MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Minimax-Optimal Policy Learning Under Unobserved Confounding

Nathan Kallus, Angela Zhou
Cornell University, New York, NY 10044, {kallus,az434}@cornell.edu

We study the problem of learning personalized decision policies from observational data while accounting for possible unobserved confounding. Previous approaches, which assume unconfoundedness, i.e., that no unobserved confounders affect both the treatment assignment as well as outcome, can lead to policies that introduce harm rather than benefit when some unobserved confounding is present, as is generally the case with observational data. Instead, since policy value and regret may not be point-identifiable, we study a method that minimizes the worst-case estimated regret of a candidate policy against a baseline policy over an uncertainty set for propensity weights that controls the extent of unobserved confounding. We prove generalization guarantees that ensure our policy will be safe when applied in practice and will in fact obtain the best-possible uniform control on the range of all possible population regrets that agree with the possible extent of confounding. We develop efficient algorithmic solutions to compute this minimax-optimal policy. Finally, we assess and compare our methods on synthetic and semi-synthetic data. In particular, we consider a case study on personalizing hormone replacement therapy based on observational data, where we validate our results on a randomized experiment. We demonstrate that hidden confounding can hinder existing policy learning approaches and lead to unwarranted harm, while our robust approach guarantees safety and focuses on well-evidenced improvement, a necessity for making personalized treatment policies learned from observational data reliable in practice.

1. Introduction

The problem of learning personalized decision policies to study "what works and for whom" in areas such as medicine, e-commerce, and civics often endeavors to draw insights from increasingly rich and plentiful observational data, such as electronic medical records (EMRs), since data from randomized controlled experiments may be scarce, costly, or unethical to acquire. A variety of methods have been proposed to address the corresponding problem of *policy learning from observational data* (Beygelzimer and Langford 2009, Dudik et al. 2014, Kallus 2017a,b, Kallus and Zhou 2018b,

Kitagawa and Tetenov 2018, Wager and Athey 2017a). These methods, as well as approaches to predict conditional-average treatment effects from observational data (Künzel et al. 2017, Nie and Wager 2017, Shalit et al. 2017, Wager and Athey 2017b), operate under the controversial assumption of *unconfoundedness*. Stated informally, such an assumption requires that the data are sufficiently informative such that there remain no unobserved confounders that jointly affect treatment assignment and individual response (Rubin 1974), effectively requiring that assignment is *as if at random* once we control for observables. This key assumption may be always made to hold *ex ante* by directly controlling the treatment assignment policy as in a randomized controlled experiment, but in other domains of key interest such as personalized medicine where EMRs are increasingly being analyzed *ex post*, unconfoundedness is an assumption that may never truly fully hold in fact. Even in randomized controlled trials, in practice, challenges such as compliance, censoring, or even site selection bias may lead to confounding.

Assuming unconfoundedness, also called *ignorability*, *conditional exogeneity*, or *selection on observables*, is controversial because it is fundamentally unverifiable since the counterfactual distribution is never identified from the data (Imbens and Rubin 2015). Thus, insights from observational studies, which passively study treatment-outcome data without directly intervening on treatment are always vulnerable to this fundamental critique. For example, studying drug efficacy by assessing outcomes of those prescribed the drug during the course of normal clinical practice may make a drug look less clinically effective if those who were prescribed the drug were sicker to begin with and therefore would have had worse outcomes regardless. Conversely, if the drug was correctly prescribed only to the patients who would most benefit from it, it may make the drug appear to be falsely effective for all patients. These issues can potentially be alleviated by controlling for more baseline factors that may have affected treatment choices but they can never really be fully eliminated in practice.

Conclusions drawn from healthcare databases such as claims data are particularly vulnerable to unobserved confounding because although they record administrative interactions and diagnostic codes, they are uninformative about medical histories, notes on patient severity, observations, and monitoring of clinical outcomes, *i.e.*, the key clinical information which may drive a physician's treatment choices. EMRs provide great promise for enabling richer personalized medicine from observational data because they record the entire patient treatment and diagnostic history, past medical history and comorbidities, as well as fine-grained information regarding patient response such as vital signs (Hoffman and Williams 2011). The growing adoption of richer EMRs can both provide higher precision for personalized treatment and render unconfoundedness more plausible, since the data includes more of the information regarding patient history and outcomes that informs physician decision-making, yet unconfoundedness, an ideal stylized assumption, still may never be fully satisfied in practice.

1.1. Unobserved Confounding: the Example of the Women's Health Initiative Parallel Clinical Trial and Observational Study

The challenges of observational data are of course not new to the modern era of data-driven decision-making, but have been widely recognized. One high-profile example is the case of the parallel Women's Health Initiative (WHI) observational study and clinical trial, which illustrates how confounding factors can lead to dramatic discrepancies in drawing clinically relevant prescriptions from randomized trial versus observational data. The parallel WHI observational study and clinical trials studied whether hormone replacement therapy (HRT) had therapeutic benefits for chronic disease prevention. While HRT was known to be clinically effective for vasomotor symptoms of menopause, earlier observational epidemiological studies additionally suggested a protective benefit against coronary heart disease (CHD) which led to the increasing clinical practice of prescribing HRT in menopause for preventive purposes (without clinical trial evidence) (Pedersen and Ottesen 2003). The parallel WHI observational study and clinical trial were designed to evaluate the efficacy of HRT in a preventive context on chronic disease, such as coronary heart disease (CHD) and breast cancer, among other clinical endpoints. Ultimately, the WHI clinical trial dramatically repudiated these purported therapeutic benefits. In fact, while the observational study suggested a protective benefit of HRT against CHD, showing a 40-50% reduction in CHD incidence, the HRT arm of the clinical trial had to be stopped early due to a dangerously elevated incidence of CHD (Prentice et al. 2005). After the WHI study, the new evidence that arose not only dramatically changed the standard of care, spurring an 80% reduction in the prescription of HRT, but also sparked a broader methodological debate about the clinical credibility of observational studies (Lawlor et al. 2004). Later in Section 7.2, we build a case-study with semi-synthetic data from the observational study and clinical trial to illustrate potential harms of policy learning from realistically confounded data. This case study, as well as others, illustrate the challenges of unobserved confounders that would continue to plague richer data-driven decision-making strategies such as personalized policy learning.

We briefly overview the range of possible unobserved confounders which were posited to reconcile the different findings from WHI. The observational study may have been confounded by plausible, well-recognized confounding phenomena, *healthy user bias* due to self-selection and *confounding by indication* due to expert-selection, which pose general challenges to the validity of research on observational health databases, and which may induce correlation in either direction between treatment selection and outcomes (Brookhart et al. 2010). Such possible confounding factors are inherent in healthcare data in which physicians determined treatment assignment to manage health outcomes in the first place. *Healthy user bias* may stem from differing lifestyle factors in the population of women self-selecting into HRT: general health-seeking behaviors correlated with selection into treatment, such as exercise or maintaining heart-healthy diets, are correlated with better expected

outcomes related to CHD on average. These same lifestyle factors tend to reduce artherosclerotic risk and risk of CHDs, but are unobserved confounders for self-enrollment into HRT. Conversely, the study may have also been *confounded by indication* or severity, where the presence of clinical activities such as prescription of HRT is correlated with, or indicates, greater initial symptom severity, which may lead to attenuation in the perceived reduction in vasomotor symptoms.

1.2. Unobserved Confounding in Other Problem Settings

We discuss the relevance of unobserved confounders in other managerial settings to highlight the broader relevance of unobserved confounding. (For discussion of causal inference models in operations management, see Ho et al. (2017).) Unobserved confounders accompany the growing use of transactional-level data either due to confounding introduced by previous managerial decisions (in analogy to provider expertise in healthcare), or private information of individuals whose interactions comprise a dataset (in analogy to self-selection). Operational decisions were historically made to improve firm outcomes: previous decisions incorporate managerial discretion or expertise that is correlated with outcomes of interest, introducing unobserved confounding. Gordon et al. (2019) find that conclusions from large advertising experiments at Facebook and observational counterparts on advertising effectiveness may differ in general, and conduct sensitivity analysis. A randomized trial of the effectiveness of search ads on eBay (Blake et al. 2015) revealed the spurious efficacy of advertising, based on observational studies of user search queries, which did not account for unobserved intent or customer loyalty.

1.3. Contributions

Because unconfoundedness may fail to hold, existing policy learning methods that operate under this assumption can lead to personalized decision policies that seek to exploit individual-level effects that are not really there, may intervene where not necessary, and may in fact lead to net harm rather than net good. Such dangers constitute obvious impediments to the use of policy learning to enhance decision making in such sensitive applications as medicine, public policy, and civics, where reliable and safe algorithms are critical to implementation. Clearly, a policy that could potentially introduce additional harm, toxicity, or risk to patients compared to current standards of care is an unacceptable replacement, and an algorithm that could potentially give rise to such a policy is unusable in medical and other sensitive settings.

To address the deficiencies of policy learning that requires untenable assumptions of unconfoundedness, in this paper we develop a framework for minimax-optimal policy learning which ensures that the personalized decision policy derived from observational data, which inevitably will have *some* unobserved confounding, will do no worse than a current policy such as the current standard of care and, in fact, will do better if the data can indeed support it. We do so by requiring that the

learned policy improve upon the baseline no matter the direction of potential unobserved confounding which generated the data. Thus, we calibrate personalized decision policies to address sensitivity to realistic violations of the unconfoundedness assumption. For the purposes of informing reliable and personalized decision-making that leverages modern machine learning, our work highlights that statistical point identification of individual-level causal effects, which previous approaches crucially rely on, may not at all be necessary for successfully learning effective policies that reliably improve on unpersonalized standards of care, but accounting for the lack of point identification is necessary.

Functionally, our approach is to optimize a policy to achieve the best worst-case improvement relative to a baseline treatment assignment policy (such as treat all or treat none), where the improvement is measured using a weighted average of outcomes and weights which take values in an uncertainty set around the *nominal*, or *observed* inverse propensity weights (IPW). This generalizes the popular class of IPW-based approaches to policy learning, which optimize an unbiased estimator for policy value under unconfoundedness (Beygelzimer and Langford 2009, Kitagawa and Tetenov 2018, Li et al. 2011, Swaminathan and Joachims 2015a,b). Unlike standard approaches, in our approach the choice of baseline is material and changes the resulting policy chosen by our method. This framing supports reliable decision-making in practice, as often a practitioner is seeking evidence of substantial improvement upon the standard of care or a default option, and/or the intervention under consideration introduces risk of toxicity or adverse effects and should not be applied without strong evidence.

Our contributions are as follows. We provide a framework for performing *minimax-optimal policy learning* that is robust in the face of unobserved confounding by using a robust optimization formulation. Our framework allows for the specification of data-driven uncertainty sets based on a sensitivity parameter describing a pointwise bound on the odds ratio between true and nominal (observed) propensities as well as uncertainty sets with a global budget-of-uncertainty parameter. Whereas previous approaches for sensitivity analysis in causal inference focus on evaluating the *range* of inferential procedures (e.g. effect estimation or hypothesis tests), we focus on the question of learning *minimax-optimal decision policies* in the presence of unmeasured confounding. Sensitivity models in causal inference introduce ambiguity sets in the space of inverse propensity weights which do not vanish with increasing data. Thus, learning decision policies under sensitivity models introduces analytical challenges in ensuring convergence. We prove a uniform convergence result both over the space of policies of restricted complexity and over the possible confounded data-generating distributions in our uncertainty set: therefore, our approach is asymptotically optimal for the population minimax regret. These results also imply an appealing improvement guarantee that shows that, up to vanishing factors that depend on the complexity of the policy class, our approach will not do worse than the baseline and, moreover, will do better, as can be easily validated

by simply evaluating the objective value of our optimization problem. Leveraging the structure of our optimization problem and characterizing the inner subproblem, we provide a set of efficient algorithms for performing robust policy optimization over parameterized policy classes and over decision trees. We assess performance on a synthetic example that illustrates the possible benefits of our approach and the effect of the uncertainty parameters. We then show, in a case study drawing on the unique simultaneous WHI observational study and clinical trial, that in regimes with realistic confounding, for a variety of possible treatment effect profiles, our approach can lead to improvement upon a baseline while learning from confounded data causes harm. This case study allows us to uniquely learn from observational data with unobserved confounding, yet assess out of sample performance on an unconfounded clinical trial.[1]

2. Problem Statement and Preliminaries

We first summarize the setup. We consider policy learning from observational data consisting of tuples of random variables $\{(X_i, T_i, Y_i) : i = 1, \dots, n\}$, comprising of covariates $X_i \in \mathcal{X}$, assigned treatment level out of m discrete treatments $T_i \in \{0, \dots, m-1\}$, and real-valued outcomes $Y_i \in \mathbb{R}$. We suppose that these constitute iid (independent and identically distributed) observations from a population and we drop subscripts to denote a generic draw from this population. We allow $m \geq 2$, so that we accommodate the case of multiple, discrete treatment levels. We let $Y_i(0), \dots, Y_i(m-1)$ denote the potential outcomes of applying each treatment option, respectively, and we assume that $Y_i = Y_i(T_i)$ so that the observed outcome corresponds to the potential outcome of the observed treatment.[2]. We let $\mathbb{E}_n$ denote the empirical expectation, i.e. taking a sample average over the data. We define the index set for treatment value t as $\mathcal{I}_t = \{i \leq n : T_i = t\}$. We use the convention that the outcomes Y_i corresponds to losses so that lower outcomes are better.

We denote the *nominal* propensity function by $\tilde{e}_t(x) = \mathbb{P}(T = t \mid X = x)$ and the nominal generalized propensity score by $\tilde{e}_{T_i}(X_i)$. This can be estimated directly from the data using a probabilistic classification model such as logistic regression or a neural network. When it is estimated, we denote

[1] The present paper builds upon an earlier paper by the authors (Kallus and Zhou 2018a). The method proposed herein is *distinct* and uses per-treatment weight normalization, which provides sharp regret bounds in both the binary- and multiple-treatment settings and enables the extension to multiple treatments. For this new method we provide new theoretical guarantees on minimax optimality, or that the policy we learn performs similarly to the one that provides the best-possible uniform control on the range of possible true regrets of the policy, and we extend previous theoretical guarantees to this new method. We provide a new generic conic-optimization-based formulation of the optimization problem that underlies the method. We provide practical tools for calibrating the sensitivity parameter in our policy learning setting. And, we introduce a new case study on hormone replacement therapy using data from the Women's Health Initiative parallel observational study and clinical trial.

[2] The equation $Y_i = Y_i(T_i)$ captures two important features. One is that the observed outcomes are consistent with the hypothetical potential outcomes. Another is that the outcome of an individual only depends on the treatment assignment of that individual and there is no interference between units. This two assumptions together are also known as the *stable unit treatment value assumption* (Rubin 1980).

the estimated nominal propensity function by $\hat{e}_t(x)$. Since we do not assume unconfoundedness, the nominal propensity is insufficient to account for confounding. We therefore additionally define the *true* propensity function as $e_t(x,y) = \mathbb{P}(T=t \mid X=x, Y(t)=y)$ and the true generalized propensity score as $e_T(X,Y)$. Note that these *cannot* be estimated from the data. Unconfoundedness (weak ignorability) is the assumption that $\tilde{e}_t(x) = e_t(x,y)$ as functions (i.e., $\mathbb{I}[T=t] \perp\!\!\!\perp Y(t) \mid X$). Here, we do *not* assume unconfoundedness and will generally have that $e_t(x) \neq e_t(x,y)$.

We consider evaluating and learning a (possibly) randomized policy mapping covariates to the probability of assigning treatment, $\pi : (t,x) \in \{0,\dots,m-1\} \times \mathcal{X} \mapsto [0,1]$, where Δ^m denotes the m-simplex. Given a policy π, we use the notation $\pi(t \mid x)$ to denote the probability π assigns to treatment t when observing covariates x. It is also convenient to also define the random treatment variable Z^π that, given X, is independent of all else, and has the distribution $\mathbb{P}(Z^\pi = t \mid X) = \pi(t \mid X)$. The policy value of π is $V(\pi) = \mathbb{E}\left[\sum_{t=0}^{m-1} \pi(t \mid X) Y(t)\right] = \mathbb{E}[Y(Z^\pi)]$. As is common for policy learning (e.g., Kallus 2017b, Wager and Athey 2017a), we focus on a restricted policy class $\Pi \subseteq [\mathcal{X} \to \Delta^m]$. Examples include deterministic linear policies, $\pi_{\alpha_{0:m-1},\beta_{0:m-1}}(t(x) \mid x) = 1$ where $t(x) \in \arg\max_{t=0,\dots,m-1} \alpha_t + \beta_t^\intercal x$; logistic policies, $\pi_{\alpha_{0:m-1},\beta_{0:m-1}}(t \mid x) \propto \exp(\alpha_t + \beta_t^\intercal x)$; or decision trees of a bounded depth, which assign any probability vector to each leaf of the tree.

3. Related Work

Our work builds upon several strands of literatures, notably policy learning from observational data as well as sensitivity analysis in causal inference.

Causal inference for personalization from observational data under unconfoundedness. The key difficulty in learning *interventional* effects from observational data is that the outcome $Y_i(T_i)$ is only observed for the treatment actually administered historically to the unit, T_i, whose assignment can itself be correlated with the potential outcomes, obfuscating differences in them. Since the data is observational and the treatment assignment procedure was not under the control of the experimenter, the distribution of covariates may be systematically different between treatment and control groups due to self selection of the individuals into treatments, medical imperatives trading off treatment risk vs. patient severity, or business imperatives to offer discounts or target advertising not completely at random. Thus, the systematic differences in covariates in the population $\mathbb{P}(X=x, Y=y \mid T=1), \mathbb{P}(X=x, Y=y \mid T=0)$, also known as *covariate shift*, make the treated and untreated populations incomparable for the purpose of assessing effect.

When *all* covariates needed to ensure unconfoundedness are assumed to be observed, *i.e*, $\tilde{e}_t(x) = e_t(x,y)$, then a variety of approaches for learning personalized intervention policies that maximize causal effect have been proposed. These fall under regression-based strategies (Bertsimas et al. 2016, Qian and Murphy 2011), reweighting-based strategies (Beygelzimer and Langford 2009, Kallus 2017a,

Kitagawa and Tetenov 2018, Swaminathan and Joachims 2015b), or doubly robust combinations thereof (Dudik et al. 2014, Wager and Athey 2017a). Regression-based strategies estimate the conditional average outcomes, $\mathbb{E}[Y(t) \mid X]$, which under unconfoundedness are equal to $\mathbb{E}[Y \mid X, T = t]$, a regression of outcome on covariates in the t-treated group. These estimates are either used directly to treat by picking the smallest value (known as *direct comparison*) or to score policies and pick the best in a restricted class (known as the *direct method*). For binary treatments, we can directly fit the difference $\mathbb{E}[Y(1) - Y(0) \mid X]$, known as the conditional average treatment effect (CATE) (Wager and Athey 2017b). If the regression functions are ill-specified, we are not guaranteed to find the best policy, even if the class is amenable to the estimation method (*e.g.*, the best linear policy does not arise from comparing the best linear CATE estimator to zero). Without unconfoundedness, the regression functions or CATE are not identifiable from the data (parametrically or non-parametrically) and these methods have no guarantees.

Reweighting-based strategies use inverse propensity weighting (IPW) (Beygelzimer and Langford 2009, Kallus 2017a, Kitagawa and Tetenov 2018, Swaminathan and Joachims 2015b) or covariate-balancing weights (Kallus 2017b) to change measure from the distribution induced by a historical logging policy to that induced by any new policy π. Specifically, these methods use the fact (Li et al. 2011) that, under unconfoundedness, $\hat{V}^{\mathrm{IPW}}(\pi)$ is unbiased for $V(\pi; \tilde{e}_T)$, where

$$\hat{V}^{\mathrm{IPW}}(\pi; \tilde{e}_T) = \sum_{t=0}^{m-1} \mathbb{E}_n \left[\frac{\pi(t \mid X) \mathbb{I}[T = t] Y}{\tilde{e}_t(X)} \right] \tag{1}$$

Optimizing $\hat{V}^{\mathrm{IPW}}(\pi)$ for deterministic policies can be phrased as a weighted classification problem (Beygelzimer and Langford 2009). Dudik et al. (2014) suggest to augment eq. (1) by using the doubly-robust estimator (Robins et al. 1994), which centers the outcomes using a regression estimate. Wager and Athey (2017a) show that since this estimate is semiparametrically efficient when using cross-fold fitting, as shown by Chernozhukov et al. (2016), this leads to better regret bounds. Since dividing by propensities can lead to extreme weights and high variance estimates, clipping the probabilities are typically necessary for good performance (Swaminathan and Joachims 2015a, Wang et al. 2017) or the use of weights that directly optimize for balance (Kallus 2017b). With or without any of those fixes, if there are unobserved confounders, then, neither a policy's value nor the optimal policy are identifiable, and any of these methods may lead to learned policies that may well introduce more harm than good. Under unconfoundedness, such reweighting-based methods are notable for being able to find best-in-class policies regardless of specification of an outcome model (or with outcome models learned at sub-parametric rates; Wager and Athey 2017a). Specifically, they focus directly on the policy learning problem rather than a prediction problem and on finding a policy that performs as the best in a given class. This leads to strong generalization guarantees (Kallus

2017b, Kitagawa and Tetenov 2018, Wager and Athey 2017a) and can also allow one to incorporate domain-specific constraints that favor simple prescriptive decision policies that are interpretable, implementable, and/or satisfy operational constraints, such as scorecards or decision-trees (Ustun and Rudin 2015). These constraints and approaches for training optimal constrained policies can be composed directly with the policy optimization problem by restricting the policy class. Because of these unique properties, our approach will also be based on a reweighting approach that directly optimizes a policy rather than a predictor.

The literature on optimal policy learning in econometrics has also considered a minimax regret criterion as summarized in Hirano and Porter (2019). Manski (2005, 2008) consider the optimal decision policy obtained by minimax regret bounds on conditional average outcomes, which arise from partial identification bounds on *arbitrary* confounding from the unidentified counterfactual probabilities: this approach is highly conservative and does not use available information on selection based on observables (namely, $\tilde{e}_t(x)$ which exists despite additional unobserved confounding.). Stoye (2009, 2012) consider minimax regret from a decision-theoretic point of view, where a closed form is available under limiting asymptotic assumptions on an experimental sampling design generating treatment assignments under a binary or Gaussian assumption on outcome models. In contrast to these lines of work, we are minimax-optimal with respect to a data-driven uncertainty set around the *estimable* inverse propensity weights, to assess *reasonable* violations of unconfoundedness, and our minimax-regret guarantees focus on uniform convergence over a policy class and a data-driven sensitivity model.

Policy improvement. A separate literature within reinforcement learning, unrelated to causal inference, considers the idea of safe policy improvement by forming an uncertainty set around the presumed unknown transition probabilities between states as in Thomas et al. (2015) or forming a trust region for safe policy exploration via concentration inequalities on estimates of policy risk as in Petrik et al. (2016). None of these consider the issue of confounding in the underlying action generation policy (the analogous propensity score) or observational data. This general approach of safely improving upon another policy using a robust or minimax formulation is related to the use of a baseline policy in our method.

Sensitivity analysis. Sensitivity analysis in causal inference tests the robustness of inferences about an average treatment effect made based on observational data to the violations of assumptions such as unconfoundedness. In contrast, our work focuses on personalized policy learning in the presence of unobserved confounding from an infinite family of potential policies. Some approaches from sensitivity analysis for assessing unconfoundedness require auxiliary data or additional structural assumptions, which we do not assume here (Imbens and Rubin 2015). Other approaches consider how large the unobserved confounding must be to invalidate the conclusions of statistical inference,

and typically consider assumptions restricting the strength of unobserved confounding, either on the selection process, or on the outcome model. For example, sensitivity analysis would assess the range of extremal p-values on the hypothesis of no effect for randomization inference, depending on the value of Γ so that consequent binary conclusions can be couched in terms of the level of unobserved confounding required to overturn a nominal conclusion (Fogarty and Small 2016, Hasegawa and Small 2017, Rosenbaum 2002). Our approach borrows the marginal sensitivity model from sensitivity analysis (Tan 2012), assuming bounds on the strength of unobserved confounding on selection into treatment, and focuses on the implications for personalized treatment decisions.

The Rosenbaum model for sensitivity analysis assesses the robustness of randomization inference to the presence of unobserved confounding by considering a uniform bound Γ on the *odds ratio* between $e_t(x,y)$ and $e_t(x,y')$, i.e., between the treatment propensities of any two units with equal covariates (Rosenbaum 2002). The closely related *marginal* sensitivity model, introduced by Tan (2012), considers a uniform bound Γ on the odds-ratio between the *nominal* propensity $e_t(x)$ and the true propensity $e_t(x,y)$. Zhao et al. (2019) provides further discussion on the relationship between the two sensitivity models. They are generally different and incomparable for equal values of Γ. The value of Γ can be calibrated against the discrepancies induced by omitting observed variables; then determining Γ can be phrased in terms of whether one thinks one has omitted a variable that could have increased or decreased the probability of treatment by as much as, say, gender or age can in the observed data (Hsu and Small 2013).

In the sampling literature, the Hájek estimator for population mean (Hájek 1971) is an extension of the classic Horvitz-Thompson estimator (Horvitz and Thompson 1952) that adds weight normalization. The objective of the minimax game we define between policy optimizer and possible confounding is a Hájek estimator for the policy value. Aronow and Lee (2012) derive sharp bounds on the estimator arising from a uniform bound on the sampling weights, showing a closed-form for the solution for a *uniform* bound on the sampling probabilities. Zhao et al. (2019) consider bounds on the Hájek estimator, but impose a parametric model on the treatment assignment probability. Miratrix et al. (2018) consider tightening the bounds from the Hájek estimator by adding shape constraints, such as log-concavity, on the cumulative distribution of outcomes. Masten and Poirier (2018) consider sup-norm bounds on propensity differences and show sharp partial identification of bounds for CATE and ATE by integrating partially identified bounds on the conditional quantile treatment effect. In contrast to the sensitivity analysis literature in causal inference, we focus on the implications of sensitivity analysis for learning a robust personalized policy function: minimax policy learning poses additional analytical challenges in ensuring convergence of data-driven robust policies.

4. Policy Learning That Is Robust to Unobserved Confounding

We now present our framework for minimax-optimal confounding-robust policy learning under unobserved confounding. Our approach minimizes a bound on policy regret against a specified baseline policy π_0, $R_{\pi_0}(\pi) = V(\pi) - V(\pi_0)$. Our bound is achieved by maximizing a reweighting-based regret estimate over an uncertainty set around the nominal propensities. This ensures that we cannot do any worse than π_0 and may in fact do better, even if the data is confounded.

The baseline policy π_0 can be any fixed policy that we want to make sure not to do worse than or deviate from unnecessarily. This is usually the current standard of care, established from prior evidence, and we would not want any algorithmic solution to personalization to introduce any harm relative to current standards. Generally, this is the policy that always assigns control, $\pi_0(0 \mid X) = 1$. Alternatively, if reliable clinical guidelines exist for some limited personalization, then $\pi_0(t \mid X)$ represents the non-constant function that encodes these.

4.1. Confounding-Robust Policy Learning by Optimizing Minimax Regret

If we had oracle access to the true inverse propensities $W_i^* = 1/e_{T_i}(X_i, Y_i)$ we could form the correct IPW estimate by replacing nominal with true propensities in eq. (1). We may go a step further and, recognizing that $\mathbb{E}[W^* \mathbb{I}[T = t]] = 1$, use the empirical sum of true propensities as a control variate by normalizing our IPW estimate by them. This gives rise to the Hájek regret estimator

$$\hat{R}^*_{\pi_0}(\pi) = \hat{R}_{\pi_0}(\pi; W^*), \qquad \text{where}$$
$$\hat{R}_{\pi_0}(\pi; W) = \sum_{t=0}^{m-1} \hat{R}^{(t)}_{\pi_0}(\pi; W), \quad \hat{R}^{(t)}_{\pi_0}(\pi; W) = \frac{\mathbb{E}_n[(\pi(t \mid X) - \pi_0(t \mid X))\mathbb{I}[T = t] YW]}{\mathbb{E}_n[W \mathbb{I}[T = t]]}$$

These estimators introduce the denominator $\mathbb{E}[W_i^* \mathbb{I}[T = t]]$ as a ratio control variate within each treatment group. It follows by Slutsky's theorem that these estimates remain consistent (*if* we know W_i^*). Note that the choice of π_0 amounts to a constant shift to $\hat{R}^*_{\pi_0}(\pi)$ and does not change which policy π minimizes the regret estimate. This will not be true of our bound, where the choice of π_0 will be material to the success of the method.

Since the oracle weights W_i^* are unknown, we instead minimize the worst-case possible value of our regret estimate, by ranging over the space of possible values for W_i^* that are consistent with the observed data and our assumptions about the confounded data-generating process. Specifically, we restrict the extent to which unobserved confounding may affect assignment probabilities.

We first consider an uncertainty set motivated by the odds-ratio bounds of the marginal sensitivity model, which restricts how far the weights can vary pointwise from the nominal propensities Tan (2012). Given a sensitivity parameter $\Gamma \geq 1$, the marginal sensitivity model posits the following restriction:

$$\Gamma^{-1} \leq \frac{(1 - \tilde{e}_T(X)) e_T(X, Y)}{\tilde{e}_T(X)(1 - e_T(X, Y))} \leq \Gamma. \tag{2}$$

The choice of Γ can be calibrated using, *e.g.*, the method of Hsu and Small (2013), and we discuss other approaches in Section 8. Note that $\Gamma = 1$ corresponds to unconfoundedness (weak ignorability) and $\Gamma = \infty$ to no restriction at all.

The restriction in eq. (2) leads to an uncertainty set for the true inverse propensity weights of each unit centered around the nominal inverse propensity weights,[3] $\tilde{W}_i = 1/\tilde{e}_{T_i}(X_i)$:

$$W^*_{1:n} \in \mathcal{W}_n^\Gamma = \left\{W \in \mathbb{R}^n \colon a_i^\Gamma \le W_i \le b_i^\Gamma,\ \ \forall i = 1, \dots, n\right\}, \quad \text{where} \tag{3}$$
$$a_i^\Gamma = 1 + \Gamma^{-1} \cdot (\tilde{W}_i - 1),\ b_i^\Gamma = 1 + \Gamma \cdot (\tilde{W}_i - 1).$$

We assume for now that $\tilde{W}_i$ is known and phrase our method in terms of it. In practice, when $\tilde{e}_t(x)$ is unknown, we suggest to estimate it (e.g., using regression) and plug in the corresponding estimates of $\tilde{W}_i$ in their place. In Section 5.3, we will show that this approach is asymptotically equivalent and provide explicit finite-sample bounds.

Given this uncertainty set, we obtain the following bound on the empirical regret Hájek estimator:

$$\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^\Gamma) = \sup_{W \in \mathcal{W}_n^\Gamma} \hat{R}_{\pi_0}(\pi; W). \tag{4}$$

We then propose to choose the policy π in our class Π to minimize this regret bound, *i.e.*, $\hat{\bar{\pi}}(\Pi, \mathcal{W}_n^\Gamma, \pi_0)$, where

$$\hat{\bar{\pi}}(\Pi, \mathcal{W}_n^\Gamma, \pi_0) \in \operatorname*{arg\,min}_{\pi \in \Pi} \hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^\Gamma) \tag{5}$$

We emphasize that different components of the framework such as weight normalization and estimation error change the population minimax-optimal policy, in contrast to the policy learning setting with unconfoundedness, where these components only affect finite-sample considerations. In particular, for our worst-case regret objective $\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^\Gamma)$, weight normalization is crucial for only enforcing robustness against *consequential* realizations of confounding that affect the *relative* weighting of outcomes. Any mode of the confounding that affects all weights similarly should have no effect on policy choice. Even if we do not know W_i^*, we know that they must satisfy the population moment conditions $\mathbb{E}[W^* \mathbb{I}[T = t]] = 1, \forall t \in \mathcal{T}$, so any realization that violates that is impossible. Moreover, different baseline policies π_0 structurally change the solution to the adversarial subproblem by shifting the contribution of the loss term $Y_i \mathbb{I}[T_i = t](\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))$ to emphasize improvement upon different baselines. In particular, if the baseline policy is in the policy class Π, it already achieves 0 regret; thus, minimizing regret necessitates learning a policy that *must* offer some benefits in terms of decreased loss regardless of confounding.

[3] The representation in eq. (3) is obtained by simply solving for $1/\tilde{e}_T$ in each of the two inequalities in eq. (2)

4.2. The Population Minimax-Optimal Policy

In the above, we showed that our approach minimizes an upper bound on an estimate for the policy regret. We can also similarly define a population-level bound and consider the population-level minimax-optimal policy. Specifically, we can translate the marginal sensitivity model, eq. (2), to an uncertainty set about the population random variable $W^* = 1/e_T(X,Y)$:

$$\mathcal{W}^\Gamma = \{W(t,x,y)\colon a_t^\Gamma(x) \le W(t,x,y) \le b_t^\Gamma(x) \ \ \forall t \le m-1, x \in \mathcal{X}, y \in \mathbb{R}\}, \ \text{ where}$$
$$a_t^\Gamma(x) = 1 + \Gamma^{-1} \cdot (1/\tilde{e}_t(x) - 1),\ b_t^\Gamma(x) = 1 + \Gamma \cdot (1/\tilde{e}_t(x) - 1).$$

Notice that $\mathcal{W}_n^\Gamma = \{(W(T_1,X_1,Y_1), \dots, W(T_n,X_n,Y_n) : W \in \mathcal{W}^\Gamma\}$ can be understood as the restriction of the above to the data. The corresponding bound on the population-level regret is $\overline{R}_{\pi_0}(\pi;\mathcal{W}^\Gamma)$, where

$$\overline{R}_{\pi_0}(\pi;\mathcal{W}) = \sup_{W\in\mathcal{W}} R_{\pi_0}(\pi;W), \qquad \text{where}$$
$$R_{\pi_0}(\pi;W) = \sum_{t=0}^{m-1} R_{\pi_0}^{(t)}(\pi;W), \quad R_{\pi_0}^{(t)}(\pi;W) = \frac{\mathbb{E}[(\pi(t \mid X) - \pi_0(t \mid X))\mathbb{I}[T=t]\, Y \cdot W(T,X,Y)]}{\mathbb{E}[\mathbb{I}[T=t]\, W(T,X,Y)]}.$$

Note that $R_{\pi_0}(\pi) = R_{\pi_0}(\pi;W^*)$. In words, $\overline{R}_{\pi_0}(\pi;\mathcal{W}^\Gamma)$ is the largest-possible true regret of π relative to π_0 over all possible distributions that agree with the observable data-generating distribution of (X,T,Y) *and* with the restrictions of the marginal sensitivity model. That is, every potentially-possible regret of π is bounded by this quantity and this quantity is also tight in that there exist distributions agreeing with the data and the assumptions that are arbitrarily close to it. The denominator in $R_{\pi_0}(\pi;W)$ ensures that we adhere to the requirement that $\mathbb{E}[\mathbb{I}[T=t]\, W^*] = 1$.[4]

In fact, the interval generated by the smallest-possible and largest-possible regret is sharp in that it is equal to the closure of all possible regrets under the marginal sensitivity model of eq. (2). We summarize this side observation as follows:

Proposition 1 **(Sharpness)**. *$\mathcal{W}^\Gamma$ is an uncertainty set for the marginal sensitivity model (eq. (2)) with parameter value Γ.*

$$\overline{\{R_{\pi_0}(\pi;W)\colon W \in \mathcal{W}^\Gamma\}} = \left[\inf_{W\in\mathcal{W}^\Gamma} R_{\pi_0}(\pi;W), \sup_{W\in\mathcal{W}^\Gamma} R_{\pi_0}(\pi;W)\right].$$

We can correspondingly conceive of what would be the *minimax-optimal* policy at the population level, *i.e.*, $\overline{\pi}^*(\Pi,\mathcal{W}^\Gamma,\pi_0)$, where

$$\overline{\pi}^*(\Pi,\mathcal{W}^\Gamma,\pi_0) \in \arg\min_{\pi\in\Pi} \overline{R}_{\pi_0}(\pi;\mathcal{W}^\Gamma) \tag{6}$$

[4] As an uncertainty set over the joint distribution $\mathbb{P}(T,X,Y(0),\dots,Y(m-1))$ this would correspond to $\{\mathbb{P}\colon \psi_t/b_t^\Gamma(x) \le \mathbb{P}(T=t \mid X=x, Y(t)=y) \le \psi_t/a_t^\Gamma(x)\ \forall t \le m-1,\ \psi \in \mathbb{R}_+^m,\ \mathbb{P} \text{ is a probability distribution}\}$.

This minimax-optimal policy is the one that would obtain the best-possible uniform control over all possible regrets under any possible realization of the true distribution of outcomes that agrees with the observable data-generating distribution of (X,T,Y) *and* with the restrictions of the marginal sensitivity model. This feature makes it an attractive target to aim for in the absence of unconfoundedness.

4.3. Extension: Budgeted Uncertainty Sets

Our approach can flexibly accommodate additional modeling assumptions beyond the odds-ratio bounds, which was motivated by sensitivity analysis. We illustrate via an example of a total-variation bounded uncertainty set how to extend our framework to accommodate additional modeling assumptions. In the subsequent sections we show that this alternative uncertainty set enjoys similar minimax optimality and tractability guarantees as the approach above.

The pointwise interval odds-ratio uncertainty set, eq. (3), might be pessimistic in ensuring robustness against every possible worst-case realization of unobserved confounding for each unit, which may be plausible under individual self-selection into treatment, whereas concerns about unobserved confounding might instead be limited to "exceptions", e.g. individuals with specific unobserved subgroup risk factors, as has also been recognized by Fogarty and Hasegawa (2019), Hasegawa and Small (2017) in the context of classic sensitivity analysis.

Specifically, we construct the uncertainty set

$$\mathcal{W}_n^{\Gamma,\Lambda} = \left\{ W \in \mathbb{R}^{\mathcal{I}_t} : \begin{array}{c} \frac{1}{|\mathcal{I}_t|} \sum_{i\in\mathcal{I}_t} |W_i - \tilde{W}_i| \le \Lambda_t \; \forall t, \\ a_i^\Gamma \le W_i \le b_i^\Gamma \; \forall i \end{array} \right\}$$

with the population counterpart,

$$\mathcal{W}^{\Gamma,\Lambda} = \left\{ W(t,x,y) : \begin{array}{c} \mathbb{E}[|W(T,X,Y) - \tilde{W}(T,X)| \,|\, T=t] \le \Lambda_t \; \forall t, \\ a_t^\Gamma(x) \le W(t,x,y) \le b_t^\Gamma(x) \; \forall t \le m-1, x \in \mathcal{X}, y \in \mathbb{R} \end{array} \right\}$$

When plugged into eq. (5), this provides an alternative policy choice criterion that is less conservative. To make the choice of parameters easier, we suggest to calibrate Λ_t as a fraction, $\rho < 1$, of the total deviation already allowed by $\mathcal{W}_n^\Gamma$. Specifically, $\Lambda_t = \rho \frac{1}{|\mathcal{I}_t|} \sum_{i\in\mathcal{I}_t} \max(\tilde{W}_i - a_i^\Gamma, b_i^\Gamma - \tilde{W}_i)$.[5]

5. Analysis, Improvement Guarantees, and Minimax Optimality

Before discussing how we actually algorithmically compute $\overline{\pi}$, we next introduce finite-sample statistical guarantees on the performance of our approach. We first prove a finite-sample *improvement* guarantee that provides that the policy we learn is assured to induce no harm, as long as the sensitivity model is well-specified. We then prove a uniform convergence result simultaneously over

[5] Enforcing the uncertainty budget separately within each treatment partition is crucial for computationally tractable policy learning and evaluation, as we discuss in Section 6.

both the space of policies, Π, and the space of possible weights that agree with our sensitivity model, $\mathcal{W}^\Gamma$. As a consequence of this uniform convergence, we obtain a bound on the minimax regret that converges to the population optimum. In our analysis, in Sections 5.1 and 5.2, we assume the nominal propensities $\tilde{e}_t(x)$ are known so that the nominal inverse weights $\tilde{W}_i$ are known. In Section 5.3, we extend all of our results to the case of *estimated* nominal propensities, where we instead plug in the estimate $\hat{e}_t(x)$ of $\tilde{e}_t(x)$. In particular, we analyze how our results change when we solve the optimization problem with some $\hat{e}_t(x)$ instead of $\tilde{e}_t(x)$, which provides a bound in terms of the estimation error, which generally vanishes as we collect more data.

For both of these bounds we assume that both outcomes and true propensities are bounded.

Assumption 1 **(Bounded outcomes)**. *Outcomes are bounded, i.e.* $|Y| \le B$.

Assumption 2 **(Overlap)**. *Strong overlap holds with respect to the true propensity: there exists* $\nu > 0$ *such that* $e_t(x,y) \ge \nu \ \forall t \in \{0,\dots,m-1\}, x \in \mathcal{X}, y \in \mathcal{Y}$

Moreover, both of these bounds depend on the flexibility of our policy class: it is critical that we search over a flexible but not completely unrestricted class in order to be assured improvement. We express the flexibility of Π using the notion of the *Vapnik-Chervonenkis (VC) major dimension*, which we define below (see Dudley 1987, p. 1309).

Definition 1. Given a ground set $\mathcal{G}$ and set of maps $\mathcal{F} \subseteq [\mathcal{G} \to \mathbb{R}]$, the *VC-major dimension* of $\mathcal{F}$ is the largest number $v \in \mathbb{N}$ such that there exists $g_1, \dots, g_v \in \mathcal{G}$ with

$$\{(\mathbb{I}[f(g_1) > \theta], \dots, \mathbb{I}[f(g_v) > \theta]) : f \in \mathcal{F}, \theta \in \mathbb{R}\} = \{0,1\}^v. \tag{7}$$

If eq. (7) holds then we say that the superlevel sets of $\mathcal{F}$ shatter $g_1, \dots, g_v$, which means that any subset of the points belong *exclusively* to some superlevel set of some $f \in \mathcal{F}$ and its complement to the corresponding sublevel set. The more complex a class is, the larger the point sets it can shatter. Thus, VC dimension is a natural expression of function class complexity or flexibility.

We will express the flexibility of Π in terms of its VC-major dimension as a set of functions from $(t,x) \in \{0,\dots,m-1\} \times \mathcal{X}$ to $[0,1]$.

Assumption 3. *The policy class* Π*, as a class of functions* $\{0,\dots,m-1\} \times \mathcal{X} \to [0,1]$*, has a finite VC-major dimension.*

Assumption 3 holds for all multi-treatment policy classes we consider, including linear, logistic, and tree policies with bounded depth. Note that our treatment differs from multi-class classifiers as we treat (t,x) as the ground set. It is nonetheless immediate to see that the VC-major dimension of both linear and logistic policies is at most $(m-1)(d+1)$. Moreover, for binary decision trees of depth no more than D, if each inner node can be a query $x_i \le \theta$ for any $i = 1,\dots,d$ and $\theta = \theta_{i1}, \dots, \theta_{iK}$ and

each leaf node is assigned its own probability vector in Δ^m, then the VC dimension of this class is at most $2^D(m-1)\log_2(dK+2)$, as can been seen by following the arguments of Golea et al. (1998) and seeing this a direct sum of 2^D leaf functions, each consisting of $D-1$ conjunctions.

5.1. Improvement Guarantee

We next prove that, if we appropriately bounded the potential hidden confounding, then the optimal value of our minimax-optimal worst-case empirical regret objective $\hat{\overline{R}}_{\pi_0}(\hat{\overline{\pi}};\mathcal{W}_n)$ (as defined in Equation (4)) is asymptotically an upper bound on the true population regret of $\hat{\overline{\pi}}$, $R_{\pi_0}(\hat{\overline{\pi}}(\Pi,\mathcal{W}_n,\pi_0))$. The result is in fact a finite-sample result that gives precisely a bound on how much the latter might exceed the former due to finite-sample errors – terms that vanish as n grows, even if there is unobserved confounding.

Our guarantee relating the sample minimax regret (defined in Equation (4)) to the population optimal regret, for any π, is then as follows:

THEOREM 1 **(Improvement bound)**. *Suppose Assumptions 1, 2, and 3 hold. Suppose, moreover, that $W^*_{1:n}\in\mathcal{W}_n$. Then, for a constant K^Π which only depends on the VC-major dimension of Π, we have that with probability at least $1-\delta$:*

$$R_{\pi_0}(\hat{\overline{\pi}}(\Pi,\mathcal{W}_n,\pi_0)) \le \hat{\overline{R}}_{\pi_0}(\hat{\overline{\pi}}(\Pi,\mathcal{W}_n,\pi_0);\mathcal{W}_n) + \frac{1}{\nu}(BK^\Pi+3)\sqrt{\frac{2\log(8m\vee 20/\delta)}{n}}. \tag{8}$$

Theorem 1 says that the *true* population regret of the policy we learn, $\hat{\overline{\pi}}(\Pi,\mathcal{W}_n,\pi_0)$, when we implement it in practice, is bounded by the objective value that the policy minimizes, plus vanishing terms. These vanishing terms, that is, the second term on the right hand side of eq. (8), vanish at a rate of $O(n^{-1/2})$ and have sub-Gaussian tails, regardless of *any* unobserved confounding. Notice that, as long as $\pi_0\in\Pi$, which can be ensured by design, then we have that our objective is nonpositive, $\hat{\overline{R}}_{\pi_0}(\pi;\mathcal{W}_n)\le 0$. Therefore, this means that we never do worse than π_0 (i.e., do harm), up to vanishing terms. Additionally, if our objective is sufficiently negative, which we can check by just evaluating it, then we are assured some strict improvement. Since we are able to guarantee this without being able to identify or estimate *any* causal effect due to the unobserved confounding, Theorem 1 exactly captures the special appeal of our approach.

Our result above is generic for any uncertainty set $\mathcal{W}_n$; it only requires that it be well-specified. Note that for both of the uncertainty sets we propose in Section 4, the specification of the population sensitivity model ($W^*\in\mathcal{W}$) implies $W^*_{1:n}\in\mathcal{W}_n$, as the latter is simply the restriction of the former to the data. In the next section we further show that we can obtain the minimax-optimal regret in these sensitivity models. These results, however, *will* depend on the uncertainty set and their complexity being manageable.

5.2. Minimax-Optimality

In the previous section we argued that our policy is assured (almost) no harm. A remaining question is whether it achieves the most improvement while doing no harm: whether or not, over all distributions that agree with our sensitivity model, it obtains the best possible uniform control on policy regret. That is, since unconfoundedness does not hold, each policy may incur a range of possible regrets, depending on the true distribution of outcomes, which we cannot pin down even with infinite data. The *best* safe policy uniformly minimizes all of these potential regrets simultaneously and is the minimax-optimal policy $\overline{\pi}^*(\Pi, \mathcal{W}, \pi_0)$ defined in eq. 6. We next show our policy is not only safe but also achieves the same uniform regret control asymptotically. In fact, we will give a finite-sample bound on our uniform regret control.

Controlling the complexity of the sensitivity model. Recall that our policy, $\hat{\overline{\pi}}(\Pi, \mathcal{W}, \pi_0)$, is defined as the minimum over π of the maximum over W of $\hat{R}_{\pi_0}(\pi; W)$. Therefore, one approach may be to establish the uniform convergence of $\hat{R}_{\pi_0}(\pi; W)$ to $R_{\pi_0}(\pi; W)$ over all policies *and* all weight functions in the sensitivity model. However, for the uncertainty sets we propose, this will fail. For example, the weight functions in $\mathcal{W}^\Gamma$ are far too many (isomorphic to all bounded functions) to expect such uniform convergence. Instead, as has been observed in similar sensitivity models with linear-fractional structure Aronow and Lee (2012), Miratrix et al. (2018), Zhao et al. (2019), we first observe that we need only consider a special subclass of weight functions, which will in fact have bounded functional complexity.

PROPOSITION 2 **(Monotone weight solution for $\mathcal{W}^\Gamma$).** *Let*

$$\overline{\mathcal{W}}^\Gamma(\pi) = \left\{ W(t,x,y) \colon \begin{array}{c} W(t,x,y) = a_t^\Gamma(x) + u(y(\pi(t \mid x) - \pi_0(t \mid x))) \cdot (b_t^\Gamma(x) - a_t^\Gamma(x)), \\ u : \mathbb{R} \to [0,1] \text{ is monotonic nondecreasing} \end{array} \right\},$$
$$\overline{\mathcal{W}}_n^\Gamma(\pi) = \{(W(T_1, X_1, Y_1), \dots, W(T_n, X_n, Y_n) : W \in \overline{\mathcal{W}}^\Gamma(\pi)\}.$$

Then, for any π,

$$\overline{R}_{\pi_0}(\pi; \mathcal{W}^\Gamma) = \sum_{t=0}^{m-1} \sup_{W \in \overline{\mathcal{W}}^\Gamma(\pi)} R_{\pi_0}^{(t)}(\pi; W), \quad \hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^\Gamma) = \sum_{t=0}^{m-1} \sup_{W \in \overline{\mathcal{W}}_n^\Gamma(\pi)} \hat{R}_{\pi_0}^{(t)}(\pi; W).$$

This result is due to the special optimization characterization we present later in Theorem 3, which uses linear-fractional optimization to show that the solution takes a monotonic, thresholding form.

COROLLARY 1. *Let $\overline{\mathcal{W}}^\Gamma = \bigcup_{\pi \in \Pi} \overline{\mathcal{W}}^\Gamma(\pi)$, $\overline{\mathcal{W}}_n^\Gamma = \bigcup_{\pi \in \Pi} \overline{\mathcal{W}}_n^\Gamma(\pi)$. Then, for any $\pi \in \Pi$,*

$$\overline{R}_{\pi_0}(\pi; \mathcal{W}^\Gamma) = \sum_{t=0}^{m-1} \sup_{W \in \overline{\mathcal{W}}^\Gamma} R_{\pi_0}^{(t)}(\pi; W), \quad \hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^\Gamma) = \sum_{t=0}^{m-1} \sup_{W \in \overline{\mathcal{W}}_n^\Gamma} \hat{R}_{\pi_0}^{(t)}(\pi; W).$$

Corollary 1 shows that, when searching for policies in Π to obtain uniform control on regret, it suffices to consider weight functions in $\overline{\mathcal{W}}^\Gamma$, which is a subset of $\mathcal{W}^\Gamma$. Again, this result crucially relies on the optimization structure of our problem.

Importantly, this subset, $\overline{\mathcal{W}}^\Gamma$, has much more structure and, in contrast to $\mathcal{W}^\Gamma$, has bounded complexity.

PROPOSITION 3. *Suppose Assumption 3 holds. Then $\overline{\mathcal{W}}^\Gamma$ has a finite VC-major dimension.*

Proposition 3 leverages the stability of VC-major classes (see Van Der Vaart and Wellner 1996, Lemma 2.6.19 and Dudley 1987, Proposition 4.2). Note that monotone functions are *not* a VC class in the usual sense of having VC subgraphs, but they are VC-hull (Giné and Nickl 2016, Example 3.6.14).

Using Corollary 1 and Proposition 3, we can obtain the following uniform convergence:

THEOREM 2. *Suppose Assumptions 1, 2, and 3 hold. Then, for a constant K^Π that depends only on the VC-major dimension of Π, we have that, with probability at least $1-\delta$:*

$$\sup_{\pi\in\Pi}\left|\hat{\overline{R}}_{\pi_0}(\pi;\mathcal{W}_n^\Gamma)-\overline{R}_{\pi_0}(\pi;\mathcal{W}^\Gamma)\right|\le 36(12+\nu^{-1})(BK^\Pi+\nu^{-1}(\Gamma-\Gamma^{-1})(K^\Pi+B+m))\sqrt{\frac{\log(15m/p)}{n}}$$

Relative to Theorem 1, the additional dependence on m arises due to the flexibility of $\overline{\mathcal{W}^\Gamma}$ where, per Proposition 2, we may effectively choose a *different* monotone function u for *each* treatment level $t=0,\dots,m-1$.

As a corollary to Theorem 2 we obtain a finite-sample bound on our minimax suboptimality, which ensures asymptotic minimax optimality:

COROLLARY 2 **(Minimax regret bounds for $\mathcal{W}^\Gamma$)**. *Suppose Assumptions 1, 2, and 3 hold. Then, with probability at least $1-\delta$, we have that*

$$\begin{aligned}&\overline{R}_{\pi_0}(\hat{\overline{\pi}}(\Pi,\mathcal{W}_n^\Gamma,\pi_0);\mathcal{W}^\Gamma)\\&\quad\le\inf_{\pi\in\Pi}\overline{R}_{\pi_0}(\pi;\mathcal{W}^\Gamma)+36(12+\nu^{-1})(BK^\Pi+\nu^{-1}(\Gamma-\Gamma^{-1})(K^\Pi+B+m))\sqrt{\frac{\log(15m/\delta)}{n}}\end{aligned}$$

It is important to note that, in contrast to Theorem 1, this result depends crucially on the structure of $\mathcal{W}^\Gamma$. The key question is how flexible is the set of worst-case weight functions for any policy.

While our budgeted uncertainty set, $\mathcal{W}^{\Gamma,\Lambda}$, is also too flexible to expect uniform convergence over it, we can make similar arguments, focusing only on the set of worst-case weights: they satisfy a nondecreasing property similar to that of Proposition 2, despite the additional constraint.

PROPOSITION 4 **(Monotone weight solution for $\mathcal{W}^{\Gamma,\Lambda}$).** *Let*

$$\overline{\mathcal{W}}^{\Gamma,\Lambda}(\pi;\mathbb{P}) = \left\{ \begin{array}{c} W(t,x,y) = a_t^\Gamma(x) + u(y(\pi(t\,|\,x) - \pi_0(t\,|\,x))) \cdot (b_t^\Gamma(x) - a_t^\Gamma(x)), \\ W(t,x,y)\colon\ u(y(\pi(t\,|\,x) - \pi_0(t\,|\,x))) : \mathbb{R} \to [0,1] \text{ is monotonic nondecreasing}, \\ \mathbb{E}_{\mathbb{P}}[|W(T,X,Y) - \tilde{W}(T,X)|\,|\,T=t] \le \Lambda_t\ \forall t \end{array} \right\},$$

$$\overline{\mathcal{W}}_n^{\Gamma,\Lambda}(\pi;\mathbb{P}) = \{(W(T_1,X_1,Y_1),\dots,W(T_n,X_n,Y_n) : W \in \overline{\mathcal{W}}^{\Gamma,\Lambda}(\pi;\mathbb{P})\}.$$

Then, for any $\pi : \mathcal{X} \to \Delta^m$,

$$\overline{R}_{\pi_0}(\pi;\mathcal{W}^{\Gamma,\Lambda}) = \sum_{t=0}^{m-1} \sup_{W \in \overline{\mathcal{W}}^{\Gamma,\Lambda}(\pi;\mathbb{P})} R_{\pi_0}^{(t)}(\pi;W), \quad \hat{\overline{R}}_{\pi_0}(\pi;\mathcal{W}_n^{\Gamma,\Lambda}) = \sum_{t=0}^{m-1} \sup_{W \in \overline{\mathcal{W}}_n^{\Gamma,\Lambda}(\pi;\mathbb{P}_n)} \hat{R}_{\pi_0}^{(t)}(\pi;W),$$

where $\mathbb{P}$ denotes the population distribution of T,X,Y and $\mathbb{P}_n$ the corresponding empirical distribution.

These arguments require proving structural properties of the optimal solution under this budgeted uncertainty set, which allow us to use the same stability arguments for various compositions of VC-major classes. We remark that the structural results for the budgeted uncertainty set are weaker than that of the unbudgeted one (Theorem 3), where we also obtain efficient algorithms. We quote the final regret bound and refer the reader to the supplement for details.

PROPOSITION 5 **(Minimax regret bounds for $\mathcal{W}^{\Gamma,\Lambda}$).** *Suppose Assumptions 1, 2, and 3 hold. Then, for a constant K^Π that depends only on the VC-major dimension of Π, we have that, with probability at least $1-\delta$:*

$$\begin{aligned} &\overline{R}_{\pi_0}(\hat{\bar{\pi}}(\Pi,\mathcal{W}_n^{\Gamma,\Lambda},\pi_0);\mathcal{W}^{\Gamma,\Lambda}) \\ &\le \inf_{\pi\in\Pi} \overline{R}_{\pi_0}(\pi;\mathcal{W}^{\Gamma,\Lambda}) + 36(12+\nu^{-1})(BK^\Pi + \nu^{-1}(\Gamma-\Gamma^{-1})((m\frac{2B\Gamma\nu^{-1}}{\min_t \Lambda_t \wedge 1} + 1)K^\Pi + B + m))\sqrt{\frac{\log(30m/\delta)}{n}} \\ &\quad + \frac{1}{n}\frac{2m\nu^{-2}B\Gamma \max_t \Lambda_t}{\min_t \Lambda_t \wedge 1} \end{aligned}$$

5.3. Estimated Propensity Scores

All of the above results are presented for the case of known nominal propensities, $\tilde{e}_t(x)$, that is, when $\mathcal{W}^\Gamma$, which is centered at the nominal inverse propensity weights $\tilde{W}_i$, is known. If, as is the case for an observational study, the nominal propensities need to be estimated from data, we optimize over $\mathcal{W}_n^\Gamma$ as an approximation to $\mathcal{W}^\Gamma$. We next show that the use of estimated nominal propensities $\hat{e}_t(x)$ results in an additive approximation error.

PROPOSITION 6 **(Bounded perturbations).** *Let $\hat{W}_i = 1/\hat{e}_{T_i}(X_i)$ and*

$$\hat{\mathcal{W}}_n^\Gamma = \left\{ W \in \mathbb{R}^n\colon \hat{a}_i^\Gamma \le W_i \le \hat{b}_i^\Gamma,\ \forall i = 1,\dots,n \right\},\ \text{where } \hat{a}_i^\Gamma = 1 + \Gamma^{-1}(\hat{W}_i - 1),\ \hat{b}_i^\Gamma = 1 + \Gamma(\hat{W}_i - 1).$$

Then, under Assumption 1, for any $\pi : \mathcal{X} \to \Delta^m$,

$$\left| \hat{\overline{R}}_{\pi_0}(\pi,\hat{\mathcal{W}}_n^\Gamma) - \hat{\overline{R}}_{\pi_0}(\pi,\mathcal{W}_n^\Gamma) \right| \le 2B(\Gamma+\Gamma^{-1})\frac{1}{n}\sum_{i=1}^n \left| \frac{1}{\hat{e}_{T_i}(X_i)} - \frac{1}{\tilde{e}_{T_i}(X_i)} \right| \tag{9}$$

Proposition 6 is a consequence of the linear-fractional optimization structure of the worst-case regret over weights in $\mathcal{W}_n^\Gamma$. The proof leverages a partial Lagrangian dual of the optimization problem and studies sensitivity to plugged-in nominal propensities in the dual. Note that by additionally assuming strong overlap in the nominal propensities, we can bound errors in the inverse propensities in terms of errors in the propensity function itself, which we can in turn bound using standard finite-sample guarantees for learning conditional expectations (Bartlett et al. 2005). Note that the bound in eq. (9) would scale as these bounds. It remains an important direction for future research to obtain bounds of the form of eq. (9) that have a multiplicative-bias property, allowing for slower-than-$n^{-1/2}$ estimation of propensities without deteriorating the overall $n^{-1/2}$ rate, as in Wager and Athey (2017a).

Since Proposition 6 holds deterministically and for all policies, including the sample-optimal policy, it immediately shows that the policy we get by optimizing our worst-case empirical regret with estimated nominal propensities, $\hat{\bar{\pi}}(\Pi, \hat{\mathcal{W}}_n, \pi_0)$, is actually near-optimal in objective relative to the worst-case empirical regret we would obtain with true nominal propensities, that is, $\hat{\bar{R}}_{\pi_0}(\pi, \mathcal{W}^\Gamma)$. Therefore, all previous results for our method similarly hold for $\hat{\bar{\pi}}(\Pi, \hat{\mathcal{W}}_n, \pi_0)$ with the addition of two times the right-hand side of eq. (9) to any previous bound. In particular, for the improvement guarantee for the case of $\mathcal{W}^\Gamma$, we need only ensure that $W^* \in \mathcal{W}^\Gamma$, which is implied by the validity of the marginal sensitivity model; we do *not* need to ensure that $W_n^* \in \hat{\mathcal{W}}_n^\Gamma$, which may be a random event depending on our estimation.

6. Algorithms for Optimizing Robust Policies

We next discuss how to algorithmically solve the policy optimization problem in eq. (5) and actually find the sample minimax-optimal policy, $\hat{\bar{\pi}}$. In the main text, we focus on differentiable parametrized policy classes, $\mathcal{F} = \{\pi_\theta(\,\cdot\,) : \theta \in \Theta\}$ such that $\pi_\theta(t \mid x)$ is differentiable with respect to θ, such as logistic policies. We will use a subgradient method to find the robust policy. In the appendix, we also discuss optimization over decision-tree based policies, using a mixed-integer optimization formulation. In both cases, our solution will depend on a characterization of the inner worst-case regret subproblem.

We first discuss how to solve the worst-case regret subproblem *for a fixed* policy, which we will then use to develop our algorithms.

6.1. Dual Formulation of Worst-Case Regret

The minimization in eq. (5) involves an inner supremum, namely $\hat{\bar{R}}_{\pi_0}(\pi; \mathcal{W}_n^\Gamma)$. Moreover, this supremum over weights W does not on the face of it appear to be a convex problem. However, a standard transformation will reveal its convexity. We next proceed to characterize this supremum,

formulate it as a linear program, and, by dualizing it, provide an efficient line-search procedure for finding the pessimal weights.

For compactness and generality, we address the optimization problem $\hat{\overline{Q}}_t(r;\mathcal{W})$ parameterized by an arbitrary reward vector $r\in\mathbb{R}^n$, where

$$\hat{\overline{Q}}_t(r;\mathcal{W}) = \max_{W\in\mathcal{W}} \frac{\sum_{i=1}^n r_i W(T_i,X_i,Y_i)\mathbb{I}[T_i=t]}{\sum_{i=1}^n W(T_i,X_i,Y_i)\mathbb{I}[T_i=t]}. \tag{10}$$

To recover $\hat{\overline{R}}_{\pi_0}(\pi;\mathcal{W}_n^\Gamma)$, we would simply compute, with $r_i=(\pi(T_i\mid X_i)-\pi_0(T_i\mid X_i))Y_i$,

$$\hat{\overline{R}}_{\pi_0}(\pi;\mathcal{W}_n^\Gamma) = \sum_{t=0}^{m-1} \hat{\overline{Q}}_t(r\,;\mathcal{W}_n^\Gamma).$$

For the remainder of this subsection, we discuss solving the program generically for the r-weighted linear fractional objective $Q(r;\mathcal{W})$, without discussion of multiple treatment partitions. In doing so, we reindex n. First we consider $\mathcal{W}_n^\Gamma$. Since $\mathcal{W}_n^\Gamma$ involves only linear constraints on W, eq. (10) for $\mathcal{W}=\mathcal{W}_n^\Gamma$ is a *linear fractional program*. We can reformulate it as a linear program by applying the Charnes-Cooper transformation (Charnes and Cooper 1962), requiring weights to sum to 1, and rescaling the pointwise bounds by a nonnegative scale factor ψ. We obtain the following equivalent linear program in a scaling factor and normalized weight variables, $\psi=\frac{1}{\sum_i^n W_i}; w=W\psi$:

$$\begin{aligned}\hat{\overline{Q}}(r;\mathcal{W}_n^\Gamma) = \max_{\psi\geq 0, w\geq 0}\quad & \sum_{i=1}^n r_i w_i \\ \text{s.t.}\quad & \sum_{i=1}^n w_i = 1 \\ & \psi a_i^\Gamma \leq w_i \leq \psi b_i^\Gamma \quad \forall\, i=1,\dots,n\end{aligned} \tag{11}$$

The dual problem to eq. (11) has dual variables $\lambda\in\mathbb{R}$ for the weight normalization constraint and $u,v\in\mathbb{R}^n_+$ for the lower bound and upper bound constraints on weights, respectively. By linear programming duality, we then have that

$$\begin{aligned}\hat{\overline{Q}}(r;\mathcal{W}_n^\Gamma) = \min_{u\geq 0, v\geq 0,\lambda}\quad & \lambda \\ \text{s.t.}\quad & -v^\intercal b^\Gamma + u^\intercal a^\Gamma \geq 0 \\ & v_i - u_i + \lambda \geq r_i \quad \forall\, i=1,\dots,n\end{aligned} \tag{12}$$

We use this to show that solving the inner subproblem requires only sorting the data and a ternary search to optimize a unimodal function. This generalizes the result of Aronow and Lee (2012) for arbitrary pointwise bounds on the weights. Crucially, the algorithmically efficient solution will allow for faster subproblem solutions when optimizing our regret bound over policies in a given policy class.

THEOREM 3 **(Normalized optimization solution)**. *Let* (i) *denote the ordering such that* $r_{(1)} \le r_{(2)} \le \cdots \le r_{(n)}$. *Then,* $\hat{\overline{Q}}(r;\mathcal{W}_n^\Gamma) = \lambda(k^*)$, *where* $k^* = \inf\{k = 1,\dots,n+1 : \lambda(k) < \lambda(k-1)\}$ *and*

$$\lambda(k) = \frac{\sum_{i<k} a_{(i)}^\Gamma r_{(i)} + \sum_{i\ge k} b_{(i)}^\Gamma r_{(i)}}{\sum_{i<k} a_{(i)}^\Gamma + \sum_{i\ge k} b_{(i)}^\Gamma}. \tag{13}$$

Specifically, we have that $\hat{\overline{Q}}(r;\mathcal{W}_n^\Gamma) = \frac{\sum_{i=1}^n r_i W_i^\dagger}{\sum_{i=1}^n W_i^\dagger}$ *where* $W_{(i)}^\dagger = a_{(i)}^\Gamma$ *if* $i < k^*$ *and* $W_{(i)}^\dagger = b_{(i)}^\Gamma$ *if* $i \ge k^*$.

Moreover, $\lambda(k)$ *is a discrete concave unimodal function.*

Next we consider a budgeted uncertainty set, $\hat{\overline{Q}}(r;\mathcal{W}_n^{\Gamma,\Lambda})$. Write an extended formulation for $\mathcal{W}_n^{\Gamma,\Lambda}$ using only linear constraints:

$$\mathcal{W}_n^{\Gamma,\Lambda} = \left\{ W \in \mathbb{R}_+^n : \exists d \text{ s.t. } \sum_{i=1}^n d_i \le \Lambda,\ d_i \ge W_i - \tilde{W}_i,\ d_i \ge \tilde{W}_i - W_i,\ a_i^\Gamma \le W_i \le b_i^\Gamma\ \forall i \right\}$$

This immediately shows that $\hat{\overline{Q}}(r;\mathcal{W}_n^{\Gamma,\Lambda})$ remains a fractional linear program. Indeed, a similar Charnes-Cooper transformation as used above yields a non-fractional linear programming formulation:

$$\begin{aligned}
\hat{\overline{Q}}(r;\mathcal{W}_n^{\Gamma,\Lambda}) = \max_{\psi>0,w\ge0,d}\ & \sum_{i=1}^n w_i r_i \\
\text{s.t.}\ & \sum_{i=1}^n d_i \le \Lambda\psi, \quad \sum_{i=1}^n w_i = 1 \\
& a_i^\Gamma \psi \le w_i \le b_i^\Gamma \psi \quad \forall\, i = 1,\dots,n \\
& d_i \ge w_i - \tilde{W}_i\psi \quad \forall\, i = 1,\dots,n \\
& d_i \ge \tilde{W}_i\psi - w_i \quad \forall\, i = 1,\dots,n
\end{aligned}$$

The corresponding dual problem is:

$$\begin{aligned}
\hat{\overline{Q}}(r;\mathcal{W}_n^{\Gamma,\Lambda}) = \min_{g\ge0,h\ge0,u\ge0,v\ge0,\nu\ge0,\lambda}\ & \lambda \\
\text{s.t.}\ & v_i - u_i + g_i - h_i + \lambda \ge r_i \quad \forall\, i = 1,\dots,n \\
& v_i \ge g_i + h_i \quad \forall\, i = 1,\dots,n \\
& -b^\intercal v + a^\intercal u - \Lambda\nu + g^\intercal\tilde{W} + h^\intercal\tilde{W} \ge 0
\end{aligned}$$

As $\hat{\overline{Q}}(r;\mathcal{W}_n^{\Gamma,\Lambda})$ remains a linear program, we can easily solve it using off-the-shelf solvers, even if it does not admit as simple of a solution as $\hat{\overline{Q}}(r;\mathcal{W}_n^\Gamma)$ does.

6.2. Optimizing Parametric and Differentiable Policies

In the main text, we consider iterative optimization to optimize over a *parametrized* policy class $\Pi = \{\pi_\theta(\cdot,\cdot) : \theta \in \Theta\}$, where the parameter space Θ is convex (usually $\Theta = \mathbb{R}^m$), and $\pi_\theta(t \mid x)$ is differentiable with respect to the parameter θ. In the appendix, we discuss global optimization approaches for policy learning, for example over the interpretable policy class of optimal trees. We

Algorithm 1 Subgradient Method

Input: step size η_0, step-schedule exponent $\kappa \in (0,1]$, initial iterate θ_0, number of iterations N

for $k = 0, \dots, N-1$ **do**:

  $\eta_k \leftarrow \eta_0 t^{-\kappa}$ ▷ Update step size

  $\ell_k, W \leftarrow \max_{W \in \mathcal{W}}, \ \in \arg\max_{W \in \mathcal{W}} \sum_{i=1}^n \frac{W_i}{\sum_{i \in \mathcal{I}_{T_i}} W_i}(\pi_{\theta_k}(T_i \mid X_i) - \pi_0(T_i \mid X_i))Y_i$ ▷ Solve inner subproblem for θ_t

  $\theta_{k+1} \leftarrow \text{Projection}_\Theta(\theta_k - \eta_k \cdot g(\theta_k; W))$ ▷ Move in subgradient direction

return $\frac{1}{n}\sum_{t=1}^N \theta_t$

suppose that $\nabla_\theta \pi_\theta(t \mid x)$ is given as an evaluation oracle. An example is logistic policies for binary treatments where it is sufficient to only parametrize for assigning $T = 1$, $\pi_{\alpha,\beta}(1 \mid X) = \sigma(\alpha + \beta^\intercal X)$ and $\Theta = \mathbb{R}^{d+1}$. Since $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, evaluating derivatives is simple.

Our gradient-based procedure leverages that we can solve the inner subproblem to full optimality in the sample. Note that $\hat{\overline{Q}}(r; \mathcal{W})$ is convex in r since it is a maximum over linear functions in r. Correspondingly, its subdifferential at r is given by the argmax set, where $\sum_{i \in \mathcal{I}_T} W$ denotes the vector of normalizing weights corresponding to the observed treatment pattern T:

$$\partial_r \hat{\overline{Q}}(r; \mathcal{W}) = \left\{ \frac{W}{\sum_{i \in \mathcal{I}_T} W} : W \in \mathcal{W} \ , \sum_{i=1}^n \frac{W_i}{\sum_{i \in \mathcal{I}^n_{T_i}} W_i} r_i \geq \hat{\overline{Q}}(r; \mathcal{W}). \right\}$$

If we set $r_i(\theta) = (\pi_\theta(T_i \mid X_i) - \pi_0(T_i \mid X_i))Y_i$, so that $\hat{\overline{Q}}(r(\theta); \mathcal{W}) = \hat{\overline{R}}_{\pi_0}(\pi_\theta(\cdot); \mathcal{W})$, then $\frac{\partial r_i(\theta)}{\partial \theta_j} = Y_i \frac{\partial \pi_\theta(T_i \mid X_i)}{\partial \theta_j}$. Although $F(\theta) := \hat{\overline{R}}_{\pi_0}(\pi_\theta; \mathcal{W})$ may not be convex in θ, this suggests a subgradient descent approach. Let

$$g(\theta; W) = \nabla_\theta \sum_{i=1}^n \frac{W_i}{\sum_{j \in \mathcal{I}_{T_i}}^n W_j}(\pi_\theta(T_i \mid X_i) - \pi_0(T_i \mid X_i))Y_i = \sum_{i=1}^n \frac{W_i}{\sum_{j \in \mathcal{I}_{T_i}}^n W_j} Y_i \nabla_\theta \pi_\theta(T_i \mid X_i)$$

Note that whenever $\partial_r \hat{\overline{Q}}(r(\theta); \mathcal{W}) = \{\frac{W}{\sum_{i \in \mathcal{I}_T} W}\}$ is a singleton then $g(\theta; W)$ is in fact a gradient of $F(\theta)$.

At each step, our algorithm starts with a current value of θ, then proceeds by finding the weights W that realize $\hat{\overline{R}}_{\pi_0}(\pi(\cdot\,; \theta)$ by using an efficient method as in the previous section, and then takes a step in the direction of $-g(\theta; W)$. Using this method, we can optimize policies, over both the unbudgeted uncertainty set $\mathcal{W}_n^\Gamma$ and the budgeted uncertainty set $\mathcal{W}_n^{\Gamma,\Lambda}$. We return the averaged θ parameter for each initialization; and we ultimately average the parameter achieving the best over multiple restarts. Our method is summarized in Alg. 1. In Section C.1 of the Appendix, we include further algorithmic refinements to this subgradient procedure that leverage the special nested structure of the uncertainty sets. We find that these refinements help empirically in stabilizing the optimization when we compute minimax-optimal policies for multiple values of Γ, as we anticipate a decision-maker would, over reasonable plausible ranges of Γ.

6.3. Optimizing Over Other Policy Classes

We next discuss how our approach can be extended to other, more general policy classes, if we have a representation of the constraint $\pi \in \Pi$ that is compatible with conic or integer programming solvers.

PROPOSITION 7. *Suppose* $\mathcal{W}_n = \{W_{1:n} : (W_i)_{i:T_i=t} \in \mathcal{W}_{n,t}\ \forall t = 0, \dots, m-1\}$ *takes a product form over treatments and that* $\mathcal{W}_{n,t}$ *is convex with a non-empty relative interior. Let* $\Pi_n = \{(\pi(T_1 \mid X_1), \dots, \pi(T_n \mid X_n)) : \pi \in \Pi\}$. *Then,*

$$\min_{\pi\in\Pi} \hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n) = \inf \left\{ \sum_{t=0}^{m-1} \lambda_t \colon p \in \Pi_n,\ (\lambda - Y_i(p_i - \pi_0(T_i \mid X_i)))_{i:T_i=t} \in \mathcal{W}_{n,t}^*\ \forall t, \right\}$$

where $S^* = \{p : u^T p \le 0\ \forall u \in S\}$ *denotes the dual cone of a set* S.

Aside from the constraint $p \in \Pi_n$, Proposition 7 provides a convex conic formulation of our optimization problem. If, as for our two proposed uncertainty sets, $\mathcal{W}_n$ is polyhedral, then this formulation is linear. The specific policy parametrization is formulated in the constraint $p \in \Pi_n$. For the case of sparse linear policies (Ustun and Rudin 2015) and fixed-depth decision trees (Bertsimas and Dunn 2017, Kallus 2017a), existing such formulations based on integer optimization exist and can be used to adapt our approach to such policy classes. In the appendix, we provide a more detailed treatment for the case of decision trees.

7. Empirical Results

In this section we present empirical results on two experiments to investigate the benefit of robustness to unobserved confounding. Our first experiment is a simple synthetic example that we use to illustrate the different methods in a controlled setting. Our second experiment develops a case-study, drawing on the data from the parallel WHI observational study and clinical trial. There, harm would be done by unwarranted aggressive intervention by personalized policy learning led astray by confounding. Our minimax-optimal approach is able to avoid such harm, and still offer improvements over baseline by personalizing care, for a variety of possible reward scalarizations of reductions in high blood pressure against known clinical benefits.

7.1. Simulated Data

7.1.1. Binary Treatments We first consider a simple linear model specification demonstrating the possible effects of significant confounding on inverse-propensity weighted estimators. We generate our data as follows, from a true propensity model based on an unobserved confounder, U, which is a function of the potential outcomes:

$$\xi \sim \text{Bern}(1/2), \quad X \sim N((2T-1)\mu_x, I_5), \quad U = \mathbb{I}[Y(1) < Y(0)]$$

$$Y(t) = \beta^\intercal x + \mathbb{I}[T=1]\beta_{treat}^\intercal x + \alpha \mathbb{I}[T=1] + \eta + \eta\xi + \epsilon$$

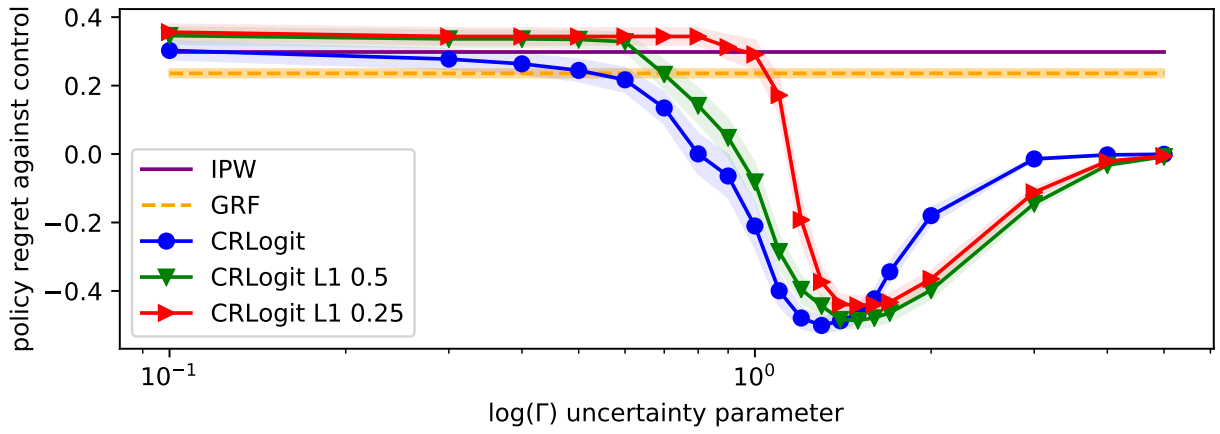


Figure 1 **Out of sample policy regret on simulated data in Sec. 7.1**

The constant treatment effect is $\alpha = 2.5$ with the linear interaction $\beta_{treat} = [-1.5, 1, -1.5, 1, 0.5]$. The covariate mean is $\mu_x = [-1, .5, -1, 0, -1]$. The noise term ξ affects outcomes with coefficients $\eta = -2, \omega = 1$, in addition to a uniform noise term $\epsilon \sim N(0,1)$ which is the same for both treatments. We let the nominal propensities be logistic in X, $\tilde{e}(X) = \sigma(\beta^\intercal X)$ with $\beta = [0, .75, -.5, 0, -1, 0]$, and we generate T_i for each unit according to the true propensity score $e(X,U)$, which we set to

$$e(X_i, U_i) = \frac{4 + 5U_i + \tilde{e}(X_i)(2 - 5U_i)}{6\tilde{e}(X_i)}.$$

In particular, this makes $e(X_i, U_i)$ realize the upper bound in eq. (2) for $\Gamma = 1.5$ when $U_i = 1$ and the lower bound otherwise. Recall $U_i = 1$ exactly when treatment with $t = 1$ is better than treatment with $t = 0$; therefore, we can interpret the confounding relationship as doctors giving the treatment option that is better for the patient, based, however, on factors that were not recorded in the data.

We compare the policies learned by a variety of methods. We consider two commonplace standard methods that assume unconfoundedness: the logistic policy minimizing the IPW estimate with nominal propensities and the direct comparison policy gotten by estimating CATE using causal forests and comparing it to zero (GRF; Wager and Athey 2017b). We compare these to two variants of our methods using the never-treat baseline policy, $\pi_0(0 \mid x) = 1$: our confounding-robust logistic policy using the unbudgeted uncertainty set (CRLogit) and our confounding-robust logistic policy using the budgeted uncertainty set (CRLogit L1) and multipliers $\rho = 0.5, 0.25$. For each of these we vary the parameter Γ in $\{0.1, 0.2, 0.3, \ldots, 1.8, 1.9, 2, 3, 4, 5\}$. For logistic policies, we run 15 random restarts of Alg. 1 with a step-schedule of $\kappa = 0.5$ and return the one with the best robust objective value, unless the best robust objective value is positive, in which case we just return π_0, which is feasible.

For each of 50 replications, we generate an observational dataset of $n = 200$ according to the above model, run each of the above mentioned methods to learn a policy, and compute the *true* value of

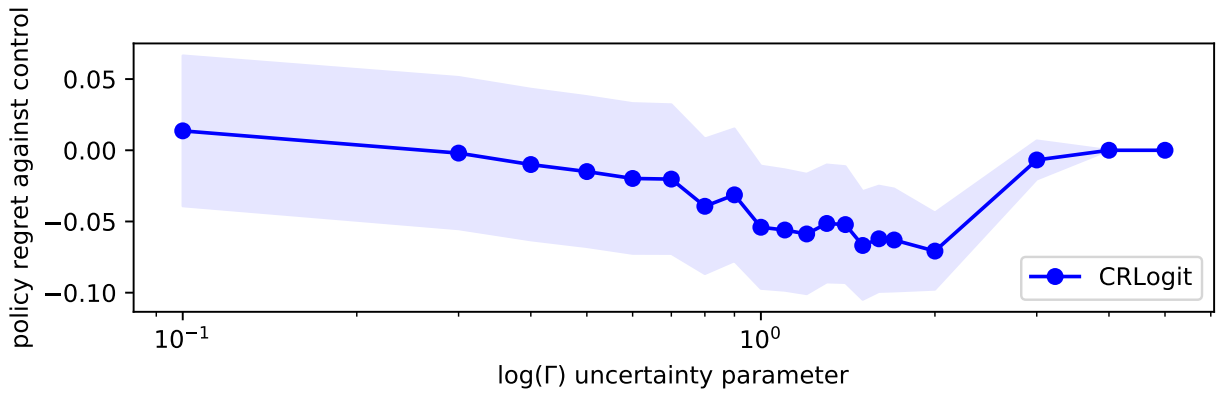


Figure 2 **Out of sample policy regret on simulated data, three treatments in Sec. 7.1**

each learned policy (by using the known counterfactuals, which we generated). We report the value as the regret relative to the value of π_0. We plot the results in Fig. 1, showing the mean regret over replications along with the standard error (shaded regions). We highlight that the worse performance of IPW and GRF does not imply an issue with the algorithms themselves but rather with relying on the assumption of unconfoundedness when it in fact fails to hold.

7.1.2. Multiple Treatments For comparison, we include an example with multiple (three) treatments. We parametrize the policy class with a multinomial logistic probability model (a direct extension of the binary treatment case), e.g. $\pi(t \mid X) = \frac{\exp(\beta_t^\top X)}{\sum_{t=0}^{m-1} \exp(\beta_t^\top X)}$. Our simulation setup is similar to the case for binary treatments. We define the outcome models. In the simulation setup, one treatment arm, $T = 1$, is high-variance due to heterogeneous treatment effects and also greater confounding (which increases variance in propensity scores). The unobserved confounding affects the high-variance treatment, $T = 1$, while now X is generated uniformly on $[-3, 3]$ for all covariates, to reduce variability.

$$\xi \sim \text{Bern}(1/2), \quad X \sim \text{Unif}(-3,3)^5, \quad U = \mathbb{I}[Y(1) < Y(0)]$$

$$Y(t) = \beta_t^\intercal x + \eta\xi + \epsilon + \sum_{t'=1}^{m-1} \mathbb{I}[t = t'](\beta_{t',\text{treat}}^\intercal x + \alpha_{t'} + \eta_{t'}\xi)$$

We parametrize the simulation by vectors of confounding effect and average treatment effect, $\eta = (0, -2, 0), \alpha = (0, 2, 0.5)$, linear effects $\beta = (0, .5, -0.5, 0, 0), \beta_0 = \vec{0}, \beta_1 = 0.75(-1., 0.5, -1., 1., 0.5), \beta_2 = \vec{0}$, and confounding effects $\beta_{0,\text{treat}} = \vec{0}, \beta_{1,\text{treat}} = (0, 1.5, -1, 0, -2), \beta_{2,\text{treat}} = (0, 0, 0.5, 0, 0.5)$. We include the results in Figure 2. We generate 50 replications from this data generating process with $n = 200$, and evaluate on a large generated test set with known counterfactuals. However, the additional parametrization (scaling with the number of treatments) leads to a noisier optimization process by the method of Alg. 1; we leave further refining the optimization for future investigation.

7.2. Assessment with Clinical Data: Women's Health Initiative Trial

We next develop a case study on the parallel Women's Health Initiative (WHI) clinical trial and observational study. We now revisit the real data, under a hypothetical scenario where treatment provides some benefit (in both the observational study and clinical trial), introducing semi-synthetic outcomes which scalarize actual clinical outcomes with a treatment effect "bonus" reflecting known ancillary benefits. We consider binary treatment vs. control, where $T = 1$ indicates treatment with hormone replacement therapy (HRT). Since we vary over a range of possible scalarizations, our focus here is not on drawing specific clinical or substantive conclusions, but rather illustrating the behavior of the method in a variety of treatment effect profiles, and illustrating that for a variety of parameters, our approach will lead to *some* degree of improvement while confounded methods would introduce harm.

7.2.1. Policy Learning and HRT We motivate our policy learning setting noting that modern clinical guidelines, included in Bakour and Williamson (2015), recognize that "when HRT is individually tailored, women gain maximum advantages and the risks are minimised." For example, heterogeneity of treatment effect in age was posited in the clinical literature[6]. For all women, the improvement of vasomotor symptoms was significant, but ultimately the greater risks of adverse events outweighed the clinical benefits for older women. Since the clinical trial itself did not include many younger women for whom treatment could be beneficial, the clinically relevant *policy learning* question is determining the optimal tailoring of targeted treatments such that the clinical benefits of HRT do not also incur substantial increase in risk of CHD and other adverse events.

7.2.2. WHI Case Study Evaluation Setup and Outcome Measures

Dataset details. We restrict attention to a complete-case subset of the WHI clinical-trial data ($n = 13594$) and a complete-case subset of the observational study ($n = 48458$), obtained after dropping the cardiac arrest covariate (which is mostly missing). $T = 1$ denotes treatment with combined estrogen-plus-progestin hormone replacement therapy. An estimate (using GRF Athey et al. (2019)) of the ATE on the blood pressure outcome, as measured on the clinical trial, is 0.64 with 0.26 standard error, while from the observational study, the estimate is -0.94 with standard error 0.38: wrongly deciding based on the observational study would introduce overall harm.

[6] Clinical explanations for heterogeneity in age suggested that estrogen may slow down *early* artherosclerosis, the formation of plaques in arteries, and have favorable endothelial effects in women with recent onset in menopause. However, unlike other options such as statin therapies which help prevent CHD at any age and stage of disease, HRT may actually worsen already-established plaques and thus increase the frequency of coronary events in older women (see Manson et al. 2013, Rossouw et al. 2013).

Outcome variable. We define our outcome variable to account for cardiovascular health as well as the clinical benefits of HRT for menopause symptoms, and we range over the potential combinations of these to study the changing behavior of our method. Specifically, letting S denote systolic blood pressure and given $\lambda < 0$, we define our outcome as

$$Y = S + T\lambda.$$

We vary λ in a grid on $[0, -1.5]$. Every λ generates a new dataset, on which we learn policies from the observational study using our framework, varying the sensitivity parameter Γ, and estimate the outcomes of these learned policies on the actual randomized WHI clinical trial data. In training on the observational dataset, nominal propensities are estimated using logistic regression. We assess our methods and appropriate baselines, which learn from the confounded observational data, and evaluate their performance on the clinical trial dataset, with constant treatment arm randomization probabilities. This demonstrates the range of possible behaviors as treatment becomes overall more or less beneficial and offers a sensitivity analysis of our method to different scalarizations of clinical benefits with the blood pressure outcome.

Clinical trial evaluation. Without access to the true counterfactual outcomes for patients, we evaluate the performance of policies out of sample by using an unnormalized Horvitz-Thompson estimator on the held-out truly-randomized data from the WHI clinical trial. As reported above, treatment was randomized at 1/2 probability; p_0, p_1 denotes the observed treatment probabilities for $T = 0, 1$. Correspondingly, our out-of-sample estimate of policy regret relative to a *control* baseline, $\pi_0(0 \mid x) = 1$, is given by[7]

$$\hat{R}^{\text{test}}_{\pi_0}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{p_{T_i}}(\mathbb{I}[T_i = 0](\pi(0 \mid X_i) - 1) + \mathbb{I}[T_i = 1]\pi(1 \mid X_i))$$

7.2.3. WHI Case Study Policy Learning Results We compare our method (CRLogit) to two benchmark methods for policy learning that do assume unconfoundedness: the logistic policy minimizing the IPW estimate with nominal propensities (IPW) and the same with policy value estimates gotten by estimating CATE using causal forests (GRF Lin; Athey et al. 2019). In Figure 3 we display a (favorable) treatment effect scalarization, $\lambda = -0.64$, where our policy, for certain values of the sensitivity parameter, indeed finds benefit, while linear policies using only estimated propensity scores or confounded outcome regression (IPW and RF lin.) still incur relative harm relative to the all-control baseline.

[7] Note that the actual realized fraction treated in the dataset are 0.502 so the estimate is also nearly equal to the corresponding Hájek estimator.

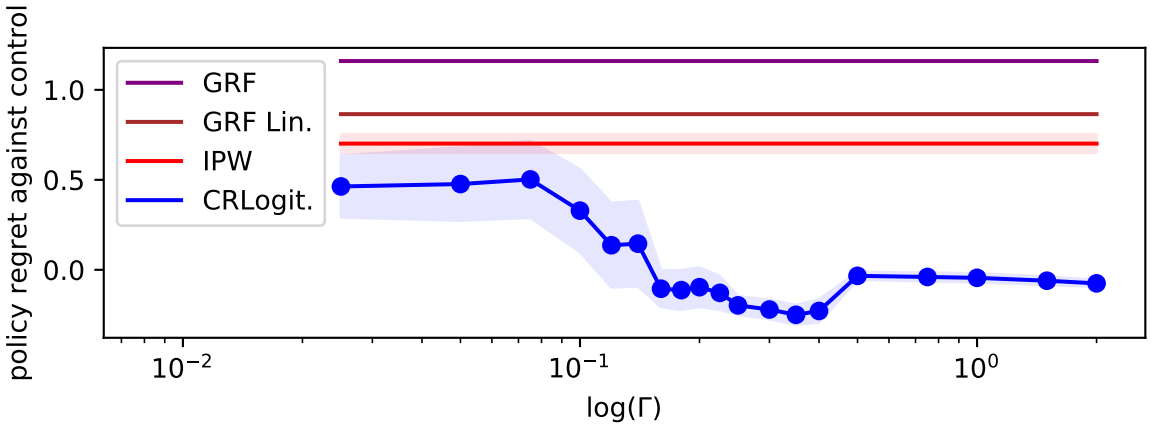


Figure 3 **Plots of out of sample regret on WHI case study data for a single treatment effect scalarization,** $\lambda = -0.64$

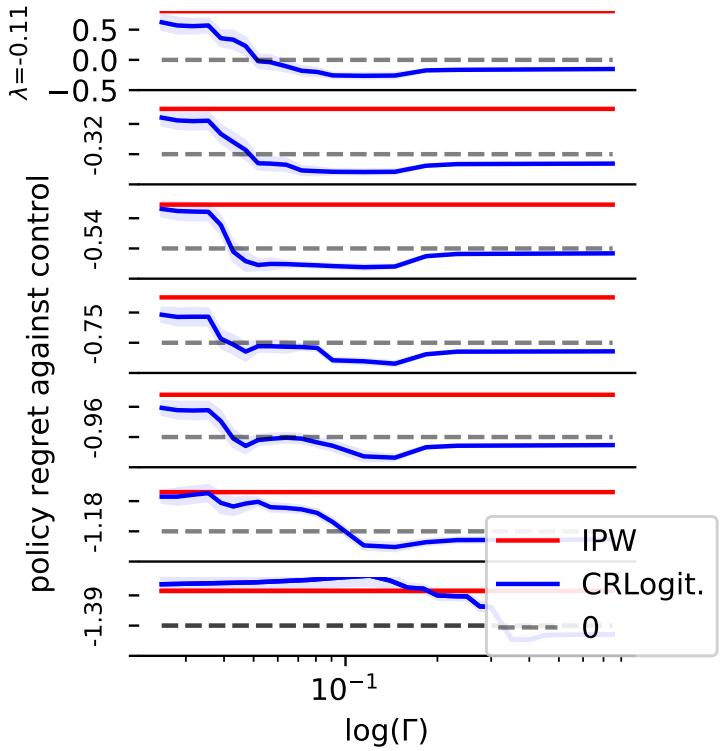
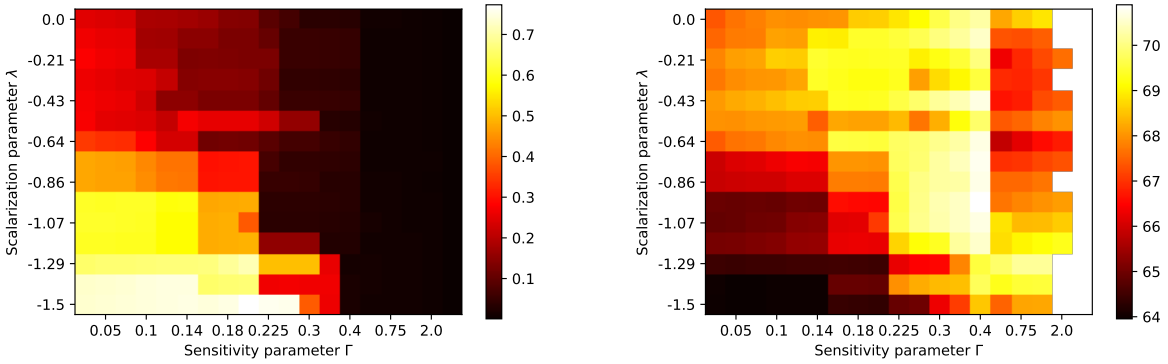


Figure 4 **Plots of out of sample regret, sensitivity analysis on scalarizations of WHI data, varying** λ

Of course, whether or not our approach finds relative improvement (or if the robust approach is overly conservative), depends on the exact treatment effect scalarization parameter λ. In Figure 4, we include a comprehensive comparison of the relative performance of our approach, IPW, and the control baseline, for a various levels of λ. (We report full numerics in Table EC.1 in the appendix.)

For moderate regions of benefit ($\lambda \in [-0.11, -1]$), confounded policies perform poorly, overtreating and inducing harm, while our approach recovers regions of improvement (where the blue line is



(a) Heatmap of percentage of patients treated with probability $\pi(1 \mid X) > 0.5$ (Sec. 7.2)

(b) Heatmap of $\mathbb{E}[\text{age} \mid \pi(1 \mid X) > 0.5]$, robust policies with scalarized WHI data (Sec. 7.2)

Figure 5 WHI Case Study: Policy properties (% of patients treated, average age of those treated)

below the dotted line of 0 regret). For larger values of treatment benefit, the improvement in regret diminishes, while a robust approaches that defaults to baseline still achieves improvement. In Figure 5a, we interpret the range of policies by plotting the percentage of individuals treated under each $\log(\Gamma)$ value on the x-axis, ranging over scalarization parameters λ (on the y-axis). In Figure 5b, we plot the average age conditional on being treated with probability greater than 0.5 under a confounding-robust policy. For regions of moderate improvement, the confounding-robust policies tend to treat younger patients on average. (Artifacts arise when assessing age conditional on treating very few people).

8. Practical Considerations in Calibrating Uncertainty Sets

In the above we demonstrated the performance of our method as we vary the parameter Γ controlling the amount of allowed confounding. For practitioners, a remaining important question is how to choose an appropriate range for Γ: we review recommendations from traditional sensitivity analysis and then propose an approach specifically designed for the policy learning problem.

8.1. Comparison to Observed Covariates and Treatment Selection

As mentioned in Sections 3 and 4.1, one broad strategy for calibrating a sensitivity model benchmarks the level of unobserved confounding relative to the informativeness of observed covariates for selection into treatment (Hsu and Small 2013). For example, we can compute the effect of omitting each observed covariate on the odds ratio of the propensity score. A decision-maker could use domain knowledge to assess whether there are plausible unobserved confounders that could have as large an effect as the observed one, suggesting a plausible upper bound on Γ. While in traditional sensitivity analysis this suggests how large Γ one should consider in testing the robustness of one's inferences, in the context of policy learning, this suggests what amount of confounding should one be concerned

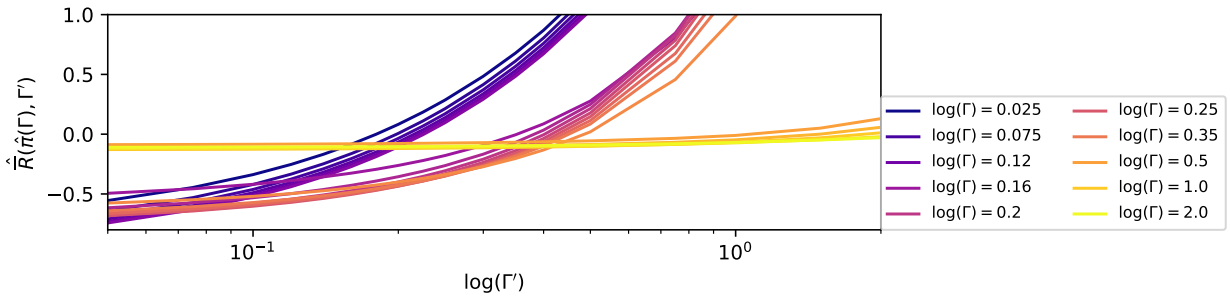


Figure 6 Calibration plot for WHI case study: for each parameter choice Γ, the curve shows the possible estimated worst-case regret when confounding may be as large as Γ', as we vary Γ'.

with protecting against to ensure no harm. In the next section, we discuss how to combine this strategy with calibration plots, which we develop, to make an informed choice about which Γ to choose for training a policy.

To illustrate this benchmarking in the WHI case study, we plot the odds ratios induced by dropping different variables in Section D in the Appendix. This shows that, aside from variables such as age that are highly predictive of treatment selection, the induced odds ratios are safely bounded by Γ somewhere in range of 0.8 to 1.2. Therefore, if we believe our omitted confounders cannot be as informative as age, we should consider the safety of our policy for confounding levels as large as Γ in the range of 0.8 to 1.2.

8.2. Calibration Plots

Next, we propose a tool to visualize the trade-offs between choosing too-high or too-low a value for Γ. Choosing too-high Γ leads to better uniform control on regret on a larger range of potential confounding, but may be conservative if the actual confounding was in fact controlled by a smaller Γ, while too-low a value of Γ achieves worse uniform control over a larger range of potential confounding. We propose to analyze this by re-evaluating, for all policies learned using some parameter Γ, that is, $\hat{\overline{R}}_{\pi_0}(\hat{\overline{\pi}}(\Pi, \mathcal{W}_n^\Gamma, \pi_0); \mathcal{W}_n^{\Gamma'})$, its corresponding *estimated* worst-case regret over a different uncertainty set $\mathcal{W}_n^{\Gamma'}$.

Specifically, we propose to visualize this in a calibration plot produced thus:

- Fix a sequence of Γ values, $\Gamma_1, \dots, \Gamma_K$.
- For every $k \in \{1, \dots, K\}$, train a confounding-robust policy under parameter Γ_k, $\hat{\overline{\pi}}_k = \hat{\overline{\pi}}(\Pi, \mathcal{W}_n^\Gamma, \pi_0)$.
- For every $k, k' \in \{1, \dots, K\}$, evaluate the minimax regret estimate under parameter $\Gamma_{k'}$, $\hat{\overline{R}}_{k,k'} = \hat{\overline{R}}_{\pi_0}(\hat{\overline{\pi}}_k; \mathcal{W}_n^{\Gamma_{k'}})$.
- For each k, plot $\hat{\overline{R}}_{k,k'}$ against Γ_k.

An example of such a calibration plot for our WHI case study is given in Figure 6.

First, this plot shows how the regret of a policy trained with one Γ may grow and possibly become positive if the true confounding may correspond to a larger $\Gamma' > \Gamma$. In the example of Figure 6, for very small $\Gamma = 1.05$, we see that the policy (which essentially assumes unconfoundedness) may incur large regret for even small values of Γ in the range of 1.1 to 1.2. Since these values are smaller than the ranges of Γ we found by considering the informativeness of observed variables, if we may have omitted a variable as important as these, we may be concerned about the safety of policies learned using such small values of Γ. Second, as we increase Γ we find that we obtain uniform control on regret even for confounding corresponding to larger Γ'. We may, however, pay in terms of performance if confounding were in fact smaller. We can assess this using the plot, which shows us the deterioration in performance for smaller levels of confounding, $\Gamma' < \Gamma$, relative to policies that are trained with lower Γ, potentially even policies that are trained assuming unconfoundedness ($\Gamma = 1$). In the example of Figure 6, we find that using $\Gamma = 1.14$ may offer safe control on regret for Γ' up to 1.2, ensuring no harm in the ranges deemed of potential concern, while it would cause only minimal inefficiencies if confounding were really smaller relative to policies that would somehow exploit this fact. Thus, calibration plots allow one to assess the trade-offs of safety and performance and choose a policy that best fits the requirements of the application domain.

9. Conclusion

In this paper, we addressed the problem of learning personalized intervention policies from observational data with unobserved confounding. Standard methods can be corrupted by this confounding and lead to harm compared to current standards of care, a concern of utmost importance in sensitive applications such as medicine, public policy, and civics. We therefore develop a framework for minimax-optimal policy learning under unobserved confounding, which optimizes personalization policies in view of possible unobserved confounding in observational data, allowing for more reliable and credible policy evaluation and learning. Our approach optimizes the minimax regret achieved by a candidate personalized decision policy against a baseline policy. We generalize the class of IPW-based estimators and construct uncertainty sets centered at the nominal IPW weights that can be calibrated by approaches for sensitivity analysis in causal inference. A future line of investigation is a confounding-robust variant based on the doubly robust estimator of policy value.

We prove a strong statistical guarantee that, if the uncertainty set is well-specified, our approach is guaranteed to do no worse than the standard of care so that it can be safely implemented, and possibly offer improvement if the data can support it. Specifically, the result proved a finite-sample guarantee that can be checked. We leverage the optimization structure of weight-normalized estimators of the policy value to perform policy optimization efficiently by subgradient descent on the robust risk and we provide uniform convergence bounds showing that our approach achieves the population

level minimax-optimal regret. Assessments on synthetic and clinical data demonstrate the benefits of minimax-optimal policy learning, which can recommend personalized treatment while maintaining strong guarantees on performance relative to baseline preferences. These tools allow an analyst to find reliable and personalized policies that can safely offer improvements even if there is unobserved confounding and to assess the different plausible levels of confounding on the performance of a robust personalized decision policies. We believe this development is absolutely crucial for the practical adoption of algorithms for personalization that work on the ever growing repositories of observational data, which are the future of algorithmic decision making due to their size and richness.

Acknowledgments

The authors thank the anonymous reviewers for their constructive inputs. This material is based upon work supported by the National Science Foundation under Grants No. 1656996 & 1846210. Angela Zhou was supported through the National Defense Science & Engineering Graduate Fellowship Program.

References

P. Aronow and D. Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 2012.

S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

S. H. Bakour and J. Williamson. Latest evidence on using hormone replacement therapy in the menopause. *The Obstetrician & Gynaecologist*, 17(1):20–28, 2015.

K. Ball and A. Pajor. The entropy of convex bodies with few extreme points. In *Proceedings of the 1989 Conference in Banach Spaces at Strob. Austria. Cambridge Univ. Press*, 1990.

P. L. Bartlett, O. Bousquet, S. Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005.

D. Bertsimas and J. Dunn. Optimal classification trees. *Machine Learning*, 2017.

D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo. Personalized diabetes management using electronic medical records. *Diabetes Care*, page dc160826, 2016.

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

T. Blake, C. Nosko, and S. Tadelis. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1):155–174, 2015.

L. Breiman, J. F. adn charles Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall, 1984.

M. A. Brookhart, T. Sturmer, R. J. Glynn, J. Rassen, and S. Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical Care*, 2010.

A. Charnes and W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 1962.

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

M. Dudik, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

R. Dudley. Universal donsker classes and metric entropy. *The Annals of Probability*, pages 1306–1326, 1987.

C. Fogarty and R. Hasegawa. An extended sensitivity analysis for heterogeneous unmeasured confounding. *The Annals of Applied Statistics*, 2019.

C. B. Fogarty and D. S. Small. Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association*, 111(516):1820–1830, 2016.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

M. Golea, P. L. Bartlett, W. S. Lee, and L. Mason. Generalization in decision trees and dnf: Does size matter? In *Advances in Neural Information Processing Systems*, pages 259–265, 1998.

B. R. Gordon, F. Zettelmeyer, N. Bhargava, and D. Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225, 2019.

J. Hájek. Comment on "an essay on the logical foundations of survey sampling, part one". In R. Holt and W. Toronto, editors, *The Foundations of Survey Sampling*, volume 236. 1971.

R. Hasegawa and D. Small. Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. *Biometrics*, 2017.

K. Hirano and J. R. Porter. Statistical decision rules in econometrics. *Handbook of Econometrics*, 7, 2019.

T.-H. Ho, N. Lim, S. Reza, and X. Xia. Om forum-causal inference models in operations management. *Manufacturing & Service Operations Management*, 19(4):509–525, 2017.

M. A. Hoffman and M. S. Williams. Electronic medical records and personalized medicine. *Human genetics*, 130(1):33–39, 2011.

D. Horvitz and D. Thompson. A generalization of sampling without replacement froma finite universe. *Journal of the American Statistical Association*, 1952.

J. Y. Hsu and D. S. Small. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 69(4):803–811, 2013.

G. Imbens and D. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

N. Kallus. Recursive partitioning for personalization using observation data. *Proceedings of the Thirty-fourth International Conference on Machine Learning*, 2017a.

N. Kallus. Balanced policy evaluation and learning. *arXiv preprint arXiv:1705.07384*, 2017b.

N. Kallus and A. Zhou. Confounding-robust policy improvement. In *Advances in neural information processing systems*, pages 9269–9279, 2018a.

N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251, 2018b.

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

S. Künzel, J. Sekhon, P. Bickel, and B. Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 2017.

D. A. Lawlor, G. D. Smith, and S. Ebrahim. Commentary: The hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *International Journal of Epidemiology*, 2004.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *Proceedings of the fourth ACM international conference on web search and data mining*, 2011.

C. Manski. *Social Choice with Partial Knoweldge of Treatment Response*. The Econometric Institute Lectures, 2005.

C. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2008.

J. E. Manson, R. T. Chlebowski, M. L. Stefanick, A. K. Aragaki, J. E. Rossouw, R. L. Prentice, G. Anderson, B. V. Howard, C. A. Thomson, A. Z. LaCroix, J. Wactawski-Wende, R. D. Jackson, M. Limacher, K. L. Margolis, S. Wassertheil-Smoller, S. A. Beresford, J. A. Cauley, C. B. Eaton, M. Gass, J. Hsia, K. C. Johnson, C. Kooperberg, L. H. Kuller, C. E. Lewis, S. Liu, L. W. Martin, J. K. Ockene, M. J. O'Sullivan, L. Powell, M. S. Simon, L. Van Horn, M. Z. Vitolins, and R. B. Wallace. The women's health initiative hormone therapy trials: Update and overview of health outcomes during the intervention and post-stopping phases, Oct 2013.

M. Masten and A. Poirier. Identification of treatment effects under conditional partial independence. *Econometrica*, 2018.

L. W. Miratrix, S. Wager, and J. R. Zubizarreta. Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika*, 2018.

X. Nie and S. Wager. Learning objectives for treatment effect estimation. 2017.

A. T. Pedersen and B. Ottesen. Issues to debate on the women's health initiative (whi) study. epidemiology or randomized clinical trials-time out for hormone replacement therapy studies? *Human Reproduction*, 2003.

M. Petrik, M. Ghavamzadeh, and Y. Chow. Safe policy improvement by minimizing robust baseline regret. *29th Conference on Neural Information Processing Systems*, 2016.

D. Pollard. *Empirical processes: theory and applications*. NSF-CBMS regional conference series in probability and statistics, 1990.

R. L. Prentice, M. Pettinger, and G. L. Anderson. Statistical issues arising in the women's health initiative. *Biometrics*, 2005.

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

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

P. Rosenbaum. *Observational Studies*. Springer Series in Statistics, 2002.

J. E. Rossouw, J. E. Manson, A. M. Kaunitz, and G. L. Anderson. Lessons learned from the women's health initiative trials of menopausal hormone therapy. *Obstetrics & Gynecology.*, 2013.

D. Rubin. Estimating causal effect of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974.

D. B. Rubin. Comments on "randomization analysis of experimental data: The fisher randomization test comment". *Journal of the American Statistical Association*, 75(371):591–593, 1980.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085, 2017.

J. Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 2009.

J. Stoye. Minimax regret treatment choice with limited validity of experiments or with covariates. *Journal of Econometrics*, 2012.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. *Proceedings of NIPS*, 2015a.

A. Swaminathan and T. Joachims. Counterfactual risk minimization. *Journal of Machine Learning Research*, 2015b.

Z. Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Associatioon*, 2012.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 2015.

A. Van Der Vaart et al. New donsker classes. *The Annals of Probability*, 24(4):2128–2140, 1996.

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

S. Wager and S. Athey. Efficient policy learning. 2017a.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017b.

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

Y.-X. Wang, A. Agarwal, and M. Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. *Proceedings of Neural Information Processing Systems 2017*, 2017.

Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

Supplemental Material for Minimax-Optimal Policy Learning Under Unobserved Confounding

Appendix A: Proofs for optimization structure

Proof of the equivalence of programs (10) *and* (11). We can easily verify that a feasible solution for one problem is feasible for the other: for a feasible solution W to (FP), we can generate a feasible solution to (LP) as $w_i = \frac{W_i}{\sum_i W_i}, \psi = \frac{1}{\sum_i W_i}$ with the same objective value. In the other direction, we can generate a feasible solution to (11) from a feasible fractional program (10) solution W, ψ if we take $W_i = \frac{w_i}{\psi}$. This solution has the same objective value since $\sum_i w_i = 1$. □

Proof of Thm. 3. We analyze the program using complementary slackness, which will yield an algorithm for finding a solution that generalizes that of Aronow and Lee (2012). At optimality only one of the primal weight bound constraints, (for nontrivial bounds $a^\Gamma < b^\Gamma$), $w_i \le \psi b_i^\Gamma$ or $\psi a_i^\Gamma \le w_i$ will be tight. For the nonbinding primal constraints, at the optimal solution, by complementary slackness the corresponding dual variable u_i or v_i will be 0. Since at least $n+1$ constraints are active in the dual, the constraint $\sum_i -b_i v_i + a_i u_i \ge 0$ is also active. So the optimal solution to the dual will satisfy:

$$\begin{aligned} &\min \lambda \\ &\text{s.t. } \lambda \ge r_i + u_i - v_i, \, \forall i \in 1, \dots, n \\ &\sum_i -b_i^\Gamma v_i + a_i^\Gamma u_i = 0 \end{aligned}$$

By non-negativity of u_i, v_i, note that $u_i > 0$ if $r_i < \lambda$ and $v_i > 0$ if $r_i > \lambda$ such that $u_i = \max(0, \lambda - r_i)$ and $v_i = \max(0, r_i - \lambda)$. Additionally, feasible objective values satisfy $\lambda \le \max_i Y_i$ and $\lambda \ge \min_i Y_i$. Let (k) denote the kth index of the increasing order statistics, an ordering where $r_{(1)} \le r_{(2)} \le \cdots \le r_{(n)}$. Then at optimality, there exists some index (k) where $Y_{(k)} < \lambda \le Y_{(k+1)}$. We can subsitute in the solution from the binding constraints $\lambda = r_i + u_i - v_i$ and obtain the following equality which holds at optimality:

$$\psi \sum_{i:(i)<(k)} a_{(i)}^\Gamma (\lambda - r_{(i)}) - \psi \sum_{i:(i)\ge(k)} b_{(i)}^\Gamma (r_{(i)} - \lambda) = 0$$

Rearranging, we have that

$$\lambda_{(k)} = \frac{\sum_{i:(i)<(k)} a_{(i)}^\Gamma r_{(i)} + \sum_{i:(i)\ge(k)} b_{(i)}^\Gamma r_{(i)}}{\sum_{i:(i)<(k)} a_{(i)}^\Gamma + \sum_{i:(i)\ge(k)} b_{(i)}}$$

Therefore, we only need to check the possible objective values $\lambda_{(k)}$ for $k = 1, \dots, n$. The primal solution is easily recovered from the dual solution: for $r_{(i)}$, take $w_{(i)} = \frac{a_{(i)}^\Gamma \mathbb{I}\{(i)\le k\} + b_{(i)}^\Gamma \mathbb{I}\{(i)>k\}}{\sum_{i:(i)<(k)} a_{(i)}^\Gamma + \sum_{i:(i)\ge(k)} b_{(i)}^\Gamma}$ and $t = \sum_{i:(i)<(k)} a_{(i)}^\Gamma + \sum_{i:(i)\ge(k)} b_{(i)}^\Gamma$. Consider the parametric restriction of the primal program, where it is parametrized by the sum of weights ψ: the value function is concave in ψ and concave in the discrete restriction of ψ to the values it takes at the solutions of $\lambda_{(k)}$, $\psi_{(k)}$, and $\psi_{(k)}$ is increasing in k. So the optimal such λ occurs with the order statistic threshold at (k) for $k^* = \inf\{k = 1, \dots, n+1 : \lambda(k+1) < \lambda(k)\}$.

Lastly, we discuss the case where Y may be discrete, or if it is distributed as a mixture of a continuous density and atoms. Our characterization of the optimization solution as monotonic and also a function of the sort order on Y implicitly assumes that outcomes Y are generated from a continuous density, so that $Y_i = Y_j$ with probability 0. Our analysis, too, requires this. We show that in the case where tiebreaking among Y is required, there is a natural lexicographic order. Let $(1),...(i)...(n)$ denote the ordering that is lexicographically increasing in $(Y_{(i)}, b_{(i)} - a_{(i)})$: when outcomes Y are discrete, the appropriate sort order includes the weights $b_{(i)} - a_{(i)}$. Denote the coefficients as r_i and $b_{(i)} - a_{(i)}$. Suppose that for a given sort order, the optimum is achieved at $\lambda(k)$. We show that the lexicographic sort order, sorting first in y and then increasing in $r\Delta$, preserves the unimodality property. Suppose y_k is the same for some interval $[k, k+j]$: we want to show that the discrepancy $\lambda(k) - \lambda(j)$ is increasing in $i, i \le j$. Denote $n(k) = \sum_{i \le k} r_i a_t(X_i) Y_i + \sum_{i \ge k+1} r_i b_t(X_i) Y_i$ and $d(k) = \sum_{i \le k} r_i a_t(X_i) + \sum_{i \ge k+1} r_i b_t(X_i)$. Then,

$$\begin{aligned}
\lambda(k) - \lambda(k+1) &= \frac{n(k)}{d(k)} - \frac{n(k) - \Delta y_{k+1} r_{k+1}}{d(k) - \Delta_{k+1} r_{k+1}} \\
&= \frac{n(k)(d(k) - \Delta_{k+1} r_{k+1}) - (n(k) - \Delta_{k+1} y_{k+1} r_{k+1}) d(k)}{d(k)(d(k) - \Delta_{k+1} r_{k+1})} \\
&= \lambda(k) \frac{\Delta_{k+1} r_{k+1}}{d(k) - r_{k+1} \Delta_{k+1}} - \frac{\Delta_{k+1} y_{k+1} r_{k+1}}{d(k) - \Delta_{k+1} r_{k+1}} \\
&= \frac{\lambda(k) - y_{k+1}}{\frac{d(k)}{r_{k+1} \Delta_{k+1}} - 1}
\end{aligned}$$

We show that if this difference changes sign, it continues to decrease: if $\lambda(k) \le \lambda(k-1)$, and if $y_k = y_{k+1}$, then $\lambda(k+1) < \lambda(k)$. By the above analysis, telescoping the finite difference $\lambda(k) - \lambda(k+1)$,

$$\lambda(k) - \lambda(j) = \sum_{i=1}^{j} \frac{\lambda(k+i) - y_{(k+i+1)}}{\frac{d(k+i)}{r_{(k+i+1)} \Delta_{(k+i+1)}} - 1} = \frac{\lambda(k) - y_{k+j}}{\frac{d(k)}{r_{k+j} \Delta_{k+j}} - 1}$$

so that where $y_{(k)} = y_{(k+j)}$, $\lambda(k) - \lambda(j)$ is decreasing as $r_{k+j} \Delta_{k+j}$ increases.

Proof of Proposition 2 We show via a similar argument to Theorem 3 that the linear program under the one-to-one change of variable $W_i = a(X_i) + (b(X_i) - a(X_i)) u_i$, where $u_i \in [0,1]$, has a similar solution structure in the variable u_i: that the optimal weights u_i^* satisfy that $u_i^* = u(Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i)))$ for some function $u : \mathcal{Y} \to [0,1]$ such that $u(u(\pi(t \mid x) - \pi_0(t \mid x)))$ is nondecreasing in $y(\pi(t \mid x) - \pi_0(t \mid x))$. Define vectors α, β such that $\alpha_i = Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))(b(X_i) - a(X_i)) \mathbb{I}[T_i = t]$ and $\beta_i = (b(X_i) - a(X_i)) \mathbb{I}[T_i = t]$, and constants $c = \sum_{i:T_i=t}^{n} a(X_i) Y_i (\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i)), d = \sum_{i:T_i=t}^{n} a(X_i)$.

$$\begin{aligned}
&\max_U \frac{\alpha^\top u + c}{\beta^\top u + d} \\
&\text{s.t. } 0 \le u \le 1
\end{aligned}$$

By applying the Charnes-Cooper transformation with $\tilde{u} = \frac{u}{\beta^\top u + d}$ and $\tilde{v} = \frac{1}{\beta^\top u + d}$, the linear-fractional program above is equivalent to the following linear program:

$$\begin{aligned}
&\max_{\tilde{u}, v} \ \alpha^\top \tilde{u} + c\tilde{v} \\
&\text{s.t. } 0 \le u \le \tilde{v} \\
&\qquad \beta^\top \tilde{u} + \tilde{v} d = 1, \tilde{v} \ge 0
\end{aligned}$$

where the solution for $\tilde{u}, \tilde{v}$ yields a solution for the original program: $\tilde{u}_i$ is such that $u_i = \frac{\tilde{u}_i}{\tilde{v}}$.

Let the dual variables $p_i \geq 0$ be associated with the primal constraints $\tilde{u}_i \leq \tilde{v}$ (corresponding to $u_i \leq 1$), $q_i \geq 0$ associated with $\tilde{u}_i \geq 0$ (corresponding to $u_i \geq 0$), and λ associated with the constraint $\beta^\top \tilde{u} + d\tilde{v} = 1$.

The dual problem is:

$$\min_{\lambda,p,q}\{\lambda \colon p - q + \lambda\beta = \alpha, -1^\top p + \lambda d \geq c, p_i \geq 0, q_i \geq 0\}$$

By complementary slackness, at most one of p_i or q_i is nonzero. For brevity, let $r_i = Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))\mathbb{I}[T_i = t]$. Rearranging the first set of equality constraints gives $p_i - q_i = \mathbb{I}[T_i = t](b(X_i) - a(X_i))(r_i) - \lambda)$, which implies that

$$p_i = \mathbb{I}[T_i = t](b(X_i) - a(X_i))\max(r_i - \lambda, 0), \qquad q_i = \mathbb{I}[T_i = t](b(X_i) - a(X_i))\max(\lambda - r_i, 0)$$

Since the constraint $-\vec{1}^\top p + \lambda d \geq c$ is tight at optimality (otherwise there exists smaller yet feasible λ that achives lower objective of the dual program),

$$\sum_i \mathbb{I}[T_i = t](b(X_i) - a(X_i))\max(r_i - \lambda, 0) = \sum_i \mathbb{I}[T_i = t]a(X_i)(r_i - \lambda)$$

This rules out both $\lambda > \max_i r_i$ and $\lambda < \min_i r_i$, thus $r_{(k)} < \lambda \leq r_{(k+1)}$ for some k where $r_{(1)}, r_{(2)}, \ldots, r_{(n)}$ are the order statistics of the sample outcomes. This means that $q_i > 0$ can happen only when $r_i \leq r_{(k)}$, i.e., $u_i = 0$; and $p_i > 0$ can happen only when $i > k+1$, i.e., $u_i = 1$. Applying this, we may rewrite the above expression to recover that the optimal λ must be one of $\lambda_{(k)}$. This proves that the structure of the optimal solution is such that there exists a nondecreasing function $u : \mathcal{R} \to [0,1]$ such that $u_i = u(Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))\mathbb{I}[T_i = t])$ attains the upper bound. □

Proof of Proposition 1, sharpness of minimax policy regret. It suffices to show that every element in the interval is achieved by some $W \in \mathcal{W}^\Gamma$ and the converse: every $W \in \mathcal{W}^\Gamma$ achieves an element of the partially identified interval. The latter follows from the definition of the endpoints as $\inf_{W \in \mathcal{W}^\Gamma} R_{\pi_0}(\pi; W)$, $\sup_{W \in \mathcal{W}^\Gamma} R_{\pi_0}(\pi; W)$: every $W' \in \mathcal{W}$ is feasible so that $R_{\pi_0}(\pi; W')$ is in the partially identified interval for every feasible W'. We then use convexity of the partially identified interval and the linear reformulation of the fractional linear program to show that every element in the interval is achieved by some $W \in \mathcal{W}$. Consider a generic element r in the partially identified interval; by convexity, it can be expressed as the convex combination of the extreme points of the interval, $r = \lambda \inf_{W \in \mathcal{W}^\Gamma} R_{\pi_0}(\pi; W) + (1-\lambda) \sup_{W \in \mathcal{W}^\Gamma} R_{\pi_0}(\pi; W)$. Let $\underline{W}^*, \overline{W}^*$ be the weight vectors achieving the supremum and infimum, respectively:

$$\underline{W}^* \in \arg\min_{W \in \mathcal{W}^\Gamma} R_{\pi_0}(\pi; W), \overline{W}^* \in \arg\max_{W \in \mathcal{W}^\Gamma} R_{\pi_0}(\pi; W)$$

We then pass to the equivalent representation of regret in terms of *normalized* weights $w(\cdot,\cdot,t) = \frac{W(\cdot,\cdot,t)}{\mathbb{E}[W(\cdot,\cdot,t)\mathbb{I}[T=t]]}$. Define the corresponding normalized weights $\underline{\tilde{w}}^*(\cdot,\cdot,t), \overline{\tilde{w}}^*(\cdot,\cdot,t)$, and analogous normalization factors $\underline{t}, \overline{t}$,

$$\underline{\tilde{w}}^*(\cdot,\cdot,t) = \frac{\underline{W}^*}{\mathbb{E}[\underline{W}^* \mid T = t]}$$

Observe that by linearity of expectation and linearity of the objective function (with respect to *normalized weights*), r is realizable by the same convex combination of the minimizing/maximizing weights:

$$\begin{aligned} r &= \lambda\mathbb{E}[Y(\pi(X) - \pi_0(X))\underline{\tilde{w}}(X,T,Y)] + (1-\lambda)\mathbb{E}[Y(\pi(X) - \pi_0(X))\overline{\tilde{w}}(X,T,Y)] \\ &= \mathbb{E}[Y(\pi(X) - \pi_0(X))(\lambda\underline{\tilde{w}} + (1-\lambda)\overline{\tilde{w}})] \end{aligned}$$

It then remains to argue that $(\lambda\underline{\tilde{w}}(\cdot,\cdot,t)+(1-\lambda)\overline{\tilde{w}}(\cdot,\cdot,t))\in\psi_t\mathcal{W}_t^\Gamma, \forall t\in[0,\dots,m-1]$, with the conic relaxation factor $t_a=\lambda\underline{\psi}_t+(1-\lambda)\overline{\psi}_t$: this follows by convexity of $\psi_t\mathcal{W}_t^\Gamma$. Whenever the uncertainty sets are linearly representable, convex combinations of elements of feasible elements of the set $\underline{\tilde{w}},\overline{\tilde{w}}$ additionally satisfy the linear inequality or equality constraints of $\phi_a\mathcal{W}$. (e.g. the equality constraint $\mathbb{E}[\mathbb{I}[T=t]\,w(\cdot,\cdot,t)]=1$.). Again by linearity, we have that the bounds constraints are satisfied with $\psi_t=\lambda\underline{\psi}_t+(1-\lambda)\overline{\psi}_t$. The claim of sharpness follows.

Appendix B: Proofs of Uniform Convergence Guarantees

B.1. Uniform convergence: tail inequalities

In this subsection, we introduce definitions and stability results from empirical process theory in order to keep the argument self-contained, and provide maximal inequalities for the function classes of interest: Π, and reparametrizations of the optimal weight functions, $\mathcal{W}^\Gamma,\overline{\mathcal{W}}^\Gamma$. We will work with the *packing and covering numbers* of Π and the spaces of weight functions, and then relate these to bounds on the VC-major dimension of the policy class. For a subset S of some metric space, the packing number $D(\epsilon,S)$ is the largest number of points we can take in S that are not within ϵ distance of one another, and the covering number $N(\epsilon,S)$ is the smallest number of points we need to take in S so that every other point is within ϵ of one of these (Pollard 1990).[8] First we introduce the stability results from empirical process theory which will yield bounds on the covering numbers for the function classes of interest. We define the class of *VC-hull* functions, broader than VC-subgraph and related to VC-major, but which result in bounded Dudley entropy integrals.

Definition EC.1 (VC-hull class). Define $\mathrm{conv}(\mathcal{F})$, the convex hull of $\mathcal{F}$:

$$\mathrm{conv}(\mathcal{F})=\{\sum_{f\in\mathcal{F}}\lambda_f f\colon f\in\mathcal{F},\sum_f\lambda_f=1,\lambda_f\geq 0,\lambda\neq 0 \text{ for finitely many } f\}$$

$\overline{\mathrm{conv}}(\mathcal{F})$ is the pointwise sequential closure of the convex hull of $\mathcal{F}$: $f\in\overline{\mathrm{conv}}(\mathcal{F})$ if there exist $f_n\in\mathrm{conv}(\mathcal{F})$ such that $f_n(x)\to f(x)$ for all x in the domain of f, as $n\to\infty$. If the class $(\mathcal{F})$ is VC-subgraph, then $\overline{\mathrm{conv}}(\mathcal{F})$ is a VC hull class of functions.

A bounded VC-major class is a VC-hull class. Since VC-hull classes are defined with respect to the sequential closure of the convex hull $(\overline{\mathrm{conv}}(\mathcal{F}))$ of another function class $\mathcal{F}$, we frequently refer to $\mathcal{F}$ as the *generating* VC-subgraph class for its corresponding VC-hull class. VC-hull classes provide a constructive definition for a VC-major class in relation to a VC-subgraph class, and satisfy the following bound on the entropy integral of the covering numbers:

Theorem EC.1 **(Theorem 2.6.9 of Van Der Vaart and Wellner (1996); Ball and Pajor (1990))**. *Suppose there is a class of functions $\mathcal{F}$, with measurable square integral envelope F with bounded second moments, that is VC-subgraph of dimension V, such that $D(\epsilon\|F\|_2,\mathcal{F},\|\cdot\|_2)\leq C\left(\frac{1}{\epsilon}\right)^V$. Then, for $\overline{\mathrm{conv}}(\mathcal{F})$, the closure of the convex hull of $\mathcal{F}$ (e.g. the VC-hull class that is generated by $\mathcal{F}$), there exists a universal constant K depending on C and V only such that*

$$\log D(\epsilon\|F\|_2,\mathcal{F},\|\cdot\|_2)\leq K\left(\frac{1}{\epsilon}\right)^{2V/(V+2)}$$

[8] The packing and covering numbers are closely related by the inequality $N(\epsilon,t_0)\leq D(\epsilon,T_0)\leq N(\epsilon/2,t_0)$.

Working with VC-major (equivalently, VC-hull) classes allows us to use stronger stability results such as the following stability result on the stability of composition of the class of monotone functions and VC-major function classes, though at the expense of introducing a universal constant in the Dudley entropy integral.

LEMMA EC.1 **(Proposition 4.2 of Dudley (1987))**. *If $\mathcal{H}$ is a VC-major class for the generating class $\mathcal{C}$, and $\mathcal{U}_{\mathrm{nd}}$ denotes the set of all nondecreasing functions from $\mathbb{R} \mapsto [0,1]$, and*

$$\mathcal{F} := \{u \circ h \colon h \in \mathcal{H}, u \in \mathcal{U}_{\mathrm{nd}}\},$$

then $\mathcal{F}$ is a major class for the monotone derived class $\mathcal{D}$ of $\mathcal{C}$. Therefore if $\mathcal{H}$ is a VC-major class, so is $\mathcal{F}$.

These stability results allow us to prove, e.g. Proposition 3, that it is sufficient to restrict to optimizing over the set of worst-case weights (with additional structure).

Proof of Proposition 3 The result follows from Proposition 2 by applying Lemma EC.1. □

We introduce the uniform convergence setup we use to provide tail inequalities. We will apply a standard chaining argument with Orlicz norms and introduce some notations from standard references, e.g. Pollard (1990), Vershynin (2018), Wainwright (2019). A function $\phi : [0,\infty) \to [0,\infty)$ is an Orlicz function if ϕ is convex, increasing, and satisfies $\phi(0) = 0, \phi(x) \to \infty$ as $x \to \infty$. For a given Orlicz function ϕ, the Orlicz norm of a random variable X is defined as $\|X\|_\phi = \inf\{t > 0 \colon \mathbb{E}[\Phi(\|X\| \mid t)] \le 1\}$. The Orlicz norm $\|Z\|_\Phi$ of random variable Z is defined by

$$\|Z\|_\Phi = \inf\{C > 0 \colon \mathbb{E}[\Phi(Z/C)] \le 1\}.$$

A constant bound on $\|Z\|_\Phi$ constrains the rate of decrease for the tail probabilities by the inequality $\mathbb{P}(|Z| \ge t) \le 1/\Phi(t/C)$ if $C = \|Z\|_\Phi$. For example, choosing the Orlicz function $\Phi(t) = \frac{1}{5}\exp(t^2)$ results in bounds by subgaussian tails decreasing like $\exp(-Ct^2)$, for some constant C.

We next introduce the tail inequalities that use a standard chaining argument to control uniform convergence over $\pi \in \Pi$ and appropriate reparametrizations of the weight functions. First we define the following function classes conditional on all the data, $(X_{1:n}, T_{1:n}, Y_{1:n})$. For π, we consider a shifted function class with an envelope function: let $f_i(\pi) = (\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))Y_i$ where

$$\mathcal{F}(X_{1:n}, T_{1:n}, Y_{1:n}) = \{(f_1(\pi), \dots, f_n(\pi)) \colon \pi \in \Pi\}.$$

Next we introduce function classes for the weight functions: the minimax-regret achieving functions $W \in \overline{\mathcal{W}}^\Gamma(\pi)$ may also be written as compositions of the nominal weight functions with a function u,

$$W \circ u(\pi) = a_t^\Gamma(x) + u(y(\pi(t \mid x) - \pi_0(t \mid x)))(b_t^\Gamma(x) - a_t^\Gamma(x)),$$

where $u \in \mathcal{U}^\Gamma(\pi)$, the class of nondecreasing functions in the index $y(\pi(t \mid x) - \pi_0(t \mid x))$ for a fixed policy π, defined as the following:

$$\mathcal{U}^\Gamma(\pi) = \left\{u(x,t,y) : \mathbb{R} \mapsto [0,1] : u(y(\pi(t \mid x) - \pi_0(t \mid x))) \text{ is monotonic nondecreasing } \right\}.$$

Analogously, we let $\overline{\mathcal{U}}^\Gamma = \cup_{\pi \in \Pi} \mathcal{U}^\Gamma(\pi)$ denote the set of nondecreasing functions on the same index, but now ranging over $\pi \in \Pi$. Clearly, optimizing over $\overline{W}^\Gamma$ is equivalent to optimizing over $\overline{\mathcal{U}}^\Gamma$:

COROLLARY EC.1. *Let* $\overline{\mathcal{U}}^\Gamma = \cup_{\pi\in\Pi}\mathcal{U}^\Gamma(\pi)$. *Then, for any* $\pi\in\Pi$,

$$\overline{R}_{\pi_0}(\pi;\overline{\mathcal{U}}^\Gamma) = \sum_{t=0}^{m-1}\sup_{u\in\overline{\mathcal{U}}^\Gamma} R^{(t)}_{\pi_0}(\pi;W(u)), \quad \hat{\overline{R}}_{\pi_0}(\pi;\overline{\mathcal{U}}^\Gamma_n) = \sum_{t=0}^{m-1}\sup_{u\in\overline{\mathcal{U}}^\Gamma_n} \hat{R}^{(t)}_{\pi_0}(\pi;W(u)).$$

This characterization is a consequence of Proposition 2 and its proof, which studies the linear change of variables from W to u.

For this section, we consider maximal inequalities for the function classes for the enveloped policy class $\mathcal{F}$ and policy-optimal weight functions. Let $\epsilon_i\in\{-1,+1\}$, be iid Rademacher variables (symmetric Bernoulli random variables with value $-1,+1$ with probability $\frac{1}{2}$), distributed independently of all else.

LEMMA EC.2 **(Uniform convergence of policy function π over envelope class $\mathcal{F}$)**. *Let* $f(\pi)\le\|F\|_2\le C$ *be a bound on the envelope function for* $f\in\mathcal{F}$. *Then for* n *large enough, there exists a universal constant* K^Π *that depends only on the VC-major dimension of* Π, *such that with probability* $>1-\delta$,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n(f_i(\pi)-\mathbb{E}[f(\pi)])\right| \le {}^9\!/_2 CK^\Pi\sqrt{\frac{\log({}^5\!/_\delta)}{n}} \tag{EC.1}$$

Proof. We first bound the deviations uniformly over the policy class and introduce the following empirical processes,

$$M=\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n(f_i(\pi)-\mathbb{E}[f(\pi)])\right|, \quad L=\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\epsilon_i f_i(\pi)\right|.$$

By a standard symmetrization argument, applying Jensen's inequality for the convex function Φ of the symmetrized process (e.g. Theorem 2.2 of Pollard (1990)), we may bound the Orlicz norm of the maxima of the empirical process by the symmetrized process, conditional on the observed data:

$$\mathbb{E}[\Phi(M)]\le\mathbb{E}[\Phi(2L)].$$

Taking Orlicz norms with $\Phi(t)=\frac{1}{5}\exp(t^2)$, we apply a tail inequality on the Orlicz norm of the symmetrized process $\Phi(2L)$, under the assumption of bounded outcomes. Applying Dudley's inequality to the symmetrized empirical process L, (e.g. Theorem 3.5 of Pollard (1990)), we have that

$$\mathbb{E}_\epsilon\left[\exp(L^2/J^2)\,|\,\mathcal{D}\right]\le 5 \quad \text{for} \quad J=9\|F\|_2\int_0^1\sqrt{\log(D(\|F\|_2\zeta,\mathcal{F}(X_{1:n})))}d\zeta. \tag{EC.2}$$

Then, applying Theorem EC.1, we have that there exists a universal constant K^Π (depending only on the VC-major dimension of the policy class), such that

$$\log D(\|F\|_2\zeta,\mathcal{F}(X_{1:n},T_{1:n}))\le K\left(\frac{1}{\zeta}\right)^{\frac{2v}{v+2}}.$$

The corresponding Dudley entropy integral is bounded by $\int_0^1\sqrt{K\left(\frac{1}{\zeta}\right)^{2v/(v+2)}}d\zeta\le\sqrt{K}\frac{v+2}{2}=K^\Pi$. By Markov's inequality, we have that

$$\mathbb{P}\left(\frac{1}{n}L>t\right)\le 5\exp(-t^2n^2/\|L\|_2^2)=5\exp(-t^2n/J^2C^2),$$

so that therefore,

$$\frac{1}{n}M\le\frac{{}^9\!/_2 CK^\Pi\sqrt{\log({}^5\!/_\delta)}}{\sqrt{n}}.$$

□

LEMMA EC.3 **(Uniform convergence of weight functions $u(y(\pi(t \mid x) - \pi_0(t \mid x)))$ over $\mathcal{U}^\Gamma(\pi)$).** *With probability $\geq 1-p$, we have that*

$$\sup_{u \in \mathcal{U}^\Gamma(\pi)} \frac{1}{n} \left| \sum_{i=1}^n u(Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))) - \mathbb{E}[u(Y(\pi(T \mid X) - \pi_0(T \mid X)))] \right| \leq 9\sqrt{\frac{2\pi \log(1/p)}{n}}$$

Proof. We define the maxima of the empirical process (and its symmetrization), H, S, for the weight function $u \in \mathcal{U}^\Gamma(\pi)$, which maximizes over u for a fixed π:

$$H = \sup_{u \in \mathcal{U}^\Gamma(\pi)} \left| \sum_{i=1}^n u(Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))) - \mathbb{E}[u(Y(\pi(T \mid X) - \pi_0(T \mid X)))] \right|$$
$$S = \sup_{u \in \mathcal{U}^\Gamma(\pi)} \left| \sum_{i=1}^n \epsilon_i u(Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))) \right|$$

Taking Orlicz norms and symmetrizing as in the proof of Lemma EC.2, we have that $\mathbb{E}[\Phi(H)] \leq \mathbb{E}[\Phi(2S)]$. We show that the entropy integral (log of the covering numbers) is bounded using the VC-hull property of the class of non-decreasing functions taking values on $[0,1]$ ultimately is VC-hull but not VC-subgraph Van Der Vaart and Wellner (1996). $\mathcal{U}^\Gamma(\pi)$ is in fact included in the symmetric convex hull of $\mathcal{I}$, $\mathcal{U}^\Gamma(\pi) \subseteq \overline{\text{conv}}(\mathcal{I})$. (This follows since taking differences of indicator thresholds recovers any interval, e.g. Example 3.6.14 of Giné and Nickl (2016)). We apply Theorem EC.1 (relating the log-covering numbers to the entropy integral for VC-hull classes and their generating VC-subgraph classes), and using a result from Sec. 3 of Van Der Vaart et al. (1996) to extract an explicit bound for this class of functions:

$$\log D(\zeta, \mathcal{U}^\Gamma(\pi)(X_{1:n}, T_{1:n}, Y_{1:n})) \leq \frac{1}{\zeta} \log(\frac{1}{\zeta}).$$

The Dudley entropy integral is in turn bounded by $\int_0^1 \sqrt{\frac{1}{\zeta} \log(\frac{1}{\zeta})} d\zeta \leq \sqrt{2\pi}$. Next, we apply Theorem 3.5 of Pollard (1990), in order to bound mgf $\mathbb{E}\left[\exp\left(\frac{1}{81} \frac{1}{2\pi} \frac{S^2}{n}\right)\right] \leq 5$ such that we can use the subgaussian tail bound $\mathbb{P}\left[\frac{1}{n} S \geq t\right] \leq 5 \exp\left(-\frac{t^2}{162\pi} n\right)$. Therefore, with probability $\geq 1-p$, we have that

$$\frac{1}{n} S \leq 9\sqrt{\frac{2\pi \log(1/p)}{n}}$$

□

An analogous result holds for the restriction of the process to a specific treatment partition t.

COROLLARY EC.2. *With probability $\geq 1-p$, we have that*

$$\sup_{u \in \mathcal{U}^\Gamma(\pi)} \frac{1}{n} \left| \sum_{i=1}^n u(Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i)))\mathbb{I}[T_i = t] - \mathbb{E}[u(Y(\pi(T \mid X) - \pi_0(T \mid X)))\mathbb{I}[T = t]] \right| \leq 9\sqrt{\frac{2\pi \log(1/p)}{n}}$$

Next, we use the previous results to obtain a uniform convergence result for the minimax weight functions when we optimize jointly over policy functions π and weight functions $u(y(\pi(t \mid x) - \pi_0(t \mid x))) \in \overline{\mathcal{U}}^\Gamma$. Under Proposition 2, the weight function class remains monotone, even under composition with a VC-major policy class: however the dimension of the resulting class is not explicit from the stability result.

LEMMA EC.4 **(Uniform convergence of $u(y(\pi(t \mid x) - \pi_0(t \mid x)))$ over $\overline{\mathcal{U}}^\Gamma$).** *With probability $\geq 1-p$, for some universal constant K, which depends only on the VC-major dimension of the composition function class in Proposition 2 v, we have that*

$$\sup_{u \in \overline{\mathcal{U}}^\Gamma} \frac{1}{n} \left| \sum_{i=1}^n u(Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))) - \mathbb{E}[u(Y(\pi(T \mid X) - \pi_0(T \mid X)))] \right| \leq 9/2\sqrt{K} \frac{v+2}{2} \sqrt{\frac{\log(1/p)}{n}}$$

Proof. We first apply Lemma EC.1 (a stability result for the composition of monotone function classes with VC-major classes) under Assumption 3. Then, applying Theorem EC.1, there exists a universal constant K^{Π}, which depends only on the VC-major dimension v of the composition function class in Proposition 2, such that:

$$\log D(\zeta, \overline{\mathcal{U}}^{\Gamma}(X_{1:n}, T_{1:n}, Y_{1:n})) \leq K \left(\frac{1}{\zeta}\right)^{\frac{2v}{v+2}}$$

The result follows by the typical chaining argument (e.g. Lemma EC.3), but instead bounding the Dudley entropy integral by $\int_0^1 \sqrt{K\,(1/\zeta)^{\frac{2v}{v+2}}}\,d\zeta \leq \sqrt{K}\frac{v+2}{2}$. □

In the main text, we encapsulate dependence on K and v as the universal constant K^{Π}.

B.2. Proof of Theorem 2

Proof of Theorem 2 The proof of uniform convergence over $\pi \in \Pi, W \in \overline{\mathcal{W}}^{\Gamma}$ follows by decomposing the regret, then applying the tail inequalities of the previous section.

Regret Decomposition The following lemma allows us to study the minimax regret via uniform convergence arguments.

Lemma EC.5.

$$\sup_{y \in S} h(y) - \sup_{y \in S} g(y) \leq \sup_{y \in S}\{h(y) - g(y)\} \tag{EC.3}$$

Proof. To see this, consider $y_1^* \in \arg\max h(y), y_2^* \in \arg\max g(y)$ and $y^* \in \arg\max h(y) - g(y)$: then

$$h(y_1^*) - g(y_2^*) \leq h(y_1^*) - g(y_1^*) \leq h(y^*) - g(y^*)$$

□

We use Proposition 2 and Lemma EC.5 in the following minimax regret decomposition where $\pi_{CR} = \frac{1}{m}$:

$$\begin{aligned}
&\sup_{\pi \in \Pi}\left\{\sup_{W \in \mathcal{W}^{\Gamma}(\pi)} \hat{R}_{\pi_0}(\pi, W) - \sup_{W \in \mathcal{W}^{\Gamma}(\pi)} R_{\pi_0}(\pi, W)\right\} \\
&\sup_{\pi \in \Pi}\left\{\sup_{W \in \overline{\mathcal{W}}^{\Gamma}(\pi)} \hat{R}_{\pi_0}(\pi, W) - \sup_{W \in \overline{\mathcal{W}}^{\Gamma}(\pi)} R_{\pi_0}(\pi, W)\right\} \\
&\leq \sup_{\pi \in \Pi, W \in \overline{\mathcal{W}}^{\Gamma}(\pi)} \hat{R}_{\pi_0}(\pi, W) - R_{\pi_0}(\pi, W) \\
&\leq \sup_{\pi \in \Pi, W \in \overline{\mathcal{W}}^{\Gamma}(\pi)} \{\hat{R}_{\pi_{\mathrm{CR}}}(\pi, W) - R_{\pi_{\mathrm{CR}}}(\pi, W)\} + \underbrace{\sup_{W \in \overline{\mathcal{W}}^{\Gamma}(\pi_0)} \{\hat{R}_{\pi_{\mathrm{CR}}}(\pi_0, W) - R_{\pi_{\mathrm{CR}}}(\pi_0, W)\}}_{③}
\end{aligned}$$

Then, using subadditivity of the supremum, that $\mathcal{W}^{\Gamma}$ is a product uncertainty set, and the elementary decomposition $\frac{a}{b} - \frac{c}{d} = a\frac{b-d}{bd} + \frac{a-c}{d}$, we further decompose the minimax regret:

$$\begin{aligned}
&\sup_{\pi \in \Pi, W \in \overline{\mathcal{W}}^{\Gamma}(\pi)} \{\hat{R}_{\pi_{\mathrm{CR}}}(\pi, W) - R_{\pi_{\mathrm{CR}}}(\pi, W)\} \\
&\leq \sup_{\pi \in \Pi, W \in \overline{\mathcal{W}}^{\Gamma}(\pi)} \left\{\sum_{t=0}^{m-1} \frac{\mathbb{E}_n[(\pi(T \mid X) - \frac{1}{m})YW\mathbb{I}[T=t]])}{E_n[W\mathbb{I}[T=t]]} - \frac{\mathbb{E}[(\pi(T \mid X) - \frac{1}{m})YW\mathbb{I}[T=t]]}{\mathbb{E}[W\mathbb{I}[T=t]]}\right\} \\
&\leq \sup_{\pi \in \Pi, W \in \overline{\mathcal{W}}^{\Gamma}(\pi)} \sum_{t=0}^{m-1} \frac{(\mathbb{E}_n - \mathbb{E})((\pi(T \mid X) - \frac{1}{m})YW\mathbb{I}[T=t])}{\mathbb{E}[\mathbb{I}\,[T=t]\,W]}
\end{aligned}$$

$$+ \sup_{\pi\in\Pi, W\in\overline{\mathcal{W}}^\Gamma(\pi)} \sum_{t=0}^{m-1} \frac{\mathbb{E}_n[(\pi(T\mid X)-\frac{1}{m})YW\mathbb{I}[T=t]])}{\mathbb{E}_n[W\mathbb{I}[T=t]]} \frac{(\mathbb{E}_n-\mathbb{E})(W\mathbb{I}[T=t])}{\mathbb{E}[\mathbb{I}[T=t]W]}$$

$$\leq \sup_{\pi\in\Pi, W\in\overline{\mathcal{W}}^\Gamma(\pi)} \underbrace{(\mathbb{E}_n-\mathbb{E})((\pi(T\mid X)-1/m)YW)}_{①} + |B| \sum_{t=0}^{m-1} \sup_{W\in\overline{\mathcal{W}}^\Gamma(1)} \underbrace{|(\mathbb{E}_n-\mathbb{E})(W\mathbb{I}[T=t])|}_{②}$$

The last inequality follows by applying submultiplicativity of the supremum (for absolute values), and since $\mathbb{E}[W\mathbb{I}[T=t]]=1$. The upper bound $\sup_{\pi\in\Pi, W\in\overline{\mathcal{W}}^\Gamma(\pi)}\left|\frac{\mathbb{E}_n[(\pi(T|X)-\frac{1}{m})YW\mathbb{I}[T=t]])}{\mathbb{E}_n[W\mathbb{I}[T=t]]}\right| \leq B$ follows since this term simply evaluates the minimax regret over $\overline{\mathcal{W}}^\Gamma(\Pi)$: due to weight normalization, it is *deterministically* bounded by B under Assumption 1. We now apply the tail inequalities of the previous section to the maximal processes of ①, ②, ③, in this order.

① *Reducing a bound on product function class to the individual function classes.* Recall the weight functions are re-parametrized with respect to u: throughout this analysis, for brevity, we denote this by $W_i(u(\pi))$:

$$W_i(u(\pi)) = a^\Gamma_{T_i}(X_i) + (b^\Gamma_{T_i}(X_i) - a^\Gamma_{T_i}(X_i))u_{T_i}(Y_i(\pi(T_i\mid X_i) - \frac{1}{m})).$$

Now define (Q,P) for the quantities for the empirical process for the product function class and its symmetrized version:

$$Q = \sup_{f\in\mathcal{F}, W\in\overline{\mathcal{W}}^\Gamma(\pi)} \left|\sum_{i=1}^n (f_i(\pi)W_i(u(\pi))) - R_{\pi_{CR}}(\pi,W))\right|,\ P = \sup_{f\in\mathcal{F}, W\in\overline{\mathcal{W}}^\Gamma(\pi)} \left|\sum_{i=1}^n \epsilon_i f_i(\pi)W_i(u(\pi)))\right|$$

By a symmetrization argument (Theorem 2.2 of Pollard (1990)), we have that

$$\mathbb{E}\Phi(Q) \leq \mathbb{E}[\Phi(2P)]$$

We now reformulate the maximal inequality over the product function class in terms of Orlicz norms on each function class $\mathcal{F}, \overline{\mathcal{W}}^\Gamma$ separately, using the fact that observation $fW = \frac{1}{4}(f+W)^2 - \frac{1}{4}(f-W)^2)$. For the weight function $W_i(u(\pi))$, we will use the contraction map $\lambda(s) = 1/2\max_x b(x) - \min_x a(x) \min(1, s^2)$. We then decompose the terms including the product of f, W to the sums of squares of f, W, optimize over $W \in \overline{\mathcal{W}}^\Gamma$ rather than $W \in \mathcal{W}^\Gamma(\pi)$, and then apply a contraction result in order to use results on convergence over $\pi\in\Pi, u\in\overline{\mathcal{U}}^\Gamma$. We next apply inequality 5.5. of Pollard (1990), which decomposes the maximal inequality over the addition of function classes, $\mathbb{E}_\epsilon\Phi(\sup_{f\in\mathcal{F}, \overline{\mathcal{W}}^\Gamma(\pi)} |\epsilon\cdot(f+W)|) \leq \frac{1}{2}\mathbb{E}_\epsilon\Phi\left(2\sup_{f\in\mathcal{F}}|\epsilon\cdot f|\right) + \frac{1}{2}\mathbb{E}_\epsilon\Phi\left(2\sup_{\overline{\mathcal{W}}^\Gamma(\pi)}|\epsilon\cdot W|\right)$.

$$\mathbb{E}[\Phi(2P)]$$

$$\leq \mathbb{E}_\epsilon\Phi\left(\sup_{f\in\mathcal{F}, \overline{\mathcal{W}}^\Gamma(\pi)} \frac{1}{2}\left|\epsilon\cdot(f(\pi) + (a^\Gamma + (b^\Gamma - a^\Gamma)u))^2\right|\right) + \mathbb{E}_\epsilon\Phi\left(\sup_{f\in\mathcal{F}, \overline{\mathcal{W}}^\Gamma(\pi)} \frac{1}{2}\left|\epsilon\cdot(f(\pi) - (a^\Gamma + (b^\Gamma - a^\Gamma)u))^2\right|\right)$$

$$\leq \mathbb{E}_\epsilon\Phi\left(\sup_{f\in\mathcal{F}, \overline{\mathcal{W}}^\Gamma(\pi)} \left|\frac{1}{2}\epsilon\cdot(f(\pi) \pm (b^\Gamma - a^\Gamma)u)^2\right|\right) + \frac{1}{2}\mathbb{E}_\epsilon\Phi\left(\sup_{f\in\mathcal{F}, \overline{\mathcal{W}}^\Gamma(\pi)} 2\left|\epsilon\cdot a^\Gamma(f(\pi) \pm (b^\Gamma - a^\Gamma)u)\right|\right)$$

$$\leq 3\mathbb{E}_\epsilon\Phi\left(8\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i f_i(\pi)\right|\right) + \frac{1}{2}\mathbb{E}_\epsilon\Phi\left(4\frac{1}{\nu}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i f_i(\pi)\right|\right)$$

$$+ 3\mathbb{E}_\epsilon\Phi\left(8\frac{1}{\nu}(\Gamma - \frac{1}{\Gamma})\sup_{u\in\overline{\mathcal{U}}^\Gamma}\left|\sum_{i=1}^n \epsilon_i u(Y_i(\pi(T_i\mid X_i) - \frac{1}{m}))\right|\right) + \frac{1}{2}\mathbb{E}_\epsilon\Phi\left(4\frac{1}{\nu^2}(\Gamma - \frac{1}{\Gamma})\sup_{u\in\overline{\mathcal{U}}^\Gamma}\left|\sum_{i=1}^n \epsilon_i u(Y_i(\pi(T_i\mid X_i) - \frac{1}{m}))\right|\right)$$

The last inequality follows from a Lipschitz contraction result (see e.g. Theorem 5.7 of Pollard (1990)). From the above decomposition, it remains to apply the tail inequalities of lemmas EC.2 to EC.4 and a contraction argument separately for the function classes on $\mathcal{F}, \overline{\mathcal{U}}^\Gamma$.

For n large enough, with probability greater than $1-p_1$, where $p_1 = \frac{p}{6}$,

$$\mathbb{E}[\Phi(2P)] \leq 18(12 + 1/\nu)(BK^{\Pi} + 2\frac{1}{\nu}(\Gamma - \frac{1}{\Gamma})K^{\Pi})\sqrt{\frac{\log(30/p)}{n}}$$

② We next bound the maximal deviations of the term

$$\sum_{t=0}^{m-1} \sup_{u \in \mathcal{U}^{\Gamma}(1)} \left| \frac{1}{n} \sum_i \mathbb{I}[T_i = t] W(t, X_i, Y_i) - \mathbb{E}[\mathbb{I}[T = t] W(t, X, Y)] \right|.$$

Note that studying uniform convergence of ②, ③, we can restrict attention to nondecreasing weights which are nondecreasing in a fixed policy, $\mathcal{U}^{\Gamma}(1)$. We apply the tail inequality of Lemma EC.3 with a contraction argument, and obtain a bound on the maxima of the absolute value deviation by an argument of Remark 8.1.5 of Vershynin (2018): note that the zero function is an element of the class of non-decreasing functions on $\mathbb{R}$, and apply Dudley's inequality to the increment process $\left|\hat{D}_t - 0\right|$. Choosing $p_2 = \frac{p}{3m}$, and taking a union bound over the event that each bound holds for each treatment partition t, we obtain the high probability bound that

$$\sum_{t=0}^{m-1} \sup_{u \in \mathcal{U}^{\Gamma}(1)} \left| \frac{1}{n} \sum_i \mathbb{I}[T_i = t] W(t, X_i, Y_i) - \mathbb{E}[\mathbb{I}[T = t] W(t, X, Y)] \right| \leq \frac{18m^{1/\nu}(\Gamma - \frac{1}{\Gamma})\sqrt{\log(15m/p)}}{\sqrt{n}}$$

③ Lastly, we bound $\sup_{u \in \mathcal{U}^{\Gamma}(\pi_0)} \left| \hat{R}_{\pi_{CR}}(\pi_0, W(u(\pi_0))) - R_{\pi_{CR}}(\pi_0, W(u(\pi_0))) \right|$, follows from the tail inequality of Lemma EC.3, such that for n large enough, , with probability greater than $1-p_2$, where $p_2 = \frac{p}{3}$,

$$\sup_{u \in \mathcal{U}^{\Gamma}(\pi_0)} \left| \hat{R}_{\pi_{CR}}(\pi_0, W(u(\pi_0))) - R_{\pi_{CR}}(\pi_0, W(u(\pi_0))) \right| \leq \frac{18B^{1/\nu}(\Gamma - \frac{1}{\Gamma})\sqrt{\log(15/p)}}{\sqrt{n}}$$

Putting together the above bounds on terms ①, ②, ③ we have that with probability $\geq 1-p$,:

$$\begin{aligned}
&\sup_{\pi \in \Pi} \left| \hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^{\Gamma}) - \overline{R}_{\pi_0}(\pi; \mathcal{W}^{\Gamma}) \right| \\
&\leq 18(12 + \nu^{-1})(BK^{\Pi} + \nu^{-1}(\Gamma - \Gamma^{-1})(2K^{\Pi} + B + m))\sqrt{\frac{\log(15m/p)}{n}}
\end{aligned}$$

□

B.3. Proof of Theorem 1

Proof of Theorem 1 We analyze uniform convergence for the true propensity weights, assumed to be in the uncertainty set, $W^* \in \mathcal{U}$. We use the tail inequalities of lemma EC.2, as well as standard Hoeffding inequalities for the sample expectations, with the *true* inverse propensity weights $W_t^*(X_i, Y_i)$. Define

$$\hat{D}_t^* = \mathbb{E}_n[(\pi(t \mid X) - 1/m) W^* \mathbb{I}[T = t]].$$

First consider an analogous regret decomposition as in the proof of Theorem 2:

$$\begin{aligned}
&\sup_{\pi \in \Pi} \{\hat{R}_{\pi_0}(\pi, W^*) - R_{\pi_0}(\pi, W^*)\} \\
&\leq \sup_{\pi \in \Pi} \{\hat{R}_{\pi_{\mathrm{CR}}}(\pi, W^*) - R_{\pi_{\mathrm{CR}}}(\pi, W^*)\} + \left(\hat{R}_{\pi_{\mathrm{CR}}}(\pi_0, W^*) - R_{\pi_{\mathrm{CR}}}(\pi_0, W^*)\right)
\end{aligned}$$

Note that the second term can be bounded by application of Hoeffding's inequality, such that with probability $\geq 1-p_3$,

$$\left|\hat{R}_{\pi_{CR}}(\pi_0, W^*) - R_{\pi_{CR}}(\pi_0, W^*)\right| \leq {}^{B}\!/_{\nu}\sqrt{\frac{\log({}^{2}\!/_{p_3})}{2n}}$$

Next, we bound the regret deviation uniformly over π:

$$\begin{aligned}
&\sup_{\pi\in\Pi}\{\hat{R}_{\pi_{\mathrm{CR}}}(\pi, W^*) - R_{\pi_{\mathrm{CR}}}(\pi, W^*)\} \\
&\quad\leq \sup_{\pi\in\Pi}\frac{1}{n}\sum_i \frac{(\pi(T_i \mid X_i) - \frac{1}{m})W_i^* Y_i}{\mathbb{E}[\hat{D}_{T_i}]} - R_{\pi_0}(\pi, W^*) + \frac{1}{n}\sum_i \frac{(\pi(T_i \mid X_i) - \frac{1}{m})W_i^* Y_i}{\mathbb{E}[\hat{D}_{T_i}]}\frac{\mathbb{E}[\hat{D}_{T_i}] - \hat{D}_{T_i}}{\hat{D}_{T_i}} \\
&\quad\leq \sup_{\pi\in\Pi}\left\{\frac{1}{n}\sum_i (\pi(T_i \mid X_i) - \frac{1}{m})W_i^* Y_i - R_{\pi_0}(\pi, W^*)\right\} + \frac{B}{\nu}\sum_i \frac{1}{n}\frac{\mathbb{E}[\hat{D}_{T_i}] - \hat{D}_{T_i}}{\hat{D}_{T_i}}
\end{aligned}$$

We apply Lemma EC.2 (e.g. a standard chaining argument with bounded envelope function $WY \leq {}^{B}\!/_{\nu}$) to bound the first term. Therefore, we have that with high probability greater than p_2, the first term is bounded by:

$$\sup_{\pi\in\Pi}\left\{\frac{1}{n}\sum_i (\pi(T_i \mid X_i) - \frac{1}{m})W_i^* Y_i - R_{\pi_0}(\pi, W^*)\right\} \leq \frac{9B}{2\nu}\sqrt{\frac{\log({}^{5}\!/_{p_2})}{n}}.$$

We then bound the second term, $\frac{B}{\nu}\sum_i \frac{1}{n}\frac{\mathbb{E}[\hat{D}_{T_i}] - \hat{D}_{T_i}}{\hat{D}_{T_i}}$: instead of summing the second term over treatments t, observe that for $n_t = \sum_i \mathbb{I}[T_i = t]$,

$${}^{B}\!/_{\nu}\sum_i \frac{1}{n}\frac{\mathbb{E}[\hat{D}_{T_i}] - \hat{D}_{T_i}}{\hat{D}_{T_i}} = {}^{B}\!/_{\nu}\frac{1}{n}\sum_{t=0}^{m-1} n_t \frac{\left|\hat{D}_t - 1\right|}{\hat{D}_t}$$

We proceed conditionally on the event that $\frac{n_t}{n} \in [\frac{1}{2}\rho_t, \frac{3}{2}\rho_t],\ \forall t \in \{0, \ldots, m-1\}$, where $\rho_t = \mathbb{P}(T = t)$ is the marginal probability of treatment. By Hoeffding's inequality, $\mathbb{P}(\left|\frac{n_t}{n} - \rho_t\right| \geq \rho_t/2) \leq 2\exp(-\frac{1}{2}\nu^2\rho_t^2 n)$, so it suffices to choose $p_4 \in [0,1]$ such that $\frac{1}{\nu}\sqrt{\frac{\log(2m/p_4)}{2n}} \leq \rho_t^2/2, \forall t \in \{0, \ldots, m-1\}$ (after taking a union bound over the m treatment groups). Next, we bound $\frac{\left|\hat{D}_t - 1\right|}{\hat{D}_t}$: by Hoeffding's inequality,

$$\mathbb{P}(\left|\hat{D}_t - 1\right| \geq \epsilon) \leq 2\exp(-2\nu^2\epsilon^2 n)$$

For $p_1 \in [0,1]$ such that $\frac{1}{\nu}\sqrt{\frac{\log(2m/p_1)}{2n}} \leq 1$ then with probability at least $1-p_1$, $\frac{1}{\hat{D}_t} \leq 2$ and $\frac{\left|(1-\hat{D}_t)\right|}{\hat{D}_t} \leq \frac{2}{\nu}\sqrt{\frac{\log({}^{2}\!/_{mp_1})}{2n}}, \forall t \in \{0, \ldots, m-1\}$ (again taking a union bound over $t \in \{0, \ldots, m-1\}$). Now combining the above tail inequalities and applying the union bound, we have that for $p_1, p_2, p_3, p_4 = \frac{\delta}{4}$ for $p > 0$, with high probability greater than $1-p$,

$$\begin{aligned}
\sup_{\pi\in\Pi}\{\hat{R}_{\pi_0}(\pi, W^*) - R_{\pi_0}(\pi, W^*)\} &\leq \frac{B}{\nu}\sqrt{\frac{\log({}^{8}\!/_{p_3})}{2n}} + 36\frac{B\sqrt{v}}{\nu}\frac{\sqrt{\log({}^{20}\!/_{p})}}{\sqrt{n}} + \frac{3}{\nu}\sqrt{\frac{\log({}^{8m}\!/_{p_1})}{2n}} \\
&\leq \frac{1}{\nu}(B(1 + \frac{9}{2}K(v+2)) + 3)\sqrt{\frac{2\log({}^{\max(8m,20)}\!/_{\delta})}{n}}
\end{aligned}$$

Lastly, the proof follows by noting that by assumption of well-specification, $W_t^* \in \mathcal{W}_t$, so there exists $\psi_t > 0, \forall t \in \mathcal{T}$ such that $\frac{W_t^*}{\psi_t} \in \mathcal{W}_t$, and we have that therefore $\hat{R}_{\pi_0}(\pi, W^*) \leq \sup_{W^* \in \mathcal{W}} \hat{R}_{\pi_0}(\pi, W^*)$. And, in the statement, we have further folded all v-dependent constants into one.

B.4. Proof of Proposition 4

Proof of Proposition 4 We prove that the budgeted uncertainty set solution has bounded entropy integral by first taking a partial Lagrangian dual with respect to the budget constraint, then invoking strong duality to study a partial maximization: we show the solution can be reparametrized to instead range over the space of nondecreasing functions on $[0,1]$, $\mathcal{U}$, for the fixed *optimal* η^*, ψ^*. We then proceed to argue that the structural result implies, using the equivalence of the linearized fractional program and the fractional program, that we may then correspondingly optimize over the values of η, ψ, and the space of nondecreasing functions. This implies that it is sufficient to restrict the optimization to the set of nondecreasing functions (which satisfy the budget constraint), and we may optimize over a set of restricted complexity. This, for example, allows us to leverage the same stability results as in the proof of Theorem 2 to obtain the same regret guarantees.

We first analyze the linearized budgeted linear program in Section 6.1 (that is, post-Charnes-Cooper transformation) for $\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda})$. Throughout, we presume that Γ is some fixed input and write a, b for a^Γ, b^Γ. We also analyze the problem within a single treatment component, and reindex $i = 1, \dots, n$ to be counting conditional on a treatment component.

$$\begin{aligned}
\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \max_{\psi \geq 0, w \geq 0, d} \quad & \sum_{i=1}^n w_i r_i \\
\text{s.t.} \quad & \sum_{i=1}^n d_i \leq \Lambda \psi, \quad \sum_{i=1}^n w_i = 1 \\
& a_i \psi \leq w_i \leq b_i \psi \quad \forall\, i = 1, \dots, n \\
& d_i \geq w_i - \tilde{W}_i \psi \quad \forall\, i = 1, \dots, n \\
& d_i \geq \tilde{W}_i \psi - w_i \quad \forall\, i = 1, \dots, n
\end{aligned}$$

In the following, we condense the linearized representation for the absolute value variable d_i and write $d_i = \left| w_i - \tilde{W}_i \psi \right|$ for brevity. First, we take the Lagrangian partial dual, dualizing the normalized budget constraint $\sum_i d_i \leq \Lambda \psi$ with Lagrange multiplier η:

$$\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \min_{\eta \geq 0} \max_{w,d,\psi} \{ \sum_i w_i r_i + \eta(\Lambda \psi - \sum_i d_i) \colon a\psi \leq w \leq b\psi, \sum_i w_i = 1, d_i = \left| w_i - \tilde{W}_i \psi \right| \}$$

We consider a partial maximization, and substitute with the transformation $u_i = \frac{w_i - a_i \psi}{\psi(b_i - a_i)}, u \in [0,1]$. Define

$$\begin{aligned}
m(u, \psi, \eta) &:= \sum_i r_i (\psi(b_i - a_i) u_i + \psi a_i) + \eta^* (\Lambda \psi - \sum_i d_i) \\
\mathcal{S}(\psi) &:= \left\{ \sum_i \psi(b_i - a_i) u_i + \psi a_i = 1, d_i = \left| w_i(u) - \tilde{W}_i \psi \right|, 0 \leq u_i \leq 1, i = 1, \dots, n \right\}
\end{aligned}$$

so that

$$\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \min_{\eta \geq 0} \max_{t > 0, u \in \mathcal{S}(\psi)} m(u, \psi, \eta)$$

By a standard min-max theorem, we may interchange the min and maximum, and by strong duality (with the Slater point of u such that $w_i(u) = \tilde{W}\psi$), there exists a saddle point pair $(u^*, \psi^*), \eta^*$ that are best-responses to each other such that $\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = m(u^*, v^*, \eta^*)$. We argue further that $\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \max_{u \in S(\psi^*)} m(u, \psi^*, \eta^*)$; e.g. we may fix a partial best response of (u, ψ^*) and η^*, and recover the optimal solution when we

optimize over u. (We show this by contradiction: Suppose not: that $\tilde{u}^* \in \arg\max_u m(u, \psi^*, \eta^*)$ is such that $m(\tilde{u}^*, \psi^*, \eta^*) > m(u^*, \psi^*, \eta^*)$. This contradicts strong duality. On the contrary, $m(\tilde{u}^*, \psi^*, \eta^*) < m(u^*, \psi^*, \eta^*)$ is not possible since u^* is feasible for ψ^*, η^* and therefore achieves a better objective value; so this contradicts definition of $\tilde{u}^* \in \arg\max_u m(u, \psi^*, \eta^*)$.) Therefore, by the preceding argument,

$$\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \max_{u \in \mathcal{S}(\psi^*)} m(u, \psi^*, \eta^*)$$

We further simplify and drop terms from the *parametric objective* $m(u, \psi^*, \eta^*)$ that are constant given η^*, ψ^* and therefore do not vary with u, such that we recover the *globally optimal* u^* by optimizing the reformulated objective $m'(u, \psi^*, \eta^*)$:

$$m'(u, \psi^*, \eta^*) := \sum_i r_i(\psi^*(b_i - a_i)u_i) - \eta^* \sum_i d_i)$$
$$u^* \in \arg\max_{u \in \mathcal{S}(\psi^*)} m'(u, \psi^*, \eta^*)$$

We next prove that we can *further* reparametrize optimization of the objective function $\overline{m}'(u, \psi^*, \eta^*)$ over $u \in \mathcal{S}(\psi^*)$ to the class of u vectors that is nondecreasing in r,

$$\mathcal{U} = \{u : \mathbb{R} \mapsto [0,1], u \text{ monotonically nondecreasing}\}.$$

We prove the following technical result, which establishes a structural result on the globally optimal $u^*(\psi, \eta)$ which establishes that is of bounded complexity.

LEMMA EC.6 **(Nondecreasing parametrization of optimal u for budgeted uncertainty set)**. *Fix $\psi, \eta \geq 0$: then the correspondingly optimal rescaled weight function $u^*(\psi, \eta)$, defined as the solution to the optimization problem,*

$$u^*(\psi, \eta) \in \arg\max\{\psi \sum_i (b_i - a_i)\left(r_i - \eta \operatorname{sgn}(w_i(u) > \tilde{W}\psi)\right) u_i : 0 \leq u \leq 1, \sum_i \psi(b_i - a_i)u_i + \sum_i a_i = 1\},$$

is non-decreasing in the coefficient index vector r. Therefore, $u^*(\psi, \eta)$ is nondecreasing in r for all ψ, η.

Proof of Lemma EC.6 By the preceding arguments, we have established the optimal subproblem solution can be written as the following program:

$$\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \max_{u \in \mathcal{S}(\psi)} \max_{\psi > 0} \min_{\eta \geq 0} \sum_i \psi(b_i - a_i)\left(r_i - \eta \operatorname{sgn}(w_i(u) > \tilde{W}\psi)\right) u_i$$
$$0 \leq u \leq 1$$
$$\sum_i \psi(b_i - a_i)u_i = 1 - \sum_i a_i$$

The idea is that given the optimal dual variable η^* and scaling factor ψ^*, the problem reduces to a similar problem as the fractional knapsack problem: it is sufficient to sort first on the multipliers r_i; then fill the knapsack lexicographically in order of distance $\left|w_i(u) - \tilde{W}\psi^*\right|$ (since the η^* penalty is fixed and identical for all units). We will prove the reparametrization over $\mathcal{S}(\psi^*) \cap \mathcal{U}$ by contradiction. Suppose not: that the optimal solution, u^* had indices i, i' such that $r_i > r_{i'}$ but $u_i < u_{i'}$. We enumerate the following cases that exhaust the possible orderings of $u_i, u_{i'}$ relative to $\tilde{W}_i\psi^*, \tilde{W}_{i'}\psi^*$:

- $w_i(u_i) < \tilde{W}_i\psi^*, w_{i'}(u_{i'}) < \tilde{W}_{i'}\psi^*$ or $w_i(u_i) > \tilde{W}_i\psi^*, w_{i'}(u_{i'}) > \tilde{W}_{i'}\psi^*$: For any same-ordered set we could generate a contradiction by increasing u_i without generating a change in sign that changes the η^* coefficient.
- $w_i(u_i) > \tilde{W}_i\psi^*, w_{i'}(u_{i'}) < \tilde{W}_{i'}\psi^*$: increasing u_i cannot change sign of η^*.
- $w_i(u_i) = \tilde{W}_i\psi^*$: We need only consider a simultaneous perturbation increasing u_i and moving $u_{i'}$ such that $d_{i'}(u_{i'})$ is decreasing; either such a perturbation increases $u_{i'}$ and overall increases the objective, or decreases $u_{i'}$ (which is offset by the increase due to $r_i > r_i'$, and offsets the increase in $d_i(u_i)$.

□

Note that the characterization of Lemma EC.6, which states that the optimal ψ, η-parametrized solution $u^*(\psi, \eta)$ is nondecreasing in r, in fact characterizes the structure of the optimal *set* of $u(\psi, \eta)$ *for all* ψ, η since the index for monotonicity, r, is independent of the parameters ψ, η. Of course, the particular optimal solution $u^*(\psi, \eta)$ may change in ψ, η. As a consequence,

$$\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \max_{u \in \mathcal{S}(\psi^*) \cap \mathcal{U}} m(u, \psi^*, \eta^*)$$

Combining this structural result with the preceding arguments, we establish that we can equivalently search over scalars $\psi, \eta > 0$, and $u \in \mathcal{S}(\psi) \cap \mathcal{U}$.

$$\begin{aligned} \hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \max_{u \in \mathcal{S}(\psi) \cap \mathcal{U}} \max_{\psi > 0} \min_{\eta \geq 0} & \sum_i \psi(b_i - a_i) \left(r_i - \eta \operatorname{sgn}(w_i(u) > \tilde{W}\psi) \right) u_i \\ & 0 \leq u \leq 1 \\ & \sum_i \psi(b_i - a_i) u_i = 1 - \sum_i a_i \end{aligned}$$

We note that by the equivalence of the linear-fractional programs and linearized program, e.g. via the primal variables $W, U = \frac{W-a}{(b-a)}$ on the one hand and the scalarized $w = W\psi, \psi = \sum_i W, u = \frac{w - a\psi}{\psi(b-a)}$ on the other hand, (and the implied transformations on d), our structural result that it is equivalent to optimize over $u(\psi, \eta)$ nondecreasing implies that $U^*(\psi, \eta) = \frac{u^*(\psi,\eta)}{\psi}, U \in [0, 1]$ is *also* a monotonically nondecreasing function in r. (Multiplying by the scalar $\psi > 0$ simply induces an isomorphism to the *same* set of monotonically nondecreasing functions in r). Using this final transformation, we show that our structural result holds also for the original primal problem.

$$\hat{\overline{Q}}(r; \mathcal{W}_n^{\Gamma,\Lambda}) = \max_{U \in \mathcal{U}} \left\{ \frac{\sum_i U_i(b_i - a_i) r_i + a_i r_i}{\sum_i U_i(b_i - a_i) + a_i} : \sum d_i(U_i) \leq \Lambda \right\}$$

To contextualize this characterization, we remark that this is weaker than Theorem 3 as this does not provide us with an algorithmic solution: nonetheless, proving this result that it is sufficient to optimize over $\mathcal{U}$, even in the primal nonconvex fractional formulation, is sufficient to establish uniform convergence. Finally, we specialize the analysis to the setting for our estimator, where $r_i = \pi(T_i \mid X_i) - \pi_0(T_i \mid X_i) Y_i$, which introduces a dependence on $\pi(X_i)$. (Note that the dependence is only on X through the function π, which is of restricted complexity.) Since we only required the VC-major property of $u(r)$, applying Lemma EC.1 is sufficient to verify that the VC-major property holds when we also range the policy $\pi \in \Pi$.

B.5. Proof of Proposition 5

Proof of Proposition 5 The proof is similar to that of Proposition 6: we study sensitivity analysis in the dual of the linearized linear program, in order to isolate an additive approximation error term of the sample budget constraint from its population counterpart; we control the latter uniformly over the space of weights by our previous tail inequality. Since we optimize in the sample based on an *sample expectation estimate* of the L1 budget constraint, we recall the definitions of $\mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P}_n)$ and $\mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P})$:

$$\mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P}_n) = \left\{ W \in \mathbb{R}_+^n : \text{ s.t. } \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} |W_i - \tilde{W}_i| \le \Lambda_t,\ a_i^\Gamma \le W_i \le b_i^\Gamma \ \forall i \right\}$$
$$\mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P}) = \left\{ W \in \mathbb{R}_+^n : \text{ s.t. } \mathbb{E}[|W(T,X,Y) - \tilde{W}(T,X)| \,|\, T=t] \le \Lambda_t,\ a_i^\Gamma \le W_i \le b_i^\Gamma \ \forall i \right\}$$

Now, we use Proposition 4 to equivalently parametrize the optimization over the set of weight functions which include the *nondecreasing* component $u(y(\pi(t \mid x) - \pi_0(t \mid x)))$, and introduce the corresponding *nondecreasing sample-budgeted* uncertainty set, $\overline{\mathcal{W}}_n^{\Gamma,\Lambda}(\mathbb{P}_n)$:

$$\overline{\mathcal{W}}^{\Gamma,\Lambda}(\pi;\mathbb{P}_n) = \left\{ W(t,x,y) : \begin{array}{c} W(t,x,y) = a_t^\Gamma(x) + u(y(\pi(t \mid x) - \pi_0(t \mid x))) \cdot (b_t^\Gamma(x) - a_t^\Gamma(x)), \\ u(y(\pi(t \mid x) - \pi_0(t \mid x))) : \mathbb{R} \to [0,1] \text{ is monotonic nondecreasing}, \\ \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} |W(T_i,X_i,Y_i) - \tilde{W}(T_i,X_i,Y_i)| \le \Lambda_t \end{array} \right\}$$

In analogy to Corollary 1, we may define the union over the policy class $\overline{\mathcal{W}}^{\Gamma,\Lambda}(\mathbb{P}_n) = \cup_{\pi \in \Pi} \overline{\mathcal{W}}^{\Gamma,\Lambda}(\pi;\mathbb{P}_n)$. The next corollary, a consequence of the nondecreasing optimal solution characterization of Proposition 4 states that we recover the optimal regret by optimizing over the restricted class of budgeted weights.

Corollary EC.3.
$$\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}^{\Gamma,\Lambda}(\mathbb{P}_n)) = \sum_{t=0}^{m-1} \sup_{W \in \overline{\mathcal{W}}^{\Gamma,\Lambda}(\mathbb{P}_n)} \hat{R}_{\pi_0}^{(t)}(\pi; W)$$

We show that $\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}^{\Gamma,\Lambda}(\mathbb{P}_n))$ and $\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P}))$ are close for the two policies of interest in the minimax regret bound: the sample-optimal $\hat{\bar{\pi}} := \hat{\bar{\pi}}(\Pi, \mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P}_n), \pi_0)$ and population-optimal $\pi^* \in \arg\inf_{\pi \in \Pi} \overline{R}_{\pi_0}(\pi; \mathcal{W}^{\Gamma,\Lambda}(\mathbb{P}))$ policies. The result will follow by applying this bound with the triangle inequality.

In the following, we denote $r_i = Y_i(\pi(T_i \mid X_i) - \pi_0(T_i \mid X_i))$ for brevity, and apply the Charnes cooper transformation. Define $\tilde{u} = u\psi$ as the corresponding transformation for u in the change of variables $W = a + (b-a)u$, and note that this preserves monotonicity of $\tilde{u}$ for all ψ. Denote the uncertainty set on w, ψ and implicitly, nondecreasing u as $\overline{\mathcal{S}}(w,\psi,\tilde{u};\pi)$:

$$\overline{\mathcal{S}}(w,\psi,\tilde{u};\pi) = \left\{ \begin{array}{c} \sum_i w_i = 1;\ \psi_{T_i} a_i^\Gamma \le w \le b_i^\Gamma \psi_{T_i}, \forall i = 1,\dots,n \\ w = a_i^\Gamma \psi_{T_i} + (b_i^\Gamma - a_i^\Gamma)\tilde{u}_i, \forall i = 1,\dots,n \\ \tilde{u}(y(\pi(t \mid x) - \pi_0(t \mid x))) \text{ monotonically nondecreasing} \\ \psi_t \ge 0, \forall t \end{array} \right\}$$
$$\overline{\mathcal{S}}(w,\psi,\tilde{u}) = \cup_{\pi \in \Pi} \overline{\mathcal{S}}(w,\psi,\tilde{u};\pi)$$

$$\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}^{\Gamma,\Lambda}(\mathbb{P}_n))$$
$$= \max \left\{ \sum_{t=0}^{m-1} \frac{\sum_i r_i W_i \mathbb{I}[T_i = t]}{\sum_i W_i \mathbb{I}[T_i = t]} : w,\psi,\tilde{u} \in \overline{\mathcal{S}} \right\}$$

$$= \max\left\{\sum_{t=0}^{m-1}\sum_i r_i w_i \mathbb{I}[T_i=t] \colon (w,\psi,\tilde{u}) \in \overline{\mathcal{S}}(w,\psi,\tilde{u}), \sum_{i\in\mathcal{I}_t}\left|w_i - \psi_{T_i}\tilde{W}_i\right| \le \psi_{T_i}\Lambda_t, \forall t\right\}$$

$$= \min_{\eta_t\ge 0,\forall t}\max_{(w,\psi,\tilde{u})\in\overline{\mathcal{S}}(w,\psi,\tilde{u})}\left\{\sum_{t=0}^{m-1}\sum_i r_i w_i\mathbb{I}[T_i=t] + \sum_{t=0}^{m-1}\eta_t(\psi_t\Lambda_t - \sum_{i\in\mathcal{I}_t}\left|w_i-\psi_t\tilde{W}_i\right|)\right\}$$

$$= \max_{w,\psi,\tilde{u}\in\overline{\mathcal{S}}(w,\psi,\tilde{u})}\left\{\sum_{t=0}^{m-1}\sum_i r_i w_i\mathbb{I}[T_i=t] + \sum_{t=0}^{m-1}\eta_t^*(\mathbb{P}_n)(\psi_t\Lambda_t - \sum_{i\in\mathcal{I}_t}\left|w_i-\psi_t\tilde{W}_i\right|)\right\}$$

for optimal dual variables $\eta_t^*(\mathbb{P}_n)$, by strong LP duality (existence of the saddle point). Similarly, for the corresponding optimal dual variable $\eta_t^*(\mathbb{P})$ for the population budget-constrained uncertainty set,

$$\hat{\overline{R}}_{\pi_0}(\pi;\mathcal{W}^{\Gamma,\Lambda}(\mathbb{P}_n))$$
$$= \max_{w,\psi,\tilde{u}\in\overline{\mathcal{S}}(w,\psi,\tilde{u})}\left\{\sum_{t=0}^{m-1}\sum_i r_i w_i\mathbb{I}[T_i=t] + \sum_{t=0}^{m-1}\eta_t^*(\mathbb{P})(\psi_t\Lambda_t - \mathbb{E}[|w(T,X,Y)-\tilde{W}(T,X)\psi_t|\,|\,T=t]\le\Lambda_t\psi_t)\right\}$$

By Lemma EC.5, we combine objectives and obtain a lower bound since we optimize over the same feasible set:

$$\overline{R}_{\pi_0}(\pi;\mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P}_n)) - \overline{R}_{\pi_0}(\pi;\mathcal{W}_n^{\Gamma,\Lambda}(\mathbb{P}))$$
$$\le \max_{w,\psi\in\overline{\mathcal{S}}}\psi_t\sum_{t=0}^{m-1}\Lambda_t(\eta_t^*(\mathbb{P}_n)-\eta_t^*(\mathbb{P})) + \max(\eta_t^*(\mathbb{P}),\eta_t^*(\mathbb{P}_n))\sum_{t=0}^{m-1}(\sum_{i\in\mathcal{I}_t}\left|w_i-\psi_t\tilde{W}_i\right| - \mathbb{E}[|w(T,X,Y)-\tilde{W}(T,X)\psi_t|\,|\,T=t])$$
$$\le \max_{w\in\overline{\mathcal{S}}(\psi^*)}\psi_t^*\sum_{t=0}^{m-1}\Lambda_t(\eta_t^*(\mathbb{P}_n)-\eta_t^*(\mathbb{P})) + \max(\eta_t^*(\mathbb{P}),\eta_t^*(\mathbb{P}_n))\sum_{t=0}^{m-1}(\sum_{i\in\mathcal{I}_t}\left|w_i-\psi_t^*\tilde{W}_i\right| - \mathbb{E}[|w(T,X,Y)-\tilde{W}(T,X)\psi_t^*|\,|\,T=t])$$

Note that $\psi_t \in \frac{1}{n_t}[\nu,1]$ by definition. Next, we argue that the optimal dual variables are bounded by first noting that the optimal primal solution is finite and bounded on $[B,-B]$ by the self-normalized property of the estimator and Assumption 1. Moreover, the constraints on W, for a fixed Γ, imply bounds on how far *feasible* W can be from their nominal values. So, we have a bound which the optimal dual variables must satisfy. Let

$$\overline{\Lambda}_t = \max(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\max(\tilde{W}_i - a_i^\Gamma, b_i^\Gamma - \tilde{W}_i), \mathbb{E}[\max(\hat{W}-a^\Gamma, b^\Gamma-\hat{W})])$$

denote the maximal total deviation of weights, induced by the uncertainty set on W^Γ. Let

$$\underline{w}_i = \psi_{T_i}(b_i^\Gamma\mathbb{I}[r_i<0] + a_i^\Gamma\mathbb{I}[r_i>0]), \overline{w}_i = \psi_{T_i}(b_i^\Gamma\mathbb{I}[r_i>0] + a_i^\Gamma\mathbb{I}[r_i<0])$$

achieve the minimal and maximal feasible rescaled primal objectives, respectively. Now, we have the bounds that $\sum_{t=0}^{m-1}\eta_t^* + \sum_i r_i\underline{w}_i \ge -B$ and $\sum_{t=0}^{m-1}\eta_t^*\overline{\Lambda}_t + \sum_i r_i\overline{w}_i \le B$ which admits a naive componentwise bound that $\eta_t^* \ge -B - \sum_i r_i\underline{w}_i, \eta_t^* \le \frac{B-\sum_i r_i\overline{w}_i}{\min_t\overline{\Lambda}_t}, \forall t$.

Therefore, since $\eta_t^* \ge 0$, we obtain the following bound:

$$\eta_t \le \max(\left|\frac{B-\sum_i r_i\overline{w}_i}{\min_t\overline{\Lambda}_t}\right|, \left|-B-\sum_i r_i\underline{w}_i\right|) \le \frac{e2\nu^{-1}B\Gamma}{\min_t\Lambda_t\wedge 1}.$$

Applying this bound on η^*:

$$2B\Gamma\nu^{-1}(\max_t 1/p_t\frac{1}{n}\sum_{t=0}^{m-1}\Lambda_t + \underbrace{\max_{w\in\overline{S}(\psi^*)}\sum_{t=0}^{m-1}(\sum_{i\in\mathcal{I}_t}\left|w_i-\psi_t\tilde{W}_i\right| - \mathbb{E}[|w(T,X,Y)-\tilde{W}(T,X)\psi_t|\,|\,T=t])}_{①})$$

It remains to study uniform convergence of ① when we optimize over w in the set of restricted complexity (recall that monotonicity over u is equivalent to monotonicity over $\tilde{u}$; or we may equivalently reparametrize in W for the fixed scaling ψ). We do so by a Lipschitz contraction argument and applying our tail inequality from Lemma EC.4. Note that the absolute value function is globally 1-Lipschitz; and the envelope function on $W(u)$ is bounded by $(b-a)u \le \nu^{-1}(\Gamma - \Gamma^{-1})$. Now, by Lipschitz contraction (Theorem 5.7 of Pollard (1990)), applying Lemma EC.4, and taking a union bound over the number of treatments, we obtain the final bound that, with high probability $\ge 1-p$,

$$\begin{aligned}&\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}^{\Gamma,\Lambda}(\mathbb{P}_n)) - \hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}^{\Gamma,\Lambda}(\mathbb{P})) \\ &\le \frac{2\nu^{-1}B\Gamma}{\min_t \Lambda_t \wedge 1}\left(\frac{\max_t {}^1\!/_{p_t} \sum_{t=0}^{m-1} \Lambda_t}{n} + 18mK^{\Pi}\nu^{-1}(\Gamma-\Gamma^{-1})\sqrt{\frac{\log({}^{5m}\!/_{p})}{n}}\right)\end{aligned}$$

The result follows by applying this bound twice, at the sample-optimal and population-optimal policies, and taking a union bound over the event of this bound holding with high probability and the previous minimax regret bound of Theorem 2, and the triangle inequality. □

B.6. Proof of Proposition 6.

Proof of Proposition 6. In the following, we first consider the optimization problem within a single treatment partition, reindexing $i = 1, \dots, n$ to enumerate the elements of a generic treatment partition. The lemma follows by applying the same analysis to each treatment partition separately. We aim to bound the approximation error incurred by optimizing over an uncertainty set derived from *estimated* propensities, $\hat{e}_t(X)$ which may differ from the oracle values $\tilde{e}_t(X)$. Recall the weight bounds derived from the oracle nominal propensities, with $\tilde{W} = {}^1\!/_{\tilde{e}_t(x)}$, are $a = 1 + \frac{1}{\Gamma}(\tilde{W}-1), b = 1 + \Gamma(\tilde{W}-1)$; for this section, we define δ_i^a, δ_i^b as the *perturbations* of the sample weights from the oracle bounds a, b:

$$\hat{\delta}_i^{a,\Gamma} = 1 + \frac{1}{\Gamma}\left({}^1\!/_{\hat{e}_{T_i}(x)} - 1\right) - a_i, \hat{\delta}_i^{b,\Gamma} = 1 + \Gamma\left({}^1\!/_{\hat{e}_{T_i}(x)} - 1\right) - b_i$$

Observe that the dual of the primal program,

$$\sup_w \left\{ \sum_i w_i y_i : \psi \cdot (a_i + \delta_i^a) \le w_i \le \psi \cdot (b_i + \delta_i^b),\ \sum_i w_i = 1 \right\} \tag{EC.4}$$

for a fixed ψ scaling, and a generic multiplier r, is the program:

$$\inf_{\lambda, u \ge 0, v \ge 0} \left\{ \lambda + \psi \cdot \left(-\sum_i (a_i + \delta_i^a) u_i + \sum_i (b_i + \delta_i^b) v_i \right) : \lambda - u_i + v_i \ge_i, \forall i = 1 \dots n \right\} \tag{EC.5}$$

and since $u, v \ge 0$, we again observe (as in the proof of Theorem 3) that by complementary slackness, $v = (r_i - \lambda)_+, u = (\lambda - r_i)_+$. We make the corresponding substitution and proceed to define the partial Lagrangian relaxation. Denote $g_{\delta_a, \delta_b}(\psi; \lambda, u, v)$, as the *objective* function with given ψ, and δ^a, δ^b perturbations to the weights:

$$\inf_{\lambda, u \ge 0, v \ge 0} g_{\delta^a, \delta^b}(\psi; \lambda, u, v) = \inf_{\lambda, u \ge 0, v \ge 0} \left\{ \lambda + \psi \cdot \left(-\sum_i (a_i + \delta_i^a)(\lambda - r_i)_+ + \sum_i (b_i + \delta_i^b)(r_i - \lambda)_+ \right) \right\}$$

As a consequence of Lemma EC.5, $|\inf f - \inf g| \le \sup|f-g|$. Furthermore, we may optimize over the restrictions of the dual variables to compact sets: since λ is a quantile of the coefficients, $\lambda \in [\min_i r_i, \max_i r_i]$. We also have that $\psi \in [\frac{1}{n}, \frac{1}{\nu n}]$ under Assumption 2 (strong overlap), and the constraint that $\sum_i w_i = 1$, so that $\psi \cdot \sum_i (b_i + \delta_i^b) \le \sum_i w_i \le \psi \cdot \sum_i (a_i + \delta_i^a)$.

We invoke strong LP duality which holds with bounded optimal value (assuming bounded outcomes); strict feasibility and boundedness implies that the problem cannot be primal infeasible or primal unbounded. Let $\{\psi^*_{a,b}, (\lambda^*_{a,b}, u^*_{a,b}, v^*_{a,b})\} \in \arg\min g_{\delta^a,\delta^b}(\psi;\lambda,u,v)$. Therefore, compactness of the feasible region gives that the optimal primal and dual variables, $\{\psi^*_{a,b}, (\lambda^*_{a,b}, u^*_{a,b}, v^*_{a,b})\}$, $\{\psi^*_{0,0}, (\lambda^*_{0,0}, u^*_{0,0}, v^*_{0,0})\}$, are also pairs of optimal best responses for the min/max partial Lagrangian duals of the perturbed and nominal problem. In the following, let $S = \{(\lambda,u,v) : \lambda \in [-B,B], u \in [0,2B], v \in [0,2B]\}$ denote the compact restriction.

$$\begin{aligned}
&\left|\overline{R}_{\pi_0}(\pi,\tilde{\mathcal{W}}) - \overline{R}_{\pi_0}(\pi,\hat{\mathcal{W}})\right| = \left|\sup_{\psi>0}\left\{\inf_{\lambda,u\ge0,v\ge0} g_{\delta^a,\delta^b}(\psi;\lambda,u,v)\right\} - \sup_{\psi'>0}\left\{\inf_{\lambda,u\ge0,v\ge0} g_{00}(\psi';\lambda,u,v)\right\}\right| \\
&= \left|\inf_{(\lambda,u,v)\in S}\left\{\sup_{\psi\in[\frac{1}{n},\frac{1}{\nu n}]} g_{\delta^a,\delta^b}(\psi;\lambda,u,v)\right\} - \inf_{(\lambda',u',v')\in S}\left\{\sup_{\psi'\in[\frac{1}{n},\frac{1}{\nu n}]} g_{00}(\psi';\lambda',u',v')\right\}\right| && \text{(EC.6)} \\
&= \left|\sup_{\psi\in[\frac{1}{n},\frac{1}{\nu n}]}\left\{\inf_{(\lambda,u,v)\in S} g_{\delta^a,\delta^b}(\psi;\lambda,u,v)\right\} - \sup_{\psi'\in[\frac{1}{n},\frac{1}{\nu n}]}\left\{\inf_{(\lambda',u',v')\in S} g_{00}(\psi';\lambda',u',v')\right\}\right| && \text{(EC.7)} \\
&\le \sup_{\psi\in\{\psi^*_{00},\psi^*_{ab}\}}\left|\inf_{(\lambda,u,v)\in S} g_{\delta^a,\delta^b}(\psi;\lambda,u,v) - \inf_{(\lambda',u',v')\in S} g_{00}(\psi;\lambda',u',v')\right| && \text{(EC.8)} \\
&\le \max_{j,k\in\{00,ab\}}\left|g_{\delta^a,\delta^b}(\psi^*_j;\lambda^*_k,u^*_k,v^*_k) - g_{00}(\psi^*_j;\lambda^*_k,u^*_k,v^*_k)\right| && \text{(EC.9)}
\end{aligned}$$

In the above, the equality of Equation (EC.6) follows since without loss of generality, we can restrict attention to bounded feasible regions for the variables. In Equation (EC.7), we swap the order of the sup and inf since strong duality holds with equality. In Equation (EC.8), restricting the supremum over ψ to the best responses ψ^*_{00}, ψ^*_{ab} doesn't change the optimal value; that λ^*, u^*, v^* and ψ^* are best responses is a consequence of von Neumann's minimax theorem, since g is bilinear in its arguments ψ and λ, u, v. Equation (EC.9) holds since $\lambda^*_{0,0}, u^*_{0,0}, v^*_{0,0}$ were optimal for $g_{0,0}$ (resp., for g_{δ^a,δ^b}) and we expand the feasible set.

Combining g_{δ^a,δ^b} and $g_{0,0}$, we can now bound the perturbation incurred based on possible values of ψ^*, λ^*:

$$\begin{aligned}
&= \max_{j,k\in\{00,ab\}}\left|\psi^*_j \cdot \left(-\sum_i \delta^a_i(\lambda^*_k - r_i)_+ + \sum_i \delta^b_i(r_i - \lambda^*_k)_+\right)\right| \\
&\le \max_{\psi\in\{\psi^*_{a,b},\psi^*_{0,0}\}} \psi\cdot(\|\delta^a\|_1 + \|\delta^b\|_1)(2\max_i r_i) \text{ since the optimal } \lambda^* \text{ is bounded} \\
&\le \frac{2\max r_i(\|\delta^a\|_1 + \|\delta^b\|_1)}{n} \\
&= (\max_i Y_i)(\Gamma + {}^1\!/\!_\Gamma)\frac{1}{n}\sum_i\left|\frac{1}{\hat{e}_t(X_i)} - \frac{1}{\tilde{e}_t(X_i)}\right|
\end{aligned}$$

The bound on the range for ψ follows since for $\psi \in \{\psi^*_{a,b}, \psi^*_{0,0}\}$, we have that $\psi \le \max\{\frac{1}{\sum_i(a_i+\delta^a_i)}, \frac{1}{\sum_i a_i}\} \le \frac{1}{n}$ since the bounds $\alpha_i + a_i$ and α_i are inverse probabilities. We simply apply the above argument for each group, under the product uncertainty set assumption. Define the treatment-conditional partial dual objective, computed for data from treatment partition $T = t$, as $g_{\delta^a,\delta^b}(\psi;\lambda,u,v;t)$. We apply the above bound for every treatment partition $T = t$, which holds *deterministically* for all π, with the multiplier $r = (\pi - \pi_0)Y$.

$\hat{\overline{R}}_{\pi_0}(\pi,\tilde{\mathcal{W}}^\Gamma_n) - \hat{\overline{R}}_{\pi_0}(\pi,\hat{\mathcal{W}}^\Gamma_n)$

$$
\begin{aligned}
&= \left| \sup_{W \in \mathcal{W}^\Gamma(\tilde{e}_T)} \sum_{t=0}^{m-1} \frac{\mathbb{E}_n[(\pi(t \mid X) - \pi_0(t \mid X))YW\mathbb{I}[T=t]]}{\mathbb{E}_n[W\mathbb{I}[T=t]]} - \sup_{W \in \mathcal{W}^\Gamma(\hat{e})} \sum_{t=0}^{m-1} \frac{\mathbb{E}_n[(\pi(t \mid X) - \pi_0(t \mid X))YW\mathbb{I}[T=t]]}{\mathbb{E}_n[W\mathbb{I}[T=t]]} \right| \\
&\leq \sum_{t=0}^{m-1} \left| \sup_{W(\cdot,\cdot;t) \in \mathcal{W}_t^\Gamma(\tilde{e}_T)} \frac{\mathbb{E}_n[(\pi(t \mid X) - \pi_0(t \mid X))YW\mathbb{I}[T=t]]}{\mathbb{E}_n[W\mathbb{I}[T=t]]} - \sup_{W(\cdot,\cdot;t) \in \mathcal{W}_t^\Gamma(\hat{e})} \frac{\mathbb{E}_n[(\pi(t \mid X) - \pi_0(t \mid X))YW\mathbb{I}[T=t]]}{\mathbb{E}_n[W\mathbb{I}[T=t]]} \right| \qquad \text{(EC.10)} \\
&= \sum_{t=0}^{m-1} \left| \sup_{\psi>0} \left\{ \inf_{\lambda,u \geq 0, v \geq 0} g_{\delta^a,\delta^b}(\psi;\lambda,u,v;t) \right\} - \sup_{\psi'>0} \left\{ \inf_{\lambda,u \geq 0, v \geq 0} g_{00}(\psi';\lambda,u,v;t) \right\} \right| \\
&= 2B(\Gamma + {}^1\!/\!_\Gamma) \frac{1}{n} \sum_{i=1}^n \left| \frac{1}{\hat{e}_{T_i}(X_i)} - \frac{1}{\tilde{e}^*_{T_i}(X_i)} \right|
\end{aligned}
$$

Here, EC.10 follows by the product set structure of the uncertainty set and application of the triangle inequality.

B.7. Proof of Proposition 7

Proof of Proposition 7. Given convex $\mathcal{S} \subseteq \mathbb{R}^n$, notice that its conic hull is $K = \{\sum_{i=1}^k \alpha_i u_i : k \in \mathbb{N}, \alpha_i \geq 0, u_i \in \mathcal{S}\} = \bigcup_{\psi \geq 0}(\psi\mathcal{S})$. Let $r \in \mathbb{R}^n$. Given that $\mathcal{S}$ has a non-empty interior, a Charnes-Cooper transformation followed by strong duality yields

$$
\sup_{u \in \mathcal{S}} \frac{\sum_{i=1}^n r_i u_i}{\sum_{i=1}^n u_i} = \sup_{u/\psi \in \mathcal{S},\, \psi \geq 0,\, \sum_{i=1}^n u_i = 1} \sum_{i=1}^n r_i u_i = \sup_{u \succeq_K 0,\, \sum_{i=1}^n u_i = 1} \sum_{i=1}^n r_i u_i = \inf_{\lambda \succeq_{K^*} r} \lambda.
$$

The statement of the proposition proceeds by applying this for each treatment level t. □

Appendix C: Optimization Algorithm Details

C.1. Subgradient Approach Refinements

We describe some additional changes to the subgradient method optimization procedure of 6.2 which improve the optimization by specializing to the unique case of our problem. Further refinements are possible with e.g. homotopy methods for LPs; we leave this to future work. In the case that we are optimizing over a series of increasing Γ parameters, $1 = \Gamma_0 < \Gamma_1 < ... < \Gamma_m$, we can use the nested property of the corresponding uncertainty sets to provide additional checks on the optimization.

1. We include a warm start for optimization for Γ_{k+1} with Γ_k as one of the random initializations: therefore we are guaranteed an initialization that does well for similar Γ.

2. For each proposed optimal policy returned by the optimization, which we denote as $\bar{\pi}(\Gamma_k)$ for a policy optimized over $W_n^{\Gamma_k}$, we check the achieved objective value of previous policies, $\bar{R}(\bar{\pi}(\Gamma_k), \Gamma_i), i < k$. If for some i, $\bar{R}(\bar{\pi}(\Gamma_k), \Gamma_i) < \bar{R}(\bar{\pi}(\Gamma_k), \Gamma_k)$, we set the policy to the previous policy, $\pi_k^* = \pi_i^*$.

We find empirically that including these refinements stabilizes the optimization when optimizing over a nested series of Γ parameters, as we anticipate a decision-maker would do in practice, given a feasible range of plausible Γ values.

C.2. Optimal Confounding-Robust Trees

We next consider the function class consisting of axis-aligned decision tree policies where each leaf is assigned a constant probability of treatment. Decision tree policies are advantageous due to their simplicity and interpretability. Our optimal confounding-robust tree (OCRT) presented below determines the best confounding-robust decision tree via global optimization using mixed-integer optimization. Our approach is to combine the dual linear program formulation of $\hat{\overline{R}}_{\pi_0}(\pi; \mathcal{W}_n^\Gamma)$ in eq. (12) with a mixed-integer formulation of this class of decision trees, following the formulation of Bertsimas and Dunn (2017), along with a special heuristic to find a good warm start.

A decision tree (with maximal splits) of a fixed depth D can be represented by an array labeled by a set of nodes, split into a set of branching nodes $\mathcal{K}_B$ and leaf nodes $\mathcal{K}_L$. The space of decision tree policies is parametrized by $\Theta = \{\{\alpha_{k_b}, \beta_{k_b}\}_{k_b \in \mathcal{K}_B}, \{c_k\}_{k \in \mathcal{K}_L}\}$, where $\alpha_{k_b}, \beta_{k_b} \in \mathbb{R}^p$ parametrize the split at branching node k_b, which directs units to the left branch if $\alpha^\intercal x < \beta$, and to the right branch otherwise. The policy assignment probability is parametrized by $c_{k_b} \in [0,1]$ for $k_b \in \mathcal{K}_L$. We consider axis-aligned splits such that α_{k_b} is a unit vector.

We let the binary assignment variables z_{ik} track assignment of data points i to leaves $k \in \mathcal{K}_L$ subject to the requirement that every instance is assigned to a leaf node according to the results of axis-aligned splits $\alpha_{k_b}^\intercal x < \beta_{k_b}$, for splits occurring at $k_b \in \mathcal{K}_B$ branch nodes. The binary variables d_k track whether a split occurs at node $k_b \in \mathcal{K}_B$. The binary variable l_k tracks whether a leaf is empty or not. The policy optimization determines both the partitions of the covariates governing assignment to terminal leaf nodes and the variables c_k for $k \in \mathcal{K}_L$ governing probability of treatment assignment in the leaf nodes. We denote $\text{par}(k)$ as the parent of node k, $A(k)$ as the set of all ancestors of node k, and the subsets $A_L(k) \cup A_R(k) = A(k)$ denote the sets of ancestor nodes where the instance was split to the left or right, respectively. In this section, we assume that the covariates are rescaled such that each covariate lies in $[0,1]$.

We introduce additional constraints to encode our dual objective in the optimal classification tree framework. We define the policy assignment probability for treatment $T = t$, $P_i^t = \sum_{k \in \mathcal{K}_L} z_{ik} c_k^t$ where c_k^t is the policy assignment probability of leaf node $k \in \mathcal{K}_L$ of assigning treatment t, and z_{ik} describes whether or not instance i is assigned to leaf node k, enforced with the additional set of auxiliary big-M constraints for the product of a binary variable and continuous variable; for each set of such product variables P_i^t.

$$
\begin{aligned}
& p_{i,k}^t \le z_{ik}; \quad p_{i,k}^t \le c_k^t; \quad p_{i,k}^t \ge c_k^t + z_{ik} - 1 && \forall i = 1, \dots, n; \forall t \in \mathcal{T}, k \in \mathcal{K}_L \\
& P_i^t = \sum_{k \in \mathcal{K}_L} p_{i,k}^t && \forall a \in \mathcal{A}, \forall i = 1, \dots, n \\
& \sum_{t=0}^{m-1} c_k^t = 1 && k \in \mathcal{K}_L \\
& p_{i,k}^t \in [0,1] && \forall t \in \mathcal{T}, \forall i = 1, \dots, n, k \in \mathcal{K}_L \\
& c_k^t \in [0,1] && \forall t \in \mathcal{T}, \forall k \in \mathcal{K}_L \\
& P_i^t \in [0,1] && \forall t \in \mathcal{T}, \forall i = 1, \dots, n
\end{aligned}
$$

The combined formulation for policy optimization with confounding-robust optimal trees is as follows:

$$\min \sum_{t=0}^{m-1} \lambda_t \tag{EC.11a}$$

$$\text{s.t. } v_i - u_i + \lambda_{T_i} \geq Y_i(P_i^{T_i} - \pi_0^{T_i}) \qquad ,\forall i \in \mathcal{I}_t \tag{EC.11b}$$

$$\sum_{i \in \mathcal{I}_t} -b_i^\Gamma v_i + a_i^\Gamma u_i \geq 0, \qquad \forall t \in \mathcal{T} \tag{EC.11c}$$

$$p_{i,k}^t \leq z_{it}; \quad p_{i,k}^t \leq c_k^t; \quad p_{i,k}^t \geq c_k^t + z_{ik} - 1 \qquad \forall t \in \mathcal{T},\, \forall i = 1, \ldots, n, k \in \mathcal{K}_L, \tag{EC.11d}$$

$$P_i^t = \sum_{k \in \mathcal{K}_L} p_{i,k}^t \qquad \forall t \in \mathcal{T},\, \forall i = 1, \ldots, n \tag{EC.11e}$$

$$\sum_{t=0}^{m-1} c_k^t = 1 \qquad k \in \mathcal{K}_L \tag{EC.11f}$$

$$c_k^t \in [0,1] \qquad \forall t \in \mathcal{T},\, \forall k \in \mathcal{K}_L \tag{EC.11g}$$

$$a_m^\intercal (x_i + \epsilon) \leq b_m + (1 - z_{ik}) \qquad \forall i = 1, \ldots, n, \forall k \in \mathcal{K}_B, \forall m \in A_L(k) \tag{EC.11h}$$

$$a_m^\intercal (x_i + \epsilon) \leq b_m - (1 + \epsilon_{max})(1 - z_{ik}) \qquad \forall i = 1, \ldots, n, \forall k \in \mathcal{K}_B, \forall m \in A_R(k) \tag{EC.11i}$$

$$\textstyle\sum_{k \in \mathcal{K}_L} z_{ik} = 1 \qquad \forall k \in \mathcal{K}_B \tag{EC.11j}$$

$$\textstyle\sum_{i=1}^n z_{ik} \geq N_{min} l_t \qquad \forall i = 1, \ldots, n \tag{EC.11k}$$

$$\textstyle\sum_{j=1}^p a_{jt} = d_t \tag{EC.11l}$$

$$0 \leq b_k \leq d_k \qquad \forall k \in \mathcal{K}_B \tag{EC.11m}$$

$$d_t \leq d_{\text{par}(k)} \qquad \forall k \in \mathcal{K}_B \setminus \{1\} \tag{EC.11n}$$

$$l_{U(k)} \geq d_{(\text{par}(k))} \qquad k \in \mathcal{K}_B \setminus 1 \tag{EC.11o}$$

$$l_k \leq d_{\text{par}(m)} \qquad \forall m \in \mathcal{T}_B, t \in [D(k_b), U(k_b)] \tag{EC.11p}$$

$$l_k \geq d_{\text{par}(t)} \qquad \forall k \in \mathcal{K}_L \tag{EC.11q}$$

$$z_{ik}, l_k \in \{0,1\} \qquad i = 1, \ldots, n, \forall k \in \mathcal{K}_L \tag{EC.11r}$$

$$a_{jk}, d_k \in \{0,1\} \qquad j = 1, \ldots, p, \forall k \in \mathcal{K}_B \tag{EC.11s}$$

$$p_{i,k}^t \in [0,1] \qquad \forall t \in \mathcal{T},\, \forall i = 1, \ldots, n, k \in \mathcal{K}_L \tag{EC.11t}$$

$$c_k^t \in [0,1] \qquad \forall t \in \mathcal{T},\, \forall k \in \mathcal{K}_L \tag{EC.11u}$$

$$P_i^t \in [0,1] \qquad \forall t \in \mathcal{T},\, \forall i = 1, \ldots, n \tag{EC.11v}$$

$$u, v \geq 0 \tag{EC.11w}$$

Constraints (EC.11e, EC.11d) set the policy assignment variable $P_i^t \in [0,1]$, which is the sum of products $p_{i,k}^t = z_{ik} c_k$ over leaf nodes. Our objective is specified via the dual formulation, and constraints (EC.11b, EC.11c) encode the constraints from the dual of the inner maximization subproblem. Constraints (EC.11h, EC.11i) enforce that if a node is in a leaf (as indicated by z_{ik}), it satisfies the splits at ancestor nodes. Constraint (EC.11j) enforces that each instance is in a leaf node, while constraint EC.11k enforces a size constraint on leaf membership. Constraints (EC.11l, EC.11n, EC.11m) enforce consistency constraints between d, indicating whether a split occurs at leaf node k, and split variables a_{jk}, b_k. $\{D(k)\}_{k \in \mathcal{K}_B}$ denotes the set of

Algorithm 2 Greedy Recursive Partitioning (**Partition**)

Input: partition $S_{L,l} = \{(X_{i_1}, T_{i_1}, Y_{i_1})\}$, depth Δ, preliminary assignment $\tau^{\Delta-1} \in [m]^n$
for $d \in [p]$ **do** (find best partition index):
    $[i] \leftarrow$ Get the sorted indices of $(\{X_{i,d}\})$
    $i_j^*, v_j^* \leftarrow$ Find the best dimension and threshold to split $x_{j^*} < x_{i_j^*, j^*}$
    $i_{j,rev}^*, v_{j,rev}^* \leftarrow$ Find the best dimension and threshold to split $x_{j^*} > x_{i_j^*, j^*}$
$j^* \leftarrow \arg\min_j i_j^*, \quad i^* \leftarrow i_{j^*}^*, \quad \theta^* \leftarrow \frac{X_{(i^*),j^*} + X_{(i^*+1),j^*}}{2}$
$\pi(X) \leftarrow x_{j^*} \le \theta^*$ if $v_j^* < v_{j,rev}^*$ else $x_{j^*} \ge \theta^*$
if (continue recursing) **then**:
    $S_L \leftarrow X_{[0:i^*]}, S_R \leftarrow X_{[i^*:|S|]}$
    update τ_0, the candidate treatment assignment
    $\hat{\Pi}_L \leftarrow$ **Partition**$(S_L, \tau', \Delta+1)$, $\hat{\Pi}_R \leftarrow$ **Partition**$(S_R, \tau_0, \Delta+1)$
return $(\pi(X)$, $\hat{\Pi}_L$, $\hat{\Pi}_R$ $)$

leaf nodes of smallest index which can be reached from splits at k, and similarly $\{U(k)\}_{k \in \mathcal{K}_B}$ denotes the set of reachable leaf nodes of largest index. Constraints (EC.11o, EC.11p, EC.11q) enforce that leaves are non-empty only if splits do occur in the relevant ancestor nodes.

For the mixed-integer linear program, we provide a warm start for the optimization via a recursive partitioning-based approach which incrementally optimizes directly the robust risk, over iterative refinements of either the constant all-treat or all-control policy, described in Sec. C.3 of the EC.

C.3. Recursive Partitioning: MIP Warm Start

We provide a heuristic recursive-partitioning based scheme for optimizing policy risk over the space of limited-depth decision trees recursively, analogous to CART's recursive partitioning approach (Breiman et al. 1984). Such an approach is used to obtain a warm start for the MIP of the optimal confounding-robust tree. The algorithm initializes by assigning the same treatment τ_0 to all, and iteratively refines the treatment assignment by recursive partitioning, seeking univariate splits which minimize the minimax risk. The candidate split threshold for each covariate is determined by iteratively re-evaluating the minimax risk for incremental changes to the policy, maintaining the invariant that the base policy is set by the leaves above a node in the tree. Using specialized data structures such as B-trees allows for $O(log(N))$ efficient updates for maintaining and updating the sorted list of multipliers $Y_i T_i (\pi_i - \pi_0)$, and manipulating pre-computed cumulative sums of the initial sorted order allows for efficient re-computation of the optimization solution. We note that such an approach is possible only for the unbudgeted uncertainty set $\mathcal{U}_n^\Gamma$, since incorporating the uncertainty budget would couple the risk across tree levels. In comparison to other approaches using tree-based approaches for estimating causal effects (Wager and Athey 2017b) or for personalization (Kallus 2017a), which consider splits based on impurities related to the expected mean squared error of causal effects on a separate sample of data from that used to estimate the causal effects within leaves, or determine the optimal treatment, our

recursive partitioning heuristic simultaneously determines the partition and the policy treatment assignment within the partition. In making greedy splits, changes in the objective function are assessed as a result of changing the policy assignment within $S_{L,l}$, and the optimal split location and sense ($\mathbb{I}[a^\intercal X < b]$ or $\mathbb{I}[a^\intercal X > b]$) are determined by changes in the policy assignment within $S_{L,l}$. However, in general, the optimal such policy assignment, determined incrementally from the assignments $\{S_{L-1,l}\}_{l\in\mathcal{T}_L}$, depends additionally on the assigned policy for other nodes at the same level as well.

Appendix D: WHI Case Study details

The selected list of covariates for personalization is as follows (using the name/description from the WHI data dictionary): Time since menopause, Systolic blood pressure (baseline), Age, BMI, BMI baseline, Diastolic blood pressure baseline, Number of falls , Number of pregnancies, Angina ever (yes), Aortic aneurysm ever (Y), Atrial fibrillation ever (Y), Bilateral oophorectomy (Y), Coronary bypass surgery ever (Y), Cigarettes/day (1-4), Cigarettes/day (15-24), Cigarettes/day (25-34), Cigarettes/day (35-44), Cigarettes/day (45+), Cigarettes/day (5-14), Cigarettes/day (<1), Diabetes treated (Y), DVT ever (Y), Ethnic Asian/Pacific Islander, Ethnic Black, Ethnic Hispanic, Ethnic Unknown, Ethnic White,High cholesterol (Y), Myocardial infarction ever (Y), Stroke ever (Y).

Table EC.1 Policy regret for WHI, under different λ scalarizations

λ	$\log(\Gamma) = 0.025$	0.05	0.075	0.1	0.12	0.14	0.16	0.18	0.2	0.225	0.25	0.3	0.35	0.4	0.5	0.75	1.0	1.5	2.0
-0.00	0.36	0.29	0.31	0.10	0.11	0.14	0.16	0.12	0.11	0.04	-0.01	-0.06	-0.07	-0.10	0.00	0.00	-0.01	-0.00	0.00
-0.11	0.41	0.32	0.28	-0.09	-0.07	-0.00	0.01	0.02	0.01	-0.02	-0.12	-0.15	-0.17	-0.19	-0.03	-0.03	-0.04	-0.05	-0.06
-0.21	0.41	0.36	0.37	0.05	0.06	-0.01	-0.03	-0.01	0.01	0.02	-0.15	-0.15	-0.18	-0.20	-0.02	-0.03	-0.03	-0.03	-0.06
-0.32	0.39	0.32	0.32	0.36	0.28	0.20	0.22	0.23	0.24	0.19	0.03	0.01	-0.02	-0.02	-0.01	-0.00	-0.00	-0.01	-0.05
-0.43	0.38	0.32	0.32	0.23	-0.03	-0.07	-0.05	-0.05	-0.03	-0.07	-0.18	-0.19	-0.21	-0.23	-0.01	0.00	0.00	-0.02	-0.05
-0.54	0.36	0.24	0.24	0.25	0.21	0.25	0.24	0.23	0.24	0.19	0.08	0.07	-0.13	-0.14	-0.02	-0.01	-0.01	-0.03	-0.03
-0.64	0.46	0.48	0.50	0.33	0.14	0.15	-0.11	-0.11	-0.10	-0.13	-0.20	-0.22	-0.25	-0.23	-0.03	-0.04	-0.04	-0.06	-0.08
-0.75	0.74	0.78	0.81	0.71	0.61	0.63	0.40	0.41	0.44	-0.08	-0.16	-0.18	-0.20	-0.19	-0.02	-0.02	-0.02	-0.03	-0.06
-0.86	0.68	0.63	0.62	0.56	0.47	0.52	0.31	0.31	0.34	-0.10	-0.11	-0.14	-0.17	-0.14	-0.01	-0.02	-0.02	-0.02	-0.05
-0.96	0.92	0.97	1.00	1.02	0.92	0.92	0.69	0.70	0.73	-0.11	-0.13	-0.16	-0.17	-0.14	-0.01	-0.02	-0.03	-0.04	-0.05
-1.07	0.83	0.85	0.96	1.01	0.96	0.99	0.76	0.79	0.60	-0.10	-0.12	-0.12	-0.13	-0.12	-0.04	-0.03	-0.04	-0.05	-0.05
-1.18	0.77	0.84	0.96	1.03	0.96	0.97	0.73	0.74	0.78	0.10	0.06	0.05	-0.12	-0.10	-0.01	-0.01	-0.00	-0.02	-0.03
-1.29	0.83	0.94	1.05	1.12	1.14	1.17	1.20	1.15	1.09	0.58	0.58	0.60	0.32	0.05	0.06	0.08	0.08	0.04	0.00
-1.39	0.88	1.04	1.17	1.28	1.33	1.38	1.16	1.17	1.19	0.24	0.23	0.23	0.24	-0.02	0.05	0.06	0.06	0.01	-0.01
-1.50	0.82	0.96	1.07	1.18	1.23	1.28	1.23	1.24	1.17	0.65	0.67	0.27	0.16	-0.05	-0.00	0.01	0.01	-0.02	-0.01

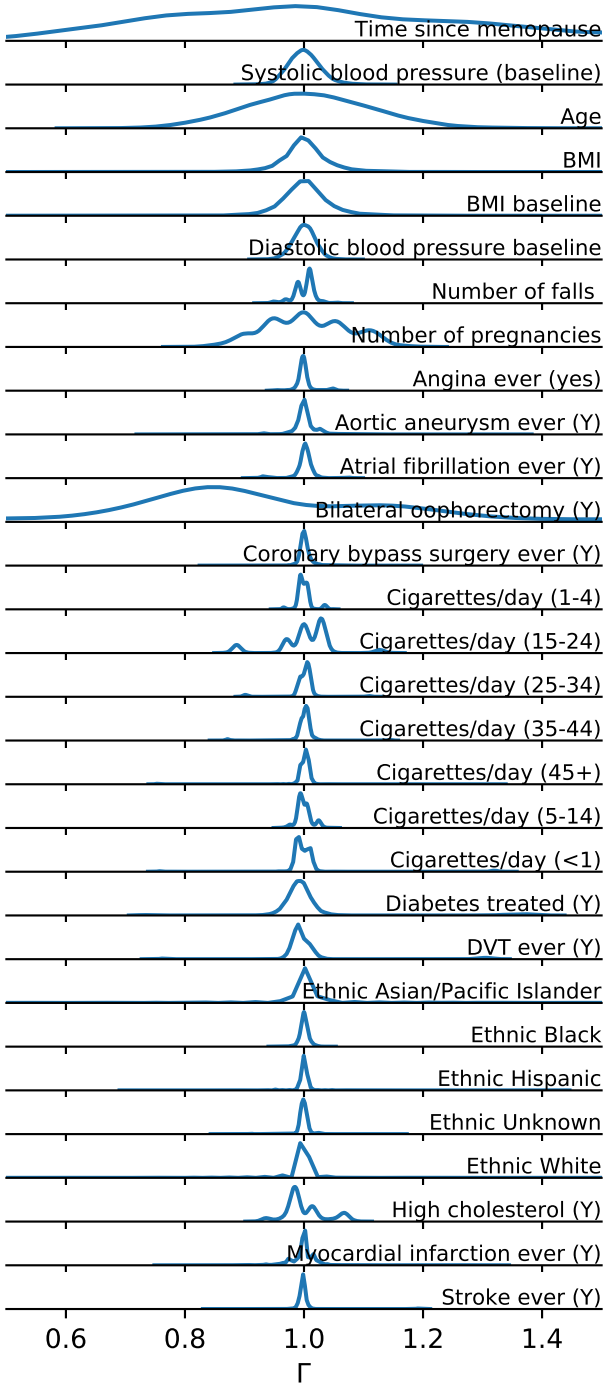


Figure EC.1 **Density comparison of odds ratios induced by training propensities with dropped covariates (one per line, in order). x-axis is the odds ratio, while y-axis (for each subplot) is a density plot; fixed y-scale $y \in [0, 10]$ for all subplots. Note that most of the probability mass is within $\Gamma \in [0.8, 1.2]$, with the exception of a few covariates with wider distributions of informativity.**