Optimal and Fair Encouragement Policy Evaluation and Learning

Angela Zhou
Department of Data Sciences and Operations
University of Southern California
zhoua@usc.edu

September 12, 2023

Abstract

In consequential domains, it is often impossible to compel individuals to take treatment, so that optimal policy rules are merely suggestions in the presence of human non-adherence to treatment recommendations. In these same domains, there may be heterogeneity both in who responds in taking-up treatment, and heterogeneity in treatment efficacy. For example, in social services, a persistent puzzle is the gap in take-up of beneficial services among those who may benefit from them the most. When in addition the decision-maker has distributional preferences over both access and average outcomes, the optimal decision rule changes. We study identification, doubly-robust estimation, and robust estimation under potential violations of positivity. We consider fairness constraints such as demographic parity in treatment take-up, and other constraints, via constrained optimization. Our framework can be extended to handle algorithmic recommendations under an often-reasonable covariate-conditional exclusion restriction, using our robustness checks for lack of positivity in the recommendation. We develop a two-stage, online learning-based algorithm for solving over parametrized policy classes under general constraints to obtain variance-sensitive regret bounds. We consider two case studies based on data from randomized encouragement to enroll in insurance and from pretrial supervised release with electronic monitoring.

1 Introduction

The intersection of causal inference and machine learning for heterogeneous treatment effect estimation can improve public health, increase revenue, and improve outcomes by personalizing treatment decisions, such as medications, e-commerce platform interactions, and social interventions, to those who benefit from it the most [Athey, 2017, Kitagawa and Tetenov, 2015, Manski, 2005, Zhao et al., 2012]. But, in many important settings, we do not have direct control over treatment, and can only optimize over *encouragements*, or *recommendations* for treatment. For example, in e-commerce, companies can rarely *compel* users to sign up for certain services, rather *nudge* or *encourage* users to sign up via promotions and offers. When we are interested in optimizing the effects of signing up – or other voluntary actions beyond a platform's control – on important final outcomes such as revenue, we therefore need to consider *fairness-constrained optimal encouragement designs*. By fairness, we refer specifically to *statistical parity* constraints which to enforce parity in performance measures, and we use the term fairness as it is often used in the algorithmic fairness literature. Often human expert oversight is required in the loop in important settings where ensuring *fairness in machine*

learning [Barocas et al., 2018] is also of interest: doctors prescribe treatment from recommendations [Lin et al., 2021], managers and workers combine their expertise to act based on decision support [Bastani et al., 2021], and in the social sector, caseworkers assign to beneficial programs based on recommendations from risk scores that support triage [De-Arteaga et al., 2020, Green and Chen, 2019, Yacoby et al., 2022].

The human in the loop requires new methodology for optimal encouragement designs because:

> *when the human in the loop makes the final prescription, algorithmic recommendations do not have direct causal effects on outcomes; they can only change the probability of treatment assignment which does have direct causal effects.*

On the other hand, this is analogous to the well-understood notion of *non-compliance/non-adherence* in randomized controlled trials in the real world [Hernán and Robins, Heard et al., 2017]. For example, patients who are prescribed treatment may not actually take medication. A common strategy is to conduct an *intention-to-treat* analysis: under assumptions of no unobserved confounders affecting treatment take-up and outcome, we may simply view encouragement as treatment.

But, in the case of prediction-informed decisions in social settings, if we are concerned about *access to the intervention* in addition to *utility of the policy over the population*, finer-grained analysis is warranted. If an outcome-optimal policy results in wide disparities in access, for example in marginalized populations not taking up incentives for healthy food due to lack of access in food deserts, or administrative burden that screens out individuals applying for social services that could benefit the most [Herd and Moynihan, 2019], this could be a serious concern for decision-makers. Indeed, a long history of welfare rights advocacy and civil rights oversight [U.S. Commission on Civil Rights, 2002] 1) is concerned about disparities in provision of resources and services in social benefits, 2) recognizes that discretionary decisions of "street-level bureaucrats" [Lipsky, 2010] may lead to disparities in access.

Even without external decision-makers screening individuals in and out, differential take-up by individuals can be a large driver of realized inequities in service delivery. The literature on administrative burden recognizes that hassles associated with social delivery of services, including those related to the social safety net, can have detrimental effects, especially on more marginalized and vulnerable populations. [Christensen et al., 2020] posits a "human capital catch-22" that certain axes of precarity such as scarcity and health both increase likelihood of requiring access to state assistance while reducing the cognitive resources required to navigate administrative burdens in service delivery. Some strategies noted for reducing administrative burden in public benefit and service programs OMB [2022] include reducing information/learning costs, which can be modeled with encouragement designs and targeted recommendations; or reducing redemption costs, which can be modeled in a setting of scarce resources with constraints and potentially personalized using the methods we develop here.

In this work, we seek optimal decision rules that optimize population-outcomes while satisfying fairness constraints, for example parity in treatment access. In contrast, previous work in algorithmic accountability primarily focuses on auditing *recommendations*, but not both the access and efficacy achieved under the final decision rule. Therefore, previous methods can fall short in mitigating potential disparities.

Our contributions are as follows: we characterize optimal and resource fairness-constrained optimal decision rules, develop statistically improved estimators and robustness checks for the setting of algorithmic recommendations with sufficiently randomized decisions. We consider two

settings: one related to encouragement designs with random allocation of encouragement, another related to algorithmic recommendations (which requires either parametric or robust extrapolation). We also develop methodology for optimizing over a constrained policy class with less conservative out-of-sample fairness constraint satisfaction by a two-stage procedure, and we provide sample complexity bounds. We assess improved recommendation rules in a stylized case study of optimizing health insurance expansion using data from the Oregon Insurance study, and another stylized case study of optimizing recommendation of supervised release based on a pretrial risk-assessment tool while reducing surveillance disparities.

2 Related Work

In the main text, we briefly highlight the most relevant methodological and substantive work and defer additional discussion to the appendix.

Optimal encouragement designs/policy learning with constraints. There is extensive literature on off-policy evaluation and learning, empirical welfare maximization, and optimal treatment regimes [Athey and Wager, 2021, Zhao et al., 2012, Manski, 2005, Kitagawa and Tetenov, 2015]. [Qiu et al., 2021] studies an optimal individualized encouragement design, though their focus is on optimal individualized treatment regimes with instrumental variables. [Kallus and Zhou, 2021a] study fairness in pricing, and some of the desiderata in that setting on revenue (here, marginal welfare) and demand (take-up) are again relevant here, but in a more general setting beyond pricing. The most closely related work in terms of problem setup is the formulation of "optimal encouragement designs" in [Qiu et al., 2021]. However, they focus on knapsack resource constraints, which have a different solution structure than fairness constraints. Their outcome models in regression adjustment are conditional on the recommended/not recommended partitions which would not allow our fairness constraints that introduce treatment- and group-dependent costs. [Sun et al., 2021] has studied uniform feasibility in constrained resource allocation, but without encouragement or fairness. [Ben-Michael et al., 2021] studies robust extrapolation in policy learning from algorithmic recommendation, but not fairness. Our later case study on supervised release appeals to a great deal of randomness in final treatment decisions for supervised release (arguably less consequential than pretrial detention, hence subject to wide discretion): our two-stage approach allows us to be less conservative in requiring robust extrapolation over only one stage.

Other causal methodology for intention-to-treat. We focus on deriving estimators for intention-to-treat analyses in view of relevant fairness constraints. Our interest is in imposing separate desiderata on treatment realizations under non-compliance; but we don't conduct instrumental variable inference: we assume unconfoundedness holds. Our analysis essentially considers simultaneously two perspectives in the constrained optimization: 1) viewing treatment as a potential outcome of a recommendation treatment, i.e. $T(R)$, and 2) an intention-to-treat stance in the causal effects of treatment on outcomes, i.e. $Y(T)$, even though treatment is not controllable. Marginally, the first perspective estimates disparities and is relevant for estimating fairness constraints, while the second is relevant for the utility objective. We also make key structural assumptions that the recommendation doesn't change the treatment probability function as a function of covariates and treatment; and that it doesn't have a direct effect on actual outcomes, only treatment does. These two perspectives are simultaneously possible when there is exogenous randomness in final treatment decisions, as in our settings considered here. Importantly, the quantities we estimate are not on joint events of take-up and final outcome utility (unlike principal stratification). Rather, we assess

personalized policies by their population-averaged utility and fairness measures.

Intention-to-treat analysis. We appeal to intention-to-treat analysis with randomness that either arises from human decision-makers or individual non-adherence/non-compliance, but we generally assume the data does not include information about the *identity* of *different* decision-makers, which is common with publicly available data. Our conditional exclusion restriction also means that certain decomposed effects are zero, so mediation analysis is less relevant. A related literature studies principal stratification Jiang et al. [2020], which is less interpretable since stratum membership is unknown. Similarly, even though encouragement effects are driven by compliers, complier-conditional analysis is less policy-relevant since complier identities are unknown. In general, our causal identification arguments are based on covariate-adjusted intention-to-treat analysis and covariate-adjusted as-treated analysis. We avoid estimation of stratum-specific effects, because if complier status is unknown, prescriptive decision rules cannot directly personalize by stratum membership.

Fair off-policy learning. We highlight some most closely related works in off-policy learning (omitting works in the sequential setting). [Metevier et al., 2019] studies high-probability fairness constraint satisfaction. [Kim et al., 2022] studies doubly-robust causal fair classification, while others have imposed deterministic resource constraints on the optimal policy formulation [Chohlas-Wood et al., 2021]. Other works study causal or counterfactual risk assessments [Mishler et al., 2021, Coston et al., 2020]. Our perspective is closer to that of off-policy learning, i.e. approximating direct control over the intervention by assuming stability in decision-maker treatment assignment probabilities. [Kallus and Zhou, 2019] studies (robust) bounds for treatment responders in binary outcome settings; this desiderata is coupled to classification notions of direct treatment. Again, our focus is on modeling the fairness implications of non-adherence. Indeed, in order to provide general algorithms and methods, we do build on prior fair classification literature. A different line of work studies "counterfactual" risk assessments which models a different concern.

3 Problem Setup

We briefly describe the problem setup. We work in the Neyman-Rubin potential outcomes framework for causal inference [Rubin, 1980]. We define the following:

- recommendation flag $R \in \{0, 1\}$, where $R = 1$ means encouraged/recommended. (We will use the terms encouragement/recommendation interchangeably).
- treatment $T(R) \in \mathcal{T}$, where $T(r) = 1$ indicates the treatment decision was 1 when the recommendation reported r.
- outcome $Y(T(R))$ is the potential outcome under encouragement r and treatment t.

Regarding fairness, we will be concerned about disparities in utility and treatment benefits (resources or burdens) across different groups, denoted $A \in \{a, b\}$. (For notational brevity, we may generically discuss identification/estimation without additionally conditioning on the protected attribute). For example, recommendations arise from binary high-risk/low-risk labels of classifiers. In practice, in consequential domains, classifier decisions are rarely automated, rather used to inform humans in the loop. The human expert in the loop decides whether or not to assign treatment. For binary outcomes, we will interpret $Y(t(r)) = 1$ as the positive outcome. When $Y \in \{0, 1\}, T \in \mathcal{T} =\in \{0, 1\}$ we may further develop analogues of fair classification criteria. We let $c(r, t, y) \colon \{0, 1\}^3 \mapsto \mathbb{R}$ denote

the cost function for $r \in \{0,1\}, t \in \mathcal{T}, y \in \{0,1\}$, which may sometimes be abbreviated $c_{rt}(y)$. We discuss identification and estimation based on the following recommendation, treatment propensity, and outcome models:

$$e_r(X,A) \coloneqq P(R=r \mid X,A), \;\; p_{t|r}(X,A) \coloneqq P(T=t \mid R=r,X,A),$$
$$\mu_{r,t}(X,A) \coloneqq \mathbb{E}[c_{rt}(Y) \mid R=r,T=t,X,A] = \mathbb{E}[c_{rt}(Y) \mid T=t,X,A] \coloneqq \mu_t(X,A) \text{ (asn.2)}$$

We are generally instead interested in *personalized recommendation rules* $\pi(r \mid X) = \pi_r(X)$ which describes the probability of assigning the recommendation r to covariates X. The average encouragement effect (AEE) is the difference in average outcomes if we refer everyone vs. no one, while the encouragement policy value $V(\pi)$ is the population expectation induced by the potential outcomes and treatment assignments realized under a recommendation policy π.

$$AEE = \mathbb{E}[Y(T(1)) - Y(T(0))], \qquad V(\pi) = \mathbb{E}[c(\pi, T(\pi), Y(\pi))].$$

Because algorithmic decision makers may be differentially responsive to recommendation, and treatment effects may be heterogeneous, the optimal recommendation rule may differ from the (infeasible) optimal treatment rule when taking constraints into account or for simpler policy classes.

Assumption 1 (Consistency and SUTVA). $Y_i = Y_i(T_i(R_i))$.

Assumption 2 (Conditional exclusion restriction). $Y(T(R)) \perp\!\!\!\perp R \mid T, X, A$.

Assumption 3 (Unconfoundedness). $Y(T(R)) \perp\!\!\!\perp T(R) \mid X, A$.

Assumption 4 (Stable responsivities under new recommendations). $P(T=t \mid R=r, X)$ remains fixed from the observational to the future dataset.

Assumption 5 (Decomposable costs). $c(r,t,y) = c_r(r) + c_t(t) + c_y(y)$

Our key assumption beyond standard causal inference assumptions is the conditional exclusion restriction assumption 2, i.e. that conditional on observable information X, the recommendation has no causal effect on the outcome beyond its effect on increasing treatment probability. This assumes that all of the covariate information that is informative of downstream outcomes is measured. Although this may not exactly hold in all applications, stating this assumption is also a starting point for sensitivity analysis under violations of it [Kallus and Zhou, 2018, Kallus et al., 2019b].

Assumption 4 is a structural assumption that limits our method to most appropriately re-optimize over small changes to existing algorithmic recommendations. This is also required for the validity of intention-to-treat analyses. For example, $p_{0|1}(x)$ (disagreement with algorithmic recommendation) could be a baseline algorithmic aversion. Not all settings are appropriate for this assumption. We don't assume micro-foundations on how or why human decision-makers were deviating from algorithmic recommendations, but take these patterns as given. One possibility for relaxing this assumption is via conducting sensitivity analysis, i.e. optimizing over unknown responsivity probabilities near known ones.

Assumption 5 is a mild assumption on modeling costs and benefits, that they are not defined on joint realizations of potential outcomes. This is relevant in practice, for example, performance management of practical systems typically compares aggregate utility and aggregate disparities rather than counterfactual individual-level utility and take-up outcomes.

We first also assume overlap in recommendations and treatment.

Assumption 6 (Overlap)**.** $\rho_r \leq e_r(X, A) \leq 1 - \rho_r$; $\rho_t \leq p_{t|r}(X, A) \leq 1 - \rho_t$ and $\rho_r, \rho_t \leq 0$

Assuming assumption 6 is like assuming we consider a randomized controlled trial with nonadherence. But later we give arguments using robustness to go beyond this, leveraging our finer-grained characterization.

Later on, we will be particularly interested in constrained formulations on the intention-to-treat effect that impose separate desiderata on outcomes under treatment, as well as treatment.

4 Method

We consider two settings: in the first, R is (as-if) randomized and satisfies overlap. Then R can be interpreted as intention to treat or prescription, whereas T is the actual realization thereof. We study identification of optimal encouragement designs with potential constraints on treatment or outcome utility patterns by group membership. We also first consider a special type of fairness constraint, resource parity, and characterize optimal decisions. In the second, R is an algorithmic recommendation that does not satisfy overlap in recommendation (but there is sufficient randomness in human decisions to satisfy overlap in treatment): we derive robustness checks in this setting.

First we discuss causal identification in optimal encouragement designs.

Proposition 1 (Regression adjustment identification)**.**

$$\mathbb{E}[c(\pi, T(\pi), Y(\pi))] = \textstyle\sum_{t\in\mathcal{T}, r\in\{0,1\}} \mathbb{E}[\pi_r(X)\mu_t(X)p_{t|r}(X)]$$

Proof of Proposition 1.

$$\begin{aligned}
\mathbb{E}[c(\pi, T(\pi), Y(\pi))] &= \textstyle\sum_{t\in\mathcal{T}, r\in\{0,1\}} \mathbb{E}[\pi_r(X)\mathbb{E}[\mathbb{I}\left[T(r) = t\right] c_{rt}(Y(t(r))) \mid R = r, X]] \\
&= \textstyle\sum_{t\in\mathcal{T}, r\in\{0,1\}} \mathbb{E}[\pi_r(X)P(T = t \mid R = r, X)\mathbb{E}[c_{rt}(Y(t(r))) \mid R = r, X]] \\
&= \textstyle\sum_{t\in\mathcal{T}, r\in\{0,1\}} \mathbb{E}[\pi_r(X)P(T = t \mid R = r, X)\mathbb{E}[c_{rt}(Y) \mid T = t, X]]
\end{aligned}$$

where the last line follows by the conditional exclusion restriction (Assumption 2) and consistency (Assumption 1). □

Resource-parity constrained optimal decision rules We consider a resource/burden parity fairness constraint:

$$V_\epsilon^* = \max_\pi \left\{\mathbb{E}[c(\pi, T(\pi), Y(\pi))] \colon \mathbb{E}[T(\pi) \mid A = a] - \mathbb{E}[T(\pi) \mid A = b] \leq \epsilon\right\} \tag{1}$$

Enforcing absolute values, etc. follows in the standard way. Not all values of ϵ may be feasible; in the appendix we give an auxiliary program to compute feasible ranges of ϵ. We first characterize a threshold solution when the policy class is unconstrained.

Proposition 2 (Threshold solutions)**.** Define

$$L(\lambda, X, A) = (p_{1|1}(X, A) - p_{1|0}(X, A)) \left\{\tau(X, A) + \frac{\lambda}{p(A)}(\mathbb{I}\left[A = a\right] - \mathbb{I}\left[A = b\right])\right\} + \lambda(p_{1|0}(X, a) - p_{1|0}(X, b))$$

Then:

$$\lambda^* \in \arg\min_\lambda \ \mathbb{E}[L(\lambda, X, A)_+], \quad \pi^*(x, u) = \mathbb{I}\{L(\lambda^*, X, u) > 0\}.$$

If instead $d(x)$ is a function of covariates x only,

$$\lambda^* \in \arg\min_{\lambda} \ \mathbb{E}[\mathbb{E}[L(\lambda, X, A) \mid X]_+], \ \ \pi^*(x) = \mathbb{I}\{\mathbb{E}[L(\lambda^*, X, A) \mid X] > 0\}.$$

Establishing this threshold structure (follows by duality of infinite-dimensional linear programming) allows us to provide a generalization bound argument.

4.1 Generalization

Proposition 3 (Policy value generalization)**.** Assume the nuisance models $\eta = [p_{1|0}, p_{1|1}, \mu_1, \mu_0]^\top, \eta \in \mathcal{F}_\eta$ are consistent and well-specified with finite VC-dimension v_η over the product function class $\mathcal{F}_\eta$. Let $\Pi = \{\mathbb{I}\{\mathbb{E}[L(\lambda, X, A; \eta) \mid X] > 0 \colon \lambda \in \mathbb{R}; \eta \in \mathcal{F}_\eta\}$.

$$\sup_{\pi \in \Pi, \lambda \in \mathbb{R}} |(\mathbb{E}_n[\pi L(\lambda, X, A)] - \mathbb{E}[\pi L(\lambda, X, A)])| = O_p(n^{-\frac{1}{2}})$$

This bound is stated for known nuisance functions: verifying stability under estimated nuisance functions further requires rate conditions.

Doubly-robust estimation We may improve statistical properties of estimation by developing *doubly robust* estimators which can achieve faster statistical convergence when both the probability of recommendation assignment (when it is random), and the probability of outcome are consistently estimated; or otherwise protect against misspecification of either model. We first consider the ideal setting when algorithmic recommendations are randomized so that $e_r(X) = P(R = r \mid X)$.

Proposition 4 (Variance-reduced estimation)**.**

$$V(\pi) = \sum_{t \in \mathcal{T}, r \in \{0,1\}} \mathbb{E}\left[\pi_r(X)\left\{\frac{\mathbb{I}\left[R = r\right]}{e_r(X)}(\mathbb{I}[T = t]c_{r1}(Y) - \mu_1(X)p_{t|r}(X)) + \mu_1(X)p_{t|r}(X)\right\}\right]$$

$$\mathbb{E}[T(\pi)] = \textstyle\sum_{r \in \{0,1\}} \mathbb{E}\left[\pi_r(X)\left\{\frac{\mathbb{I}[R=r]}{e_r(X)}(T(r) - p_{1|r}(x)) + p_{1|r}(x)\right\}\right]$$

Although similar characterization appears in [Qiu et al., 2021] for the doubly-robust policy value alone, note that doubly-robust versions of the constraints we study would result in differences in the Lagrangian so we retain the full expression rather than simplifying. For example, for regression adjustment, Proposition 9 provides interpretability on how constraints affect the optimal decision rule. In the appendix we provide additional results describing extensions of Proposition 8 with improved estimation.

4.2 Robust estimation with treatment overlap but not recommendation overlap

When recommendations are e.g. the high-risk/low-risk labels from binary classifiers, we may not satisfy the overlap assumption, since algorithmic recommendations are deterministic functions of covariates. However, note that identification in Proposition 1 requires only SUTVA and consistency, and the exclusion restriction assumption. Additional assumptions may be required to extrapolate $p_{t|r}(X)$ beyond regions of common support. On the other hand, supposing that positivity held with respect to T given covariates X, given unconfoundedness, our finer-grained approach can be beneficial because we only require robust extrapolation of $p_{t|r}(X)$, response to recommendations, rather than the outcome models $\mu_t(X)$.

We first describe what can be done if we allow ourselves parametric extrapolation on $p_{1|1}(X)$, treatment responsivity. In the case study later on, the support of $X \mid R = 1$ is a superset of the support of $X \mid R = 0$ in the observational data. Given this, we derive the following alternative identification based on marginal control variates (where $p_t = P(T = t \mid X)$ marginalizes over the distribution of R in the observational data):

Proposition 5 (Control variate for alternative identification)**.** Assume that $Y(T(r)) \perp T(r) \mid R = r, X$.

$$V(\pi) = \sum_{t \in \mathcal{T}, r \in \{0,1\}} \mathbb{E}\left[\left\{c_{rt}(Y(t)) \frac{\mathbb{I}\left[T = t\right]}{p_t(X)} + \left(1 - \frac{\mathbb{I}\left[T = t\right]}{p_t(X)}\right) \mu_t(X)\right\} p_{t|r}(X)\right]$$

On the other hand, parametric extrapolation is generally unsatisfactory because conclusions will be driven by model specification rather than observed data. Nonetheless, it can provide a starting point for robust extrapolation of structurally plausible treatment response probabilities. Note that our two-stage decomposition means that we impose this robust extrapolation on treatment response probabilities. But, sufficient randomness in the human decision-maker could satisfy overlap in the space of treatments, which is sufficient given our assumptions on the conditional exclusion restriction.

Robust extrapolation under violations of overlap We next describe methods for robust extrapolation under structural assumptions about smoothness of outcome models. Define the regions of no-overlap as the following: let $\mathcal{X}_r^{\text{nov}} = \{x : P(R = r \mid x) = 0\}$; on this region we do not jointly observe all potential values of (t, r, x), and let $\mathcal{X}^{\text{nov}} = \bigcup_r \mathcal{X}_r^{\text{nov}}$. Correspondingly, define the overlap region as $\mathcal{X}^{\text{ov}} = (\mathcal{X}^{\text{nov}})^c$. We consider uncertainty sets for ambiguous treatment recommendation probabilities. For example, one plausible structural assumption is *monotonicity*, that is, making an algorithmic recommendation can only increase the probability of being treated.

We define the following uncertainty set:

$$\mathcal{U}_{q_{t|r}} \coloneqq \left\{q_{1|r}(x') \colon q_{1|r}(x) \geq p_{1|r}(x),\ \forall x \in \mathcal{X}_r^{\text{nov}}\ \textstyle\sum_{t \in \mathcal{T}} q_{t|r}(x) = 1, \forall x, r\right\}$$

We could assume uniform bounds on unknown probabilities, or more refined bounds, such as Lipschitz-smoothness with respect to some distance metric d, or boundedness.

$$\begin{aligned}
\mathcal{U}_{\text{lip}} &\coloneqq \left\{q_{1|r}(x') \colon d(q_{1|r}(x'), p_{1|r}(x)) \leq L d(x', x),\ (x', x) \in (\mathcal{X}^{\text{nov}} \times \mathcal{X}^{\text{nov}})\right\} \\
\mathcal{U}_{\text{bnd}} &\coloneqq \left\{q_{1|r}(x') \colon \underline{b}(x) \leq q_{1|r}(x') \leq \overline{b}(x)\right\}
\end{aligned}$$

Define $V_{ov}(\pi) \coloneqq \sum_{t \in \mathcal{T}, r \in \{0,1\}} \mathbb{E}[\pi_r(X) p_{t|r}(X) \mu_t(X) \mathbb{I}\{X \in \mathcal{X}^{\text{ov}}\}]$. Let $\mathcal{U}$ denote the uncertainty set including any custom constraints, e.g. $\mathcal{U} = \mathcal{U}_{q_{t|r}} \cap \mathcal{U}_{\text{lip}}$. Then we may obtain robust bounds by optimizing over regions of no overlap:

$$\begin{aligned}
\overline{V}(\pi) &\coloneqq V_{ov}(\pi) + \overline{V}_{nov}(\pi), \\
\text{where } \overline{V}_{nov}(\pi) &\coloneqq \max_{q_{tr}(X) \in \mathcal{U}} \left\{\textstyle\sum_{t \in \mathcal{T}, r \in \{0,1\}} \mathbb{E}[\pi_r(X) \mu_t(X) q_{tr}(X) \mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\}]\right\}
\end{aligned}$$

In the specialized, but practically relevant case of binary outcomes/treatments/recommendations, we obtain the following simplifications for bounds on the policy value, and the minimax robust policy that optimizes the worst-case overlap extrapolation function. In the special case of constant uniform bounds, it is equivalent (in the case of binary outcomes) to consider marginalizations:

Lemma 1 (Binary outcomes, constant bound)**.** *Let* $\mathcal{U}_{cbnd} \coloneqq \left\{q_{t|r}(x') \colon \underline{B} \leq q_{1|r}(x') \leq \overline{B}\right\}$ *and* $\mathcal{U} = \mathcal{U}_{q_{t|r}} \cap \mathcal{U}_{cbnd}$*. Define* $\beta_{t|r} \coloneqq \mathbb{E}[q_{t|r}(X, A) \mid T = t]$*.If* $T \in \{0, 1\}$*,*

$$\overline{V}_{no}(\pi) = \sum_{t \in \mathcal{T}, r \in \{0,1\}} \mathbb{E}[c^*_{rt} \beta_{t|r} \mathbb{E}[Y \pi_r(X) \mid T = t] \mathbb{I}\{X \in \mathcal{X}^{nov}\}]],$$

$$\textit{where } c^*_{rt} = \begin{cases} \overline{B}\mathbb{I}\left[t = 1\right] + \underline{B}\mathbb{I}\left[t = 0\right] & \textit{if } \mathbb{E}[Y \pi_r(X) \mid T = t] \geq 0 \\ \overline{B}\mathbb{I}\left[t = 0\right] + \underline{B}\mathbb{I}\left[t = 1\right] & \textit{if } \mathbb{E}[Y \pi_r(X) \mid T = t] < 0 \end{cases}.$$

We consider the case of continuous-valued outcomes in the example setting of the simple resource-parity constrained program of **??**. We first study simple uncertainty sets, like intervals, to deduce insights about the robust policy. In the appendix we include a more general reformulation for polytopic uncertainty sets.

Proposition 6 (Robust linear program)**.** Suppose $r, t \in \{0, 1\}$, and $q_{r1}(\cdot, u) \in \mathcal{U}_{bnd}, \forall r, u$. Define

$$\tau(x, a) \coloneqq \mu_1(x, a) - \mu_0(x, a), \quad \Delta B_r(x, u) \coloneqq (\overline{B}_r(x, u) - \underline{B}_r(x, u)),$$
$$B^{\text{mid}}_r(x, u) \coloneqq \underline{B}_r(x, u) + \frac{1}{2}\Delta B_r(x, u), \quad c_1(\pi) \coloneqq \textstyle\sum_r \mathbb{E}[\tau \pi_r B^{\text{mid}}],$$
$$\mathbb{E}[\Delta_{ov} T(\pi)] \coloneqq \mathbb{E}[T(\pi)\mathbb{I}\{X \in \mathcal{X}^{\text{ov}}\} \mid A = a] - \mathbb{E}[T(\pi)\mathbb{I}\{X \in \mathcal{X}^{\text{ov}}\} \mid A = b].$$

Then the robust linear program is:

$$\begin{aligned} &\max V_{ov}(\pi) + \mathbb{E}[\mu_0] + c_1(\pi) - \tfrac{1}{2}\textstyle\sum_r \mathbb{E}[|\tau| \, \pi_r \Delta B_r(X, A) \mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\}] \\ &\text{s.t. } \textstyle\sum_r \{\mathbb{E}[\pi_r \overline{B}_r(X, A)\mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\} \mid A = a] - \mathbb{E}[\pi_r \underline{B}_r(X, A)\mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\} \mid A = b]\} + \Delta^T_{ov}(\pi) \leq \epsilon \end{aligned}$$

5 Additional fairness constraints and policy optimization

We previously discussed policy optimization, over unrestricted decision rules, given estimates. We now introduce general methodology to handle 1) optimization over a policy class of restricted functional form and 2) more general fairness constraints. We first introduce the fair-classification algorithm of [Agarwal et al., 2018], describe our extensions to obtain variance-sensitive regret bounds and less conservative policy optimization (inspired by a regularized ERM argument given in [Chernozhukov et al., 2019]), and then provide sample complexity analysis.

Algorithm and setup We first describe the reductions-based approach for fair classification of Agarwal et al. [2018] before describing our adaptation for constrained policy learning, and localized two-stage variance reduction. They consider classification (i.e. loss minimization) under fairness constraints that can be represented generically as a linear program. In the following, to be consistent with standard form for linear programs, note that we consider costs Y so that we can phrase the saddle-point as minimization-maximization. The $|\mathcal{K}|$ linear constraints and J groups (values of protected attribute A) are summarized via a coefficient matrix $M \in \mathbb{R}^{K \times J}$, which multiplies a vector of constraint moments $h_j(\pi), j \in [J]$ (with J the number of groups), $O = (X, A, R, T, Y)$ denoting our data observations, and d the constraint constant vector:

$$\begin{aligned} &h_j(\pi) = \mathbb{E}\left[g_j(O, \pi(X)) \mid \mathcal{E}_j\right] \quad \text{for } j \in J, \\ &Mh(\pi) \leq d \end{aligned}$$

Algorithm 1 REDFAIR($\mathcal{D}, g, \mathcal{E}, M, d$)

Input: $\mathcal{D} = \{(X_i, R_i, T_i, Y_i, A_i)\}_{i=1}^n$, $g, \mathcal{E}, M, \widehat{d}$, B, accuracy ν, α, stepsize ω, initialization $\theta_1 = 0 \in \mathbb{R}^{|\mathcal{K}|}$
for iteration $t = 1, 2, \ldots$ **do**
  Set $\lambda_{t,k} = B\frac{\exp\{\theta_k\}}{1+\sum_{k' \in \mathcal{K}} \exp\{\theta_{k'}\}}$ for all $k \in \mathcal{K}$,
  $\beta_t \leftarrow \text{BEST}_\beta(\lambda_t)$,
  $\widehat{Q}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t \beta_{t'}$
  $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t \lambda_{t'}$,
  $\bar{L} \leftarrow L(\hat{Q}_t, \text{BEST}_\lambda(\hat{Q}_t)), \underline{L} \leftarrow (\text{BEST}_\beta(\hat{\lambda}_t), \hat{\lambda}_t)$,
  $\nu_t \leftarrow \max\{L(\hat{Q}_t, \hat{\lambda}_t) - \underline{L}, \quad \bar{L} - L(\hat{Q}_t, \hat{\lambda}_t)\}$, If $\nu_t \leq \nu$ then return $(\hat{Q}_t, \hat{\lambda}_t)$
  $\theta_{t+1,i} = \theta_t + \log(1 - \omega(M\hat{\mu}(h_t) - \hat{c})_j), \forall i$
end for

The elements of $h_j(\pi)$ are average functionals, for example the average treatment takeup in group j. Importantly, the moment function g_j depends on π while the conditioning event $\mathcal{E}_j$ cannot depend on π. Many important fairness constraints can nonetheless be written in this framework, such as burden/resource parity, parity in true positive rates, but not measures such as calibration whose conditioning event does depend on π. (See Appendix B.2 for examples omitted for brevity).

Our objective function is the policy value $V(\pi)$. (Later this is linearized, as in [Agarwal et al., 2018] by optimizing over distributions over policies). We further consider a convexification of Π via randomized policies $Q \in \Delta(\Pi)$, where $\Delta(\Pi)$ is the set of distributions over Π, i.e. a randomized classifier that samples a policy $\pi \sim Q$. Therefore, our target estimand is the optimal distribution Q over policies π that maximizes the objective value $V(Q)$ subject to the fairness constraints encoded in $Mh(Q) \leq d$:

$$\min_{Q \in \Delta(\Pi)} \{V(Q) : \quad Mh(Q) \leq d\}$$

Next we discuss the cost-weighted classification reduction of off-policy learning [Zhao et al., 2012], which we use to solve constrained off-policy learning via [Agarwal et al., 2014].

Weighted classification reduction and off-policy estimation. There is a well-known reduction of optimizing the zero-one loss for policy learning to weighted classification. Note that the reductions approach of [Agarwal et al., 2014] works with the Lagrangian relaxation which only further introduces datapoint-dependent additional weights. Notationally, in this section, for policy optimization, $\pi \in \{-1, +1\}, T \in \{-1, +1\}$ (for notational convenience alone). We consider parameterized policy classes so that $\pi(x) = \pi(1 \mid x) = \text{sign}(f_\beta(x))$ for some index function f depending on a parameter $\beta \in \mathbb{R}^d$. Consider the centered regret $J(\pi) = \mathbb{E}[Y(\pi)] - \frac{1}{2}\mathbb{E}[\mathbb{E}[Y \mid R = 1, X] + \mathbb{E}[Y \mid R = 0, X]]$. We summarize different estimation strategies via the score function $\psi_{(\cdot)}(O)$, where $(\cdot) \in \{DM, IPW, DR\}$: the necessary property is that $\mathbb{E}[\psi \mid X] = \mathbb{E}[Y \mid R = 1, X] - \mathbb{E}[Y \mid R = 0, X]$. The specific functional forms of these different estimators are as follows, where $\mu_r^R(X) = \mathbb{E}[Y \mid R = r, X]$:

$$\psi_{DM} = (p_{1|1}(X) - p_{1|0}(X))(\mu_1(X) - \mu_0(X)), \psi_{IPW} = \frac{RY}{e_R(X)}, \psi_{DR} = \psi_{DM} + \psi_{IPW} + \frac{R\mu^R(X)}{e_R(X)}.$$

Therefore the centered regret can be reparametrized via the parameter β as: $J(\beta) = J(\text{sgn}(f_\beta(\cdot))) = \mathbb{E}[\text{sgn}(f_\beta(X))\{\psi\}]$. We can apply the standard reduction to cost-sensitive classification since $\psi_i \,\text{sgn}(f_\beta(X_i)) = |\psi_i|(1 - 2\mathbb{I}[\text{sgn}(f_\beta(X_i)) \neq \text{sgn}(\psi_i)])$. Then we can use surrogate losses for the

Algorithm 2 Two-stage localized fair classification via reductions

1: Randomly split the data into two folds $\mathcal{D}_1, \mathcal{D}_2$.

2: Obtain $\hat{Q}_1^*$ and the index set of binding constraints $\hat{\mathcal{I}}_1$ by learning nuisances η_1 and running Algorithm 1 on $\mathcal{D}_1$ with REDFAIR$(\mathcal{D}_1, h, \mathcal{E}, M, d; \eta_1)$.

3: $\hat{\sigma}_j^2 \leftarrow \hat{\mathrm{Var}}(g_j(O, \hat{Q}_1)\mathbb{I}[\mathcal{E}_j]/p_j),\ \forall j$
$\hat{d} \leftarrow d + 2\sum_{j\in\mathcal{J}} |M_{k,j}|\, \hat{\sigma}_j^2 n^{-\alpha}$

4: Augment additional constraints with ϵ_n-policy-value and constraint slices relative to $\hat{\pi}_1$: define an augmented system (where subscripting by $\hat{\mathcal{I}}_1$ subindexes the corresponding matrix or vector):

$$\tilde{h}_{j'}(Q) = \mathbb{E}_{n_1}[\{g_{j'}(O; \hat{Q}_1) - g_{j'}(O; Q)\} \mid \mathcal{E}_j],\ \ \forall j' \in \hat{\mathcal{I}}_1,$$
$$\tilde{h}^v(Q) = \mathbb{E}_{n_1}[v_{DR}(O; \hat{Q}_1, \eta_1) - v_{DR}(O; Q, \eta_1)]$$
$$\tilde{M} = [M; M_{\hat{\mathcal{I}}_1}, \vec{1}], \tilde{h} = [h, \tilde{h}, \tilde{h}^v]^\top, \tilde{d} = [\hat{d}, \epsilon_n\vec{1}, \epsilon_n]^\top, \tilde{\mathcal{E}} = [\mathcal{E}, \mathcal{E}_{\hat{\mathcal{I}}_1}, \emptyset]^\top$$

5: Solve $\min_{Q\in\Delta(\Pi)}\{\hat{V}(Q)\colon\ \tilde{M}\tilde{h}(Q) \le \tilde{d}\}$.
Obtain $\hat{Q}_2^*$ by running Algorithm 1 on $\mathcal{D}_2$ with REDFAIR$(\mathcal{D}, \tilde{g}, \tilde{\mathcal{E}}, \tilde{M}, \tilde{d}, \eta_2)$.

zero-one loss. Although many functional forms for $\ell(\cdot)$ are Fisher-consistent, one such choice of ℓ is the logistic (cross-entropy) loss given below:

$$\mathbb{E}[|\psi|\, \ell(f_\beta(X), \mathrm{sgn}(\psi))], \qquad l(g, s) = 2\log(1 + \exp(g)) - (s + 1) \tag{2}$$

Optimization. Ultimately, the optimization is solved using sampled and estimated moments. Define the integrand of the constrained, weighted empirical risk minimization as $v_{(\cdot)}(O; \pi_\beta, \eta) = \left|\psi_{(\cdot)}(O; \eta)\right| \ell(f_\beta(X), \mathrm{sgn}(\psi_{(\cdot)}(O; \eta)))$. Our estimate of the objective function is therefore

$$V_{(\cdot)}(Q) = \mathbb{E}[\left|\psi_{(\cdot)}\right| \ell(f_\beta, \mathrm{sgn}(\psi_{(\cdot)}))] = \mathbb{E}_{\pi_\beta\sim Q}[v_{(\cdot)}(O; \pi_\beta, \eta)].$$

Note that for the rest of our discussions of algorithms for constrained policy optimization, we overload notation and use $V_{(\cdot)}(Q)$ to refer to policy *regret*, as above. The optimal policies are the same for regret vs. value. We obtain the sample estimator $\hat{V}_{(\cdot)}(Q)$ and sample constraint moments $\hat{h}(Q)$ analogously. We also add a feasibility margin ϵ_k which depends on concentration of the estimated constraints, so the sampled constraint vector is $\hat{d}_k = d_k + \epsilon_k$, for all k. We seek an approximate saddle point so that the constrained solution is equivalent to the Lagrangian,

$$\hat{L}(Q, \lambda) = \hat{V}(Q) + \lambda^\top(M\hat{h}(Q) - \hat{d}), \qquad \min_{Q\in\Delta(\Pi)}\{\hat{V}(Q)\colon M\hat{h}(Q) \le \hat{d}\} = \min_{Q\in\Delta(\Pi)} \max_{\lambda\in\mathbb{R}_+^K} \hat{L}(Q, \lambda).$$

We simultaneously solve for an approximate saddle point over the B-bounded domain of λ:

$$\min_{Q\in\Delta} \max_{\lambda\in\mathbb{R}_+^{|\mathcal{X}|}, \|\lambda\|_1\le B} \hat{L}(Q, \lambda), \qquad \max_{\lambda\in\mathbb{R}_+^{|\mathcal{X}|}, \|\lambda\|_1\le B} \min_{Q\in\Delta} \hat{L}(Q, \lambda)$$

[Agarwal et al., 2018, Theorem 3] gives out-of-sample generalization guarantees on the policy value and constraint violation achieved by the approximate saddle point output by the algorithm. The analysis is generic under rate assumptions on uniform convergence of policy and constraint values, summarized in the following assumption.

Assumption 7 (Rate assumption on policy and constraint values.)**.** There exists $C, C' \geq 0$ and $\alpha \leq 1/2$ such that $\sup_{Q \in \Delta(\Pi)}\{V(Q;\eta) - \hat{V}(Q;\hat{\eta})\} \leq Cn^{-\alpha}$ and $\varepsilon_k = C' \sum_{j \in \mathcal{J}} |M_{k,j}|\, n_j^{-\alpha}$, where n_j is the number of data points that fall in $\mathcal{E}_j$.

Such a rate α follows from standard analyses in causal inference, depending on the exact assumptions and estimators, and is used to set the constraint violation feasibility margin $\epsilon_k = O(n^{-\alpha})$. Next we summarize the optimization algorithm.

We play a no-regret (second-order multiplicative weights [Cesa-Bianchi et al., 2007, Steinhardt and Liang, 2014], a slight variant of Hedge/exponentiated gradient [Freund and Schapire, 1997]) algorithm for the $\lambda-$player, while using best-response oracles for the $Q-$player. Full details are in Algorithm 1. Given λ_t, $\text{BEST}_\beta(\lambda_t)$ computes a best response over Q; since the worst-case distribution will place all its weight on one classifier, this step can be implemented by a reduction to cost-sensitive/weighted classification [Beygelzimer and Langford, 2009, Zhao et al., 2012], which we describe in further detail below. Computing the best response over $\text{BEST}_\lambda(\hat{Q}_t))$ selects the most violated constraint. Further details are in Appendix B.2.

Two-stage variance-constrained algorithm. We seek to improve upon this procedure so that we may obtain regret bounds on policy value and fairness constraint violation that exhibit more favorable dependence on the maximal variance over small-variance *slices* near the optimal policy, rather than worst-case constants over all policies [Chernozhukov et al., 2019, Athey and Wager, 2021]. Further, the constraint feasibility slacks were set via generalization bounds that typically depend on worst-case constants: so adapting to estimates of the variance can achieve tighter fairness control.

These challenges motivate the two-stage procedure, described formally in Algorithm 2 and verbally here. We adapt an out-of-sample regularization scheme developed in [Chernozhukov et al., 2019], which recovers variance-sensitive regret bounds via a small modification to an empirical risk minimization procedure (and by extension, policy learning). We split the data into two subsets $\mathcal{D}_1, \mathcal{D}_2$, and first learn nuisance estimators $\hat{\eta}_1$ from $\mathcal{D}_1$ (possibly with further sample-splitting) for use in our policy value and constraint estimates. We run Algorithm 1 ($\text{REDFAIR}(\mathcal{D}_1, h, \mathcal{E}, M, d; \hat{\eta}_1)$) on data from $\mathcal{D}_1$ to obtain an estimate of the optimal policy distribution $\hat{Q}_1$, and the constraint variances at $\hat{Q}_1$. We also identify the binding constraints from the first stage via the index set $\hat{I}_1$. Next, we *augment* the constraint matrix with additional constraints that require feasible policies for the second-stage policy distribution to achieve ϵ_n close policy value and constraint moment values relative to $\hat{Q}_1$. Since errors concentrate quickly, this can be viewed as variance regularization. And, we set the constraint slacks $\hat{d} \leftarrow d + 2\sum_{j \in \mathcal{J}} |M_{k,j}|\, \hat{\sigma}_j^2 n^{-\alpha}$ in the second stage using estimated variance constants from $\hat{Q}_1$. This results in tighter control of fairness constraints. The second stage solves for an approximate saddle point of the augmented system, with objective function and constraints evaluated on $\mathcal{D}_2$ and returns $\hat{Q}_2$.

Next, we provide a generalization bound on the out-of-sample performance of the policy returned by the two-stage procedure. Importantly, because of our two stage procedure, the regret of the policy depends on the worst-case variance of near-optimal policies (rather than all policies). Define the function classes

$$\mathcal{F}_\Pi = \{v_{DR}(\cdot, \pi; \eta) \colon \pi \in \Pi, \eta \in \mathcal{F}_\eta\}, \qquad \mathcal{F}_j = \{g_j(\cdot, \pi; \eta) \colon \pi \in \Pi, \eta \in \mathcal{F}_\eta\}$$

and the empirical entropy integral $\kappa(r, \mathcal{F}) = \inf_{\alpha \geq 0}\{4\alpha + 10\int_\alpha^r \sqrt{\frac{\mathcal{H}_2(\epsilon, \mathcal{F}, n)}{n}} d\epsilon\}$ where $H_2(\epsilon, \mathcal{F}, n)$ is the L_2 empirical entropy, i.e. log of the $\|\cdot\|_2$ ϵ-covering number. We make a mild assumption of a learnable function class (bounded entropy integral) [Van Der Vaart et al., 1996]. Many standard

function classes used in machine learning such as linear models, polynomials, kernel regression, and neural networks satisfy this assumption [Wainwright, 2019].

Assumption 8. The function classes $\mathcal{F}_\Pi, \{\mathcal{F}_j\}_{j\in\mathcal{J}}$ satisfy that for any constant $r, \kappa(r, \mathcal{F}) \to 0$ as $n \to \infty$. The function classes $\{\mathcal{F}_j\}_{j\in\mathcal{J}}$ comprise of L_j-Lipschitz contractions of π.

We will assume that we are using doubly-robust/orthogonalized estimation as in proposition 4, hence state results depending on estimation error of nuisance vector η. The next theorem summarizes the out-of-sample performance of the distribution over policies output by the two-stage algorithm of Algorithm 2, $\hat{Q}_2$.

Theorem 1 (Variance-Based Oracle Policy Regret)**.** *Suppose that the mean-squared-error of the nuisance estimates is upper bounded w.p.* $1-\delta/2$ *by* $\chi^2_{n,\delta}$*, over the randomness of the nuisance sample:* $\max_l\{\mathbb{E}[(\hat{\eta}_l - \eta_l)^2]\}_{l\in[L]} \coloneqq \chi^2_n$.

Let $v^0_{DR}(O;Q)$ *denote evaluation with true nuisance functions* η_0*; define* $r = \sup_{Q\in\mathcal{Q}} \sqrt{\mathbb{E}\left[v^0_{DR}(O;Q)^2\right]}$ *and* $\epsilon_n = \Theta\left(\kappa\left(r, \mathcal{F}_\Pi\right) + r\sqrt{\frac{\log(1/\delta)}{n}}\right)$*. Moreover, denote an* ϵ*-regret slice of the policy space:*

$$\mathcal{Q}_*(\epsilon) = \left\{Q \in \Delta[\Pi] : V(Q^0_*) - V(Q) \le \epsilon,\ \ h(Q^0_*) - h(Q) \le d + \epsilon\right\}.$$

Let $\tilde{\epsilon}_n = O(\epsilon_n + \chi^2_{n,\delta})$ *and denote the variance of the difference between any two policies in an* ϵ_n*-regret slice, evaluated at the true nuisance quantities:*

$$\bar{\sigma}^2_{\mathcal{D}_2} = \sup\left\{\mathrm{Var}\left(v^0_{DR}(O;Q) - v^0_{DR}\left(O;Q'\right)\right) : Q, Q' \in \mathcal{Q}_*(\tilde{\epsilon}_n)\right\}.$$

(Define $\bar{\sigma}^2_{k,\mathcal{D}_2}$ *analogously for the variance of constraint moments). Then, letting* $\gamma(Q) \coloneqq Mh(Q)$ *denote the constraint values, the policy distribution* Q_2 *returned by the out-of-sample regularized ERM satisfies w.p.* $1-\delta$ *over the randomness of* S*:*

$$V(\hat{Q}_2) - V(Q^*) = O(\kappa(\bar{\sigma}_{\mathcal{D}_2}, \mathrm{conv}(\mathcal{F}_\Pi)) + \bar{\sigma}_{\mathcal{D}_2} n^{-\frac{1}{2}}\sqrt{\log(3/\delta)} + \chi^2_{n,\delta})$$
$$(\gamma_k(\hat{Q}_2) - d_k) - (\gamma_k(Q^*) - d_k) = O(\kappa(\bar{\sigma}_{k,\mathcal{D}_2}, \mathrm{conv}(\mathcal{F}_j)) + \bar{\sigma}_{k,\mathcal{D}_2} n^{-\frac{1}{2}}\sqrt{\log(3/\delta)} + \chi^2_{n,\delta})$$

Note that the benefits of double robustness arise since the additional estimation error is an additive term that is quadratic in the root mean-squared error of the nuisance estimation instead of linear; this reflects the rate-double-robustness benefits of orthogonalized estimation. The specific benefits of the two-stage approach are that 1) the constants are improved from an absolute, structure-agnostic bound to depending on the variance of low-regret policies, which also reflects improved variance from using doubly-robust estimation as in proposition 4, and 2) less-conservative out-of-sample fairness constraint satisfaction.

6 Case Studies

We apply our methods to datasets in application areas relevant to our setting. First we consider a setting with bona-fide randomized encouragements, building on the Oregon Health Insurance Study. Next we consider a setting with more complex data, supervised release and final release recommendation decisions in a setting without overlap in algorithmic recommendation (but much randomness in observed treatments).

6.1 Oregon Health Insurance Study

The Oregon Health Insurance Study [Finkelstein et al., 2012] is an important study on the causal effect of expanding public health insurance on healthcare utilization, outcomes, and other outcomes. It is based on a randomized controlled trial made possible by resource limitations, which enabled the use of a randomized lottery to expand Medicaid eligibility for low-income uninsured adults. Outcomes of interest included health care utilization, financial hardship, health, and labor market outcomes and political participation.

Because the Oregon Health Insurance Study expanded access to *enroll* in Medicaid, a social safety net program, the effective treatment policy is in the space of *encouragement* to enroll in insurance (via access to Medicaid) rather than direct enrollment. This encouragement structure is shared by many other interventions in social services that may invest in nudges to individuals to enroll, tailored assistance, outreach, etc., but typically do not automatically enroll or automatically initiate transfers. Indeed this so-called *administrative burden* of requiring eligible individuals to undergo a costly enrollment process, rather than automatically enrolling all eligible individuals, is a common policy design lever in social safety net programs. Therefore we expect many beneficial interventions in this consequential domain to have this encouragement structure.

We preprocess the data by partially running the Stata replication file, obtaining a processed data file as input, and then selecting a subset of covariates that could be relevant for personalization. These covariates include household information that affected stratified lottery probabilities, socioeconomic demographics, medical status and other health information.

In the notation of our framework, the setup of the optimal/fair encouragement policy design question is as follows:

- X: covariates (baseline household information, socioeconomic demographics, health information)
- A: race (non-white/white), or gender (female/male)

  These protected attributes were binarized.
- R: encouragement: lottery status of expanded eligibility (i.e. invitation to enroll when individual was previously ineligible to enroll)
- T: whether the individual is enrolled in insurance ever

  Note that for $R = 1$ this can be either Medicaid or private insurance while for $R = 0$ this is still well-defined as this can be private insurance.
- Y: number of doctor visits

  This outcome was used as a measure of healthcare utilization. Overall, the study found statistically significant effects on healthcare utilization. An implicit assumption is that increased healthcare utilization leads to better health outcomes.

We subsetted the data to include complete cases only (i.e. without missing covariates). We learned propensity and treatment propensity models via logistic regression for each group, and used gradient-boosted regression for the outcome model. We first include results for regression adjustment identification. One potential concern is the continued use of the healthcare utilization variable as an outcome measure. From a methodological angle, it displays heterogeneity in treatment effects.

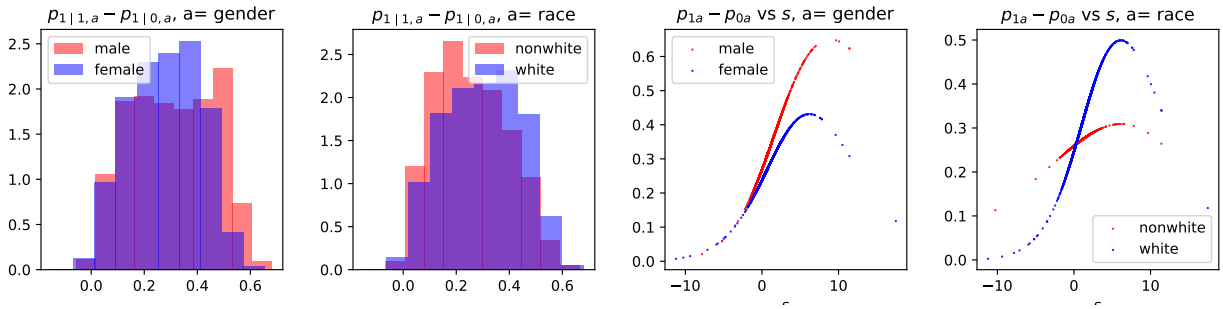

Figure 1: Distribution of lift in treatment probabilities $p_{1|1,a} - p_{1|0,a} = P(T = 1 \mid R = 1, A = a, X) - P(T = 1 \mid R = 0, A = a, X)$, and plot of $p_{1|1,a} - p_{1|0,a}$ vs. τ.

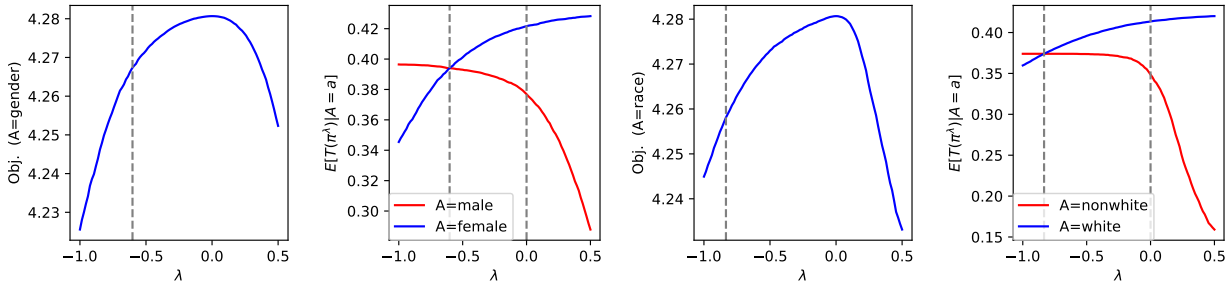

Figure 2: Policy value $V(\pi^\lambda)$, treatment value $\mathbb{E}[T(\pi^\lambda) \mid A = a]$, for A = race, gender.

From the substantive angle, healthcare utilization remains a proxy outcome measure for other health measures, and interpreting increases in healthcare utilization as beneficial is justified primarily by assuming that individuals were constrained by the costs of uninsured healthcare previously, so that increases in healthcare utilization reflect that access to insurance increases in access to care.

In Figure 1 we plot descriptive statistics. We include histograms of the treatment responsivity lifts $p_{1|1a}(x, a) - p_{1|0a}(x, a)$. We see some differences in distributions of responsivity by gender and race. We then regress treatment responsivity on the outcome-model estimate of τ. We find substantially more heterogeneity in treatment responsivity by race than by gender: whites are substantially more likely to take up insurance when made eligible, conditional on the same expected treatment effect heterogeneity in increase in healthcare utilization. (This is broadly consistent with health policy discussions regarding mistrust of the healthcare system).

Next we consider imposing treatment parity constraints on an unconstrained optimal policy (defined on these estimates). In Figure 2 we display the objective value, and $\mathbb{E}[T(\pi) \mid A = a]$, for gender and race, respectively, enumerated over values of the constraint. We use costs of 2 for the number of doctors visits and 1 for enrollment in Medicaid (so $\mathbb{E}[T(\pi) \mid A = a]$ is on the scale of probability of enrollment). These costs were chosen arbitrarily. Finding optimal policies that improve disparities in group-conditional access can be done with relatively little impact to the overall objective value. These group-conditional access disparities can be reduced from 4 percentage points (0.04) for gender and about 6 percentage points (0.06) for race at a cost of 0.01 or 0.02 in objective value (twice the number of doctors' visits). On the other hand, relative improvements/compromises in access

value for the "advantaged group" show different tradeoffs. Plotting the tradeoff curve for race shows that, consistent with the large differences in treatment responsivity we see for whites, improving access for blacks. Looking at this disparity curve given λ however, we can also see that small values of λ can have relatively large improvements in access for blacks before these improvements saturate, and larger λ values lead to smaller increases in access for blacks vs. larger decreases in access for whites.

6.2 Decision-Making Framework for Electronic Monitoring case study.

Another case study is on a dataset of judicial decisions on *supervised* release based on risk-score-informed recommendations of supervised release under an electronic-monitoring program [Office of the Chief Judge, 2019a]. The PSA-DMF (Public Safety Assessment Decision Making Framework) uses a prediction of failure to appear for a future court data to inform pretrial decisions, including our focus on supervised release with electronic monitoring, where judges make the final decision [psa, 2016]. Despite a large literature on algorithmic fairness of pretrial risk assessment, to the best of our knowledge, it is unclear what empirical evidence justifies release recommendation matrices that have been used in practice to recommend supervised release. There are current policy concerns about disparities in increasing use of supervised release given mixed evidence on outcomes [Office of the Chief Judge, 2019a, Gross]; e.g. Safety and Justice Challenge [2022] concludes "targeted efforts to reduce racial disparities are necessary". We focus on a publicly-available dataset from Cook County which includes information about defendant characteristics, algorithmic recommendation for electronic monitoring, detention/release/supervised release decisions, and failure to appear and other outcomes [Office of the Chief Judge, 2019b]. The data were initially used to assess bail reform [Office of the Chief Judge, 2019a], though some line-level data is aggregated for privacy.

Finally, we note that we focus on supervised release in the setting of this program which enrolls defendants in electronic monitoring. "Supervised release" in general is a broad term that can encompass substantially very different programs, from more limiting electronic monitoring vs. referrals and enrollment in supportive services elsewhere. For example, the expansion of supervised release via access to supportive services and caseworkers [Akinnibi and Holder, 2023] has recently been touted as a factor in enabling New York's bail reform, and hence the release (supervised or unsupervised) of more defendants. Still, the broad structure we outline here of algorithmic recommendation informing (but not compelling) a final judicial decision with wide discretion, with concerns about disparities in both take-up and outcomes, is shared across supervised release, even though the programs themselves can be very different in different jurisdictions. Therefore, we also expect that fairness concerns will defer depending on the exact program implementation.

- $Z \in \{0, 1\}$: released population (with or without conditions)

  All of the analysis occurs in the (XZ, AZ, RZ, YZ) strata, i.e. among the released population only. For brevity we drop the Z designation in describing the data below.

- X: covariates (age, top charge category, PSA FTA/NCA score bucket and flag, top charge category)

- A: race (non-white/white), or gender (female/male)

  These protected attributes were reported as binarized.

- R: algorithmic recommendation: a recommendation from the PSA-DMF matrix for supervised release (at any intensity of supervision conditions)
- T: whether the individual is released under supervision (at any intensity of supervision conditions)
- Y: failure to appear ($Y = 1$)

Caveats re: data issues In this initial case study, we work with publicly available data [Office of the Chief Judge, 2019b]. First, before describing the analysis, we acknowledge important data issues. These issues are similar to those that arise in data from the criminal justice system [Bao et al., 2021], as well as additional particularities with our particular data source. Because of these issues, this section should be thought of as exploratory. In future work we will seek more granular data with additional robustness checks to support substantive conclusions.

Our analysis proceeds conditional on the non-detained population. This could make sense in a setting where decision-making frameworks for supervised release are unlikely to change judicial decisions to detain or release: our results apply to marginal defendants. Covariate levels (including PSA scores) were discretized for privacy. Moreover, the recorded final supervision decision does not include intensity, but different intensities are recommended in the data, which we collapse into a single level. The PSA-DMF is an algorithmic recommendation so here we are appealing to overlap in treatment recommendations, but using parametric extrapolation in responsivity. So, we are assuming randomness in treatment assignment that arises either from quasi-random assignment to judges or noise/variability in judicial decisions. We strongly appeal to this interpretation of randomness in treatment assignment in the conceptualization of a causal effect of treatment with supervised release. Other accounts and conceptualizations of judicial decision-making could instead argue that judicial decisions such as conditional release are by their very nature discretionary, and do not admit valid counterfactuals. We instead appeal to a hypothetical randomized experiment (if unethical) where individuals could conceivably be randomized into supervised release or not. Finally, unconfoundedness is likely untrue, but sensitivity analysis could address this in ways quite similar to those studied previously in the literature [Kallus et al., 2019b, Kallus and Zhou, 2021b, 2018].

Again, we caution that this analysis is an exploratory exercise to illustrate the relevance of the methods. Future work will use proprietary data for a higher-fidelity empirical study.

Analysis Next in Figure 3 we provide descriptive information illustrating heterogeneity (including by protected attribute) in adherence and effectiveness. We observe wide variation in judges assigning supervised release beyond the recommendation. We use logistic regression to estimate outcome models and treatment response models. The first figure shows estimates of the causal effect for different groups, by gender (similar heterogeneity for race). The outcome is failure to appear, so negative scores are beneficial. The second figure illustrates the difference in responsiveness: how much more likely decision-makers are to assign treatment when there is vs. isn't an algorithmic recommendation to do so. The last figure plots a logistic regression of the lift in responsiveness on the causal effect $\tau(x, a) = \mu_1(x, a) - \mu_0(x, a)$. We observe disparities in how responsive decision-makers are conditional on the same treatment effect efficacy. This is importantly not a claim of animus because decision-makers didn't have access to causal effect estimates. Nonetheless, disparities persist.

In Figure 4 we highlight results from constrained policy optimization. The first two plots in each set illustrate the objective function value and $A = a$ average treatment cost, respectively; for A being

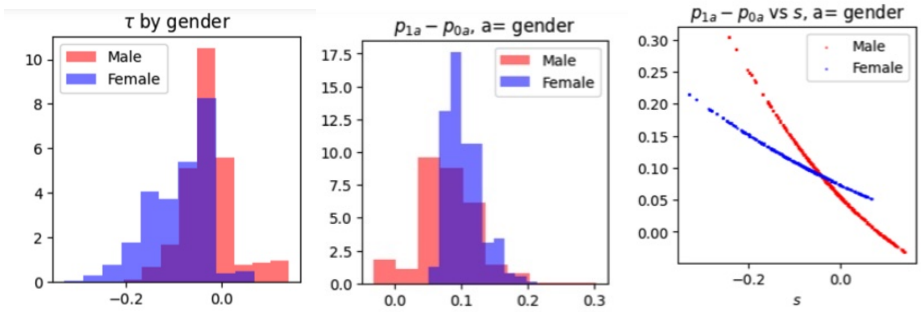

Figure 3: Distribution of treatment effect by gender, lift in treatment probabilities $p_{11a} - p_{01a} = P(T = 1 \mid R = 1, A = a, X) - P(T = 1 \mid R = 0, A = a, X)$, and plot of $p_{11a} - p_{01a}$ vs. τ.

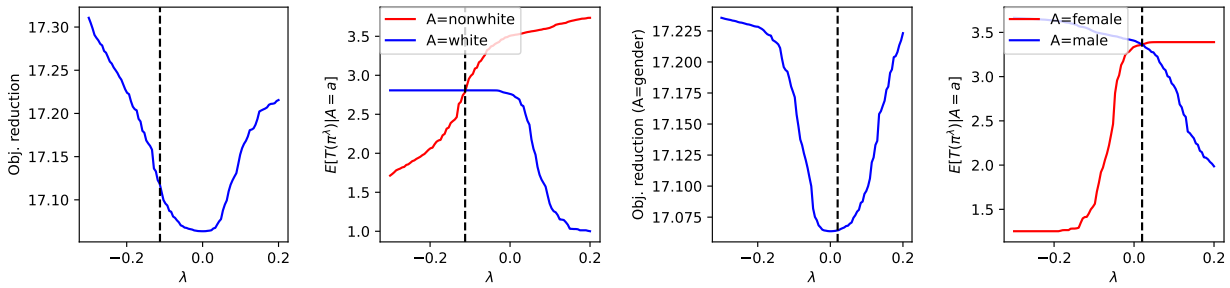

Figure 4: Policy value $V(\pi^\lambda)$, treatment value $\mathbb{E}[T(\pi^\lambda) \mid A = a]$, for A = race, gender.

race (nonwhite/white) or gender (female/male), respectively. We use costs of 100 for $Y = 1$ (failure to appear, 0 for $Y = 0$, and 20 when $T = 1$ (set arbitrarily). On the x-axis we plot the penalty λ that we use to assess the solutions of Proposition 9. The vertical dashed line indicates the solution achieving $\epsilon = 0$, i.e. parity in treatment take-up. Near-optimal policies that reduce treatment disparity can be of interest due to advocacy concerns about how the expansion of supervised release could increase the surveillance of already surveillance-burdened marginalized populations. We see that indeed, for race, surveillance-parity constrained policies can substantially reduce disparities for nonwhites while not increasing surveillance on whites that much: the red line decreases significantly at low increase to the blue line (and low increases to the objective value). On the other hand, for gender, the opportunity for improvement in surveillance disparity is much smaller. See the appendix for further experiments and computational details.

References

Public safety assessment decision making framework - cook county, il [effective march 2016]. https://news.wttw.com/sites/default/files/article/file-attachments/PSA%20Decision%20Making%20Framework.pdf, 2016.

Dec 2022. URL https://www.whitehouse.gov/wp-content/uploads/2022/12/BurdenReductionStrategies.pdf.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and

simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

F. Akinnibi and S. Holder. America is the world leader in locking people up. one city found a fix. `https://www.bloomberg.com/news/features/2023-08-30/nyc-s-cash-bail-reform-program-is-working-but-caseworkers-need-help`, 2023. [Accessed 08-09-2023].

D. Arnold, W. Dobbie, and C. S. Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932, 2018.

D. Arnold, W. Dobbie, and P. Hull. Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9):2992–3038, 2022.

S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 2017.

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

M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, and S. Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

H. Bastani, O. Bastani, and W. P. Sinchaisri. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*, 2021.

E. Ben-Michael, D. J. Greiner, K. Imai, and Z. Jiang. Safe policy learning through extrapolation: Application to pre-trial risk assessment. *arXiv preprint arXiv:2109.11679*, 2021.

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138, 2009.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

V. Chernozhukov, M. Demirer, G. Lewis, and V. Syrgkanis. Semi-parametric efficient policy learning with continuous actions. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Chohlas-Wood, M. Coots, H. Zhu, E. Brunskill, and S. Goel. Learning to be fair: A consequentialist approach to equitable decision-making. *arXiv preprint arXiv:2109.08792*, 2021.

J. Christensen, L. Aarøe, M. Baekgaard, P. Herd, and D. P. Moynihan. Human capital and administrative burden: The role of cognitive resources in citizen-state interactions. *Public Administration Review*, 80(1):127–136, 2020.

A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 582–593, 2020.

M. De-Arteaga, R. Fogliato, and A. Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

J. L. Doleac and M. T. Stevenson. Algorithmic risk assessments in the hands of humans. *Salem Center*, 2020.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106, 2012.

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

B. Green and Y. Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 90–99, 2019.

B. Green and Y. Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW2):1–33, 2021.

T. Gross. Letter Regarding Electronic Monitoring in Illinois — Community Renewal Society — communityrenewalsociety.org. `https://www.communityrenewalsociety.org/blog/letter-regarding-electronic-monitoring-in-illinois`. [Accessed 08-09-2023].

K. Heard, E. O'Toole, R. Naimpally, and L. Bressler. Real world challenges to randomization and their solutions. *Boston, MA: Abdul Latif Jameel Poverty Action Lab*, 2017.

P. Herd and D. P. Moynihan. *Administrative burden: Policymaking by other means*. Russell Sage Foundation, 2019.

M. A. Hernán and J. M. Robins. Causal inference.

K. Imai, Z. Jiang, J. Greiner, R. Halen, and S. Shin. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *arXiv preprint arXiv:2012.02845*, 2020.

Z. Jiang, S. Yang, and P. Ding. Multiply robust estimation of causal effects under principal ignorability. *arXiv preprint arXiv:2012.01615*, 2020.

N. Kallus and A. Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, pages 9269–9279, 2018.

N. Kallus and A. Zhou. Assessing disparate impact of personalized interventions: identifiability and bounds. *Advances in neural information processing systems*, 32, 2019.

N. Kallus and A. Zhou. Fairness, welfare, and equity in personalized pricing. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 296–314, 2021a.

N. Kallus and A. Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021b.

N. Kallus, X. Mao, and A. Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019a.

N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2281–2290, 2019b.

K. Kim, E. Kennedy, and J. Zubizarreta. Doubly robust counterfactual classification. *Advances in Neural Information Processing Systems*, 35:34831–34845, 2022.

T. Kitagawa and A. Tetenov. Empirical welfare maximization. 2015.

W. Lin, S.-H. Kim, and J. Tong. Does algorithm aversion exist in the field? an empirical analysis of algorithm use determinants in diabetes self-management. *An Empirical Analysis of Algorithm Use Determinants in Diabetes Self-Management (July 23, 2021). USC Marshall School of Business Research Paper Sponsored by iORB, No. Forthcoming*, 2021.

M. Lipsky. *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation, 2010.

L. Liu, Z. Shahn, J. M. Robins, and A. Rotnitzky. Efficient estimation of optimal regimes under a no direct effect assumption. *Journal of the American Statistical Association*, 116(533):224–239, 2021.

J. Ludwig and S. Mullainathan. Fragile algorithms and fallible decision-makers: lessons from the justice system. *Journal of Economic Perspectives*, 35(4):71–96, 2021.

K. Lum, E. Ma, and M. Baiocchi. The causal impact of bail on case outcomes for indigent defendants in new york city. *Observational Studies*, 3(1):38–64, 2017.

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

A. Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

B. Metevier, S. Giguere, S. Brockman, A. Kobren, Y. Brun, E. Brunskill, and P. S. Thomas. Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems*, 32, 2019.

A. Mishler, E. H. Kennedy, and A. Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

Office of the Chief Judge. Bail reform in cook county: An examination of general order 18.8a and bail in felony cases. 2019a.

Office of the Chief Judge. Bail reform. 2019b. URL `https://www.cookcountycourt.org/HOME/Bail-Reform`.

H. Qiu, M. Carone, E. Sadikova, M. Petukhova, R. C. Kessler, and A. Luedtke. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191, 2021.

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.

Safety and C. f. C. I. Justice Challenge. Expanding supervised release in new york city. 2022. URL `https://safetyandjusticechallenge.org/resources/expanding-supervised-release-in-new-york-city/`.

A. Shapiro. On duality theory of conic linear problems. *Semi-Infinite Programming: Recent Advances*, pages 135–165, 2001.

J. Steinhardt and P. Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International conference on machine learning*, pages 1593–1601. PMLR, 2014.

H. Sun, E. Munro, G. Kalashnov, S. Du, and S. Wager. Treatment allocation under uncertain costs. *arXiv preprint arXiv:2103.11066*, 2021.

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

U.S. Commission on Civil Rights. A new paradigm for welfare reform: The need for civil rights enforcement. 2002.

A. W. Van Der Vaart, J. A. Wellner, A. W. van der Vaart, and J. A. Wellner. *Weak convergence*. Springer, 1996.

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

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

Y. Yacoby, B. Green, C. L. Griffin Jr, and F. Doshi-Velez. "if it didn't happen, why would i change my decision?": How judges respond to counterfactual explanations for the public safety assessment. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 219–230, 2022.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

A Additional discussion

A.1 Additional related work

Principal stratification and mediation analysis in causal inference [Liu et al., 2021] studies an optimal test-and-treat regime under a no-direct-effect assumption, that assigning a diagnostic test has no effect on outcomes except via propensity to treat, and studies semiparametric efficiency using Structural Nested-Mean Models. Though our exclusion restriction is also a no-direct-effect assumption, our optimal treatment regime is in the space of recommendations only as we do not have control over the final decision-maker, and we consider generally nonparametric models.

We briefly go into more detail about formal differences, due to our specific assumptions, that delineate the differences to mediation analysis. Namely, our conditional exclusion restriction implies that $Y_{1T_0} = Y_{T_0}$ and that $Y_{0T_1} = Y_{1T_1}$ (in mediation notation with $T_r = T(r)$ in our notation), so that so-called *net direct effects* are identically zero and the *net indirect effect* is the treatment effect (also called average encouragement effect here).

Human-in-the-loop in consequential domains. There is a great deal of interest in designing algorithms for the "human in the loop" and studying expertise and discretion in human oversight in consequential domains [De-Arteaga et al., 2020]. On the algorithmic side, recent work focuses on frameworks for learning to defer or human-algorithm collaboration. Our focus is *prior* to the design of these procedures for improved human-algorithm collaboration: we primarily hold fixed current human responsiveness to algorithmic recommendations. Therefore, our method can be helpful for optimizing local nudges. Incorporating these algorithmic design ideas would be interesting directions for future work.

Empirical literature on judicial discretion in the pretrial setting. Studying a slightly different substantive question, namely causal effects of pretrial decisions on later outcomes, a line of work uses individual judge decision-makers as a leniency instrumental variable for the treatment effect of (for example, EM) on pretrial outcomes [Arnold et al., 2022, 2018, Lum et al., 2017]. And, judge IVs rely on quasi-random assignment of individual judges. We focus on the prescriptive question of optimal recommendation rules in view of patterns of judicial discretion, rather than the descriptive question of causal impacts of detention on downstream outcomes.

A number of works have emphasized the role of judicial discretion in pretrial risk assessments in particular [Green and Chen, 2021, Doleac and Stevenson, 2020, Ludwig and Mullainathan, 2021]. In contrast to these works, we focus on studying decisions about electronic monitoring, which is an intermediate degree of decision lever to prevent FTA that nonetheless imposes costs. [Imai et al., 2020] study a randomized experiment of provision of the PSA and estimate (the sign of) principal causal effects, including potential group-conditional disparities. They are interested in a causal effect on the principal stratum of those marginal defendants who would not commit a new crime if recommended for detention. [Ben-Michael et al., 2021] study policy learning in the absence of positivity (since the PSA is a deterministic function of covariates) and consider a case study on determining optimal recommendation/detention decisions; however their observed outcomes are downstream of judicial decision-making. Relative to their approach, we handle lack of overlap via an exclusion restriction so that we only require ambiguity on *treatment responsivity models* rather than causal outcome models.

B Additional discussion on method

B.1 Additional discussion on constrained optimization

Feasibility program We can obtain upper/lower bounds on ϵ in order to obtain a feasible region for ϵ by solving the below optimization over maximal/minimal values of the constraint:

$$\overline{\epsilon}, \underline{\epsilon} \in \max_{\pi} / \min_{\pi} \mathbb{E}[T(\pi) \mid A = a] - \mathbb{E}[T(\pi) \mid A = b] \tag{3}$$

$$V_{\epsilon}^{*} = \max_{\pi} \{\mathbb{E}[c(\pi, T(\pi), Y(\pi))] \colon \mathbb{E}[T(\pi) \mid A = a] - \mathbb{E}[T(\pi) \mid A = b] \leq \epsilon\} \tag{4}$$

B.2 Additional discussion on Algorithm 2 (general algorithm)

B.2.1 Additional fairness constraints and examples in this framework

In this section we discuss additional fairness constraints and how to formulate them in the generic framework. Much of this discussion is quite similar to [Agarwal et al., 2018] (including in notation) and is included in this appendix for completeness only. We only additionally provide novel identification results for another fairness measure on causal policies in Appendix B.2.2, concrete discussion of the reduction to weighted classification, and provide concrete descriptions of the causal fairness constraints in the more general framework.

We first discuss how to impose the treatment parity constraint. This is similar to the demographic parity example in Agarwal et al. [2018], with different coefficients, but included for completeness. (Instead, recommendation parity in $\mathbb{E}[\pi \mid A = a]$ is indeed nearly identical to demographic parity.)

Example 1 (Writing treatment parity in the general constrained classification framework.)**.** We write the constraint

$$\mathbb{E}[T(\pi) \mid A = a] - \mathbb{E}[T(\pi) \mid A = b] \tag{5}$$

in this framework as follows:

$$\mathbb{E}[T(\pi) \mid A = a] = \mathbb{E}[\pi_1(X)(p_{1|1}(X, A) - p_{1|0}(X, A)) + p_{1|0}(X, A) \mid A = a]$$

For each $u \in \mathcal{A}$ we enforce that

$$\textstyle\sum_{r \in \{0,1\}} \mathbb{E}\left[\pi_r(X) p_{1|r}(X, A) \mid A = u\right] = \sum_{r \in \{0,1\}} \mathbb{E}\left[\pi_r(X, A) p_{1|r}(X, A)\right]$$

We can write this in the generic notation given previously by letting $\mathcal{J} = \mathcal{A} \cup \{\circ\}$,

$$g_j(O, \pi(X); \eta) = \pi_1(X)(p_{1|1}(X, A) - p_{1|0}(X, A)) + p_{1|0}(X, A), \forall j.$$

We let the conditioning events $\mathcal{E}_a = \{A = a\}, \mathcal{E}_\circ = \{\text{True}\}$, i.e. conditioning on the latter is equivalent to evaluating the marginal expectation. Then we express Equation (5) as a set of equality constraints $h_a(\pi) = h_\circ(\pi)$, leading to pairs of inequality constraints,

$$\begin{Bmatrix} h_u(\pi) - h_\circ(\pi) \leq 0 \\ h_\circ(\pi) - h_u(\pi) \leq 0 \end{Bmatrix}_{u \in \mathcal{A}}$$

The corresponding coefficients of M over this enumeration over groups ($\mathcal{A}$) and epigraphical enforcement of equality ($\{+,-\}$) equation (1), gives $\mathcal{K} = \mathcal{A}\times\{+,-\}$ so that $M_{(a,+),a'} = \mathbf{1}\{a'=a\}$, $M_{(a,+),\star} = -1$, $M_{(a,-),a'} = -\mathbf{1}\{a'=a\}$, $M_{(a,-),\star} = 1$, and $\mathbf{d} = \mathbf{0}$. Further we can relax equality to small amounts of constraint relaxation by instead setting $d_k > 0$ for some (or all) k.

Next, we discuss a more complicated fairness measure. We first discuss identification and estimation before we also describe how to incorporate it in the generic framework.

B.2.2 Responder-dependent fairness measures

We consider a responder framework on outcomes (under our conditional exclusion restriction). Because the contribution to the AEE is indeed from the responder strata, this corresponds to additional estimation of the responder stratum.

We enumerate the four possible realizations of potential outcomes (given any fixed recommendation) as $(Y(0(r)), Y(1(r)) \in \{0,1\}^2$. We call units with $(Y(0(r)), Y(1(r))) = (0,1)$ responders, $(Y(0(r)), Y(1(r))) = (1,0)$ anti-responders, and $Y(0(r)) = Y(1(r))$ non-responders. Such a decomposition is general for the binary setting.

Assumption 9 (Binary outcomes, treatment)**.**

$$T, Y \in \{0,1\}$$

Assumption 10 (Monotonicity)**.**

$$Y(T(1)) \geq Y(T(0))$$

Importantly, the conditional exclusion restriction of Assumption 2 implies that responder status is independent of recommendation. Conditional on observables, whether a particular individual is a responder is independent of whether someone decides to treat them when recommended. In this way, we study responder status analogous to its use elsewhere in disparity assessment in algorithmic fairness [Imai et al., 2020, Kallus et al., 2019a]. Importantly, this assumption implies that the conditioning event (of being a responder) is therefore independent of the policy π, so it can be handled in the same framework. s

We may consider reducing disparities in resource expenditure given responder status.

We may be interested in the probability of receiving treatment assignment given responder status.

Example 2 (Fair treatment expenditure given responder status)**.**

$$\mathbb{E}[T(\pi) \mid Y(1(R)) > Y(0(R)), A = a] - \mathbb{E}[T(\pi) \mid Y(1(R)) > Y(0(R)), A = b] \leq \epsilon$$

We can obtain identification via regression adjustment:

Proposition 7 (Identification of treatment expenditure given responder status)**.** Assume Assumptions 9 and 10.

$$P(T(\pi) = 1 \mid A = a, Y(1(\pi)) > Y(0(\pi))) = \frac{\sum_r \mathbb{E}[\pi_r(X) p_{1|r}(X,A)(\mu_1(X,A) - \mu_0(X,A)) \mid A = a])}{\mathbb{E}[(\mu_1(X,A) - \mu_0(X,A)) \mid A = a]}$$

Therefore this can be expressed in the general framework.

Example 3 (Writing treatment responder-conditional parity in the general constrained classification framework.)**.** For each $u \in \mathcal{A}$ we enforce that

$$\frac{\sum_r \mathbb{E}[\pi_r(X) p_{1|r}(X,A)(\mu_1(X,A)-\mu_0(X,A))|A=u])}{\mathbb{E}[(\mu_1(X,A)-\mu_0(X,A))|A=u]} = \frac{\sum_r \mathbb{E}[\pi_r(X) p_{1|r}(X,A)(\mu_1(X,A)-\mu_0(X,A))])}{\mathbb{E}[(\mu_1(X,A)-\mu_0(X,A))]}$$

We can write this in the generic notation given previously by letting $\mathcal{J} = \mathcal{A} \cup \{\circ\}$,

$$g_j(O, \pi(X); \eta) = \frac{\{\pi_1(X)(p_{1|1}(X,A) - p_{1|0}(X,A)) + p_{1|0}(X,A)\}(\mu_1(X,A) - \mu_0(X,A))}{\mathbb{E}[(\mu_1(X,A) - \mu_0(X,A)) \mid A = a]}, \forall j.$$

Let $\mathcal{E}_a^j = \{A = a_j\}, \mathcal{E}_\circ = \{\text{True}\}$, and we express Equation (5) as a set of equality constraints of the above moment $h_a(\pi) = h_\circ(\pi)$, leading to pairs of inequality constraints,

$$\begin{Bmatrix} h_u(\pi) - h_\circ(\pi) \leq 0 \\ h_\circ(\pi) - h_u(\pi) \leq 0 \end{Bmatrix}_{u \in \mathcal{A}}$$

The corresponding coefficients of M proceed analogously as for treatment parity.

B.2.3 Best-response oracles

Best-responding classifier π, given λ: $\text{BEST}_\pi(\lambda)$ The best-response oracle, given a particular λ value, optimizes the Lagrangian given π:

$$\begin{aligned} L(\pi, \lambda) &= \hat{V}(\pi) + \lambda^\top (M\hat{h}(\pi) - \hat{d}) \\ &= \hat{V}(\pi) - \lambda^\top \hat{d} + \sum_{k,j} \frac{M_{k,j}\lambda_k}{p_j} \mathbb{E}_n \left[g_j(O, \pi) 1\left\{O \in \mathcal{E}_j\right\} \right]. \end{aligned}$$

Best-responding Lagrange multiplier λ, given π: $\text{BEST}_{\boldsymbol{\lambda}}(Q)$ is the best response of the Λ player. It can be chosen to be either 0 or put all the mass on the most violated constraint. Let $\gamma(Q) := Mh(Q)$ denote the constraint values, then $\text{BEST}_{\boldsymbol{\lambda}}(Q)$ returns $\begin{cases} \mathbf{0} & \text{if } \widehat{\gamma}(Q) \leq \widehat{\mathbf{c}} \\ B\mathbf{e}_{k^*} & \text{otherwise, where } k^* = \arg\max_k \left[\widehat{\gamma}_k(Q) - \widehat{c}_k\right] \end{cases}$

B.2.4 Weighted classification reduction

There is a well-known reduction of optimizing the zero-one loss for policy learning to weighted classification. A cost-sensitive classification problem is

$$\arg\min_{\pi_1} \sum_{i=1}^n \pi_1(X_i) C_i^1 + (1 - \pi_1(X_i)) C_i^0$$

The weighted classification error is $\sum_{i=1}^n W_i 1\left\{h(X_i) \neq Y_i\right\}$ which is an equivalent formulation if $W_i = \left|C_i^0 - C_i^1\right|$ and $Y_i = 1\left\{C_i^0 \geq C_i^1\right\}$.

The reduction to weighted classification is particularly helpful since taking the Lagrangian will introduce datapoint-dependent penalties that can be interpreted as additional weights. We can consider the centered regret $J(\pi) = \mathbb{E}[Y(\pi)] - \frac{1}{2}\mathbb{E}[\mathbb{E}[Y \mid R = 1, X] + \mathbb{E}[Y \mid R = 0, X]]$. Then

$$J(\theta) = J(\text{sgn}(g_\theta(\cdot))) = \mathbb{E}[\text{sgn}(g_\theta(X))\{\psi\}]$$

where ψ can be one of, where $\mu_r^R(X) = \mathbb{E}[Y \mid R = r, X]$,

$$\psi_{DM} = (p_{1|1}(X) - p_{1|0}(X))(\mu_1(X) - \mu_0(X)), \psi_{IPW} = \frac{RY}{e_R(X)}, \psi_{DR} = \psi_{DM} + \psi_{IPW} + \frac{R\mu^R(X)}{e_R(X)}$$

We can apply the standard reduction to cost-sensitive classification since $\psi_i \operatorname{sgn}(g_\theta(X_i)) = |\psi_i| (1 - 2\mathbb{I}[\operatorname{sgn}(g_\theta(X_i)) \neq \operatorname{sgn}(\psi_i)])$. Then we can use surrogate losses for the zero-one loss,

$$L(\theta) = \mathbb{E}[|\psi| \, \ell(g_\theta(X), \operatorname{sgn}(\psi))]$$

Although many functional forms for $\ell(\cdot)$ are Fisher-consistent, the logistic (cross-entropy) loss will be particularly relevant: $l(g, s) = 2\log(1 + \exp(g)) - (s + 1)g$.

Example 4 (Treatment parity, continued (weighted classification reduction))**.** The cost-sensitive reduction for a vector of Lagrange multipliers can be deduced by applying the weighted classification reduction to the Lagrangian:

$$L(\beta) = \mathbb{E}\left[|\tilde{\psi}^\lambda| \ell\left(f_\beta(X), \operatorname{sgn}(\tilde{\psi}^\lambda)\right)\right], \qquad \text{where } \tilde{\psi}^\lambda = \psi + \frac{\lambda_A}{p_A}(p_{1|1} - p_{1|0}) - \sum_{a \in \mathcal{A}} \lambda_a.$$

where $p_a := \hat{P}(A = a)$ and $\lambda_a := \lambda_{(a,+)} - \lambda_{(a,-)}$, effectively replacing two non-negative Lagrange multipliers by a single multiplier, which can be either positive or negative.

Example 5 (Responder-conditional treatment parity, continued)**.** The Lagrangian is $L(\beta) = \mathbb{E}\left[|\tilde{\psi}^\lambda| \ell\left(f_\beta(X), \operatorname{sgn}(\tilde{\psi}^\lambda)\right)\right]$ with weights:

$$\tilde{\psi}^\lambda = \psi + \frac{\lambda_A}{p_A} \frac{(p_{1|1} - p_{1|0})(\mu_1 - \mu_0)}{\mathbb{E}_n[(\mu_1(X, A) - \mu_0(X, A)) \mid A = a]} - \sum_{a \in \mathcal{A}} \lambda_a.$$

where $p_a := \hat{P}(A = a)$ and $\lambda_a := \lambda_{(a,+)} - \lambda_{(a,-)}$.

B.3 Proofs

Proof of Proposition 7.

$$\begin{aligned}
&P(T(\pi) = 1 \mid A = a, Y(1(\pi)) > Y(0(\pi))) \\
&= \frac{P(T(\pi) = 1, Y(1(r)) > Y(0(r)) \mid A = a)}{P(Y(1(\pi)) > Y(0(\pi)) \mid A = a)} && \text{by Bayes' rule} \\
&= \frac{P(T(\pi) = 1, Y(1) > Y(0) \mid A = a)}{P(Y(1) > Y(0) \mid A = a)} && \text{by Assumption 2} \\
&= \frac{\sum_r \mathbb{E}[\mathbb{E}[\pi_r(X)\mathbb{I}[T(r) = 1]\,\mathbb{I}[Y(1) > Y(0)] \mid A = a, X]])}{P(Y(1) > Y(0) \mid A = a)} && \text{by iter. exp} \\
&\frac{\sum_r \mathbb{E}[\pi_r(X) p_{1|r}(X, A)(\mu_1(X, A) - \mu_0(X, A)) \mid A = a])}{\mathbb{E}[(\mu_1(X, A) - \mu_0(X, A)) \mid A = a]} && \text{by Proposition 1}
\end{aligned}$$

□

C Proofs

C.1 Proofs for generalization under unconstrained policies

Proposition 8 (Policy value generalization)**.** Assume the nuisance models $\eta = [p_{1|0}, p_{1|1}, \mu_1, \mu_0, e_r(X)]^\top, \eta \in \mathcal{F}_\eta$ are consistent and well-specified with finite VC-dimension v_η over the product function class H. We provide a proof for the general case, including doubly-robust estimators, which applies to the statement of Proposition 8 by taking $\eta = [p_{1|0}, p_{1|1}, \mu_1, \mu_0]$.

Let $\Pi = \{\mathbb{I}\{\mathbb{E}[L(\lambda, X, A; \eta) \mid X] > 0 \colon \lambda \in \mathbb{R}; \eta \in \mathcal{F}_\eta\}$.

$$\sup_{\pi \in \Pi, \lambda \in \mathbb{R}} |(\mathbb{E}_n[\pi L(\lambda, X, A)] - \mathbb{E}[\pi L(\lambda, X, A)])| = O_p(n^{-\frac{1}{2}})$$

The generalization bound allows deducing risk bounds on the out-of-sample value:

Corollary 2.

$$\mathbb{E}[L(\hat{\lambda}, X, A)_+] \leq \mathbb{E}[L(\lambda^*, X, A)_+] + O_p(n^{-\frac{1}{2}})$$

Proof of Proposition 8. We study a general Lagrangian, which takes as input pseudo-outcomes $\psi^{t|r}(O; \eta), \psi^{y|t}(O; \eta), \psi^{1|0,\Delta A}$ where each satisfies that

$$\begin{aligned}
\mathbb{E}[\psi^{t|r}(O; \eta) \mid X, A] &= p_{1|1}(X, A) - p_{1|0}(X, A) \\
\mathbb{E}[\psi^{y|t}(O; \eta) \mid X, A] &= \tau(X, A) \\
\mathbb{E}[\psi^{1|0,\Delta A} \mid X] &= p_{1|0}(X, a) - p_{1|0}(X, b)
\end{aligned}$$

We make high-level stability assumptions on pseudooutcomes ψ relative to the nuisance functions η (these are satisfied by standard estimators that we will consider):

Assumption 11. $\psi^{t|r}, \psi^{y|t}, \psi^{1|0,\Delta A}$ respectively are Lipschitz contractions with respect to η and bounded

We study a generalized Lagrangian of an optimization problem that took these pseudooutcome estimates as inputs:

$$L(\lambda, X, A; \eta) = \psi_{t|r}(O; \eta) \left\{ \psi_{y|t}(O; \eta) + \frac{\lambda}{p(A)} (\mathbb{I}[A = a] - \mathbb{I}[A = b]) \right\} + \lambda(\psi^{1|0,\Delta A}(O; \eta))$$

We will show that

$$\sup_{\pi \in \Pi, \lambda \in \mathbb{R}} |(\mathbb{E}_n[\pi L(\lambda, X, A)] - \mathbb{E}[\pi L(\lambda, X, A)])| = O_p(n^{-\frac{1}{2}})$$

which, by applying the generalization bound twice gives that

$$\mathbb{E}_n[\pi L(\lambda, X, A)] = \mathbb{E}[\pi L(\lambda, X, A)]) + O_p(n^{-\frac{1}{2}})$$

Write Lagrangian as

$$\max_\pi \min_\lambda = \min_\lambda \max_\pi = \min_\lambda \mathbb{E}[L(O, \lambda; \eta)_+]$$

Suppose the Rademacher complexity of η_k is given by $\mathcal{R}(H_k)$, so that [Bartlett and Mendelson, 2002, Thm. 12] gives that the Rademacher complexity of the product nuisance class H is therefore

$\sum_k \mathcal{R}(H_k)$. The main result follows by applying vector-valued extensions of Lipschitz contraction of Rademacher complexity given in Maurer [2016]. Suppose that $\psi^{t|r}, \psi^{y|t}, \psi^{1|0,\Delta A}$ are Lipschitz with constants $C^L_{t|r}, C^L_{y|t}, C^L_{1|0,\Delta A}$.

We establish VC-properties of

$$\mathcal{F}_{L_1}(O_{1:n}) = \{(g_\eta(O_1), g_\eta(O_i), \dots g_\eta(O_n))\colon \eta \in \mathcal{F}_\eta\}\,, \text{ where } g_\eta(O) = \psi_{t|r}(O;\eta)\psi_{y|t}(O;\eta)$$

$$\mathcal{F}_{L_2}(O_{1:n}) = \{(h_\eta(O_1), h_\eta(O_i), \dots h_\eta(O_n))\colon \eta \in \mathcal{F}_\eta\}\,, \text{ where } h_\eta(O) = \psi_{t|r}(O;\eta)\frac{\lambda}{p(A)}(\mathbb{I}\left[A=a\right] - \mathbb{I}\left[A=b\right])$$

$$\mathcal{F}_{L_3}(O_{1:n}) = \{(m_\eta(O_1), m_\eta(O_i), \dots m_\eta(O_n))\colon \eta \in \mathcal{F}_\eta\}\,, \text{ where } m_\eta(O) = \lambda(\psi^{1|0,\Delta A}(O;\eta))$$

and the function class for the truncated Lagrangian,

$$\mathcal{F}_{L_+} = \{\{(g_\eta(O_i) + h_\eta(O_i) + m_\eta(O_i))_+\}_{1:n}\colon g \in \mathcal{F}_{L_1}(O_{1:n}), h \in \mathcal{F}_{L_2}(O_{1:n}), m \in \mathcal{F}_{L_3}(O_{1:n}), \eta \in \mathcal{F}_\eta\}$$

[Maurer, 2016, Corollary 4] (and discussion of product function classes) gives the following: Let $\mathcal{X}$ be any set, $(x_1, \dots, x_n) \in \mathcal{X}^n$, let F be a class of functions $f : \mathcal{X} \to \ell_2$ and let $h_i : \ell_2 \to \mathbb{R}$ have Lipschitz norm L. Then

$$\mathbb{E}\sup_{\eta\in H}\sum_i \epsilon_i\psi_i\left(\eta\left(O_i\right)\right) \le \sqrt{2}L\mathbb{E}\sup_{\eta\in H}\sum_{i,k}\epsilon_{ik}\eta\left(O_i\right) \le \sqrt{2}L\sum_k \mathbb{E}\sup_{\eta_k\in H_k}\sum_i \epsilon_i\eta_k\left(O_i\right) \tag{6}$$

where ϵ_{ik} is an independent doubly indexed Rademacher sequence and $f_k\left(x_i\right)$ is the k-th component of $f\left(x_i\right)$.

Applying Equation (6) to each of the component classes $\mathcal{F}_{L_1}(O_{1:n}), \mathcal{F}_{L_2}(O_{1:n}), \mathcal{F}_{L_3}(O_{1:n})$, and Lipschitz contraction [Bartlett and Mendelson, 2002, Thm. 12.4] of the positive part function $\mathcal{F}_{L_+}$, we obtain the bound

$$\sup_{\lambda,\eta}|\mathbb{E}_n[L(O,\lambda;\eta)_+] - \mathbb{E}[L(O,\lambda;\eta)_+]| \le \sqrt{2}(C^L_{t|r}C^L_{y|t} + C^L_{t|r}B_{p_a}B + BC^L_{1|0,\Delta A})\sum_k \mathcal{R}(H_k)$$

□

Proposition 9 (Threshold solutions)**.** Define

$$L(\lambda, X, A) = (p_{1|1}(X,A) - p_{1|0}(X,A))\left\{\tau(X,A) + \frac{\lambda}{p(A)}(\mathbb{I}\left[A=a\right] - \mathbb{I}\left[A=b\right])\right\} + \lambda(p_{1|0}(X,a) - p_{1|0}(X,b))$$

$$\lambda^* \in \arg\min_\lambda\ \mathbb{E}[L(\lambda,X,A)_+],\ \ \pi^*(x,u) = \mathbb{I}\{L(\lambda^*,X,u) > 0\}$$

If instead $d(x)$ is a function of covariates x only,

$$\lambda^* \in \arg\min_\lambda\ \mathbb{E}[\mathbb{E}[L(\lambda,X,A)\mid X]_+],\ \ \pi^*(x) = \mathbb{I}\{\mathbb{E}[L(\lambda^*,X,A)\mid X] > 0\}$$

Proof of Proposition 9. The characterization follows by strong duality in infinite-dimensional linear programming [Shapiro, 2001]. Strict feasibility can be satisfied by, e.g. solving eq. (3) to set ranges for ϵ. □

C.2 Proofs for robust characterization

Proof of Proposition 5.

$$
\begin{aligned}
V(\pi) &= \sum_{t\in\mathcal{T},r\in\{0,1\}} \mathbb{E}[\pi_r(X)\mathbb{E}[c_{rt}(Y(t))\mathbb{I}\left[T(r)=t\right] \mid R=r,X]] \\
&= \sum_{t\in\mathcal{T},r\in\{0,1\}} \mathbb{E}[\pi_r(X)\mathbb{E}[c_{rt}(Y(t)) \mid R=r,X]P(T(r)=t \mid R=r,X)] && \text{unconf.} \\
&= \sum_{t\in\mathcal{T},r\in\{0,1\}} \mathbb{E}[\pi_r(X)\mathbb{E}[c_{rt}(Y(t)) \mid X]P(T(r)=t \mid R=r,X)] && \text{Assumption 2 (ER)} \\
& && (7) \\
&= \sum_{t\in\mathcal{T},r\in\{0,1\}} \mathbb{E}\left[\pi_r(X)\mathbb{E}\left[c_{rt}(Y(t))\frac{\mathbb{I}\left[T(r)=t\right]}{p_t(X)} \mid X\right]P(T(r)=t \mid R=r,X)\right] && \text{unconf.} \\
&= \sum_{t\in\mathcal{T},r\in\{0,1\}} \mathbb{E}\left[\pi_r(X)\left\{\mathbb{E}\left[c_{rt}(Y(t))\frac{\mathbb{I}\left[T(r)=t\right]}{p_t(X)}+\left(1-\frac{T}{p_t(X)}\right)\mu_t(X) \mid X\right]p_{t|r}(X)\right\}\right] && \text{control variate} \\
&= \sum_{t\in\mathcal{T},r\in\{0,1\}} \mathbb{E}\left[\pi_r(X)\left\{\left\{c_{rt}(Y(t))\frac{\mathbb{I}\left[T(r)=t\right]}{p_t(X)}+\left(1-\frac{T}{p_t(X)}\right)\mu_t(X)\right\}p_{t|r}(X)\right\}\right] && \text{(LOTE)}
\end{aligned}
$$

where $p_t(X) = P(T=t \mid X)$ (marginally over R in the observational data) and (LOTE) is an abbreviation for the law of total expectation. □

Proof of Lemma 1.

$$
\begin{aligned}
\overline{V}_{no}(\pi) &\coloneqq \max_{q_{tr}(X)\in\mathcal{U}}\left\{\sum_{t\in\mathcal{T},r\in\{0,1\}}\mathbb{E}[\pi_r(X)\mu_t(X)q_{tr}(X)\mathbb{I}\left[X\in\mathcal{X}^{\text{nov}}\right]]\right\} \\
&= \max_{q_{tr}(X)\in\mathcal{U}}\left\{\sum_{t\in\mathcal{T},r\in\{0,1\}}\mathbb{E}[\pi_r(X)\mathbb{E}[Y \mid T=t,X]q_{tr}(X)\mathbb{I}\left[X\in\mathcal{X}^{\text{nov}}\right]]\right\}
\end{aligned}
$$

Note the objective function can be reparametrized under a surjection of $q_{t|r}(X)$ to its marginalization, i.e. marginal expectation over a $\{T=t\}$ partition (equivalently $\{T=t,A=a\}$ partition for a fairness-constrained setting).

Define

$$\beta_{t|r}(a) \coloneqq \mathbb{E}[q_{t|r}(X,A) \mid T=t,A=a], \beta_{t|r} \coloneqq \mathbb{E}[q_{t|r}(X,A) \mid T=t]$$

Therefore we may reparametrize $\overline{V}_{no}(\pi)$ as an optimization over constant coefficients (bounded by B):

$$= \max \left\{ \sum_{t\in\mathcal{T}, r\in\{0,1\}} \mathbb{E}[\{c_t \beta_{t|r}\}\pi_r(X)\mathbb{E}[Y \mid T=t, X]\mathbb{I}\left[X \in \mathcal{X}^{\text{nov}}\right]]] \colon \underline{B} \le c_1 \le \overline{B}, c_0 = 1 - c_1 \right\}$$

$$= \max \left\{ \sum_{t\in\mathcal{T}, r\in\{0,1\}} \mathbb{E}[\{c_t \beta_{t|r}\}\mathbb{E}[Y\pi_r(X) \mid T=t]\mathbb{I}\left[X \in \mathcal{X}^{\text{nov}}\right]]] \colon \underline{B} \le c_1 \le \overline{B}, c_0 = 1 - c_1 \right\} \qquad \text{(LOTE)}$$

$$= \sum_{t\in\mathcal{T}, r\in\{0,1\}} \mathbb{E}[c^*_{rt} \beta_{t|r}\mathbb{E}[Y\pi_r(X) \mid T=t]\mathbb{I}\left[X \in \mathcal{X}^{\text{nov}}\right]]]$$

$$\text{where } c^*_{rt} = \begin{cases} \overline{B}\mathbb{I}\left[t=1\right] + \underline{B}\mathbb{I}\left[t=0\right] & \text{if } \mathbb{E}[Y\pi_r(X) \mid T=t] \ge 0 \\ \overline{B}\mathbb{I}\left[t=0\right] + \underline{B}\mathbb{I}\left[t=1\right] & \text{if } \mathbb{E}[Y\pi_r(X) \mid T=t] < 0 \end{cases}$$

□

Proof of proposition 6.

$$\max_{\pi} \mathbb{E}[c(\pi, T(\pi), Y(\pi))\mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\}] + \mathbb{E}[c(\pi, T(\pi), Y(\pi))\mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\}] \tag{8}$$

$$\mathbb{E}[T(\pi)\mathbb{I}\{X \in \mathcal{X}^{\text{ov}}\} \mid A=a] - \mathbb{E}[T(\pi)\mathbb{I}\{X \in \mathcal{X}^{\text{ov}}\} \mid A=b] \tag{9}$$

$$+ \mathbb{E}[T(\pi)\mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\} \mid A=a] - \mathbb{E}[T(\pi)\mathbb{I}\{X \in \mathcal{X}^{\text{nov}}\} \mid A=b] \le \epsilon, \forall q_{r1} \in \mathcal{U} \tag{10}$$

Define

$$g_r(x,u) = (\mu_{r1}(x,u) - \mu_{r0}(x,u))$$

then we can rewrite this further and apply the standard epigraph transformation:

$$\max t$$

$$t - \int_{x\in\mathcal{X}^{\text{nov}}} \sum_{u\in\{a,b\}} \sum_{r\in\{0,1\}} \{g_r(x,u)\pi_r(x,u)f(x,u)\}q_{r1}(x,u)\}dx \le V_{ov}(\pi) + \mathbb{E}[\mu_0], \forall q_{r1} \in \mathcal{U}$$

$$\int_{x\in\mathcal{X}^{\text{nov}}} \{f(x \mid a)(\sum_r \pi_r(x,a)q_{r1}(x,a)) - f(x \mid b)(\sum_r \pi_r(x,b)q_{r1}(x,b))\} + \mathbb{E}[\Delta_{ov}T(\pi)] \le \epsilon, \forall q_{r1} \in \mathcal{U}$$

Project the uncertainty set onto the direct product of uncertainty sets:

$$\max t$$

$$t - \int_{x\in\mathcal{X}^{\text{nov}}} \sum_{u\in\{a,b\}} \sum_{r\in\{0,1\}} \{g_r(x,u)\pi_r(x,u)f(x,u)\}q_{r1}(x,u)\}dx \le V_{ov}(\pi) + \mathbb{E}[\mu_0], \forall q_{r1} \in \mathcal{U}_\infty$$

$$\int_{x\in\mathcal{X}^{\text{nov}}} \{f(x \mid a)(\sum_r \pi_r(x,a)q_{r1}(x,a)) - f(x \mid b)(\sum_r \pi_r(x,b)q_{r1}(x,b))\} + \mathbb{E}[\Delta_{ov}T(\pi)] \le \epsilon, \forall q_{r1} \in \mathcal{U}_\in$$

Clearly robust feasibility of the resource parity constraint over the interval is obtained by the highest/lowest bounds for groups a, b, respectively:

$$\max t$$

$$t - \int_{x\in\mathcal{X}^{\text{nov}}} \sum_{u\in\{a,b\}} \sum_{r\in\{0,1\}} \{g_r(x,u)\pi_r(x,u)f(x,u)\}q_{r1}(x,u)\}dx \le V_{ov}(\pi) + \mathbb{E}[\mu_0], \forall q_{r1} \in \mathcal{U}_\infty$$

$$\int_{x\in\mathcal{X}^{\text{nov}}} \{f(x \mid a)(\sum_r \pi_r(x,a)\overline{B}_r(x,a)) - f(x \mid b)(\sum_r \pi_r(x,b)\underline{B}_r(x,u))\} + \mathbb{E}[\Delta_{ov}T(\pi)] \le \epsilon$$

We define

$$\delta_{r1}(x,u) = \frac{2(q_{r1}(x,u) - \underline{B}_r(x,u))}{\overline{B}_r(x,u) - \underline{B}_r(x,u)} - (\overline{B}_r(x,u) - \underline{B}_r(x,u)),$$

then

$$\{\underline{B}_r(x,u) \le q_{r1}(x,u) \le \overline{B}_r(x,u)\} \implies \{\|\delta_{r1}(x,u)\|_\infty \le 1\}$$

and

$$q_{r1}(x,u) = \underline{B}_r(x,u) + \frac{1}{2}(\overline{B}_r(x,u) - \underline{B}_r(x,u))(\delta_{r1}(x,u) + 1).$$

For brevity we denote $\Delta B = (\overline{B}_r(x,u) - \underline{B}_r(x,u))$, so

$$\max t$$

$$t + \max_{\substack{\|\delta_{r1}(x,u)\|_\infty \le 1 \\ r\in\{0,1\}, u\in\{a,b\}}} \left\{ - \int_{x\in\mathcal{X}^{\text{nov}}} \sum_{u\in\{a,b\}} \sum_{r\in\{0,1\}} \{g_r(x,u)\pi_r(x,u)f(x,u)\}\frac{1}{2}\Delta B(x,u)\delta_{r1}(x,u)dx \right\} - c_1(\pi) \le V_{ov}(\pi) +$$

$$\int_{x\in\mathcal{X}^{\text{nov}}} \{f(x \mid a)(\sum_r \pi_r(x,a)\overline{B}_r(x,a)) - f(x \mid b)(\sum_r \pi_r(x,b)\underline{B}_r(x,u))\} + \mathbb{E}[\Delta_{ov}T(\pi)] \le \epsilon,$$

where

$$c_1(\pi) = \int_{x\in\mathcal{X}^{\text{nov}}} \sum_{u\in\{a,b\}} \sum_{r\in\{0,1\}} \{g_r(x,u)\pi_r(x,u)f(x,u)\}(\underline{B}_r(x,u) + \frac{1}{2}(\overline{B}_r(x,u) - \underline{B}_r(x,u)))dx$$

This is equivalent to:

$$\max t$$

$$t + \int_{x\in\mathcal{X}^{\text{nov}}} \sum_{u\in\{a,b\}} \sum_{r\in\{0,1\}} |-g_r(x,u)\pi_r(x,u)f(x,u)| \frac{1}{2}\Delta B(x,u)dx - c_1(\pi) \le V_{ov}(\pi) + \mathbb{E}[\mu_0]$$

$$\int_{x\in\mathcal{X}^{\text{nov}}} \{f(x \mid a)(\sum_r \pi_r(x,a)\overline{B}_r(x,a)) - f(x \mid b)(\sum_r \pi_r(x,b)\underline{B}_r(x,u))\} + \mathbb{E}[\Delta_{ov}T(\pi)] \le \epsilon$$

Undoing the epigraph transformation, we obtain:

$$\max V_{ov}(\pi) + \mathbb{E}[\mu_0] + c_1(\pi) - \int_{x\in\mathcal{X}^{\text{nov}}} \sum_{u\in\{a,b\}} \sum_{r\in\{0,1\}} |-g_r(x,u)\pi_r(x,u)f(x,u)| \frac{1}{2}\Delta B(x,u)dx$$

$$\int_{x\in\mathcal{X}^{\text{nov}}} \{f(x \mid a)(\sum_r \pi_r(x,a)\overline{B}_r(x,a)) - f(x \mid b)(\sum_r \pi_r(x,b)\underline{B}_r(x,u))\} + \mathbb{E}[\Delta_{ov}T(\pi)] \le \epsilon$$

and simplifying the absolute value:

$$\max V_{ov}(\pi) + \mathbb{E}[\mu_0] + c_1(\pi) - \int_{x \in \mathcal{X}^{\text{nov}}} \sum_{u \in \{a,b\}} \sum_{r \in \{0,1\}} |g_r(x,u)\pi_r(x,u)f(x,u)| \frac{1}{2}\Delta B(x,u)dx$$

$$\int_{x \in \mathcal{X}^{\text{nov}}} \{f(x \mid a)(\sum_r \pi_r(x,a)\overline{B}_r(x,a)) - f(x \mid b)(\sum_r \pi_r(x,b)\underline{B}_r(x,u))\} + \mathbb{E}[\Delta_{ov}T(\pi)] \leq \epsilon$$

□

C.3 Proofs for general fairness-constrained policy optimization algorithm and analysis

We begin by summarizing some notation that will simplify some statements. Define, for observation tuples $O \sim (X, A, R, T, Y)$, the value estimate $v(Q; \eta)$ given some pseudo-outcome $\psi(O; \eta)$ dependent on observation information and nuisance functions η. (We often suppress notation of η for brevity). We let estimators sub/super-scripted by 1 denote estimators from the first dataset.

$$
\begin{aligned}
v_{(\cdot)}(O; Q, \eta) &= \mathbb{E}_{\pi \sim Q}[v_{(\cdot)}(O; \pi, \eta)] \\
v_{(\cdot)}(Q) &= \mathbb{E}[v_{(\cdot)}(Q)] \\
\hat{V}_1^{(\cdot)}(Q) &= \mathbb{E}_{n_1}[v_{(\cdot)}(Q)] \\
g_j(O; Q) &= \mathbb{E}_{\pi \sim Q}[g_j(O; \pi) \mid O, \mathcal{E}_j] \\
h_j(Q) &= \mathbb{E}[g_j(O; Q) \mid \mathcal{E}_j] \\
\hat{h}_j^1(Q) &= \mathbb{E}_{n_1}[g_j(O; Q) \mid \mathcal{E}_j]
\end{aligned}
$$

C.3.1 Preliminaries: results from other works used without proof

Theorem 3 (Theorem 3, [Agarwal et al., 2018] (saddle point generalization bound for **??**))**.** *Let* $\xi := \max_h \|M\hat{\mu}(h) - \hat{c}\|_\infty$. *Suppose Assumption 7 holds for* $C' \geq 2C + 2 + \sqrt{\ln(4/\delta)/2}$, *where* $\delta > 0$. *Let* $Q^\star$ *minimize* $V(Q)$ *subject to* $Mh(Q) \leq c$. *Then* **??** *with* $\nu \propto n^{-\alpha}, B \propto n^\alpha$ *and* $\omega \propto \xi^{-2} n^{-2\alpha}$ *terminates in* $O\left(\xi^2 n^{4\alpha} \ln |\mathcal{K}|\right)$ *iterations and returns* $\hat{Q}$. *If* $np_j^\star \geq 8 \log(2/\delta)$ *for all* j, *then with probability at least* $1 - (|\mathcal{J}| + 1)\delta$ *then for all* k, $\hat{Q}$ *satisfies:*

$$
\begin{aligned}
V(\hat{Q}) &\leq V\left(Q^\star\right) + \widetilde{O}\left(n^{-\alpha}\right) \\
\gamma_k(\widehat{Q}) &\leq d_k + \frac{1 + 2\nu}{B} + \sum_{j \in \mathcal{J}} |M_{k,j}| \, \widetilde{O}\left(\left(np_j^\star\right)^{-\alpha}\right)
\end{aligned}
$$

The proof of [Agarwal et al., 2018, Thm. 3] is modular in invoking Rademacher complexity bounds on the objective function and constraint moments, so that invoking standard Rademacher complexity bounds for off-policy evaluation/learning [Athey and Wager, 2021, Swaminathan and Joachims, 2015] yields the above statement for $V(\pi)$ (and analogously, randomized policies by [Bartlett and Mendelson, 2002, Thm. 12.2] giving stability for convex hulls of policy classes).

More specifically, we use standard local Rademacher complexity bounds.

Definition 1 (Local Rademacher Complexity)**.** The local Rademacher complexity for a generic $f \in \mathcal{F}$ is:

$$
\mathcal{R}(r, \mathcal{F}) = \mathbb{E}_{\epsilon, X_{1:n}}\left[\sup_{f \in \mathcal{F}: \|f\|_2 \leq r} \frac{1}{n} \sum_{i=1}^n \epsilon_i f\left(X_i\right)\right]
$$

The following is a generic concentration inequality for local Rademacher complexity over some radius r; see Wainwright [2019] for more background.

Lemma 2 (Lemma 5 of [Chernozhukov et al., 2019]/Lemma 4, Foster and Syrgkanis [2019])**.** *Consider any* $Q^* \in \mathcal{Q}$. *Assume that* $v(\pi)$ *is L-Lipschitz in its first argument with respect to the* ℓ_2 *norm and let:*

$$
Z_n(r) = \sup_{Q \in \mathcal{Q}} \{|\mathbb{E}_n[\hat{v}(Q) - \hat{v}(Q^*)] - \mathbb{E}[v(Q) - v(Q^*)]| : \mathbb{E}[(v(Q) - v(Q^*))^2]^{\frac{1}{2}} \leq r\}
$$

Then for some constant C_3:

$$Z_n(r) \leq C_3 \left(\mathcal{R}\left(r, \mathcal{Q} - Q^*\right) + r\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \right)$$

Lemma 3 (Concentration of conditional moments ([Agarwal et al., 2018, Woodworth et al., 2017])**.** *For any $j \in \mathcal{J}$, with probability at least $1-\delta$, for all Q,*

$$\left| \widehat{h}_j(Q) - h_j(Q) \right| \leq 2\mathcal{R}_{n_j}(\mathcal{Q}) + \frac{2}{\sqrt{n_j}} + \sqrt{\frac{\ln(2/\delta)}{2n_j}}$$

If $np_j^\star \geq 8\log(2/\delta)$, then with probability at least $1-\delta$, for all Q,

$$\left| \widehat{h}_j(Q) - h_j(Q) \right| \leq 2\mathcal{R}_{np_j^\star/2}(\mathcal{Q}) + 2\sqrt{\frac{2}{np_j^\star}} + \sqrt{\frac{\ln(4/\delta)}{np_j^\star}}$$

Lemma 4 (Orthogonality (analogous to [Chernozhukov et al., 2019] (Lemma 8), others))**.** *Suppose the nuisance estimates satisfy a mean-squared-error bound*

$$\max_l \{\mathbb{E}[(\hat{\eta}_l - \eta_l)^2]\}_{l \in [L]} \coloneqq \chi_n^2$$

Then w.p. $1-\delta$ over the randomness of the policy sample,

$$V(Q_0) - V(\hat{Q}) \leq O(R_{n,\delta} + \chi_n^2)$$

C.4 Adapted lemmas

In this subsection we collect results similar to those that have appeared previously, but that require substantial additional argumentation in our specific saddle point setting.

The next lemma establishes the variance of small-regret policies with estimated vs. true nuisances is close, up to nuisance estimation error.

Lemma 5 (Feasible vs. oracle nuisances in low-variance regret slices (Chernozhukov et al. [2019], Lemma 9))**.** *Suppose that the mean squared error of the nuisance estimates is upper bounded w.p. $1-\delta$ by $\chi_{n,\delta}^2$ and suppose $\chi_{n,\delta}^2 \leq \epsilon_n$. Then:*

$$V_2^0 = \sup_{Q, Q' \in \mathcal{Q}_*\left(\epsilon_n + 2\chi_{n,\delta}^2\right)} \mathrm{Var}\left(v_{DR}^0(x; Q) - v_{DR}^0\left(x; Q'\right)\right)$$

Then $V_2 \leq V_2^0 + O\left(\chi_{n,\delta}\right)$.

C.5 Proof of Theorem 1

Proof of Theorem 1. We first study the meta-algorithm with "oracle" nuisance functions $\eta = \eta_0$. For brevity below we notationally suppress the dependence of v on observation O.

Define

$$\Pi_2(\epsilon_n) = \left\{\pi \in \Pi : \mathbb{E}_{n_1}[v(\pi;\eta_0) - v(\hat{Q}_1;\eta_0)] \le \epsilon_n, \mathbb{E}_{n_1}\left[g_j(O;\pi,\eta_0) - g_j\left(O;\hat{Q}_1,\eta_0\right) \mid \mathcal{E}_j\right] \le \epsilon_n, j \in \hat{\mathcal{I}}_1\right\}$$
$$\mathcal{Q}_2(\epsilon_n) = \{Q \in \Delta(\Pi_2(\epsilon_n))\}$$
$$\mathcal{Q}^*(\epsilon_n) = \{Q \in \Delta(\Pi) : \mathbb{E}[(v(Q;\eta_0) - v(Q^*;\eta_0)] \le \epsilon_n, \mathbb{E}[g_j(O;Q,\eta_0) \mid \mathcal{E}_j] - \mathbb{E}[g_j(O;Q^*,\eta_0) \mid \mathcal{E}_j] \le \epsilon_n\}$$

In the following, we suppress notational dependence on η_0.

Note that $\hat{Q}_1 \in \mathcal{Q}_2(\epsilon_n)$.

Step 1: First we argue that w.p. $1 - \delta/6$, $Q^* \in \mathcal{Q}_2$.

Invoking Theorem 3 on the output of the first stage of the algorithm, yields that with probability $1 - \frac{\delta}{6}$ over the randomness in $\mathcal{D}_1$, by choice of $\epsilon_n = \bar{O}(n^{-\alpha})$),

$$V(\hat{Q}_1) \le V(Q^*) + \epsilon_n/2$$
$$\gamma_k(\hat{Q}_1) \le d_k + \sum_{j \in \mathcal{J}} |M_{k,j}| \widetilde{O}\left((np_j^*)^{-\alpha}\right) \le d_k + \epsilon_n/2 \quad \text{for all } k$$

Further, by Lemma 2,

$$\sup_{Q \in \mathcal{Q}} |\mathbb{E}_{n_1}[(v(Q) - v(Q^*))] - \mathbb{E}[(v(Q) - v(Q^*))]| \le \epsilon_n/2$$
$$\sup_{Q \in \mathcal{Q}} |\mathbb{E}_{n_1}[(g(O;Q) - g(O;Q^*))] - \mathbb{E}[(g(O;Q) - g(O;Q^*))]| \le \epsilon_n/2$$

Therefore, with high probability on the good event, $Q^* \in \mathcal{Q}_2$.

Step 2: Again invoking Theorem 3, this time on the output of the second stage of the algorithm with function space Π_2 (hence implicitly $\mathcal{Q}_2$), and conditioning on the "good event" that $Q^* \in \mathcal{Q}_2$, we obtain the bound that with probability $\ge 1 - \delta/3$ over the randomness of the second sample $\mathcal{D}_2$,

$$V(\hat{Q}_2) \le V(Q^*) + \epsilon_n/2$$
$$\gamma_k(\hat{Q}_2) \le \gamma_k(Q^*) + \epsilon_n/2$$

Step 3: empirical small-regret slices relate to population small-regret slices, and variance bounds

We show that if $Q \in \mathcal{Q}_2$, then with high probability $Q \in \mathcal{Q}_2^0$ (defined on small population value- and constraint-regret slices relative to $\hat{Q}_1$ rather than small empirical regret slices)

$$\mathcal{Q}_2^0 = \{Q \in \text{conv}(\Pi) : \left|V(Q) - V(\hat{Q}_1)\right| \le \epsilon_n/2, \mathbb{E}[g_j(O;Q) - g_j(O;\hat{Q}_1)) \mid \mathcal{E}_j] \le \epsilon_n, \forall j\}$$

so that w.h.p. $\mathcal{Q}_2 \subseteq \mathcal{Q}_2^0$.

Note that for $Q \in \mathcal{Q}$, w.h.p. $1 - \delta/6$ over the first sample, we have that

$$\sup_{Q \in \mathcal{Q}} \left|\mathbb{E}_n[v(Q) - v(\hat{Q}_1)] - \mathbb{E}[v(Q) - v(\hat{Q}_1)]\right| \le 2 \sup_{Q \in \mathcal{Q}} |\mathbb{E}_n[v(Q)] - \mathbb{E}[v(Q)]| \le \epsilon,$$
$$\sup_{Q \in \mathcal{Q}} \left|\mathbb{E}_{n_1}[g_j(O;Q) - g_j(O;\hat{Q}_1) \mid \mathcal{E}_j] - \mathbb{E}[g_j(O;Q) - g_j(O;\hat{Q}_1) \mid \mathcal{E}_j]\right|$$
$$\le 2 \sup_{Q \in \mathcal{Q}} |\mathbb{E}_{n_1}[g_j(O;Q) \mid \mathcal{E}_j] - \mathbb{E}[g_j(O;Q) \mid \mathcal{E}_j]| \le \epsilon, \forall j$$

The second bound follows from [Bartlett and Mendelson, 2002, Theorem 12.2] (equivalence of Rademacher complexity over convex hull of the policy class) and linearity of the policy value and constraint estimators in π, and hence Q.

On the other hand since Q_1 achieves low policy regret, the triangle inequality implies that we can contain the true policy by increasing the error radius. That is, for all $Q \in \mathcal{Q}_2$, with high probability $\geq 1 - \delta/3$:

$$|\mathbb{E}[(v(Q) - v(Q^*))]| \leq \left|\mathbb{E}[(v(Q) - v(\hat{Q}_1))]\right| + \left|\mathbb{E}[(v(\hat{Q}_1) - v(Q^*))]\right| \leq \epsilon_n$$
$$|\mathbb{E}[g_j(O;Q) - g_j(O;Q^*) \mid \mathcal{E}_j]| \leq \left|\mathbb{E}[g_j(O;Q) - g_j(O;\hat{Q}_1) \mid \mathcal{E}_j]\right| + \left|\mathbb{E}[g_j(O;\hat{Q}_1) - g_j(O;Q^*) \mid \mathcal{E}_j]\right| \leq \epsilon_n$$

Define the space of distributions over policies that achieve value and constraint regret in the population of at most ϵ_n :

$$\mathcal{Q}_*(\epsilon_n) = \{Q \in \mathcal{Q} \colon V(Q) - V(Q^*) \leq \epsilon_n,\ \ \mathbb{E}[g_j(O;Q) - g_j(O;Q^*) \mid \mathcal{E}_j] \leq \epsilon_n, \forall j\},$$

so that on that high-probability event,

$$\mathcal{Q}_2^0(\epsilon_n) \subseteq \mathcal{Q}_*(\epsilon_n). \tag{11}$$

Then on that event with probability $\geq 1 - \delta/3$,

$$\begin{aligned} r_2^2 &= \sup_{Q \in \mathcal{Q}_2} \mathbb{E}[(v(Q) - v(Q^*))^2] \leq \sup_{Q \in \mathcal{Q}^*(\epsilon_n)} \mathbb{E}[(v(Q) - v(Q^*))^2] \\ &= \sup_{Q \in \mathcal{Q}^*(\epsilon_n)} \mathrm{Var}(v(Q) - v(Q^*)) + \mathbb{E}[(v(Q) - v(Q^*))]^2 \\ &\leq \sup_{Q \in \mathcal{Q}^*(\epsilon_n)} \mathrm{Var}(v(Q) - v(Q^*)) + \epsilon_n^2 \end{aligned}$$

Therefore:

$$r_2 \leq \sqrt{\sup_{Q \in \mathcal{Q}_*(\epsilon_n)} \mathrm{Var}\left(v(Q) - v(Q^*)\right)} + 2\epsilon_n = \sqrt{V_2} + 2\epsilon_n$$

Combining this with the local Rademacher complexity bound, we obtain that:

$$\mathbb{E}[v(\hat{Q}_2) - v(Q^*)] = O\left(\kappa\left(\sqrt{V_2} + 2\epsilon_n, \mathcal{Q}_*\left(\epsilon_n\right)\right) + \sqrt{\frac{V_2 \log(3/\delta)}{n}}\right)$$

These same arguments apply for the variance of the constraints

$$V_2^j = \sup\left\{\mathrm{Var}\left(g_j(O;Q) - g_j\left(O;Q'\right)\right) : Q, Q' \in \mathcal{Q}_*(\tilde{\epsilon}_n)\right\}$$

□

C.6 Proofs of auxiliary/adapted lemmas

Proof of Lemma 5. The proof is analogous to that of [Chernozhukov et al., 2019, Lemma 9] except for the step of establishing that $\pi_* \in \mathcal{Q}^0_{\epsilon_n + O(\chi^2_{n,\delta})}$: in our case we must establish relationships between saddlepoints under estimated vs. true nuisances. We show an analogous version below.

Define the saddle points to the following problems (with estimated vs. true nuisances):

$$(Q^*_{0,0}, \lambda^*_{0,0}) \in \arg\min_Q \max_\lambda \mathbb{E}[v_{DR}(Q;\eta_0)] + \lambda^\top(\gamma_{DR}(Q;\eta_0) - d) \coloneqq L(Q,\lambda;\eta_0,\eta_0) \coloneqq L(Q,\lambda),$$
$$(Q^*_{\eta,0}, \lambda^*_{\eta,0}) \in \arg\min_Q \max_\lambda \mathbb{E}[v_{DR}(Q;\eta)] + \lambda^\top(\gamma_{DR}(Q;\eta_0) - d),$$
$$(Q^*, \lambda^*) \in \arg\min_Q \max_\lambda \mathbb{E}[v_{DR}(Q;\eta)] + \lambda^\top(\gamma_{DR}(Q;\eta) - d).$$

We have that:

$$\begin{aligned}
&\mathbb{E}[v_{DR}(Q^*)] \le L(Q^*,\lambda^*;\eta,\eta) + \nu \\
&\le L(Q^*,\lambda^*;\eta,\eta_0) + \nu + \chi^2_{n,\delta} \\
&\le L(Q^*,\lambda^*;\eta,\eta_0) + \nu + \chi^2_{n,\delta} && \text{by Lemma 4} \\
&\le L(Q^*,\lambda^*_{\eta,0};\eta,\eta_0) + \nu + \chi^2_{n,\delta} && \text{by saddlepoint prop.} \\
&\le L(Q^*_{\eta,0},\lambda^*_{\eta,0};\eta,\eta_0) + \left|L(Q^*_{\eta,0},\lambda^*_{\eta,0};\eta,\eta_0) - L(Q^*,\lambda^*_{\eta,0};\eta,\eta_0)\right| + \nu + \chi^2_{n,\delta} && \text{triangle ineq.} \\
&\le L(Q^*_{\eta,0},\lambda^*_{\eta,0};\eta,\eta_0) + \epsilon_n + \nu + \chi^2_{n,\delta} && \text{assuming } \epsilon_n \ge \chi^2_{n,\delta} \\
&\le \mathbb{E}[v_{DR}(Q^*_{\eta,0};\eta)] + \epsilon_n + 2\nu + \chi^2_{n,\delta} && \text{apx. complementary slackness} \\
&\le \mathbb{E}[v_{DR}(Q^*_{0,0};\eta)] + \epsilon_n + 2\nu + \chi^2_{n,\delta} && \text{suboptimality}
\end{aligned}$$

Hence

$$\mathbb{E}[v_{DR}(Q^*;\eta)] - \mathbb{E}[v_{DR}(Q^*_{0,0};\eta)] \le \epsilon_n + 2\nu + \chi^2_{n,\delta}.$$

We generally assume that the saddlepoint suboptimality ν is of lower order than ϵ_n (since it is under our computational control).

Applying Lemma 4 gives;

$$V(Q^*) - V(Q^*_{0,0}) \le \epsilon_n + 2\nu + 2\chi^2_{n,\delta}.$$

Define policy classes with respect to small-population regret slices (with a nuisance-estimation enlarged radius):

$$\mathcal{Q}^0(\epsilon) = \{Q \in \Delta(\Pi) \colon V(Q^*_0) - V(Q) \le \epsilon, \gamma(Q^*_0) - \gamma(Q) \le \epsilon\}$$

Then we have that

$$V_2^{obj} \le \sup_{Q \in \mathcal{Q}^0(\epsilon_n)} \mathrm{Var}(v_{DR}(O;\pi) - v_{DR}(O;\pi^*)),$$

where we have shown that $\pi^* \in \mathcal{Q}^0(\epsilon + 2\nu + 2\chi^2_{n,\delta})$.

Following the rest of the argumentation in [Chernozhukov et al., 2019, Lemma 9] from here onwards gives the result, i.e. studying the case of estimated nuisances with our Lemma 5 and Lemma 4. □