combes, R. T. d. (2019). Safe policy improvement with an estimated baseline policy. *arXiv preprint arXiv:1909.05236*.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. (2015). High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388. PMLR.

van der Hoeven, D., Koolen, W. M., Cesa-Bianchi, N., and Zhivotovskiy, N. (2022). A regret–variance trade-off in online learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vidyasagar, M. (2013). *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media.

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

Whitehouse, J., Austern, M., and Syrgkanis, V. (2025). Inference on optimal policy values and other irregular functionals via smoothing. *arXiv preprint arXiv:2507.11780*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Policy Learning with Abstention: Supplementary Material

A Missing Proofs from Section 3

Additional notation. Let d denote the VC dimension of Π and set

$$\alpha \;:=\; \sqrt{\frac{d\log\frac{n}{d} \;+\; \log\frac{1}{\delta}}{n}}.$$

We use $\mathbf{c}$ to denote absolute constants.

Recall that $\widehat{\pi}$ is the policy selected by the first EWM step (Step 3 in Algorithm 1); $\widehat{\Pi}$ is the set of "almost" optimal policies (Step 6); $\widetilde{\Pi}$ is the set of abstaining policies constructed from $\widehat{\Pi}$ (Step 9); elements of $\widetilde{\Pi}$ are denoted by π', and $\widetilde{\pi}$ is the final abstaining policy returned by Algorithm 1 (Step 10). The class $\widetilde{\Pi}$ is formed by projecting disagreements of policies in $\widehat{\Pi}$ with the fixed reference $\widehat{\pi} \in \Pi$. Notationally, we write ϕ for a generic element of $\widetilde{\Pi}$ and ϕ^* for a maximizer of $V^{(p)}$. For brevity, we also use π' to denote the abstaining policy obtained from a given π via its disagreement with $\widehat{\pi}$; e.g., π_1', π_2' are the abstaining policies constructed from π_1, π_2, respectively.

For $f : \mathcal{X} \to \mathbb{R}$,

$$\|f\|_{P_X,q} \;:=\; \Big(\mathbb{E}_{P_X}\big|f(X)\big|^q\Big)^{1/q}.$$

For $h : \mathcal{X} \times \{0,1\} \times [0,1] \to \mathbb{R}$,

$$\|h\|_{P_Z,q} \;:=\; \Big(\mathbb{E}_{P_Z}\big|h(Z)\big|^q\Big)^{1/q}, \qquad Z = (X, D, Y).$$

On samples $Z_{1:n}$ the empirical L^2 norm is denoted as:

$$\|h\|_{L^2(Z_{1:n})} \;:=\; \Big(\tfrac{1}{n}\sum_{i=1}^{n} h(Z_i)^2\Big)^{1/2},$$

abbreviated $\|h\|_n$ when clear. Finally, note that for for binary policies,

$$\mathbb{E}\big[\mathbb{1}\{\pi_1(X) \neq \pi_2(X)\}\big] \;=\; \|\pi_1 - \pi_2\|_{P_X,1}.$$

VC subgraph dimension. Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. The *VC subgraph dimension* of $\mathcal{F}$ is the VC dimension of the family of subgraphs

$$\mathsf{Subgraph}(\mathcal{F}) \;:=\; \big\{\,\{(x,t) \in \mathcal{X} \times \mathbb{R} :\; t < f(x)\} :\; f \in \mathcal{F}\,\big\}.$$

Equivalently, for any finite set $\{(x_i,t_i)\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$, we say $\mathsf{Subgraph}(\mathcal{F})$ *shatters* this set if for every labeling $b \in \{0,1\}^n$ there exists $f \in \mathcal{F}$ such that

$$\mathbf{1}\{t_i < f(x_i)\} = b_i, \quad i = 1, \dots, n.$$

The VC subgraph dimension $\mathrm{VC}_{\mathrm{sub}}(\mathcal{F})$ is the largest n for which some n-point set in $\mathcal{X} \times \mathbb{R}$ is shattered (or $+\infty$ if no finite maximum exists). For $\{0,1\}$–valued classes, $\mathrm{VC}_{\mathrm{sub}}(\mathcal{F})$ coincides with the usual VC dimension.

First, we restate that a standard result regarding the VC subgraph dimension (see Lemma A.1 of Kitagawa and Tetenov (2018)):

Lemma A.1. *Let Π be a class of binary functions with VC dimension d, and define the real-valued class*

$$\mathcal{F}_\Pi \;=\; \Big\{\, z \mapsto \pi(x)\,\gamma_1(z) \;+\; \big(1 - \pi(x)\big)\,\gamma_2(z) \;:\; \pi \in \Pi \Big\},$$

where γ_1, γ_2 are fixed real-valued functions and x denotes the covariate component of z. Then the VC subgraph dimension of $\mathcal{F}_\Pi$ is at most d.

A.1 Concentration Inequalities

This section uses standard techniques from Empirical Process Theory, in particular we use localized Rademacher complexity to obtain Bernstein-type concentration for the IPW estimator.

Lemma A.2. *Let Π be a class of binary functions with VC dimension d, and define*

$$\mathcal{F}_\Pi \;=\; \Big\{ z \mapsto \pi(x)\,\gamma_1(z) + \big(1-\pi(x)\big)\,\gamma_2(z) \;:\; \pi \in \Pi \Big\},$$

where γ_1, γ_2 are a.s. bounded with $|\gamma_1(Z)|, |\gamma_2(Z)| \le B$. Fix $f_{\pi^*} \in \mathcal{F}_\Pi$. For any $f_\pi \in \mathcal{F}_\Pi$ constructed from π as above, with probability at least $1-\delta$,

$$\big|\, \mathbb{E}[f_{\pi^*}(Z) - f_\pi(Z)] \;-\; \mathbb{E}_n[f_{\pi^*}(Z) - f_\pi(Z)] \,\big| \;\le\; \boldsymbol{c}\Big(\alpha\sqrt{B\,\mathbb{E}|f_{\pi^*} - f_\pi|} + B\alpha^2\Big),$$

where $\alpha \;\coloneqq\; \sqrt{\frac{d\log\frac{n}{d} + \log\frac{1}{\delta}}{n}}$.

Proof. We begin by constructing the star convex function class

$$\mathcal{G} = \left\{ \beta\,\frac{f_{\pi^*}(z) - f_\pi(z)}{2\,B} : f_\pi \in \mathcal{F}_\Pi\,,\; \beta \in [0,1] \right\}.$$

Next, we define the critical radius as

$$r(n,\delta) = \inf\left\{ r \ge 0 : \mathbb{P}\left(\sup_{g\in\mathcal{G}, \mathbb{E}g^2 \le r^2} |(\mathbb{E} - \mathbb{E}_n)g| \le cr^2 \right) \ge 1-\delta \right\}$$

where c is an absolute constant that will be chosen later. In particular, for such a choice of $r(n,\delta)$, with probability at least $1-\delta$ we have

$$|(\mathbb{E} - \mathbb{E}_n)g| \le c\left(r(n,\delta)\sqrt{\mathbb{E}g^2} + r(n,\delta)^2 \right). \tag{6}$$

The inequality Equation (6) holds trivially for any g satisfying $\mathbb{E}g^2 < r(n,\delta)^2$. Now, if $\mathbb{E}g^2 \ge r(n,\delta)^2$, define

$$g' \coloneqq \frac{r(n,\delta)\; g}{\sqrt{\mathbb{E}g^2}},$$

so that $\mathbb{E}g'^2 < r(n,\delta)^2$. Applying the inequality to g' and then scaling back, we obtain

$$|(\mathbb{E} - \mathbb{E}_n)g| \le c\; r(n,\delta)\sqrt{\mathbb{E}g^2},$$

This gives us the bound in Equation (6).

Next, observe that since $|g(Z)| \in [0,1]$, it follows that

$$\mathbb{E}g^2 \le \mathbb{E}\,|g|\,.$$

By a standard symmetrization argument, we obtain the inequality

$$\mathbb{E}\left[\sup_{\mathbb{E}g^2 \le r^2} |(\mathbb{E} - \mathbb{E}_n)g| \right] \le 2\mathbb{E}[\mathcal{R}(Z_{1:n}, \mathcal{G})], \tag{7}$$

where the Empirical Rademacher complexity is defined by

$$\mathcal{R}(Z_{1:n}, \mathcal{G}) \coloneqq \mathbb{E}_{\epsilon_{1:n}}\left[\sup_{\mathbb{E}g^2 \le r^2} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i g(Z_i) \right| \Big| Z_{1:n} \right].$$

We now introduce

$$M = \sup_{\substack{f,g\in\mathcal{G} \\ \mathbb{E}f^2,\mathbb{E}g^2\le r^2}} \|f-g\|_{L^2(Z_{1:n})},$$

with

$$\|f-g\|^2_{L^2(Z_{1:n})} := \frac{\sum_{i=1}^n (f(Z_i)-g(Z_i))^2}{n}.$$

The following version of Dudley's inequality will be used:

$$\mathcal{R}(Z_{1:n},\mathcal{G}) \le 4\gamma + \frac{12}{\sqrt{n}}\int_{\gamma}^{M}\sqrt{\log N\left(\varepsilon,\mathcal{G},\|\cdot\|_{L^2(Z_{1:n})}\right)}d\varepsilon \qquad \forall\gamma>0. \tag{8}$$

To bound the covering numbers, we use the fact that the covering numbers of a star-shaped hull $\mathcal{G}$ of a class $\mathcal{F}$ (of VC dimension d) satisfy (Bartlett and Mendelson, 2006):

$$N\left(\varepsilon,\mathcal{G},\|\cdot\|_{L^2(Z_{1:n})}\right) \le N\left(\varepsilon,\mathcal{F},\|\cdot\|_{L^2(Z_{1:n})}\right)\left(1+\left\lceil\frac{2}{\varepsilon}\right\rceil\right) \le e(d+1)\left(\frac{2e}{\varepsilon}^2\right)^d\left(1+\left\lceil\frac{2}{\varepsilon}\right\rceil\right) \le e(d+1)\left(\frac{2e}{\varepsilon}^2\right)^{d+1}.$$

The second-to-last inequality follows from standard bounds on the covering numbers of VC classes. Substituting $\gamma = d/n$ into Equation (8), we get

$$\begin{aligned}
\mathcal{R}(Z_{1:n},\mathcal{G}) &\le \frac{4d}{n} + \frac{12}{\sqrt{n}}\int_{d/n}^{M}\sqrt{(d+1)\log\frac{2e}{\varepsilon}+\log(d+1)}\ d\varepsilon \\
&\le \frac{4d}{n} + \frac{12}{\sqrt{n}}\cdot\sqrt{(d+1)\log\frac{2en}{d}+\log(d+1)}\ \int_{d/n}^{M} d\varepsilon \\
&\le \frac{\mathbf{c}\,d}{n} + \frac{\mathbf{c}\,M}{\sqrt{n}}\sqrt{d\log\frac{n}{d}}.
\end{aligned} \tag{9}$$

We also have

$$\begin{aligned}
\mathbb{E}M &= \mathbb{E}\sqrt{\frac{\sup_{\mathbb{E}g^2\le r^2}\sum_i g^2(Z_i)}{n}} \\
&\le \sqrt{\frac{\mathbb{E}\sup_{\mathbb{E}g^2\le r^2}\sum_i g^2(Z_i)}{n}} && \text{(Jensen's inequality)} \\
&\le \sqrt{\frac{\mathbb{E}\sup_{\mathbb{E}g^2\le r^2}\sum_i (g^2(Z_i)-\mathbb{E}g^2(Z_i))}{n}+r^2} \\
&\le \sqrt{\frac{2\ \mathbb{E}\sup_{\mathbb{E}g^2\le r^2}\sum_i \epsilon_i g^2(Z_i)}{n}+r^2} && \text{(via Symmetrization)} \\
&\le \sqrt{\frac{2\ \mathbb{E}\sup_{\mathbb{E}g^2\le r^2}\sum_i \epsilon_i g(Z_i)}{n}+r^2} && \text{(via Contraction Lemma (Shalev-Shwartz and Ben-David, 2014))} \\
&\le \sqrt{\frac{2\ \mathbb{E}\sup_{\mathbb{E}g^2\le r^2}\sum_i \epsilon_i g(Z_i)}{n}}+r \\
&= \sqrt{\frac{2\ \mathbb{E}\ \mathcal{R}(Z_{1:n},\mathcal{G})}{n}}+r.
\end{aligned}$$

Taking expectation in Equation (9) we have

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(Z_{1:n},\mathcal{G})] &\le \frac{\mathbf{c}\,d}{n} + \frac{\mathbf{c}\,\mathbb{E}M}{\sqrt{n}}\sqrt{d\log\frac{n}{d}} \\
&\le \mathbf{c}\,r\sqrt{\frac{d\log\frac{n}{d}}{n}} + \mathbf{c}\sqrt{\frac{\mathbb{E}\mathcal{R}(Z_{1:n},\mathcal{G})}{n}}\sqrt{\frac{d\log\frac{n}{d}}{n}} + \frac{\mathbf{c}\,d}{n}
\end{aligned}$$

The inequality above, together with Equation (7), implies that

$$\mathbb{E}\left[\sup_{\mathbb{E}g^2 \le r^2} |(\mathbb{E}-\mathbb{E}_n)g|\right] \le 2\ \mathbb{E}[\mathcal{R}(Z_{1:n},\mathcal{G})] \le \mathbf{c}\left(r\sqrt{\frac{d\log\frac{n}{d}}{n}} + \frac{d\log\frac{n}{d}}{n}\right)$$

Finally, by applying Talagrand's inequality (Theorem 3.27 of Wainwright (2019)), we obtain

$$\mathbb{P}\left(\sup_{\mathbb{E}g^2 \le r^2} |(\mathbb{E}-\mathbb{E}_n)g| \ge \mathbf{c}\left(r\sqrt{\frac{d\log\frac{n}{d}+\log\frac{1}{\delta}}{n}} + \frac{d\log\frac{n}{d}+\log\frac{1}{\delta}}{n}\right)\right) \le \delta.$$

Now we invoke the Peeling Argument (Wainwright, 2019). For $k = 0, 1, \ldots, K := \lceil \log_2(n) \rceil$, define shells $\mathcal{G}_k = \{g \in \mathcal{G} : 2^{-(k+1)} < \mathbb{E}g^2 \le 2^{-k}\}$. Apply the tail bound to each $\mathcal{G}_k$ with confidence levels $\delta_k = \delta/(K+1)$ and union bound over k. On the resulting $1-\delta$ event we have, simultaneously for all $g \in \mathcal{G}$,

$$|(\mathbb{E}-\mathbb{E}_n)g| \ \le\ \mathbf{c}\left(\alpha\sqrt{\mathbb{E}g^2} + \alpha^2\right), \quad \alpha := \sqrt{\frac{d\log\frac{en}{d}+\log\frac{1}{\delta}}{n}}.$$

Taking $r(n,\delta) \asymp \alpha$ gives the desired localized bound. Finally, since $g = (f_{\pi^*} - f_\pi)/(2B)$ and $\mathbb{E}g^2 = \frac{\mathbb{E}\,(f_{\pi^*}-f_\pi)^2}{(2B)^2} \le \frac{\mathbb{E}|f_{\pi^*}-f_\pi|}{2B}$,

$$\left|(\mathbb{E}-\mathbb{E}_n)\big(f_{\pi^*} - f_\pi\big)\right| \ \le\ \mathbf{c}\left(\alpha\sqrt{B\ \mathbb{E}|f_{\pi^*}-f_\pi|} + B\alpha^2\right),$$

which is exactly the stated inequality. □

A.2 Proof of Theorem 3.1

For $z = (x, y, d)$, recall the *normalized IPW score*

$$f_\pi(z) \ :=\ \kappa\left(\pi(x)\,\frac{yd}{p_o(x)} \ +\ \big(1-\pi(x)\big)\,\frac{y(1-d)}{1-p_o(x)}\right),$$

so that $f_\pi(Z) \in [0,1]$ almost surely (by $Y \in [0,1]$ and $p_o(x) \in [\kappa, 1-\kappa]$), and write $\mathbb{E}[\cdot]$ for population expectation and $\mathbb{E}_n[\cdot]$ for the empirical average over the sample.

Corollary A.3. *For any policy π, with probability at least $1-\delta$, we have*

$$|V(\pi^*) - V(\pi) - (V_n(\pi^*) - V_n(\pi))| \le \frac{\boldsymbol{c}}{\kappa}\left(\alpha\sqrt{\mathbb{E}\,|f_\pi - f_{\pi^*}|} + \alpha^2\right).$$

Proof. We invoke Lemma A.2 with $\gamma_1(z) = \frac{yd}{p_o(x)}$ and $\gamma_2(z) = \frac{y(1-d)}{1-p_o(x)}$ and setting $B = \frac{1}{\kappa}$ gives us the stated bound. □

Corollary A.4. *With probability at least $1-\delta$, for any $\pi_1, \pi_2 \in \Pi$ the following holds for some constant $\boldsymbol{c}$:*

$$\Big|\mathbb{E}_n\big|f_{\pi_1} - f_{\pi_2}\big| - \mathbb{E}\big|f_{\pi_1} - f_{\pi_2}\big|\Big| \le \boldsymbol{c}\left(\alpha\sqrt{\mathbb{E}\big|f_{\pi_1} - f_{\pi_2}\big|} + \alpha^2\right). \tag{10}$$

Proof. We invoke Lemma A.2. In particular, note that

$$|(f_{\pi_1} - f_{\pi_2})(x,y,d)| = \mathbb{1}\{\pi_1(x) \ne \pi_2(x)\}\left|\frac{\kappa\ y\,d}{p_o(x)} + \frac{\kappa y\,(1-d)}{1-p_o(x)}\right|.$$

Moreover, the class

$$\mathcal{H} = \{x \to \mathbb{1}\{\pi_1(x) \ne \pi_2(x)\} : \pi_1, \pi_2 \in \Pi\}$$

has VC dimension is at most $10d$ Bousquet and Zhivotovskiy (2021); Vidyasagar (2013). We now invoke Lemma A.2 with $\Pi = \mathcal{H}$, $\gamma_1(z) = \left|\frac{\kappa\ y\,d}{p_o(x)} + \frac{\kappa y\,(1-d)}{1-p_o(x)}\right|$, $\gamma_2(z) = 0$ and $B = 2$. □

Using Corollary A.3 and Corollary A.4 we have

Corollary A.5. *For any policy π, with probability at least $1-\delta$, we have*

$$|V(\pi^*) - V(\pi) - (V_n(\pi^*) - V_n(\pi))| \leq \frac{\boldsymbol{c}}{\kappa}\left(\alpha\sqrt{\mathbb{E}_n|f_\pi - f_{\pi^*}|} + \alpha^2\right).$$

Lemma A.6. *Let Π be a class of binary-valued functions with VC dimension d, and fix $\widehat{\pi} \in \Pi$. Suppose we construct a $\{0,1,*\}$-valued policy class $\widetilde{\Pi}$*

$$\pi'(x) = \begin{cases} \widehat{\pi}(x) & \text{if } \ \widehat{\pi}(x) = \pi(x) \\ * & \text{otherwise} \end{cases} \qquad \widetilde{\Pi} = \{\pi' : \forall \pi \in \widehat{\Pi}\}$$

Further, suppose $\mathcal{D}(\Pi) = \sup_{\pi_1,\pi_2\in\Pi}\|\pi_1 - \pi_2\|_{P_X,1}$ Then, with probability at least $1-\delta$, for every $\pi' \in \widetilde{\Pi}$ we have

$$\left|V^{(p)}(\phi^*) - V^{(p)}(\pi') - \left(V_n^{(p)}(\phi^*) - V_n^{(p)}(\pi')\right)\right| \lesssim \sqrt{\mathcal{D}(\pi)}\ \frac{\alpha}{\kappa} + \frac{\alpha^2}{\kappa}.$$

where $\phi^* = \arg\max_{\pi'\in\widetilde{\Pi}} V^{(p)}(\pi')$.

Proof. Define the function class induced by the disagreements between $\widehat{\pi}$ and any $\pi \in \Pi$

$$\mathcal{F}_\Pi = \Bigg\{(x,d,y) \to \mathbb{1}\{\pi(x) \neq \widehat{\pi}(x)\}\left(\frac{1}{2}\left(\frac{\kappa y d}{p_o(x)} + \frac{\kappa y(1-d)}{1-p_o(x)}\right) + p\right)$$
$$+ \mathbb{1}\{\pi(x) = \widehat{\pi}(x)\}\left(\widehat{\pi}(x)\frac{\kappa y d}{p_o(x)} + (1-\widehat{\pi}(x))\frac{\kappa y(1-d)}{1-p_o(x)}\right) : \pi \in \Pi\Bigg\}.$$

Notice that the class $\mathcal{H} = \{\mathbb{1}\{\pi(x) = \widehat{\pi}(x)\} : \pi \in \Pi\}$ has VC dimension d ($\widehat{\pi}$ is a fixed function). We invoke Lemma A.2 with $\gamma_1(z) = \frac{1}{2}\left(\frac{\kappa y d}{p_o(x)} + \frac{\kappa y(1-d)}{1-p_o(x)}\right) + p$ and $\gamma_2(z) = \widehat{\pi}(x)\frac{\kappa y d}{p_o(x)} + (1-\widehat{\pi}(x))\frac{\kappa y(1-d)}{1-p_o(x)}$ and $B = 2$. Moreover, we have for any $f_{\pi_1}, f_{\pi_2} \in \mathcal{F}_\Pi$ we have

$$\begin{aligned} \mathbb{E}\,|f_{\pi_1} - f_{\pi_2}| &\leq 2\ \mathbb{E}[\mathbb{1}\{\pi_1 \neq \pi_2\}] && \left(\text{since } \tfrac{\kappa y d}{p_o(x)},\ \tfrac{\kappa y(1-d)}{1-p_o(x)},\ p \leq 1\right) \\ &= 2\ \|\pi_1 - \pi_2\|_{P_X,1} \leq 2\mathcal{D}(\Pi) \end{aligned}$$

Plugging this upper bound in the bound in Lemma A.2 gives us the required bound. □

Lemma A.7. *With probability greater than $1-\delta$ any $\pi \in \widehat{\Pi}$ satisfies*

$$V(\pi^*) - V(\pi) \lesssim \frac{1}{\kappa}\left(\alpha\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha^2\right)$$

where $\mathcal{D}(\widehat{\Pi}) = \sup_{\pi_1,\pi_2}\|\pi_1 - \pi\|_{P_X,1}$.

Proof. For each $\pi \in \widehat{\Pi}$, we obtain

$$\begin{aligned} V(\pi^*) - V(\pi) &= V(\pi^*) - V(\pi) - (V_n(\pi^*) - V_n(\pi)) + (V_n(\pi^*) - V_n(\pi)) \\ &\lesssim \frac{\alpha}{\kappa}\sqrt{\mathbb{E}\,|f_{\pi^*} - f_\pi|} + \frac{\alpha^2}{\kappa} + (V_n(\pi^*) - V_n(\pi)) && \text{(via Corollary A.3)} \\ &\lesssim \frac{\alpha}{\kappa}\sqrt{\mathbb{E}\,|f_{\pi^*} - f_\pi|} + \frac{\alpha^2}{\kappa} + (V_n(\widehat{\pi}) - V_n(\pi)) && \text{(since } \widehat{\pi} \text{ is the empirical maximizer)} \\ &\lesssim \tfrac{1}{\kappa}\left(\alpha\sqrt{\mathbb{E}\,|f_{\pi^*} - f_\pi|} + \alpha\sqrt{\mathbb{E}_n|f_{\widehat{\pi}} - f_\pi|} + \alpha^2\right) \\ &\lesssim \tfrac{1}{\kappa}\left(\alpha\sqrt{\mathbb{E}\,|f_{\pi^*} - f_\pi|} + \alpha\sqrt{\mathbb{E}\,|f_{\widehat{\pi}} - f_\pi|} + \alpha^2\right) && \text{(via Corollary A.4)} \\ &\lesssim \tfrac{1}{\kappa}\left(\alpha\sqrt{\|\pi^* - \pi\|_{P_X,1}} + \alpha\sqrt{\|\widehat{\pi} - \pi\|_{P_X,1}} + \alpha^2\right) \\ &\lesssim \tfrac{1}{\kappa}\left(\alpha\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha^2\right). \end{aligned}$$

□

Proof of Theorem 3.1 . We first show that $\pi^* \in \widehat{\Pi}$. With probability at least $1-\delta$, we have

$$\begin{aligned} V_n(\widehat{\pi}) - V_n(\pi^*) &\leq V_n(\widehat{\pi}) - V_n(\pi^*) - (V(\widehat{\pi}) - V(\pi^*)) \\ &\leq \alpha\sqrt{\mathbb{E}\left|f_{\pi^*} - f_{\widehat{\pi}}\right|} + \alpha^2 && \text{(via Corollary A.3)} \\ &\lesssim \alpha\sqrt{\mathbb{E}_n\left|f_{\pi^*} - f_{\widehat{\pi}}\right|} + \alpha^2 \quad . && \text{(via Corollary A.4)} \end{aligned}$$

Hence, $\pi^* \in \widehat{\Pi}$ for an appropriate constant in Step 6 in Algorithm 1.

For the second EWM on $\widetilde{\Pi}$, let $\phi^* \in \widetilde{\Pi}$ be the policy that maximizes the true value $V^{(p)}$, and let ϕ be an abstaining policy constructed from π ($\phi(x) = *$ if $\mathbb{1}\{\pi(x) \neq \widehat{\pi}(x)\}$ and $\widehat{\pi}(x)$ otherwise) where π satisfies

$$\|\pi - \widehat{\pi}\|_{P_X,1} \geq \frac{\mathcal{D}(\widehat{\Pi})}{2}.$$

Such a π exists by the definition of $\mathcal{D}(\widehat{\Pi})$ and the triangle inequality. Since π and $\widehat{\pi}$ are binary, by the definition of f_π we have

$$\mathbb{E}\left|f_\pi - f_{\widehat{\pi}}\right| = \kappa\mathbb{E}\left|(\pi(X) - \widehat{\pi}(X))\left(\frac{YD}{p_o(x)} - \frac{Y(1-D)}{1-p_o(x)}\right)\right| = \kappa\mathbb{E}\left[\mathbb{1}\{\pi(X) \neq \widehat{\pi}(X)\}\left(\frac{YD}{p_o(x)} + \frac{Y(1-D)}{1-p_o(x)}\right)\right].$$

Thus,

$$\begin{aligned} V^{(p)}(\phi^*) &\geq V^{(p)}(\phi) \\ &= \mathbb{E}\left[\left(\frac{\pi(X)+\widehat{\pi}(X)}{2}\right)\frac{YD}{p_o(x)} + \left(1 - \frac{\pi(X)+\widehat{\pi}(X)}{2}\right)\frac{Y(1-D)}{1-p_o(x)} + p\cdot\mathbb{1}\{\pi(X) \neq \widehat{\pi}(X)\}\right] \\ &\geq \mathbb{E}\left[\left(\frac{\pi(X)+\widehat{\pi}(X)}{2}\right)\frac{YD}{p_o(x)} + \left(1 - \frac{\pi(X)+\widehat{\pi}(X)}{2}\right)\frac{Y(1-D)}{1-p_o(x)}\right] + p\cdot\|\pi - \widehat{\pi}\|_{P_X,1} \\ &= \frac{V(\widehat{\pi}) + V(\pi)}{2} + \frac{p}{2}\mathcal{D}(\widehat{\Pi}) \end{aligned}$$

Further using Lemma A.7 we have

$$V^{(p)}(\phi^*) - V(\pi^*) \gtrsim p\,\mathcal{D}(\widehat{\Pi}) - \frac{\alpha}{\kappa}\sqrt{\mathcal{D}(\widehat{\Pi})} + \frac{\alpha^2}{\kappa}.$$

Since $\widetilde{\pi}$ maximizes $V_n^{(p)}$, using Lemma A.6 we get

$$V^{(p)}(\widetilde{\pi}) - V^{(p)}(\phi^*) \gtrsim -\frac{\alpha\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha^2}{\kappa},$$

which implies

$$V^{(p)}(\widetilde{\pi}) - V(\pi^*) \gtrsim p\,\mathcal{D}(\widehat{\Pi}) - \frac{\alpha}{\kappa}\sqrt{\mathcal{D}(\widehat{\Pi})} + \frac{\alpha^2}{\kappa}. \tag{11}$$

Minimizing the quadratic with respect to $\sqrt{\mathcal{D}(\widehat{\Pi})}$,

$$V(\pi^*) - V^{(p)}(\widetilde{\pi}) \lesssim \frac{\alpha^2}{\kappa^2 p},$$

Finally, substituting $\alpha := \sqrt{\frac{d\log\frac{n}{d} + \log\frac{1}{\delta}}{n}}$ gives us the desired regret bound. □

Proof of Proposition 3.3. Substituting $p = 0$ in Equation (11) and noting that $\mathcal{D}(\widehat{\Pi}) \leq \mathbf{c}$ for some constant $\mathbf{c}$, proves the result. □

Algorithm 4 Unkown Propensities: DR learner.

1: **Input:** Samples $\{(X_i, D_i, Y_i)\}_{i=1}^n$, policy class Π, overlap κ, confidence $\delta \in (0,1)$, bonus p, VC dimension d, Propensity estimate $\hat{p}$, Outcome regression estimate $\hat{g}$.
2: **Set:** $\alpha_{\mathrm{DR}} \leftarrow \sqrt{\frac{d \log \frac{n}{d} + \log \frac{1}{\delta}}{n}} + \mathrm{Err}_{\mathrm{DR}}$.
3: **Split:** Partition the samples into two sets of size $n/2$: $\mathcal{D}_1, \mathcal{D}_2$.
4: **EWM:** $\widehat{\pi} = \arg\max_{\pi \in \Pi} V_{\mathrm{DR},n}(\pi)$ (computed on $\mathcal{D}_1$).
5: **Select near-optimal policies:**

$$\widehat{\Pi} \leftarrow \Big\{\pi \in \Pi :\ V_{\mathrm{DR},n}(\widehat{\pi}) - V_{\mathrm{DR},n}(\pi) \le \frac{\mathbf{c}}{\kappa}\Big(\alpha_{\mathrm{DR}}^2 + \alpha_{\mathrm{DR}}\sqrt{\mathbb{E}_n\,|\widehat{\pi} - \pi|}\Big)\Big\}.$$

6: **Abstention projection:** For each $\pi \in \widehat{\Pi}$, define

$$\pi'(X) = \begin{cases} \pi(X), & \text{if } \pi(X) = \widehat{\pi}(X), \\ *, & \text{otherwise}, \end{cases} \qquad \widetilde{\Pi} \leftarrow \{\,\pi' : \pi \in \widehat{\Pi}\,\}.$$

7: **EWM with abstention:** $\widetilde{\pi} \in \arg\max_{\pi \in \widetilde{\Pi}} V_{\mathrm{DR},n}^{(p)}(\pi)$ (evaluate $V_{\mathrm{DR},n}^{(p)}$ on $\mathcal{D}_2$).
8: **Return** $\widetilde{\pi}$.

B Doubly Robust Abstention: Additional Details and Proofs

Algorithm 4 is identical in structure to the known–propensity version, with two key edits.

(i) *Scores:* every occurrence of the outcome Y in the value computations is replaced by the DR pseudo–outcome $\widehat{\varphi}$ defined below, and we optimize the DR objectives $V_{\mathrm{DR},n}(\cdot)$ and $V_{\mathrm{DR},n}^{(p)}(\cdot)$ in Steps 4 and 7.

(ii) *Radius:* the selection radius incorporates nuisance error,

$$\alpha_{\mathrm{DR}} \ \leftarrow\ \sqrt{\frac{d \log \frac{n}{d} + \log \frac{1}{\delta}}{n}} \ +\ \mathrm{Err}_{\mathrm{DR}},$$

so the near–optimal set in Step 5 uses α_{DR} and the disagreement between with the EWM $\mathbb{E}_n|\widehat{\pi} - \pi|$. Where $\mathrm{Err}_{\mathrm{DR}}$ is any upper bound on product error in the nuisance: $\mathrm{Err}_{\mathrm{DR}} \ \ge\ \mathbb{E}\Big[(\widehat{p}(X) - p_o(X))^2 \sum_{d=0}^{1} (\widehat{g}(d,X) - g_o(d,X))^2\Big]^{1/2}$. In practice, we simply increase the constant in Step 6 of Algorithm 1; when the nuisance product error is $o_p(n^{-1/2})$, this yields the same rate as Theorem 3.1. Such product rates are a common assumption in the doubly robust literature (Foster and Syrgkanis, 2023) and hold for several function classes as discussed in (Foster and Syrgkanis, 2023, Section 5 and Appendix E). Moreover, a natural choice of $\mathrm{Err}_{\mathrm{DR}}$ is the L_4 error rates:

$$\|\widehat{p}(X) - p_o(X)\|_{P_X,4} \cdot \sum_{d \in \{0,1\}} \|\widehat{g}(d,X) - g_o(d,X)\|_{P_X,4}.$$

The minimax L_4 error rates are well-understood for many nonparametric classes of interest, such as smooth classes (Stone, 1980, 1982), Holder classes (Lepskii, 1992), Besov classes (Donoho and Johnstone, 1998) and convex function classes (Guntuboyina and Sen, 2015).

Pseudo–outcomes and DR objectives. For $d \in \{0,1\}$ and nuisance estimates $\widehat{g}(d,x)$, $\widehat{p}(x)$, define

$$\widehat{\varphi}(x,d,y) \ :=\ \widehat{g}(d,x) \ +\ \Big(\tfrac{d \cdot D}{\widehat{p}(x)} + \tfrac{(1-d)\cdot(1-D)}{1-\widehat{p}(x)}\Big)\big(y - \widehat{g}(d,x)\big).$$

Under $g_o(\cdot,x) \in [0,1]$ and $p_o(x) \in [\kappa, 1-\kappa]$, the DR value functionals are

$$\begin{aligned} V_{\mathrm{DR}}(\pi) &= \mathbb{E}\big[\pi(X)\widehat{\varphi}(X,1,Y) + (1-\pi(X))\widehat{\varphi}(X,0,Y)\big], \\ V_{\mathrm{DR}}^{(p)}(\pi) &= \mathbb{E}\Big[\mathbb{1}\{\pi \neq *\}\ (\cdot) + \mathbb{1}\{\pi = *\}\ \Big(\tfrac{\widehat{\varphi}(X,1,Y) + \widehat{\varphi}(X,0,Y)}{2} + p\Big)\Big], \end{aligned}$$

Unlike the IPW case, $V_{\mathrm{DR},n}$ and $V_{\mathrm{DR},n}^{(p)}$ are, in general, *biased* plug–ins; we control this bias by introducing $\mathrm{Err}_{\mathrm{DR}}$ (or an upper bound on it). The concentration lemmas mirror the known–propensity proofs. The resulting bounds (e.g., Lemmas B.1 and B.2 and their abstention analog Lemma B.5) yield the same n, d, δ–dependence as in the IPW case, up to constants. The bias induced by estimating nuisances in $V_{\mathrm{DR},n}$ and $V^{(p)}_{\mathrm{DR},n}$ is quantified by Lemmas B.4 and B.6, which bound the drift from V and $V^{(p)}$ in terms of $\mathrm{Err}_{\mathrm{DR}}$ and the policy disagreement.

B.0.1 Proof of Theorem 3.5

Lemma B.1. *For any policy π, with probability at least $1-\delta$,*

$$|V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\pi) - (V_{\mathrm{DR},n}(\pi^*) - V_{\mathrm{DR},n}(\pi))| \le \frac{\boldsymbol{c}}{\kappa}\left(\alpha\sqrt{\mathbb{E}\,|\pi-\pi^*|} + \alpha^2\right).$$

Proof. Define the DR score with fixed nuisances (or computed on a disjoint fold):

$$f^{\mathrm{DR}}_{\pi}(Z) \;\coloneqq\; \pi(X)\,\widehat{\varphi}(X,1,Y) + \big(1-\pi(X)\big)\,\widehat{\varphi}(X,0,Y),$$

so that $V_{\mathrm{DR}}(\pi) = \mathbb{E}[f^{\mathrm{DR}}_{\pi}(Z)]$ and $V_{\mathrm{DR},n}(\pi) = \mathbb{E}_n[f^{\mathrm{DR}}_{\pi}(Z)]$. Under bounded outcomes $Y \in [0,1]$ and strict overlap $p_o(X) \in [\kappa, 1-\kappa]$, each pseudo-outcome is a.s. bounded:

$$\big|\widehat{\varphi}(X,1,Y)\big|,\ \big|\widehat{\varphi}(X,0,Y)\big| \;\le\; \frac{C}{\kappa} \quad \text{a.s.}$$

set $B \coloneqq 2/\kappa$. Apply Lemma A.2 to the class

$$\mathcal{F}_{\Pi} \;=\; \Big\{ z \mapsto \pi(x)\,\widehat{\varphi}(x,1,y) + \big(1-\pi(x)\big)\,\widehat{\varphi}(x,0,y) :\; \pi \in \Pi \Big\},$$

with $\gamma_1 = \widehat{\varphi}(\cdot,1,\cdot)$, $\gamma_2 = \widehat{\varphi}(\cdot,0,\cdot)$ and the fixed $f^{\mathrm{DR}}_{\pi^*}$. We obtain, with probability at least $1-\delta$,

$$\Big|\,\mathbb{E}\big[f^{\mathrm{DR}}_{\pi^*} - f^{\mathrm{DR}}_{\pi}\big] - \mathbb{E}_n\big[f^{\mathrm{DR}}_{\pi^*} - f^{\mathrm{DR}}_{\pi}\big]\,\Big| \;\le\; \mathbf{c}\left(\alpha\sqrt{B\,\mathbb{E}\big|f^{\mathrm{DR}}_{\pi^*} - f^{\mathrm{DR}}_{\pi}\big|} + B\alpha^2\right).$$

Since $\big|f^{\mathrm{DR}}_{\pi^*} - f^{\mathrm{DR}}_{\pi}\big| \le B\,\big|\pi^* - \pi\big|$ pointwise, we have $\mathbb{E}|f^{\mathrm{DR}}_{\pi^*} - f^{\mathrm{DR}}_{\pi}| \le B\,\mathbb{E}|\pi^* - \pi|$, and the RHS becomes

$$\alpha\sqrt{B^2\,\mathbb{E}|\pi^* - \pi|} + B\alpha^2 \lesssim \frac{\mathbf{c}}{\kappa}\left(\alpha\sqrt{\mathbb{E}|\pi^* - \pi|} + \alpha^2\right),$$

which is the claim. □

Lemma B.2. *With probability at least $1-\delta$, for any $f_{\pi_1}, f_{\pi_2} \in \mathcal{F}_{\Pi}$ the following inequalities hold:*

$$\Big|\,\mathbb{E}_n\big|\pi_1 - \pi_2\big| - \mathbb{E}\big|\pi_1 - \pi_2\big|\,\Big| \le \boldsymbol{c}\left(\alpha\sqrt{\mathbb{E}\big|\pi_1 - \pi_2\big|} + \alpha^2\right),$$
$$\Big|\,\mathbb{E}_n\big|\pi_1 - \pi_2\big| - \mathbb{E}\big|\pi_1 - \pi_2\big|\,\Big| \le \boldsymbol{c}\left(\alpha\sqrt{\mathbb{E}_n\big|\pi_1 - \pi_2\big|} + \alpha^2\right).$$

Proof. Apply Lemma A.2 to the class

$$\mathcal{F}^{\mathrm{id}}_{\Pi} \;=\; \Big\{ z \mapsto \pi(x)\cdot 1 + \big(1-\pi(x)\big)\cdot 0 :\; \pi \in \Pi \Big\},$$

with $\gamma_1 \equiv 1$, $\gamma_2 \equiv 0$, hence $B = 1$ and $f_\pi(z) = \pi(x)$. Fix f_{π^*} with $\pi^* = \pi_2$ and take f_π with $\pi = \pi_1$; then

$$|f_{\pi_1}(Z) - f_{\pi_2}(Z)| \;=\; |\pi_1(X) - \pi_2(X)| \;=\; \mathbb{1}\{\pi_1(X) \neq \pi_2(X)\}\,.$$

The two displayed bounds follow directly from Lemma A.2 (first with population square-root term, then with the empirical one). □

Lemma B.3. *For any policy π, with probability at least $1-\delta$, we have*

$$|V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\pi) - (V_{\mathrm{DR},n}(\pi^*) - V_{\mathrm{DR},n}(\pi))| \le \frac{\boldsymbol{c}}{\kappa}\left(\alpha\sqrt{\mathbb{E}_n\,|\pi-\pi^*|} + \alpha^2\right).$$

Proof. Combine Lemma B.1 with Lemma B.2 by replacing the population square-root term $\sqrt{\mathbb{E}|\pi - \pi^*|}$ in Lemma B.1 with its empirical counterpart via the second inequality of Lemma B.2. □

Recall the definition,

$$\mathrm{Err}_{\mathrm{DR}} \;\coloneqq\; \left\{\mathbb{E}\Big[\big(\hat{p}(X)-p_o(X)\big)^2 \sum_{d\in\{0,1\}} \big(\hat{g}(d,X)-g_o(d,X)\big)^2\Big]\right\}^{1/2}.$$

Lemma B.4. *For any two binary policies $\pi_1, \pi_2 \in \Pi$,*

$$\big(V_{\mathrm{DR}}(\pi_1) - V_{\mathrm{DR}}(\pi_2)\big) \;-\; \big(V(\pi_1)-V(\pi_2)\big) \;\le\; 2\,\kappa^{-1}\,\mathrm{Err}_{\mathrm{DR}}\,\|\pi_1-\pi_2\|_{P_X,2}.$$

Proof. By the definitions of V_{DR} and V,

$$\begin{aligned}
V_{\mathrm{DR}}(\pi) &= \mathbb{E}\Big[\pi(X)\Big(\hat{g}(1,X) + \tfrac{\mathbb{1}\{D=1\}}{\hat{p}(X)}\big(Y-\hat{g}(1,X)\big)\Big) \\
&\qquad + \;(1-\pi(X))\Big(\hat{g}(0,X) + \tfrac{\mathbb{1}\{D=0\}}{1-\hat{p}(X)}\big(Y-\hat{g}(0,X)\big)\Big)\Big] \\
&= \mathbb{E}\Big[\pi(X)\Big(\hat{g}(1,X) + \tfrac{p_o(X)}{\hat{p}(X)}\big(g_o(1,X)-\hat{g}(1,X)\big)\Big) \\
&\qquad + \;(1-\pi(X))\Big(\hat{g}(0,X) + \tfrac{1-p_o(X)}{1-\hat{p}(X)}\big(g_o(0,X)-\hat{g}(0,X)\big)\Big)\Big], \\
V(\pi) &= \mathbb{E}\big[\pi(X)g_o(1,X) + (1-\pi(X))g_o(0,X)\big].
\end{aligned}$$

Subtracting the two displays and taking the difference between π_1 and π_2 gives

$$\begin{aligned}
&\big(V_{\mathrm{DR}}(\pi_1) - V_{\mathrm{DR}}(\pi_2)\big) - \big(V(\pi_1)-V(\pi_2)\big) \\
&= \mathbb{E}\Big[(\pi_1-\pi_2)(X)\Big(1-\tfrac{p_o(X)}{\hat{p}(X)}\Big)\big(g_o(1,X)-\hat{g}(1,X)\big) \\
&\qquad + (\pi_2-\pi_1)(X)\Big(1-\tfrac{1-p_o(X)}{1-\hat{p}(X)}\Big)\big(g_o(0,X)-\hat{g}(0,X)\big)\Big].
\end{aligned}$$

Taking absolute values and using the triangle inequality,

$$\begin{aligned}
\cdots &\le \mathbb{E}\Big[|\pi_1-\pi_2|(X)\,\Big|1-\tfrac{p_o(X)}{\hat{p}(X)}\Big|\,\big|g_o(1,X)-\hat{g}(1,X)\big|\Big] \\
&\quad + \mathbb{E}\Big[|\pi_1-\pi_2|(X)\,\Big|1-\tfrac{1-p_o(X)}{1-\hat{p}(X)}\Big|\,\big|g_o(0,X)-\hat{g}(0,X)\big|\Big].
\end{aligned}$$

By strict overlap, $\big|1-\frac{p_o}{\hat{p}}\big| \le \kappa^{-1}|\hat{p}-p_o|$ and $\big|1-\frac{1-p_o}{1-\hat{p}}\big| \le \kappa^{-1}|\hat{p}-p_o|$. Hence,

$$\begin{aligned}
\cdots &\le \kappa^{-1}\,\mathbb{E}\Big[|\pi_1-\pi_2|(X)\,|\hat{p}-p_o|\,\big(|g_o(1,X)-\hat{g}(1,X)| + |g_o(0,X)-\hat{g}(0,X)|\big)\Big] \\
&\le \kappa^{-1}\,\|\pi_1-\pi_2\|_{P_X,2}\,\Big(\mathbb{E}\big[(\hat{p}-p_o)^2\big(|g_o(1,X)-\hat{g}(1,X)| + |g_o(0,X)-\hat{g}(0,X)|\big)^2\big]\Big)^{1/2} \\
&\le 2\,\kappa^{-1}\,\|\pi_1-\pi_2\|_{P_X,2}\,\mathrm{Err}_{\mathrm{DR}},
\end{aligned}$$

using Cauchy–Schwarz and $(a+b)^2 \le 2(a^2+b^2)$. This proves the claim. □

Lemma B.5. *Let Π be a class of binary-valued functions with VC dimension d, and fix $\widehat{\pi}\in\Pi$. Construct the $\{0,1,*\}$-valued class*

$$\pi'(x) \;=\; \begin{cases} \widehat{\pi}(x), & \textit{if } \widehat{\pi}(x)=\pi(x), \\ *, & \textit{otherwise}, \end{cases} \qquad \widetilde{\Pi} \;=\; \{\pi' : \pi\in\Pi\}.$$

Let $\mathcal{D}(\Pi) \coloneqq \sup_{\pi_1,\pi_2\in\Pi}\|\pi_1-\pi_2\|_{P_X,1}$. Then, with probability at least $1-\delta$, for every $\pi'\in\widetilde{\Pi}$,

$$\big|\,V^{(p)}_{\mathrm{DR}}(\phi^*) - V^{(p)}_{\mathrm{DR}}(\pi') \;-\; \big(V^{(p)}_{\mathrm{DR},n}(\phi^*) - V^{(p)}_{\mathrm{DR},n}(\pi')\big)\,\big| \;\lesssim\; \frac{1}{\kappa}\Big(\alpha\sqrt{\mathcal{D}(\Pi)} + \alpha^2\Big),$$

where $\phi^* \in \arg\max_{\pi'\in\widetilde{\Pi}} V^{(p)}_{\mathrm{DR}}(\pi')$.

Proof. Define the function class induced by the $\widetilde{\pi}$ (disagreement w.r.t. $\widehat{\pi}$):

$$\mathcal{F}_{\Pi}^{\mathrm{DR}} = \Big\{ (x,d,y) \mapsto \mathbb{1}\{\pi(x) \neq \widehat{\pi}(x)\}\Big(\tfrac{1}{2}\big(\widehat{\varphi}(x,1,y)+\widehat{\varphi}(x,0,y)\big)+p\Big) \\ + \mathbb{1}\{\pi(x)=\widehat{\pi}(x)\}\left(\widehat{\pi}(x)\widehat{\varphi}(x,1,y)+(1-\widehat{\pi}(x))\widehat{\varphi}(x,0,y)\right) : \ \pi \in \Pi \Big\},$$

Since $\widehat{\pi}$ is fixed, the indicator class $\mathcal{H} = \{\mathbb{1}\{\pi(x)=\widehat{\pi}(x)\} : \pi \in \Pi\}$ has VC dimension d. Under bounded outcomes and strict overlap, the DR scores are a.s. bounded by a constant factor of $1/\kappa$, so for any $f_{\pi_1}, f_{\pi_2} \in \mathcal{F}_{\Pi}^{\mathrm{DR}}$,

$$\mathbb{E}\big|f_{\pi_1}-f_{\pi_2}\big| \ \leq \ \mathbf{c}\ \mathbb{E}\big[\mathbb{1}\{\pi_1(X)\neq\pi_2(X)\}\big] \ = \ \mathbf{c}\ \|\pi_1-\pi_2\|_{P_X,1} \ \leq \ \mathbf{c}\ \mathcal{D}(\Pi),$$

Applying Lemma A.2 to $\mathcal{F}_{\Pi}^{\mathrm{DR}}$ (with the fixed comparator f_{ϕ^*} and boundedness constant folded into $1/\kappa$) yields, uniformly over $\pi' \in \widetilde{\Pi}$,

$$\big|\,\mathbb{E}f_{\phi^*}-\mathbb{E}f_{\pi'} \ - \ (\mathbb{E}_n f_{\phi^*}-\mathbb{E}_n f_{\pi'})\,\big| \ \lesssim \ \frac{1}{\kappa}\Big(\alpha\sqrt{\mathbb{E}|f_{\phi^*}-f_{\pi'}|}+\alpha^2\Big).$$

Using $\mathbb{E}|f_{\phi^*}-f_{\pi'}| \lesssim \mathcal{D}(\Pi)$ and identifying $\mathbb{E}f_{\pi'} = V_{\mathrm{DR}}^{(p)}(\pi')$, $\mathbb{E}_n f_{\pi'} = V_{\mathrm{DR},n}^{(p)}(\pi')$ completes the proof. □

Lemma B.4 implies that any near-optimal policy under V_{DR} must also be near-optimal under V.

Lemma B.6. *Let π_1', π_2' be abstaining policies obtained using disagreement with respect to a fixed reference $\widehat{\pi}$, from binary policies $\pi_1, \pi_2 \in \Pi$. Then*

$$\Big|\big(V_{\mathrm{DR}}^{(p)}(\pi_1')-V_{\mathrm{DR}}^{(p)}(\pi_2')\big) \ - \ \big(V^{(p)}(\pi_1')-V^{(p)}(\pi_2')\big)\Big| \ \leq \ \kappa^{-1}\,\mathrm{Err}_{\mathrm{DR}}\ \|\pi_1-\pi_2\|_{P_X,2}.$$

Proof. For any abstaining policy π', define

$$\Delta^{(p)}(\pi') \ \coloneqq \ V_{\mathrm{DR}}^{(p)}(\pi')-V^{(p)}(\pi').$$

Condition on $X=x$. The DR–minus–truth discrepancy at x is a linear combination of the two nuisance errors with coefficients determined by the action selected by $\pi'(x)$:

$$\Delta^{(p)}(\pi' \mid X=x) = \begin{cases} \Big(1-\frac{p_o(x)}{\hat{p}(x)}\Big)\big(g_o(1,x)-\hat{g}(1,x)\big), & \pi'(x)=1,\\ \Big(1-\frac{1-p_o(x)}{1-\hat{p}(x)}\Big)\big(g_o(0,x)-\hat{g}(0,x)\big), & \pi'(x)=0,\\ \frac{1}{2}\Big[\Big(1-\frac{p_o(x)}{\hat{p}(x)}\Big)\big(g_o(1,x)-\hat{g}(1,x)\big)+\Big(1-\frac{1-p_o(x)}{1-\hat{p}(x)}\Big)\big(g_o(0,x)-\hat{g}(0,x)\big)\Big], & \pi'(x)=*. \end{cases}$$

By strict overlap,

$$\Big|1-\tfrac{p_o}{\hat{p}}\Big|, \ \Big|1-\tfrac{1-p_o}{1-\hat{p}}\Big| \ \leq \ \kappa^{-1}\,|\hat{p}-p_o|.$$

Hence, for any x,

$$\big|\Delta^{(p)}(\pi_1' \mid X=x)-\Delta^{(p)}(\pi_2' \mid X=x)\big| \ \leq \ \kappa^{-1}\,|\hat{p}(x)-p_o(x)|\,\tfrac{1}{2}\sum_{d\in\{0,1\}}\big|g_o(d,x)-\hat{g}(d,x)\big| \ \cdot \ \mathbf{1}\{\pi_1'(x)\neq\pi_2'(x)\}.$$

Taking expectations and applying Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}\Big[\big|\Delta^{(p)}(\pi_1')-\Delta^{(p)}(\pi_2')\big|\Big] &\leq \kappa^{-1}\,\mathbb{E}\Big[\mathbb{1}\{\pi_1'\neq\pi_2'\}\ |\hat{p}-p_o|\ \tfrac{1}{2}\sum_d(|g_o(d,X)-\hat{g}(d,X)|)\Big]\\ &\leq \kappa^{-1}\,\|\mathbb{1}\{\pi_1'\neq\pi_2'\}\|_{P_X,2}\ \Big(\mathbb{E}\big[(\hat{p}-p_o)^2\big(\tfrac{1}{2}\sum_d|g_o(d,X)-\hat{g}(d,X)|\big)^2\big]\Big)^{1/2}\\ &\leq \kappa^{-1}\,\|\mathbb{1}\{\pi_1'\neq\pi_2'\}\|_{P_X,2}\ \mathrm{Err}_{\mathrm{DR}}.\end{aligned}$$

Finally, by the disagreement–projection construction, $\mathbb{1}\{\pi_1'\neq\pi_2'\} = \mathbb{1}\{\pi_1\neq\pi_2\} = |\pi_1-\pi_2|$. This yields the stated bound. □

Lemma B.7. *Fix a reference policy $\widehat{\pi} \in \Pi$ and any binary policy $\pi \in \Pi$. Let π' be the abstention projection of π w.r.t. $\widehat{\pi}$, i.e.,*

$$\pi'(x) = \begin{cases} \pi(x), & \pi(x) = \widehat{\pi}(x), \\ *, & \textit{otherwise}. \end{cases}$$

Then

$$\Big| \big(V_{\mathrm{DR}}^{(p)}(\pi') - V_{\mathrm{DR}}(\pi^*) \big) \; - \; \big(V^{(p)}(\pi') - V(\pi^*) \big) \Big| \; \leq \; \kappa^{-1} \, \mathrm{Err}_{\mathrm{DR}} \Big(\|\pi - \pi^*\|_{P_X,2} + \|\widehat{\pi} - \pi^*\|_{P_X,2} \Big).$$

Proof. Define the DR–truth discrepancies $\Delta^{(p)}(\tilde{\pi}) := V_{\mathrm{DR}}^{(p)}(\tilde{\pi}) - V^{(p)}(\tilde{\pi})$ and $\Delta(\tilde{\pi}) := V_{\mathrm{DR}}(\tilde{\pi}) - V(\tilde{\pi})$. Introduce the projection of π^* onto $\{0, 1, *\}$ via $\widehat{\pi}$, denoted $\pi^{*'}$, by

$$\pi^{*'}(x) = \begin{cases} \pi^*(x), & \pi^*(x) = \widehat{\pi}(x), \\ *, & \text{otherwise}. \end{cases}$$

Add and subtract $V_{\mathrm{DR}}^{(p)}(\pi^{*'})$ and $V^{(p)}(\pi^{*'})$:

$$\big(V_{\mathrm{DR}}^{(p)}(\pi') - V_{\mathrm{DR}}(\pi^*)\big) - \big(V^{(p)}(\pi') - V(\pi^*)\big) = \underbrace{\Big(\Delta^{(p)}(\pi') - \Delta^{(p)}(\pi^{*'})\Big)}_{\text{(A)}} + \underbrace{\Big(\Delta^{(p)}(\pi^{*'}) - \Delta(\pi^*)\Big)}_{\text{(B)}}.$$

For (A), apply Lemma B.6 to the pair $(\pi', \pi^{*'})$, which are the abstention projections of (π, π^*) w.r.t. the same $\widehat{\pi}$:

$$\text{(A)} \; \leq \; \kappa^{-1} \, \mathrm{Err}_{\mathrm{DR}} \, \|\pi - \pi^*\|_{P_X,2}.$$

For (B), $\pi^{*'}$ and π^* differ only on $\{x : \pi^*(x) \neq \widehat{\pi}(x)\}$. We again invoke Lemma B.6 but we set π^* at the reference policy (denoted as $\widehat{\pi}$ in Lemma B.6), $\pi_1 = \widehat{\pi}$, $\pi_2 = \pi^*$ and note that π^* never disagrees with itself and hence $(V^{(p)}(\pi_2') = V(\pi_2))$. Hence we get,

$$\text{(B)} \; \leq \; \kappa^{-1} \, \mathrm{Err}_{\mathrm{DR}} \, \|\widehat{\pi} - \pi^*\|_{P_X,2}.$$

Summing the two bounds gives the claim. □

Lemma B.8. *With probability at least $1 - \delta$, the optimal policy π^* belongs to $\widehat{\Pi}$ constructed in Step 5.*

Proof. Recall $\widehat{\pi} = \arg\max_{\pi \in \Pi} V_{\mathrm{DR},n}(\pi)$. By Lemma B.1 and Lemma B.4, for any $\pi \in \Pi$,

$$\begin{aligned} V(\pi^*) - V(\widehat{\pi}) &\leq \big(V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\widehat{\pi})\big) + \mathbf{c} \, \kappa^{-1} \Big(\alpha \sqrt{\mathbb{E}|\pi - \pi^*|} + \alpha^2\Big) \\ &\leq \big(V_{\mathrm{DR},n}(\pi^*) - V_{\mathrm{DR},n}(\widehat{\pi})\big) + \mathbf{c} \, \kappa^{-1} \Big(\alpha \sqrt{\mathbb{E}|\pi - \pi^*|} + \alpha^2\Big) + \mathbf{c} \, \kappa^{-1} \, \mathrm{Err}_{\mathrm{DR}} \, \|\pi^* - \pi\|_{P_X,2}. \end{aligned}$$

Since π, π^* are binary, $\|\pi^* - \pi\|_{P_X,2} = \sqrt{\mathbb{E}|\pi - \pi^*|}$. Hence

$$V(\pi^*) - V(\widehat{\pi}) \leq \big(V_{\mathrm{DR},n}(\pi^*) - V_{\mathrm{DR},n}(\widehat{\pi})\big) + \mathbf{c} \, \kappa^{-1} \Big((\mathrm{Err}_{\mathrm{DR}} + \alpha) \sqrt{\mathbb{E}|\pi - \pi^*|} + \alpha^2\Big). \tag{12}$$

Because $V(\pi^*) - V(\widehat{\pi}) \geq 0$, rearranging (12) gives

$$\begin{aligned} V_{\mathrm{DR},n}(\widehat{\pi}) - V_{\mathrm{DR},n}(\pi^*) &\lesssim \kappa^{-1} \Big((\mathrm{Err}_{\mathrm{DR}} + \alpha) \sqrt{\mathbb{E}|\pi - \pi^*|} + \alpha^2\Big) \\ &\lesssim \kappa^{-1} \Big((\mathrm{Err}_{\mathrm{DR}} + \alpha) \sqrt{\mathbb{E}_n|\pi - \pi^*|} + \mathrm{Err}_{\mathrm{DR}}^2 + \alpha^2\Big) \qquad \text{(by Lemma B.2)} \\ &\lesssim \kappa^{-1} \Big(\alpha_{\mathrm{DR}} \sqrt{\mathbb{E}_n|\pi - \pi^*|} + \alpha_{\mathrm{DR}}^2\Big), \end{aligned}$$

where $\alpha_{\mathrm{DR}} = \sqrt{\frac{d \log \frac{n}{d} + \log \frac{1}{\delta}}{n}} + \mathrm{Err}_{\mathrm{DR}}$. Choosing the constant in Step 5 accordingly shows that $\pi^* \in \widehat{\Pi}$. □

Lemma B.9. *With probability at least* $1-\delta$*, every* $\pi \in \widehat{\Pi}$ *satisfies*

$$V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\pi) \;\lesssim\; \frac{1}{\kappa}\Big(\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha_{\mathrm{DR}}^2\Big),$$

where $\mathcal{D}(\widehat{\Pi}) := \sup_{\pi_1,\pi_2\in\widehat{\Pi}} \|\pi_1 - \pi_2\|_{P_X,1}$.

Proof. Fix $\pi \in \widehat{\Pi}$. By Lemma B.1, on a $1-\delta$ event,

$$V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\pi) \;\le\; \big(V_{\mathrm{DR},n}(\pi^*) - V_{\mathrm{DR},n}(\pi)\big) + \frac{\mathbf{c}}{\kappa}\Big(\alpha\sqrt{\mathbb{E}|\pi-\pi^*|} + \alpha^2\Big).$$

Since $\widehat{\pi}$ maximizes $V_{\mathrm{DR},n}$, $V_{\mathrm{DR},n}(\pi^*) - V_{\mathrm{DR},n}(\widehat{\pi}) \le 0$, hence

$$V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\pi) \;\le\; \big(V_{\mathrm{DR},n}(\widehat{\pi}) - V_{\mathrm{DR},n}(\pi)\big) + \frac{\mathbf{c}}{\kappa}\Big(\alpha\sqrt{\mathbb{E}|\pi-\pi^*|} + \alpha^2\Big).$$

By the definition of $\widehat{\Pi}$ (Step 5),

$$V_{\mathrm{DR},n}(\widehat{\pi}) - V_{\mathrm{DR},n}(\pi) \;\le\; \frac{\mathbf{c}}{\kappa}\Big(\alpha_{\mathrm{DR}}^2 + \alpha_{\mathrm{DR}}\sqrt{\mathbb{E}_n|\widehat{\pi}-\pi|}\Big).$$

Combining the last two displays,

$$V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\pi) \;\le\; \frac{\mathbf{c}}{\kappa}\Big(\alpha_{\mathrm{DR}}^2 + \alpha_{\mathrm{DR}}\sqrt{\mathbb{E}_n|\widehat{\pi}-\pi|} + \alpha\sqrt{\mathbb{E}|\pi-\pi^*|} + \alpha^2\Big).$$

Because $\pi,\widehat{\pi},\pi^* \in \widehat{\Pi}$, we have $\mathbb{E}|\widehat{\pi}-\pi| \le \mathcal{D}(\widehat{\Pi})$ and $\mathbb{E}|\pi-\pi^*| \le \mathcal{D}(\widehat{\Pi})$. Using Lemma B.2, the fact that $|\widehat{\pi}-\pi| \in [0,1]$, and the bound $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, we obtain

$$\sqrt{\mathbb{E}_n|\widehat{\pi}-\pi|} \;\le\; \sqrt{\mathbb{E}|\widehat{\pi}-\pi|} + \mathbf{c}\;\alpha \;\le\; \sqrt{\mathcal{D}(\widehat{\Pi})} + \mathbf{c}\;\alpha.$$

Since $\alpha \le \alpha_{\mathrm{DR}}$, we conclude

$$V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}(\pi) \;\lesssim\; \frac{1}{\kappa}\Big(\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha_{\mathrm{DR}}^2\Big),$$

as claimed. □

Proof of Theorem 3.5. Let $\phi^* \in \widetilde{\Pi}$ be the policy that maximizes $V_{\mathrm{DR}}^{(p)}(\cdot)$. We follow a similar proof idea to the known propensity proof. Let ϕ be a policy such that $\|\pi - \widehat{\pi}\|_{P_X,1} \ge \frac{\mathcal{D}(\widehat{\Pi})}{2}$. We have that

$$\begin{aligned}
V_{\mathrm{DR}}^{(p)}(\phi^*) &\ge V_{\mathrm{DR}}^{(p)}(\phi) \\
&= \mathbb{E}\Bigg[\frac{\pi(X)+\widehat{\pi}(X)}{2}\hat{\varphi}(X,1,Y) \\
&\qquad + \left(1 - \frac{\pi(X)+\widehat{\pi}(X)}{2}\right)\hat{\varphi}(X,0,Y) \\
&\qquad + p\mathbb{E}[\mathbb{1}\{\pi(X) \neq \widehat{\pi}(X)\}]\Bigg] \\
&\ge \frac{1}{2}\big(V_{\mathrm{DR}}^{(p)}(\pi) + V_{\mathrm{DR}}^{(p)}(\widehat{\pi})\big) + p\,\mathbb{E}\,|\pi(X) - \widehat{\pi}(X)| \\
&= \frac{1}{2}\big(V_{\mathrm{DR}}(\pi) + V_{\mathrm{DR}}(\widehat{\pi})\big) + p\,\mathbb{E}\,|\pi(X) - \widehat{\pi}(X)| \qquad \text{(since } \pi,\widehat{\pi} \text{ are binary)} \\
&\ge V_{\mathrm{DR}}(\pi^*) + \frac{p}{2}\mathcal{D}(\widehat{\pi}) - \mathbf{c}\,\kappa^{-1}\left(\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha_{\mathrm{DR}}^2\right)
\end{aligned}$$

Where the last inequality uses the fact that $\pi^* \in \widehat{\Pi}$ (Lemma B.8) and Lemma B.9. Moreover since, $\widetilde{\pi}$ maximizes $V_{\mathrm{DR},n}^{(p)}$, via Lemma B.5 we have

$$V_{\mathrm{DR}}^{(p)}(\widetilde{\pi}) \gtrsim V^{(p)}(\phi^*) - \kappa^{-1}\left(\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha_{\mathrm{DR}}^2\right)$$

Combining with the second last inequality we get

$$V_{\mathrm{DR}}(\pi^*) - V_{\mathrm{DR}}^{(p)}(\widetilde{\pi}) \lesssim \mathbf{c}\,\kappa^{-1}\left(\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha_{\mathrm{DR}}^2\right) - \frac{p}{2}\mathcal{D}(\widehat{\Pi})$$

Now we finally invoke Lemma B.7 to get

$$\begin{aligned} V(\pi^*) - V^{(p)}(\widetilde{\pi}) &\lesssim \mathbf{c}\,\kappa^{-1}\left(\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha_{\mathrm{DR}}^2\right) + \mathrm{Err}_{\mathrm{DR}}\left(\|\pi - \pi^*\|_{P_X,2} + \|\widehat{\pi} - \pi^*\|_{P_X,2}\right) - \frac{p}{2}\mathcal{D}(\widehat{\Pi}) \\ &\lesssim \mathbf{c}\,\kappa^{-1}\left(\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} + \alpha_{\mathrm{DR}}^2\right) + \mathrm{Err}_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})} - \frac{p}{2}\mathcal{D}(\widehat{\Pi}) \end{aligned}$$

Finally, absorbing the second term into $\alpha_{\mathrm{DR}}\sqrt{\mathcal{D}(\widehat{\Pi})}$, and maximizing the the quadratic wrt $\sqrt{\mathcal{D}(\widehat{\Pi})}$ we get

$$V(\pi^*) - V^{(p)}(\widetilde{\pi}) \lesssim \frac{\alpha_{\mathrm{DR}}^2}{\kappa^2\, p}.$$

□

C Missing proofs from Section 4

C.1 Proofs of Theorems 4.1 and 4.2

Recall that the procedure has two phases. In Phase 1 we run Algorithm 1 as a black box to obtain an abstaining policy $\widetilde{\pi}$. In Phase 2 we learn a (non-abstaining) policy on the region where $\widetilde{\pi}$ abstains, $\mathcal{X}_{rem} \subseteq \mathcal{X}$, either by EWM (when the combinatorial diameter is finite) or via a CATE oracle; denote this policy by ϕ. The final policy is the mixture

$$\pi_{\mathrm{final}}(x) = \begin{cases} \phi(x), & x \in \mathcal{X}_{rem}, \\ \widetilde{\pi}(x), & x \notin \mathcal{X}_{rem}. \end{cases}$$

For any policy π, let $V(\pi \mid \mathcal{X}_{rem})$ and $V(\pi \mid \mathcal{X}_{rem}^c)$ denote its value contributions on the abstention and non-abstention regions, respectively:

$$\begin{aligned} V(\pi; \mathcal{X}_{rem}) &\coloneqq \mathbb{E}\left[\mathbb{1}\{X \in \mathcal{X}_{rem}\}\left(\pi(X)Y(1) + (1-\pi(X))Y(0)\right)\right], \\ V(\pi; \mathcal{X}_{rem}^c) &\coloneqq \mathbb{E}\left[\mathbb{1}\{X \notin \mathcal{X}_{rem}\}\left(\pi(X)Y(1) + (1-\pi(X))Y(0)\right)\right], \end{aligned}$$

so that $V(\pi) = V(\pi; \mathcal{X}_{rem}) + V(\pi; \mathcal{X}_{rem}^c)$.

Hence the value of the final policy is

$$\begin{aligned} V(\pi_{\mathrm{final}}) &= V(\widetilde{\pi}; \mathcal{X}_{rem}^c) + V(\phi; \mathcal{X}_{rem}) \\ &= \underbrace{V(\widetilde{\pi}; \mathcal{X}_{rem}^c) + V(\pi^B; \mathcal{X}_{rem})}_{\text{(A)}} + \underbrace{V(\phi; \mathcal{X}_{rem}) - V(\pi^B; \mathcal{X}_{rem})}_{\text{(B)}}. \end{aligned}$$

Where π^B is the Bayes optimal policy. Further, we can write (A) as

$$\begin{aligned} \text{(A)} &= \mathbb{E}\Big[\mathbb{1}\{\widetilde{\pi}(X) \neq *\}\left(\widetilde{\pi}(X)Y(1) + (1-\widetilde{\pi}(X))Y(0)\right)\Big] + \mathbb{E}\Big[\mathbb{1}\{\widetilde{\pi}(X) = *\}\left(\pi^B(X)Y(1) + (1-\pi^B(X))Y(0)\right)\Big] \\ &= \mathbb{E}\Big[\mathbb{1}\{\widetilde{\pi}(X) \neq *\}\left(\widetilde{\pi}(X)Y(1) + (1-\widetilde{\pi}(X))Y(0)\right)\Big] + \mathbb{E}\left[\mathbb{1}\{\widetilde{\pi}(X) = *\}\left(\frac{Y(1)+Y(0)}{2} + \frac{|\tau_o(X)|}{2}\right)\right] \\ &\qquad\qquad \text{(since } \pi_B(x) = \mathbb{1}\{\tau_o(x) > 0\}\text{)} \\ &\geq \mathbb{E}\Big[\mathbb{1}\{\widetilde{\pi}(X) \neq *\}\left(\widetilde{\pi}(X)Y(1) + (1-\widetilde{\pi}(X))Y(0)\right)\Big] + \mathbb{E}\left[\mathbb{1}\{\widetilde{\pi}(X) = *\}\left(\frac{Y(1)+Y(0)}{2} + \frac{h}{2}\right)\right] \\ &\qquad\qquad \text{(margin condition (3))} \\ &= V^p(\widetilde{\pi}) \end{aligned}$$

Where $V^{(p)}$ is the abstention value with $p = \frac{h}{2}$.

Hence, we can now write the regret of π_{final} as

$$\begin{aligned} V(\pi^*) - V(\pi_{\text{final}}) &\leq V(\pi^*) - V^{(p)}(\widetilde{\pi}) - \text{(B)} \\ &\lesssim \frac{d \log \frac{n}{d} + \log \frac{1}{\delta}}{p\, n\, \kappa^2} - \text{(B)} && \text{(via Theorem 3.1)} \end{aligned}$$

In the remainder of this section, we will focus on proving bounds for (B). Conditioning on the first phase makes $\mathcal{X}_{rem}$ fixed (measurable w.r.t. $\mathcal{D}_1 \cup \mathcal{D}_2$), so analysis on the third split (i.e. bounds on (B)) treats it as nonrandom. Finally, the third split has size m ($m = n/3$) and with $\rho := \mathbb{P}(X \in \mathcal{X}_{rem})$ and $N := \sum_{i=1}^{m} \mathbb{1}\{X_i \in \mathcal{X}_{rem}\} \sim \text{Bin}(m, \rho)$, applying chernoff bound we get

$$N \;\geq\; m\rho \;-\; \sqrt{2\, m\rho\, \log \tfrac{1}{\delta}} \quad \text{with probability at least } 1 - \delta.$$

If $\rho < \frac{8 \log(1/\delta)}{m}$, we obtain a trivial bound on (B):

$$\begin{aligned} -(B) = V(\pi^B; \mathcal{X}_{rem}) - V(\phi; \mathcal{X}_{rem}) &= \mathbb{E}\Big[\mathbb{1}\{X \in \mathcal{X}_{rem}\}\ \big(\pi^B(X) - \phi(X)\big)\big(Y(1) - Y(0)\big)\Big] \\ &= \mathbb{E}\Big[\big(\pi^B(X) - \phi(X)\big)\big(Y(1) - Y(0)\big) \;\big|\; X \in \mathcal{X}_{rem}\Big]\ \mathbb{P}(X \in \mathcal{X}_{rem}) \\ &\leq \mathbb{P}(X \in \mathcal{X}_{rem}) && \text{(bounded } Y) \\ &= \;\rho \;\leq\; \frac{8 \log(1/\delta)}{m} \;\lesssim\; \frac{8 \log(1/\delta)}{n}, \end{aligned}$$

in which case the conclusions of Theorems 4.1 and 4.2 follow immediately.

Otherwise, when $\rho \geq \frac{8 \log(1/\delta)}{m}$, the Chernoff bound implies that with probability at least $1 - \delta$,

$$N \;\geq\; \frac{m\rho}{2} = \frac{n}{6}\, \mathbb{P}\{X \in \mathcal{X}_{rem}\}. \tag{13}$$

Proof of Theorem 4.1 By finite combinatorial diameter D, the set of points on which policies in Π can disagree has size at most D; since $\mathcal{X}_{rem}$ collects (projected) disagreements w.r.t. $\widehat{\pi}$, we have $|\mathcal{X}_{rem}| \leq D$, hence the number of labelings on $\mathcal{X}_{rem}$ is $2^{|\mathcal{X}_{rem}|} \leq 2^D$, so we may run EWM over the finite class Π_{rem} (all $2^{|\mathcal{X}_{rem}|}$ labelings) and, in particular, the Bayes rule restricted to $\mathcal{X}_{rem}$, $\pi^B(x) = \mathbb{1}\{\tau_o(x) > 0\}$, belongs to Π_{rem}.

$$\begin{aligned} V(\pi^B; \mathcal{X}_{rem}) - V(\phi; \mathcal{X}_{rem}) &= \mathbb{E}\left[\mathbb{1}\{X \in \mathcal{X}_{rem}\}\ \big(\pi^B(X) - \phi(X)\big)\big(Y(1) - Y(0)\big)\right] \\ &= \mathbb{E}\left[\big(\pi^B(X) - \phi(X)\big)\big(Y(1) - Y(0)\big) \;\big|\; X \in \mathcal{X}_{rem}\right] \mathbb{P}(X \in \mathcal{X}_{rem}) \end{aligned}$$

Hence we can invoke Theorem 2.3 of Kitagawa and Tetenov (2018), which establishes fast rates for policy learning under a (soft) margin condition. Our setting imposes a hard margin, which is a special case of theirs (formally, take the soft–margin parameter $\alpha \to \infty$). In fact, for a finite policy class, one may use the Bernstein inequality for each policy and union bound, instead of using uniform bounds for VC classes. We obtain using (13):

$$\mathbb{E}\left[\big(\pi^B(X) - \phi(X)\big)\big(Y(1) - Y(0)\big) \;\big|\; X \in \mathcal{X}_{rem}\right] \lesssim \frac{\log \frac{2^D}{\delta}}{\kappa^2\, h}\ \frac{1}{n\, \mathbb{P}\{X \in \mathcal{X}_{rem}\}}$$

Hence, we finally get,

$$V(\pi^B; \mathcal{X}_{rem}) - V(\phi; \mathcal{X}_{rem}) \lesssim \frac{D + \log \frac{1}{\delta}}{\kappa^2\, h\, n}$$

Proof of Theorem 4.2 We start from the decomposition over the abstention region:

$$V(\pi^B;\mathcal{X}_{rem}) - V(\phi;\mathcal{X}_{rem}) = \mathbb{E}\Big[\mathbb{1}\{X\in\mathcal{X}_{rem}\}\,\big(\pi^B(X)-\phi(X)\big)\big(Y(1)-Y(0)\big)\Big]$$
$$= \mathbb{E}\Big[\big(\pi^B(X)-\phi(X)\big)\big(Y(1)-Y(0)\big)\;\Big|\; X\in\mathcal{X}_{rem}\Big]\,\mathbb{P}(X\in\mathcal{X}_{rem}). \tag{14}$$

Since $\pi^B(x)=\mathbb{1}\{\tau_o(x)>0\}$ and $\phi(x)=\mathbb{1}\{\widehat{\tau}(x)>0\}$, we can write

$$\mathbb{E}\Big[\big(\pi^B(X)-\phi(X)\big)\big(Y(1)-Y(0)\big)\;\Big|\; X\in\mathcal{X}_{rem}\Big] = \mathbb{E}\Big[\mathbb{1}\{\mathrm{sign}(\tau_o(X))\neq\mathrm{sign}(\widehat{\tau}(X))\}\,|\tau_o(X)|\;\Big|\;X\in\mathcal{X}_{rem}\Big].$$

By the hard margin assumption $|\tau_o(X)|\geq h$, whenever the signs of τ_o and $\widehat{\tau}$ disagree, they are bound to be separated by at least h gap.

$$\mathbb{E}\left[\mathbb{1}\{\mathrm{sign}(\tau_o(X))\neq\mathrm{sign}(\widehat{\tau}(X))\}\,|\tau_o(X)|\;\middle|\;X\in\mathcal{X}_{rem}\right]$$
$$= \mathbb{E}\left[\mathbb{1}\{\mathrm{sign}(\tau_o(X))\neq\mathrm{sign}(\widehat{\tau}(X))\}\;\mathbb{1}\{|\widehat{\tau}(X)-\tau_o(X)|\geq h\}\;|\tau_o(X)|\;\middle|\;X\in\mathcal{X}_{rem}\right]$$

Further we write

$$\mathbb{E}\left[\mathbb{1}\{\mathrm{sign}(\tau_o(X))\neq\mathrm{sign}(\widehat{\tau}(X))\}\,\mathbb{1}\{|\widehat{\tau}(X)-\tau_o(X)|\geq h\}\;|\tau_o(X)|\;\middle|\;X\in\mathcal{X}_{rem}\right]$$
$$\leq \mathbb{E}\left[\mathbb{1}\{|\widehat{\tau}(X)-\tau_o(X)|\geq h\}\;|\widehat{\tau}(X)-\tau_o(X)|\;\middle|\;X\in\mathcal{X}_{rem}\right]$$

To bound the last expectation, use the elementary inequality $\mathbb{1}\{|u|\geq a\}\;|u|\;\leq\;\frac{u^2}{a}$ for all $u\in\mathbb{R},\; a>0$.

$$\mathbb{E}\left[\mathbb{1}\{|\widehat{\tau}(X)-\tau_o(X)|\geq h\}\;|\widehat{\tau}(X)-\tau_o(X)|\;\middle|\;X\in\mathcal{X}_{rem}\right] \leq \frac{2}{h}\,\mathbb{E}\left[(\widehat{\tau}(X)-\tau_o(X))^2\;\middle|\;X\in\mathcal{X}_{rem}\right]$$
$$\lesssim \frac{c_\delta}{h}\,(n\;\mathbb{P}(X\in\mathcal{X}_{rem}))^{-2\beta} \tag{15}$$

The last inequality follows via (13) and CATE Oracle rate. Next, we bound the mass of the abstention region. By construction of $\mathcal{X}_{rem}$ there exists $\pi\in\widehat{\Pi}$ such that $\mathbb{1}\{X\in\mathcal{X}_{rem}\}=\mathbb{1}\{\pi(X)\neq\widehat{\pi}(X)\}$.

$$\begin{aligned}
\mathbb{P}\left(X\in\mathcal{X}_{rem}\right) &= \mathbb{E}\left[\mathbb{1}\{\pi(X)\neq\widehat{\pi}(X)\}\right]\\
&= \mathbb{E}\left[\mathbb{1}\big\{\pi(X)\neq\pi^B(X)\big\}\right]+\mathbb{E}\left[\mathbb{1}\big\{\widehat{\pi}(X)\neq\pi^B(X)\big\}\right]\\
&\leq \frac{1}{h}\left(E\left[\mathbb{1}\big\{\pi(X)\neq\pi^B(X)\big\}\,h\right]+\mathbb{E}\left[\mathbb{1}\big\{\widehat{\pi}(X)\neq\pi^B(X)\big\}\,h\right]\right)\\
&\leq \frac{1}{h}\left(E\left[\mathbb{1}\big\{\pi(X)\neq\pi^B(X)\big\}\,|\tau_o(X)|\right]+\mathbb{E}\left[\mathbb{1}\big\{\widehat{\pi}(X)\neq\pi^B(X)\big\}\,|\tau_o(X)|\right]\right)\\
&= \frac{1}{h}\left(V(\pi^B)-V(\pi)+V(\pi^B)-V(\widehat{\pi})\right)\\
&\lesssim \frac{1}{h}\left(V(\pi^*(X))-V(\pi^B(X))+\frac{\alpha}{\kappa}\right)
\end{aligned}$$

using $|\tau_o|\geq h$ and that $\widehat{\pi},\pi$ belong in the almost optimal policy set $\widehat{\Pi}$ (so $V(\pi^*)-V(\widehat{\pi})\lesssim\alpha/\kappa$ with $\alpha=\frac{1}{\kappa}\sqrt{\frac{d\log\frac{n}{d}+\log\frac{1}{\delta}}{n}}$.

Finally, insert (15) into (14) and expand the right–hand side using the bound on $\mathbb{P}(X\in\mathcal{X}_{rem})$. Collecting the terms arising from (i) $V^{(p)}$ regret, (ii) the oracle CATE estimation on $\mathcal{X}_{rem}$ and the bound on $\mathbb{P}(X\in\mathcal{X}_{rem})$ yields the three contributions $\mathrm{Reg}_1,\mathrm{Reg}_2,\mathrm{Reg}_3$ stated in the theorem.

C.2 Proof of Proposition 4.3

We show that the policy abstention setting that we introduced has a natural relationship with distributionally robust policy learning. Specifically, consider a setting where the true data distribution of $(Y(0), Y(1)) \mid X$ is different from our training data. This would happen when our observational data is outdated and cannot reflect the true effect of the current treatment of interest. For instance, we may want to find an optimal subpopulation for up-to-date vaccination, but we only have outcome data for an older version of the vaccine, that might be close to but different from the potential outcomes of the latest vaccine.

While deterministic policies are optimal without such an outcome distribution shift, they can be problematic otherwise. Intuitively, deterministic policies tend to "put all eggs in one basket", rendering them more vulnerable to any potential systematic shift of the potential outcomes. To mathematically formulate this intuition, we assume the true potential outcome distribution lies in some W_1-ball of the distribution that generates our observations, *i.e.*

$$\mathbb{P}_{\texttt{test}} \in \mathcal{P}_\alpha(\mathbb{P}_{\texttt{train}}) := \{\mathbb{P} : W_1(\mathbb{P}, \mathbb{P}_{\texttt{train}}) \le \alpha\}$$

and we would like to maximize the worst case value of a policy induced by the ambiguity set $\mathcal{P}_\alpha(\mathbb{P}_{\texttt{train}})$.

Proof of Proposition 4.3. Fix X and write $\mu_d(x) = \mathbb{E}[Y(d) \mid X = x]$, $d \in \{0,1\}$, and define, for a policy $\tilde{\pi} : X \to \{0, 1, *\}$, the conditional reward at x by

$$\begin{aligned} r_{\tilde{\pi}}(x) &= \mathbf{1}\{\tilde{\pi}(x) \neq *\}[\tilde{\pi}(x)\mu_1(x) + (1 - \tilde{\pi}(x))\mu_0(x)] \\ &\quad + \mathbf{1}\{\tilde{\pi}(x) = *\} \cdot 0.5(\mu_0(x) + \mu_1(x)); \end{aligned}$$

let $P_\alpha(P_{\text{train}}) = \{P : W_1(P, P_{\text{train}}) \le \alpha\}$ with the W_1 ball taken on the $(Y(0), Y(1))$ coordinates and the X-marginal fixed at P_X^*, so $V(\pi) = \mathbb{E}_P[r_\pi(X)]$ and

$$\min_{P \in P_\alpha} V(\pi) = \mathbb{E}_{P_X^*}\left[\min_{P_X(x)} \mathbb{E}_P[r_\pi(X) \mid X = x]\right].$$

By Kantorovich–Rubinstein duality for the ℓ_1 ground metric on $(Y(0), Y(1))$, for any affine $g(u, v) = au + bv$ we have

$$\inf_{W_1(P, P_{\text{train}}) \le \alpha} \mathbb{E}_P[g(Y(0), Y(1))] = \mathbb{E}_{P_{\text{train}}}[g] - \alpha\|(a, b)\|_\infty.$$

Apply this pointwise at x with g chosen according to $\pi(x)$, this becomes

$$\pi(x)\mathbb{E}[Y(1) \mid X = x] + (1 - \pi(x))\mathbb{E}[Y(0) \mid X = x] - \alpha \max\{\pi(x), 1 - \pi(x)\}.$$

Since this is a linear function on $[0, 0.5]$ and $[0.5, 1]$, it must attains maximum value in $\{0, 0.5, 1\}$. If $\pi(x) = 1$ then $g(u, v) = v$ and $\|(0,1)\|_\infty = 1$, if $\pi(x) = 0$ then $g(u, v) = u$ and $\|(1,0)\|_\infty = 1$, and if $\pi(x) = 0.5$ then $g(u, v) = 0.5(u + v)$ and $\|(0.5, 0.5)\|_\infty = 0.5$; hence for every x,

$$\begin{aligned} \min_{P_X(x)} \mathbb{E}_P[r_\pi(X) \mid X = x] &= \mathbf{1}\{\pi(x) \neq 0.5\}[\pi(x)\mu_1(x) + (1 - \pi(x))\mu_0(x) - \alpha] \\ &\quad + \mathbf{1}\{\pi(x) = 0.5\}[0.5(\mu_0(x) + \mu_1(x)) - \alpha/2] \\ &= \mathbf{1}\{\tilde{\pi}(x) \neq 0.5\}[\tilde{\pi}(x)\mu_1(x) + (1 - \tilde{\pi}(x))\mu_0(x) - \alpha] \\ &\quad + \mathbf{1}\{\tilde{\pi}(x) = *\}[0.5(\mu_0(x) + \mu_1(x)) - \alpha/2] \end{aligned}$$

Taking expectation over X yields

$$\begin{aligned} \min_{P \in P_\alpha} V(\pi) &= \mathbb{E}[\mathbf{1}\{\tilde{\pi}(X) \neq *\}(\tilde{\pi}\mu_1 + (1 - \tilde{\pi})\mu_0) + \mathbf{1}\{\tilde{\pi}(X) = *\} \cdot 0.5(\mu_0 + \mu_1)] \\ &\quad - \alpha \cdot \mathbb{P}(\tilde{\pi}(X) \neq *) - (\alpha/2) \cdot \mathbb{P}(\tilde{\pi}(X) = *). \end{aligned}$$

Comparing with the abstention objective

$$V^{(p)}(\pi) = \mathbb{E}[\mathbf{1}\{\tilde{\pi} \neq *\}(\pi\mu_1 + (1 - \pi)\mu_0) + \mathbf{1}\{\tilde{\pi} = *\}(0.5(\mu_0 + \mu_1) + p)]$$

at $p = \alpha/2$ gives $\min_{P \in P_\alpha} V(\pi) = V^{(\alpha/2)}(\pi) - \alpha$, because the difference at each x is α in both the non-abstain case (no bonus vs $-\alpha$) and the abstain case ($+\alpha/2$ vs $-\alpha/2$). $\square$

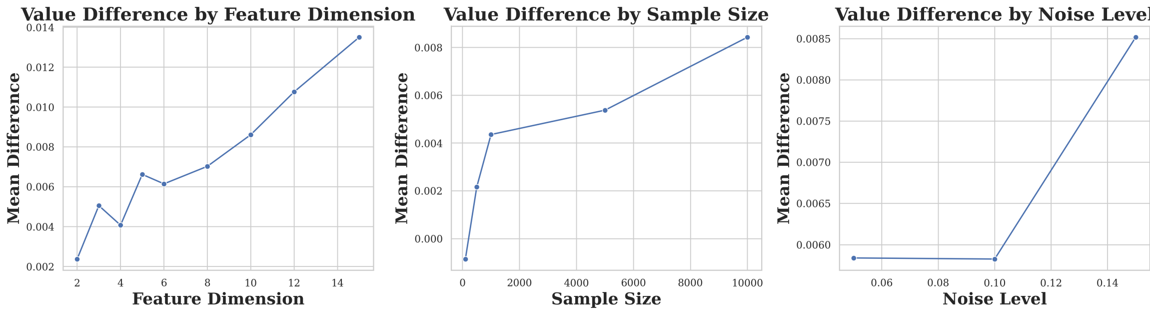

Figure 3: Mean value difference under abstention.

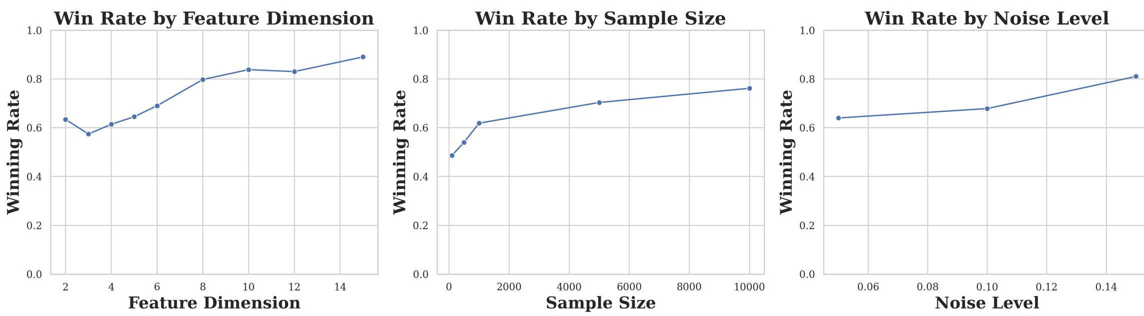

Figure 4: Winning rate of Algorithm 1 compared with EWM.

D Experiment Details

In this section, we provide further details of the experiments.

D.1 Policy Learning with Abstention

Experimental setting. We study Algorithm 1 on synthetic policy learning problems with known ground truth. For each problem instance, X are drawn i.i.d. (standard normal), treatment is assigned with an X-dependent logistic propensity in $[0.1, 0.9]$, and observed outcomes are $Y = Y(D) + \epsilon$ that lie in $[0, 1]$. We consider three reward regimes (linear, nonlinear, complex), multiple noise levels, and a range of feature dimensions and sample sizes.

Policy class and hyperparameters. The base class Π contains simple threshold policies (including linear-threshold with intercept). We follow Algorithm 1 and select an empirical-welfare maximizer from Π. Algorithm parameters are set to be $\kappa = 0.1$, confidence $\delta = 0.05$, and abstention bonus $p = 0.05$.

Evaluation protocol. For each configuration we report: (i) Monte-Carlo ground-truth $V(p)(\pi)$ computed with many draws; (ii) IPW estimates on observed data; (iii) abstention rate. We run multiple replications per configuration to produce reliable results.

From Figures 3 and 4, we can see that under all parameter configurations, abstention beats EWM in most instances, with positive mean difference relative to EWM, indicating its superiority.

D.2 Safe Policy Learning

We evaluate safe policy improvement methods that must avoid degrading a fixed baseline ω. Experiments are conducted under varying reward variance and baseline optimality gap.

Data-generating process. For each dataset we sample $X \sim U[0,1]^d (d = 5)$. The CATE is chosen to be $\tau(X) = 2(X_1 + X_2 - 1)$; and the potential outcomes are $Y(0) = X_3 + \epsilon, Y(1) = Y(0) + \tau(X), \epsilon \sim N(0, \sigma^2)$; treatment $D \sim \text{Bernoulli}(p(X))$, with $p(X)$ either constant or logistic in X, clipped to $[0.1, 0.9]$. The true optimal policy is $\pi^*(x) = 1\{\tau(X) > 0\}$.

Baselines. We compare with several baseline algorithms for safe policy learning: (i) Safe EWM: direct EWM followed by comparing its estimated value with that of the baseline policy, and (ii) two versions (t-test LCB and clipped-CI) of the HCPI algorithms proposed in Thomas et al. (2015).

For each parameter value and sample size $n \in \{200, 500, 1000, 2000, 5000\}$, we run 500 independent replications. For each method we compute ground-truth value using $Y(0), Y(1)$ and report mean true-value gain $V(\pi_{chosen}) - V(\omega)$, rate of mistake $V(\pi_{chosen}) < V(\omega)$, and rate of improvement $P[V(\pi_{chosen}) > V(\omega)]$. Results are aggregated across runs and visualized as in Figures 1 and 2 respectively.