Generalizing Off-Policy Learning under Sample Selection Bias

Abstract

Learning personalized decision policies that generalize to the target population is of great relevance. Since training data is often not representative of the target population, standard policy learning methods may yield policies that do not generalize target population. To address this challenge, we propose a novel framework for learning policies that generalize to the target population. For this, we characterize the difference between the training data and the target population as a sample selection bias using a selection variable. Over an uncertainty set around this selection variable, we optimize the minimax value of a policy to achieve the best worst-case policy value on the target population. In order to solve the minimax problem, we derive an efficient algorithm based on a convex-concave procedure and prove convergence for parametrized spaces of policies such as logistic policies. We prove that, if the uncertainty set is well-specified, our policies generalize to the target population as they can not do worse than on the training data. Using simulated data and a clinical trial, we demonstrate that, compared to standard policy learning methods, our framework improves the generalizability of policies substantially.

1 INTRODUCTION

Learning personalized policies has become integral to modern decision-making in a variety of domains such as medicine [Hill and Su, 2013], and public policy [Kube et al., 2019]. Since in these domains exploration is costly or otherwise infeasible, many methods have been proposed for off-policy learning, i. e., policy learning from existing data [e. g., Athey and Wager, 2021, Beygelzimer and Langford, 2009, Dudík et al., 2014, Kallus, 2018].

A major challenge in off-policy learning is the generalizability of policies. Generalizability is concerned with whether a policy learned on the data for training (i. e., training data) is also effective in the target population. Standard methods for policy learning yield policies that are effective on the target population, if, and only if, the training data is representative of the target population [e. g., Beygelzimer and Langford, 2009, Dudík et al., 2014]. However, this may not hold true in practice [e. g., Buchanan et al., 2018, Cole and Stuart, 2010, Downs and Black, 1998, Flores et al., 2021, Norris et al., 2001, Rothwell, 2005]. For instance, a review of HIV/AIDS clinical trials found that women are largely underrepresented in these trials [Gandhi et al., 2005, Greenblatt, 2011], so that data from these trials is not representative of the actual target population (i.e., the population of HIV-positive patients in the USA). Therefore, when data from such trials is used to derive policies, standard methods for policy learning may not generalize to the target population. As such, these policies may be ineffective or even harmful on the target population and, therefore, not relevant in practice.

In this paper, we develop a framework for learning policies from training data that generalize to the target population.[1] For this, we characterize the difference between training data and target population as a sample selection bias using an unknown selection variable [e. g., Cortes et al., 2008, Manski, 1989]. If we had oracle access to the true selection variable, we could re-weight the data accordingly in order to obtain the value of a policy on the target population. Since, in practice, the true selection variable is unknown, the value of a policy on the target population is not identifiable from training data. Instead, we derive bounds on the odds-ratio of the selection probability, which yields an uncertainty set around the true selection probabilities. Then, our framework optimizes the minimax value of a policy to achieve the best worst-case policy value on the target population. We prove that, if the uncertainty set is well-specified, our framework

[1] Code available at `github.com/anonymous/GeneralOPL` (link anonymized for peer-review; code for review in the supplements.)

yields policies that do not do worse on the target population than the worst-case policy value estimated from the training data. As such, these policies can generalize to the target population. In order to efficiently optimize the minimax value of a policy, we show that it can be written as a difference of convex functions (DC) program. Then, by leveraging the structure of the adversarial subproblem, we develop a tailored **mini**max **c**onvex-**c**oncave **p**rocedure (MMCCP). We prove that MMCCP converges for certain parameterized spaces of policies such as logistic policies. Using synthetic data and a clinical trial, we demonstrate that standard policy learning methods generalize poorly, while our framework improves the generalizability of policies substantially. As such, our framework enables to learn reliable policies that can be implemented in the target population.

2 PRELIMINARIES

In this section, we describe the setup, formulate the problem of generalizing policies, and discuss related work.

2.1 SETUP

We consider a binary treatment $T \in \{0, 1\}$, covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$, and the outcome $Y \in \mathbb{R}$. We use the convention that lower outcomes are preferred. Using the Neyman-Rubin potential outcomes framework [Rubin, 2005], let $Y(0), Y(1)$ be the potential outcomes for each of the treatments. Further, let a *policy* π be a map from the covariates to the probability of treatment assignment, i. e., $\pi : \mathcal{X} \to [0, 1]$. Then, the *policy value* of π is given by

$$V(\pi) = \mathbb{E}[Y^\pi] = \mathbb{E}[\pi(X)Y(1) + (1 - \pi(X))Y(0)]. \tag{1}$$

The objective of *policy learning* is to find a policy $\overline{\pi}$ in a policy class Π that minimizes the policy value, i. e., $\overline{\pi} \in \arg\min_{\pi \in \Pi} V(\pi)$.

We make the following three standard assumptions [Rubin, 1974]: (i) consistency (i. e., $Y = Y(T)$); (ii) positivity (i. e., $0 < p(T = 1 \mid X = x) < 1$ for all x); and (iii) strong ignorability (i. e., $Y(0), Y(1) \perp\!\!\!\perp T \mid X$). Then we can identify the policy value in (1) in terms of the observed data (X, T, Y).

2.2 PROBLEM FORMULATION

Suppose we are interested in learning a policy that minimizes the policy value under the target distribution $(X, T, Y) \sim \mathbb{P}$, i. e., $V_{\text{Target}}(\pi)$. However, we are *not* given data from the target distribution, but only data from a (potentially different) training distribution $(X, T, Y) \sim \mathbb{P}_{\text{Train}}$.

Standard policy learning methods assume that the training and target distributions are identical. However, even in carefully designed clinical trials, the subjects in the trial are often not representative of the target population, i. e., $\mathbb{P}_{\text{Train}} \neq \mathbb{P}$ [e. g., Buchanan et al., 2018, Cortes et al., 2008, Downs and Black, 1998, Flores et al., 2021, Gandhi et al., 2005, Greenblatt, 2011, Rothwell, 2005]. Hence, standard methods for policy learning yield policies that minimize the policy value on the training data, i. e., $V_{\text{Train}}(\pi) = \mathbb{E}_{\text{Train}}[Y^\pi]$. However, since the policy value depends on the underlying data distribution, these policies may *not* minimize the policy value on the target population, i. e., $V_{\text{Target}}(\pi) = \mathbb{E}[Y^\pi]$. This can be seen when writing $\mathbb{E}[Y^\pi]$ in terms of the distribution $\mathbb{P}_{\text{Train}}$ using a change of probability measure:

$$\mathbb{E}[Y^\pi] = \mathbb{E}_{\text{Train}}[R\, Y^\pi], \tag{2}$$

where the random variable $R = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}_{\text{Train}}$ is the *Radon-Nikodým derivative*,[2] also know as *density ratio*. As a direct implication, if $\mathbb{P}_{\text{Train}} \neq \mathbb{P}$ and, thus, $R \neq 1$, it follows that

$$\mathbb{E}[Y^\pi] \neq \mathbb{E}_{\text{Train}}[Y^\pi]. \tag{3}$$

In other words, a policy learned from training data using standard methods may not generalize to the target population, and, as such, may be of little help in practice.

In this paper, we consider the realistic setting in which the training data is *not* representative of the target population. We propose a framework for learning policies that generalize to the target population only given data from the training distribution, i. e., $\{(X_i, T_i, Y_i)\}_{i=1}^n \sim \mathbb{P}_{\text{Train}}$.

2.3 RELATED WORK

Despite the vast literature on off-policy learning, less work considers the problem of learning policies that generalize to the target population. Below, we summarize works on off-policy learning and works on external validity in causality, which is closely related to generalizability.

Off-policy learning. Off-policy learning methods can be broadly divided into three categories: (i) Direct methods estimate the outcome functions $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$ and plug them into (1), i. e., $\hat{V}^{\text{DM}}(\pi) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\hat{\mu}_1(X_i) + (1 - \pi(X_i))\hat{\mu}_0(X_i)$ [e. g., Bennett and Kallus, 2020]. Direct methods are known to be weak against model misspecification with regards to $\mu_t(x)$. (ii) Weighting methods re-weight the outcome data such that it looks as if it were generated by the policy that is evaluated [e. g., Bottou et al., 2013, Horvitz and Thompson, 1952, Kallus, 2018, Li et al., 2011]. A common choice for weights are the normalized inverse propensity weights [Swaminathan and Joachims, 2015], i. e., $\hat{V}^{\text{NIPW}}(\pi) = \frac{1}{n}\sum_{i=1}^n 2W_i^{\text{IPW}}(1 - 2T_i)(1 - T_i - \pi(X_i))Y_i/(\frac{1}{n}\sum_{j=1}^n W_j^{\text{IPW}})$, where

[2] The standard assumption that $\mathbb{P}$ is absolute continuous with respect to $\mathbb{P}_{\text{Train}}$, i. e., $\mathbb{P} \ll \mathbb{P}_{\text{Train}}$, is made in order to ensure that the Radon-Nikodým derivative is well-defined.

$W_i^{\mathrm{IPW}} = 1/((1-2T_i)(1-T_i-\pi^b(X_i)))$ and $\pi^b(x) = \mathbb{P}(T=1 \mid X=x)$ is the so-called behavior policy, which was used to generate the training data. (iii) Doubly robust methods combine direct and weighting methods typically using the augmented inverse propensity weight estimator [Athey and Wager, 2021, Dudík et al., 2014, Thomas and Brunskill, 2016]. When the direct estimate of $\hat{\mu}_t$ is biased, the doubly robust method weights the residuals by the inverse propensity weights in order to remove the bias, i. e., $\hat{V}^{\mathrm{DR}}(\pi) = \hat{V}^{\mathrm{DM}}(\pi) + \frac{1}{n}\sum_{i=1}^n W_i^{\mathrm{IPW}}(1-2T_i)(1-T_i-\pi(X_i))(Y_i-\hat{\mu}_{T_i}(X_i))$.

The above methods have become the standard for off-policy learning. Despite their widespread use, the above methods implicitly assume that the training data, which is used to learn the policy, is representative of the target population. As such, when the training data is *not* representative of the target population, we cannot rely on the above methods, as policies may not generalize to the target population.

Distributionally robust optimization. A related, yet fundamentally different idea is distributionally robust optimization (DRO) [e. g., Duchi and Namkoong, 2018], which studies robustness towards distributional shifts. DRO has found application in off-policy learning by optimizing worst-case policy values [Si et al., 2020] and individualized treatment rules [Zhao et al., 2019b, Mo et al., 2020]. While generalizability is related to DRO, since the difference between training and target distribution can be seen as a distributional shift, it is fundamentally different, as DRO allows for arbitrary changes in distribution. In contrast, generalizability considers a training distribution that is, potentially, not representative of the target population, but derived from the target population. That is, generalizability considers differences in the distributions the arise from an unknown selection mechanism into the training data. Moreover, DRO and its applications require the decision-maker to quantify the distance between training and target distribution in terms of some divergence measure (typically the Kullback-Leibler divergence), which may be notoriously difficult for domain experts such as clinicians. In contrast, our approach allows for user-friendly and intuitive calibration of the parameters involved in the uncertainty set due to recognizing that the differences arises from an unknown selection mechanism.

External validity in causality. External validity in causal inference is concerned with whether causal effect estimates obtained from a study sample are also valid for the target population. A common approach to address this is to reweight the data with the inverse of a subject's probability to be selected into the study sample [e. g., Buchanan et al., 2018, Cole and Stuart, 2010, Dahabreh et al., 2019, Imai et al., 2013, Stuart et al., 2011]. This idea has been extended to a doubly robust method for off-policy learning [Uehara et al., 2020]. Predominantly used in economics, the Heckman correction is another technique that is also based on a subject's selection probability [Heckman, 1979]. However, in order to estimate these selection probabilities, all existing approaches assume that data from both the study sample *and* the target population is given. In practice, however, we are only given data from the study sample and not from the target population. Other approaches include approximations of the bias arising from the difference in the study sample and target population by using weights that do not depend on the selection variable [Andrews and Oster, 2017], by bounding the weights directly [Aronow and Lee, 2013], or, in addition, by constraining the shape of the population outcome distribution [Miratrix et al., 2018].

Different to the above approaches and more practically, we do not assume that we have access to samples from the target population and, therewith, we cannot estimate a subject's selection probability. As a remedy, we present our framework for learning generalizable policies in the following.

3 GENERALIZING OFF-POLICY LEARNING UNDER SAMPLE SELECTION BIAS

In this section, we introduce our framework for learning policies that generalize to the target population. For this, we first characterize the difference between the training distribution $\mathbb{P}_{\text{Train}}$ and the target distribution $\mathbb{P}$ as a sample selection bias (Section 3.1). Then, based on this, we derive an uncertainty set and optimize the minimax policy value to achieve the best worst-case policy value (Section 3.2). We prove that policies learned in this way do not do worse on the target population than the worst-case policy value and, as such, can be generalized to the target population (Section 3.3).

3.1 SAMPLE SELECTION BIAS

In this section, we characterize the difference between the training distribution $\mathbb{P}_{\text{Train}}$ and the target distribution $\mathbb{P}$ as a sample selection bias using a selection variable [e. g., Cortes et al., 2008, Manski, 1989]. This then allows us to characterize the Radon-Nikodým derivative $R = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}_{\text{Train}}$ in (2) in terms of the selection variable.

We represent the selection bias with a selection variable $S \in \{0,1\}$. If, for a subject, $S=1$, the subject is included in the training data, and, if $S=0$, the subject is excluded from the training data. As a result, we can write the training distribution in terms of the target distribution:

$$\mathbb{P}_{\text{Train}}(\cdot) = \mathbb{P}(\cdot \mid S=1). \tag{4}$$

Based on this, we characterize the Radon-Nikodým derivative, which enables us to write the policy value on the target population in terms of the selection variable S and the training distribution $\mathbb{P}_{\text{Train}}$.

Proposition 1. *Under the sample selection bias, we can write the Radon-Nikodým derivative $R = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}_{\text{Train}}$ as*

$$R = \frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1 \mid X,T,Y)}, \tag{5}$$

and, therefore, we can write the policy value on the target population as

$$V_{\textit{Target}}(\pi) = \mathbb{E}_{\text{Train}}\left[\frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1 \mid X,T,Y)} Y^{\pi}\right]. \tag{6}$$

See Appendix A.1 for a proof. If, hypothetically, we observed S, we could estimate $R = \mathbb{P}(S=1)/\mathbb{P}(S=1 \mid X,T,Y)$[3] and re-weight the data accordingly in order to obtain the policy value on the target population. However, we never observe the selection variable S, since we only observe the training data for which $S=1$. This renders the selection variable S unidentifiable from the training data. Instead, we use an uncertainty set over which we optimize the minimax policy value on the target population.

3.2 LEARNING GENERALIZABLE POLICIES BY OPTIMIZING MINIMAX POLICY VALUE

We derive an uncertainty set around $R = \mathbb{P}(S=1)/\mathbb{P}(S=1 \mid X,T,Y)$ over which we maximize the policy value to obtain the worst-case policy value. Then, our framework optimizes the minimax policy value to achieve the best worst-case policy value on the target population.

If we had oracle access to the true Radon-Nikodým derivative $R_i^* = R_i^*(X_i, T_i, Y_i)$, we could estimate the policy value on the target population using Proposition 1, that is, by re-weighting the data with R^*. This often leads to high variance estimates due to probabilities close to zero. As a remedy, since $\mathbb{E}[R^*] = 1$, we use the empirical sum of the true Radon-Nikodým derivatives as a control variate to normalize the estimate. This gives rise to the following Hajek estimator for the policy value on the target population $V_{\text{Target}}(\pi)$:

$$\hat{V}^*_{\text{Target}}(\pi) = \frac{\sum_{i=1}^n R_i^* \psi_i(\pi)}{\sum_{i=1}^n R_i^*}, \tag{7}$$

where $\psi_i(\pi)$ corresponds to one of the three standard methods for policy learning: direct, weighting, and doubly robust methods. Formally, $\psi_i(\pi)$ is either $\psi_i^{\text{DM}}(\pi)$, $\psi_i^{\text{NIPW}}(\pi)$, or $\psi_i^{\text{DR}}(\pi)$ given as:

$$\psi_i^{\text{DM}}(\pi) = \pi(X_i)\mu_1(X_i) + (1-\pi(X_i))\mu_0(X_i), \tag{8}$$

$$\psi_i^{\text{NIPW}}(\pi) = \frac{2W_i^{\text{IPW}}(1-2T_i)}{\frac{1}{n}\sum_{j=1}^n W_j^{\text{IPW}}}(1-T_i-\pi(X_i))Y_i, \tag{9}$$

$$\psi_i^{\text{DR}}(\pi) = \psi_i^{\text{DM}}(\pi) + W_i^{\text{IPW}}(1-2T_i)(1-T_i-\pi(X_i))(Y_i-\mu_{T_i}(X_i)). \tag{10}$$

The outcome functions $\mu_t(x)$ and the weights W^{IPW} need to be estimated from data. Any $\psi(\pi)$ in (8), (9), or (10) can be chosen for estimating the policy value as long as the data is re-weighted with the Radon-Nikodým derivative R^*.

Since the true R^* is unknown, we instead derive a worst-case policy value on the target population. This allows to ensure that our policy does not do worse than expected once it is implemented in the target population. For this, we maximize (7) over an uncertainty set around R^*. We consider an uncertainty set motivated by sensitivity analysis in causality [e. g., Kallus et al., 2019, Kallus and Zhou, 2018, Rosenbaum, 2002, Zhao et al., 2019a], which restricts by how much $\mathbb{P}(S=1 \mid X,T,Y)$ can vary from $\mathbb{P}(S=1)$ via the odds-ratio characterization as follows:

$$\frac{1}{\Gamma} \leqslant \frac{\mathbb{P}(S=1)(1-\mathbb{P}(S=1 \mid X,T,Y))}{\mathbb{P}(S=1 \mid X,T,Y)(1-\mathbb{P}(S=1))} \leqslant \Gamma, \tag{11}$$

where $\Gamma \geqslant 1$. For $\Gamma = 1$, we have equal probability of selection, i. e., $\mathbb{P}(S=1 \mid X,T,Y) = \mathbb{P}(S=1)$ and, thus, no difference between the training data and the target population. Larger values of Γ allow for larger variation in the probabilities of selection. The bounded odds-ratio in (11) immediately yields an uncertainty set for the Radon-Nikodým derivative:

$$\mathcal{R} = \{R \in \mathbb{R}_+^n : l \leqslant R_i \leqslant u,\ \forall i\}, \tag{12}$$

$$\text{where } l = \frac{1-\mathbb{P}(S=1)+\Gamma\mathbb{P}(S=1)}{\Gamma}, \tag{13}$$

$$u = \Gamma(1-\mathbb{P}(S=1)) + \mathbb{P}(S=1). \tag{14}$$

The uncertainty set $\mathcal{R}$ includes all Radon-Nikodým derivatives R that satisfy the odds-ratio restriction in (11). For a given policy, we seek the maximum policy value on the target population among all Radon-Nikodým derivatives in the uncertainty set. This yields the following worst-case policy.

Definition 1. *(Worst-case policy value.) The worst-case policy value on the target population under the bounded odds-ratio with parameter Γ is given by*

$$\overline{V}_{\textit{Target}}(\pi;\mathcal{R}) = \max_{R\in\mathcal{R}} \frac{\sum_{i=1}^n R_i \psi_i(\pi)}{\sum_{i=1}^n R_i}, \tag{15}$$

where $\psi_i(\pi)$ corresponds to either (8), (9), or (10).

[3] Under the standard assumption $\mathbb{P} \ll \mathbb{P}_{\text{Train}}$, the selection variable satisfies positivity, i. e., $\mathbb{P}(S=1 \mid X,T,Y) > 0$. Therefore, the ratio $\mathbb{P}(S=1)/\mathbb{P}(S=1 \mid X,T,Y)$ is well-defined.

Then, we seek the optimal policy in a policy class Π, which minimizes the worst-case policy value on the target population, i. e.,

$$\overline{\pi}(\Pi, \mathcal{R}) \in \arg\min_{\pi \in \Pi} \overline{V}_{\text{Target}}(\pi; \mathcal{R}). \tag{16}$$

In particular, a policy learned with our framework generalizes to the target population, since it does not do worse on the target population than the worst-case policy value estimated using the training data. For this, a decision-maker only has to quantify the population selection probability, i. e., $\mathbb{P}(S = 1)$ and appropriately choose the maximum deviation from it via Γ. We discuss data-driven approaches to choose these quantities in Section 3.4. We derive a tailored convex-concave procedure for optimizing (16) in Section 4.

3.3 THEORETICAL GUARANTEES FOR GENERALIZABILITY

We prove that, if the Radon-Nikodým is appropriately bounded, the worst-case policy value, $\overline{V}_{\text{Target}}(\pi; \mathcal{R})$, is asymptotically an upper bound for the true policy value on the target population, $V_{\text{Target}}(\pi)$. As such, a policy learned with our framework does not do worse on the target population than the worst-case policy value. Similar to [Athey and Wager, 2021], we express the flexibility of a policy class Π using the notion of the Rademacher complexity, i. e., $\mathcal{R}_n(\Pi)$.[4]

Theorem 1. *(Generalization bound.) Suppose the true Radon-Nikodým derivative is appropriately bounded, i. e., $R^* \in \mathcal{R}$ and, therefore, $l \leqslant R_i^* \leqslant u$, and we have bounded outcomes, i. e., $|Y| < C$. Then, for a constant K^ψ depending on $\psi(\pi)$ and for some $\delta > 0$, we have that,*

$$V_{Target}(\pi) \leqslant \overline{V}_{Target}(\pi; \mathcal{R}) + 2C \frac{u}{l} K_\psi \left(\mathcal{R}_n(\Pi) + \sqrt{\frac{18 \log(4/\delta)}{n}} \right), \tag{17}$$

with probability at least $1 - \delta$ and for any $\pi \in \Pi$. □

See Appendix A.2 for a proof. All policy classes we consider have $\sqrt{n}$-vanishing Rademacher complexity, i. e., $\mathcal{R}_n(\Pi) = \mathcal{O}(n^{-1/2})$. Therefore, Theorem 1 proves that, asymptotically, $\overline{V}_{\text{Target}}(\pi)$ is an upper bound for $V_{\text{Target}}(\pi)$. This guarantees that $\overline{\pi}(\Pi, \mathcal{R})$ from (16) does not do worse on the target population than the worst-case policy value, which is calculated using training data. In particular, since $\overline{\pi}(\Pi, \mathcal{R})$ minimizes the right hand side of (17), $\overline{\pi}(\Pi, \mathcal{R})$ is the best policy that guarantees to generalize to the target population. Our bound in Theorem 1 holds without complete knowledge of the selection variable and proves that our framework yields policies that generalizes to the target population.

[4]The empirical Rademacher complexity of a policy class Π is defined as $\mathcal{R}_n(\Pi) = \frac{1}{2^n} \Sigma_{\sigma \in \{-1,+1\}^n} \sup_{\pi \in \Pi} |\frac{1}{n} \sum_{i=1}^n \sigma_i \pi(X_i)|$.

Note that in Theorem 1, we use the true nuisance functions instead of estimates, since it has been shown that this does not affect the leading term in the convergence rate of the policy value (see Athey and Wager [2021]; Sec. 3.1, Sec. 3.2, and Lemma 4). This holds true if the nuisance functions have finite second moment and we use consistent estimators for the nuisance functions and L^2 errors decay with $1/n^\zeta$, where ζ depends on the nuisance functions. Hence, to provide a generalization bound on the policy value, it is enough to consider the true nuisance functions as we did in Theorem 1.

3.4 CALIBRATION OF Γ AND $\mathbb{P}(S = 1)$

In this section, we discuss two approaches to calibrate the parameters Γ and $\mathbb{P}(S = 1)$ in (11), which are context-dependent: (i) Practitioner calibration with domain knowledge and (ii) data-driven calibration.

(i) Practitioner calibration: This approach is based on domain knowledge of practitioners about variables that impact selection into training data. First, $\mathbb{P}(S = 1)$, the population probability of inclusion, needs to be quantified. If the study is randomized, a value $\approx 1/2$ is reasonable. Second, Γ, the largest deviation from $\mathbb{P}(S = 1)$, needs to be quantified. Our framework allows a practitioner-friendly choice of calibration parameters. In fact, both questions may be simply answered using domain knowledge.

(ii) Data-driven calibration: Although our framework enables practitioners to choose appropriate calibration parameters, we provide a fully data-driven approach for calibrating Γ and $\mathbb{P}(S = 1)$. To this end, we consider a setting in which samples from *one* of the covariates of the target population are provided. This is reasonable, since we often have limited understanding of the target population and, for instance, know covariates such as the distribution of gender or age in the target population. Once we are given one covariate, e. g., x_{age}, we proceed in two steps: (1) For calibrating $\mathbb{P}(S = 1)$, we approximate $\mathbb{P}(S = 1 \mid X, Y, T)$ via an estimate of $\mathbb{P}(S = 1 \mid x_{\text{age}})$ and, based on this, we approximate $\mathbb{P}(S = 1)$ by averaging over x_{age}, i. e., $\frac{1}{n} \sum_{i=1}^n \mathbb{P}(S_i = 1 \mid x_{\text{age},i})$. (2) For calibrating Γ, we take the maximum of the odds-ratio in (11) with the above estimates for $\mathbb{P}(S = 1)$ and $\mathbb{P}(S = 1 \mid x_{\text{age}})$ plugged in, which yields a value for Γ. We use this data-driven calibration procedure in our experiments (Section 5).

In case the uncertainty regarding the calibration parameters remains high, large values for Γ can be chosen, yielding a wide uncertainty set and conservative policies.

4 OPTIMIZING GENERALIZABLE POLICIES

In this section, we derive an efficient algorithm for optimizing the minimax policy value in (16). For this, we consider

a parameterized policy class $\Pi = \{\pi(\cdot, \theta) : \theta \in \Theta\}$ and the minimax problem

$$\min_{\theta \in \Theta} \max_{R \in \mathcal{R}} \frac{\sum_{i=1}^{n} R_i \psi_i(\theta)}{\sum_{i=1}^{n} R_i}, \tag{MMP}$$

where $\psi_i(\theta)$ denotes $\psi_i(\pi(\cdot, \theta))$ and corresponds to either (8), (9), or (10). The above minimax problem is non-trivial, since it is in general non-convex in θ. We first derive a closed-form solution of the worst-case policy value subproblem (Section 4.1). Then, based on this, we develop a tailored convex-concave procedure that solves (MMP) (Section 4.2).

4.1 CLOSED-FORM SOLUTION OF WORST-CASE POLICY VALUE

The solution of (MMP) involves the worst-case policy value subproblem in (15). We derive a closed-form solution of the subproblem and the corresponding Radon-Nikodým derivative at the optimal solution in Theorem 2.

Theorem 2. *(Closed-form solution of worst-case policy value.) Let (i) denote the ordering such that $\psi_{(1)}(\theta) \leqslant \ldots \leqslant \psi_{(n)}(\theta)$. Then, an optimal solution of the worst-case policy value subproblem (15) is given by*

$$\overline{V}_{\textit{Target}}(\pi; \mathcal{R}) = \frac{l \sum_{i=1}^{k^*} \psi_{(i)}(\theta) + u \sum_{i=k^*+1}^{n} \psi_{(i)}(\theta)}{lk^* + u(n - k^*)}, \tag{18}$$

with

$$k^* = \inf \Big\{ k \in \{0, \ldots, n\} : \tag{19}$$

$$\frac{l \sum_{i=1}^{k} \psi_{(i)}(\theta) + u \sum_{i=k+1}^{n} \psi_{(i)}(\theta)}{lk + u(n - k)} \leqslant \psi_{(k+1)}(\theta) \Big\}. \tag{20}$$

The Radon-Nikodým derivative at optimal solution is given by $R_{(i)} = l\mathbb{1}\{(i) \leqslant k^*\} + u\mathbb{1}\{(i) > k^*\}$.

See Appendix A.3 for a proof. Theorem 2 is appealing for two reasons: (i) We prove that $\overline{V}_{\text{Target}}(\pi; \mathcal{R})$ is efficiently solved by a linear search over the sorted data. (ii) We prove that the worst-case policy value is given by a maximum over a finite set, which we use in the following section to show that the minimax problem can be written as a difference-of-convex functions (DC) problem. Based on this, we develop a convex-concave procedure to efficiently solve the minimax problem in (MMP).

4.2 MINIMAX CONVEX-CONCAVE PROCEDURE

In this section, we develop the **mi**ni**m**ax **c**onvex-**c**oncave **p**rocedure (MMCCP) to efficiently solve the minimax problem (MMP). For this, we derive a DC-representation of the worst-case policy value based on its closed-form solution in Theorem 2. For this, the following assumptions are made.

Assumption 1. *The set Θ is nonempty, compact, and convex. Furthermore, π is a DC-function in θ, i. e., $\pi(X, \theta) = \tilde{g}(X, \theta) - \tilde{h}(X, \theta)$, where $\tilde{g}$ and $\tilde{h}$ are convex in θ, and differentiable.*

Note that Assumption 1 is very general as the class of DC-functions is very rich. For instance, it includes all twice continuously differentiable functions [Horst and Thoai, 1999]. We later show that Assumption 1 is fulfilled for the established policy class of logistic policies. First, we show that $\psi_i(\theta)$ can be written as a DC-function.

Lemma 1. *(DC-representation of $\psi_i(\theta)$.) Under Assumption 1, $\psi_i(\theta)$ is a DC-function in θ, i. e.,*

$$\psi_i(\theta) = g_i(\theta) - h_i(\theta), \tag{21}$$

where g_i and h_i are convex in θ.

See Appendix A.4 for a proof. Now, using Lemma 1 and Theorem 2, we prove that the worst-case policy value can be written as a DC-function.

Theorem 3. *(DC-representation of worst-case policy value.) Under Assumption 1, the worst-case policy value $\overline{V}_{\textit{Target}}(\pi; \mathcal{R})$ is a DC-function in θ, i. e.,*

$$\overline{V}_{\textit{Target}}(\pi; \mathcal{R}) = g(\theta) - h(\theta), \tag{22}$$

where $g(\theta)$ and $h(\theta)$ are convex and given by

$$g(\theta) = \max_{R \in \mathcal{R}} \frac{\sum_{i=1}^{n} R_i \psi_i(\theta)}{\sum_{i=1}^{n} R_i} + \sum_{i=1}^{n} h_i(\theta) c_i, \tag{23}$$

$$h(\theta) = \sum_{i=1}^{n} h_i(\theta) c_i, \tag{24}$$

with g_i and h_i from Lemma 1, and non-negative constants c_i for all i.

See Appendix A.5 for a proof. Finally, with the DC-representation of the worst-case policy value in Theorem 3, we can write the original minimax problem in (MMP) as a DC-program, i. e.,

$$\min_{\theta \in \Theta} g(\theta) - h(\theta), \tag{25}$$

where $g(\theta)$ and $h(\theta)$ are convex and given in Theorem 3. Hence, we can solve the minimax problem via a convex-concave procedure [Sriperumbudur and Lanckriet, 2009, Yuille and Rangarajan, 2003]. This yields our tailored MMCCP for solving (MMP) as outlined in Algorithm 1. Next, we prove that the sequence $(\theta^k)_{k \in \mathbb{N}}$ generated by MMCCP yields monotonically decreasing worst-case policy values and converges under mild assumptions.

Theorem 4. *(Theoretical Analysis of MMCCP.) Suppose the outcomes are bounded, i. e., $|Y| < C$, and Assumption 1 holds. Then, the following holds true:*

Algorithm 1 MMCCP

Input: Initial theta θ^0, convergence tolerance δ_{tol}
Set $k \leftarrow 0$
repeat
Solve the *convex* problem:
$\theta^{k+1} \in \arg\min_{\theta\in\Theta} \max_{R\in\mathcal{R}} \frac{\sum_{i=1}^n R_i \psi_i(\theta)}{\sum_{i=1}^n R_i} + \sum_{i=1}^n c_i(h_i(\theta) - \langle \theta, \nabla h_i(\theta^k)\rangle)$
Set $k \leftarrow k+1$
until $\|\theta^k - \theta^{k-1}\| < \delta_{\text{tol}}$

1. *The sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by MMCCP satisfies the monotonic descent property, i. e., for all $k \in \mathbb{N}$,*

$$\max_{R\in\mathcal{R}} \frac{\sum_{i=1}^n R_i \psi_i(\theta^{k+1})}{\sum_{i=1}^n R_i} \leqslant \max_{R\in\mathcal{R}} \frac{\sum_{i=1}^n R_i \psi_i(\theta^{k})}{\sum_{i=1}^n R_i}. \tag{26}$$

2. *If $\tilde{g}$ and $\tilde{h}$ from Assumption 1 are strongly convex,[5] then every limit point θ^* of $(\theta^k)_{k\in\mathbb{N}}$ is a stationary point[6] of (MMP). Furthermore, it holds: $\lim_{k\to\infty} \|\theta^{k+1} - \theta^k\| = 0$.*

See Appendix A.6 for a proof. To summarize, we develop a tailored convex-concave procedure that efficiently solves (MMP). This is only possible since we proved that the worst-case policy value has a DC-representation (see Theorem 3). In particular, our algorithm can be used on a rich class of policies and converges under mild assumptions. We now demonstrate that Assumption 1 holds for an established parameterized policy class which we use in our experiments.

Logistic policies: Logistic policies are defined by $\pi(X, \theta) = \sigma(\theta^\intercal X)$, where $\sigma(z) = 1/(1 + e^{-z})$. To find a DC-representation, it is sufficient to decompose $\sigma(z)$. Hence, we set $z = \theta^\intercal X$ and write

$$\tilde{g}_{\log}(z) = \begin{cases} \frac{1}{4}z + \frac{1}{2}, & \text{if } z \geqslant 0, \\ \frac{1}{2}\tanh(\frac{1}{2}z) + \frac{1}{2}, & \text{else,} \end{cases}, \tag{27}$$

$$\tilde{h}_{\log}(z) = \begin{cases} \frac{1}{4}z - \frac{1}{2}\tanh(\frac{1}{2}z), & \text{if } z \geqslant 0, \\ 0, & \text{else.} \end{cases} \tag{28}$$

It is straightforward to check that both functions are convex. They can be made strongly convex by adding $\frac{\lambda}{2}z^2$ to both functions. Since $\tilde{g}_{\text{Log}}$ and $\tilde{h}_{\text{Log}}$ are differentiable, Assumption 1 is fulfilled and, hence, MMCCP converges for logistic policies. In Appendix C, we show that Assumption 1 also holds for linear policies. In addition, logistic policies also satisfy the generalization bound in Theorem 1, since they have $\sqrt{n}$-vanishing Rademacher complexity. This can be seen by using that σ is Lipschitz together with the Rademacher bound for linear classes [Maurer, 2006] and the scalar concentration inequality [Maurer, 2016].

[5] A function f is strongly convex, if $\rho(f) > 0$, where $\rho(f)$ is the modulus of strong convexity of a convex function f, which is defined as $\rho(f) = \sup\{\rho \geqslant 0 : f(\cdot) - \frac{\rho}{2}\|\cdot\|_2^2 \text{ is convex}\}$.

[6] Note that the objective function is in general not differentiable, see Appendix D. Hence, we consider stationary points in the context of convex analysis, i. e., $0 \in \partial g(\theta^*) \cap \partial h(\theta^*)$, where ∂ denotes the subgradient.

5 EXPERIMENTS

In this section, we compare several policy learning methods to policies learned with our framework on the example of logistic policies. We demonstrate that our framework generalizes substantially better to the target population.

5.1 SIMULATION STUDY

We first consider a simulation study to demonstrate the effect of unrepresentative training data. For this, we consider the following data-generating process for the target population:

$$\mathbf{X} \sim \mathcal{N}_5(\mu, \mathbf{I}_5), \quad T \mid \mathbf{X} \sim \text{Bern}(1/2), \tag{29}$$

$$Y \mid (\mathbf{X}, T) = m(\mathbf{X}) + T \cdot C(\mathbf{X}) + \epsilon, \tag{30}$$

where $m(\mathbf{X}) = \beta_0^\intercal \mathbf{X} + 3\xi$, $C(\mathbf{X}) = 5/2 + \beta_1^\intercal \mathbf{X} - 4\xi$, $\xi \sim \text{Bern}(1/2)$, and $\epsilon \sim \mathcal{N}(0, 1)$. The covariate mean is $\mu = [-1, 1/2, -1, 0, -1]$ and the outcome means are $\beta_0 = [0, 3/4, -1/2, 0, -1]$ and $\beta_1 = [-3/2, 1, -3/2, 1, 1/2]$, respectively. We sample from the target population using the following selection variable

$$S \sim \text{Bern}\Big(\frac{1}{2} + \frac{0.95}{2}\tanh(-10\, C(\mathbf{X}))\Big), \tag{31}$$

which yields training data that is unrepresentative for the tail of the covariate distribution.[7] As baselines, we consider three established policy learning methods: the direct method (**DM**), the normalized inverse propensity weights method (**NIPW**), and the doubly robust method (**DR**). We compare these established methods against our generalizable methods with each of the three $\psi(\theta)$ in (8), (9), and (10): the worst-case policy value obtained with the direct method (**GenDM**), obtained with the normalized IPW (**GenNIPW**), and with the doubly robust method (**GenDR**). We use kernel and logistic regression for estimating $\mu_t(x)$ and W^{IPW}. The parameter $\mathbb{P}(S = 1)$ is chosen by the data-driven calibration in Section 3.4 and Γ is varied across $\{1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 3.0, \ldots, 10.0\}$. Details on implementation of MMCCP are in Appendix D.

We present the results for the different values of Γ in Figure 1. Specifically, we show by how much our methods improve over the policy regret, $V_{\text{Target}}(\hat{\pi}) - V_{\text{Target}}(\pi^*)$, of the corresponding baseline policy (i. e., DM, NIPW, and DR) when tested on the target population. Our methods achieve lower policy regrets on the target population across

[7] Note that we multiply the second term in (31) with $\frac{0.95}{2}$ to ensure that the selection probability remains strictly positive.

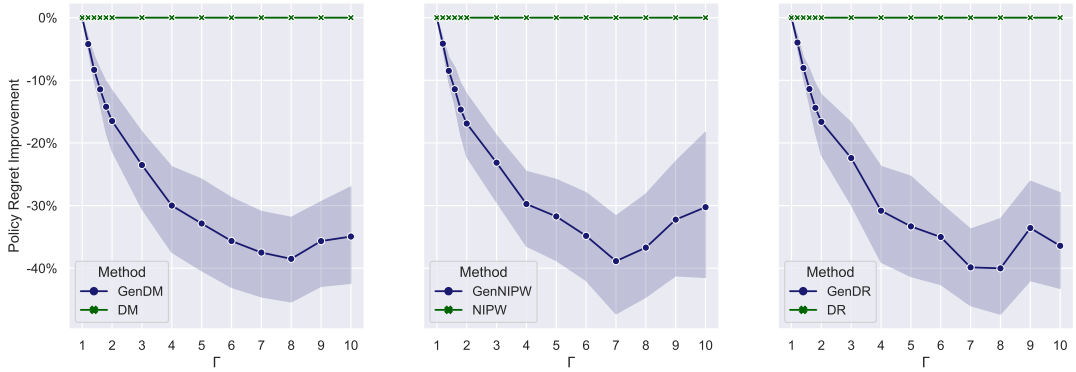

Figure 1: Policy regret improvement of our methods (blue) over the baseline methods (green) on the target population for different values of Γ. Compared to the baseline methods (i. e., DM, NIPW, and DR), our methods (i. e., GenDM, GenNIPW, and GenDR) show superior generalizability and, as such, improve the policy regret by up to 40 % at the true $\Gamma^* = 8$. Lower is better.

all methods and across all values of Γ. Specifically, relative to the policy regret of the baseline policy (green line), our methods (blue line) improve the policy regret on the target population by up to 40 %. By construction, for $\Gamma = 1$ (left end of plots), our methods resemble the baseline methods and yield the same policy regret. When we increase Γ, our policies achieve substantial improvements of the policy regret on the target population over the baselines. The best policy regret on the target population is achieved for $\Gamma = 8$, which is consistent with the simulation specifications, as the true $\Gamma^* = 8$. For $\Gamma = 8$, relative to the baseline policies, our methods improve the policy regret by up to 40 %. This demonstrates that policies learned with our framework generalize substantially better to the target population.

5.2 EXPERIMENTS ON CLINICAL TRIAL DATA

We evaluate our methods using the AIDS Clinical Trial Group (ACTG) study 175 [Hammer et al., 1996], which is particularly suited for evaluating our framework. This is because HIV-positive females tend to be underrepresented, which makes these studies not representative of the target population (i. e., the HIV-positive population in the USA) [Gandhi et al., 2005, Greenblatt, 2011]. In fact, in the ACTG 175 study, only 5.8 % of the patients are female, whereas HIV-positive females are more common in the USA population. The outcome Y is the difference between the cluster of differentiation 4 (CD4) cell counts at the beginning of the study and the CD4 counts after 20 ± 5 weeks. The average treatment effects on the male and female subgroups are -8.97 and -1.39, respectively, suggesting a large discrepancy in treatment effects between both subgroups. We consider two treatment arms: one treatment arm for both zidovudine (ZDV) and zalcitabine (ZAL) ($T = 1$) vs. one treatment arm for ZDV only ($T = 0$), comprising $1,056$ patients in total. We consider 12 covariates (details on the covariates are in Appendix B). Again, we compare our methods against the established baseline methods. This is a real-world clinical trial and, hence, we cannot access the true policy values on

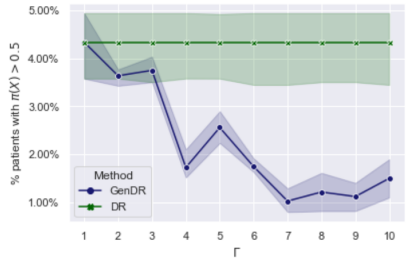

Figure 2: Percentage of patients with $\pi(X) > 0.5$ for our GenDR policy method. Fewer patients are treated for increasing Γ.

the target population. However, we investigate the behavior of our policies by studying the percentage of patients that are treated (i. e., $\pi(X) > 0.5$) for varying Γ. For our GenDR, the result is presented in Figure 2. The results for GenDM and GenNIPW are in Appendix E. We find that, compared to the baseline policy, our policy treats fewer patients for increasing Γ. This seems reasonable, since females are underrepresented and have a lower average treatment effect. Specifically, the baseline policy tends to treat more patients, since there are more patients in the study that benefit from the treatment. However, in the target population (with a greater proportion of females), fewer patients are expected to benefit (due to the lower treatment effect in the female subgroup). Our policy accounts for the underrepresentation of females and, as such, tends to treat fewer patients. This result indicates the potential of our framework for learning policies that generalize to the target population.

6 CONCLUSION

We propose a novel framework for learning policies that generalize to the target population by optimizes the minimax policy value on the target population. We prove that our framework yields policies that do not do worse on the target population than the worst-case policy value. We solve the minimax problem via a tailored convex-concave procedure for which we prove convergence for parametrized spaces of policies. Experiments demonstrate the benefit of learning generalizable policies using our framework.

References

I. Andrews and E. Oster. Weighting for external validity. Technical report, National Bureau of Economic Research, 2017.

P. M. Aronow and D. K. Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1), 2013.

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

A. Bennett and N. Kallus. Efficient policy learning from surrogate-loss classification reductions. In *International Conference on Machine Learning*. PMLR, 2020.

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

A. L. Buchanan, M. G. Hudgens, S. R. Cole, K. R. Mollan, P. E. Sax, E. S. Daar, A. A. Adimora, J. J. Eron, and M. J. Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 181(4):1193, 2018.

A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186, 1962.

S. R. Cole and E. A. Stuart. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American Journal of Epidemiology*, 172(1): 107–115, 2010.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*. Springer, 2008.

I. J. Dahabreh, S. E. Robertson, E. J. Tchetgen, E. A. Stuart, and M. A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

S. H. Downs and N. Black. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health*, 52(6):377–384, 1998.

J. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

M. Dudík, D. Erhan, J. Langford, L. Li, et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

L. E. Flores, W. R. Frontera, M. P. Andrasik, C. Del Rio, A. Mondríguez-González, S. A. Price, E. M. Krantz, S. A. Pergam, and J. K. Silver. Assessment of the inclusion of racial/ethnic minority, female, and older individuals in vaccine clinical trials. *JAMA Network Open*, 4(2): e2037640, 2021.

Mo. Gandhi, N. Ameli, P. Bacchetti, G. B. Sharp, A. L. French, M. Young, S. J. Gange, K. Anastos, S. Holman, A. Levine, et al. Eligibility criteria for hiv clinical trials and generalizability of results: the gap between published reports and study protocols. *Aids*, 19(16):1885–1896, 2005.

R. M. Greenblatt. Priority issues concerning hiv infection among women. *Women's Health Issues*, 21(6):S266–S271, 2011.

S. M. Hammer, D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, pages 153–161, 1979.

J. L. Hill and Y. S. Su. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Annals of Applied Statistics*, 7(3):1386–1420, 2013.

R. Horst and N. V. Thoai. Dc programming: overview. *Journal of Optimization Theory and Applications*, 103(1): 1–43, 1999.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

K. Imai, M. Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

N. Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, 2018.

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

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

A. Kube, S. Das, and P. J. Fowler. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *AAAI Conference on Artificial Intelligence*, 2019.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 2013.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *International Conference on Web Search and Data Mining*, 2011.

C. F. Manski. Anatomy of the selection problem. *Journal of Human resources*, pages 343–360, 1989.

A. Maurer. The rademacher complexity of linear transformation classes. In *International Conference on Computational Learning Theory*. Springer, 2006.

A. Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

L. W. Miratrix, S. Wager, and J. R. Zubizarreta. Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika*, 105(1):103–114, 2018.

W. Mo, Z. Qi, and Y. Liu. Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, pages 1–16, 2020.

S. L. Norris, M. M. Engelgau, and K. V. Narayan. Effectiveness of self-management training in type 2 diabetes: a systematic review of randomized controlled trials. *Diabetes Care*, 24(3):561–587, 2001.

D. N. Phan, H. M. Le, and H. A. Le Thi. Accelerated difference of convex functions algorithm and its application to sparse binary logistic regression. In *International Joint Conference on Artificial Intelligence*, 2018.

P. R. Rosenbaum. Overt bias in observational studies. In *Observational Studies*, pages 71–104. Springer, 2002.

P. M. Rothwell. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

D. B. Rubin. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

N. Si, F. Zhang, Z. Zhou, and J. Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning*, 2020.

B. K. Sriperumbudur and G. R. Lanckriet. On the convergence of the concave-convex procedure. In *Advances in Neural Information Processing Systems*, 2009.

E. A. Stuart, S. R. Cole, C. P. Bradshaw, and P. J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2): 369–386, 2011.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015.

P. D. Tao and L. T. H. An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

M. Uehara, M. Kato, and S. Yasui. Off-policy evaluation and learning for external validity under a covariate shift. In *Advances in Neural Information Processing Systems*, 2020.

A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

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

Y.-Q. Zhao, D. Zeng, C. M. Tangen, and M. L. LeBlanc. Robustifying trial-derived optimal treatment rules for a target population. *Electronic Journal of Statistics*, 13(1): 1717, 2019b.

A MATHEMATICAL APPENDIX

A.1 PROOF OF PROPOSITION 1

We know that, with the Radon-Nikodým derivative $R = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}_{\text{Train}}$,

$$V(\pi) = \mathbb{E}[Y(\pi)] = \mathbb{E}_{\text{Train}}[R\,Y^{\pi}]. \tag{31}$$

where

$$R = \frac{\mathrm{d}\mathbb{P}(X,T,Y)}{\mathrm{d}\mathbb{P}_{\text{Train}}(X,T,Y)} = \frac{\mathrm{d}\mathbb{P}(X,T,Y)}{\mathrm{d}\mathbb{P}(X,T,Y \mid S=1)} \tag{32}$$

$$= \frac{\mathrm{d}\mathbb{P}(X,T,Y)}{\mathrm{d}\mathbb{P}(X,T,Y,S=1)}\mathbb{P}(S=1) = \frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1 \mid X,T,Y)}. \tag{33}$$

□

Remark 1. With the above result for the Radon-Nikodým derivative, we can see the effect of the selection variable S: If S does not depend on X, T, and Y, then $R = 1$. Therefore, $\mathbb{P}$ would be identical to $\mathbb{P}_{\text{Train}}$ and, as a consequence, the policy value on the target population, i. e., $V_{\text{Target}}(\pi)$, would coincide with the policy value on the training data, i. e., $\mathbb{E}_{\text{Train}}[Y^{\pi}]$. If, however, S depends on X, T, and Y, then the policy value on the target population does not coincide with the policy value on the training data and, therefore, $V_{\text{Target}}(\pi) \neq \mathbb{E}_{\text{Train}}[Y^{\pi}]$.

A.2 PROOF OF THEOREM 1

Let $Z = \frac{1}{n}\sum_{i=1}^{n} R_i^*$. Then,

$$V_{\text{Target}}(\pi) \leqslant \hat{V}^*_{\text{Target}}(\pi) + \sup_{\pi\in\Pi}\left|\hat{V}^*_{\text{Target}}(\pi) - V_{\text{Target}}(\pi)\right|, \tag{34}$$

and

$$\sup_{\pi\in\Pi}\left|\hat{V}^*_{\text{Target}}(\pi) - V_{\text{Target}}(\pi)\right| \tag{35}$$

$$= \sup_{\pi\in\Pi}\left|\frac{\frac{1}{n}\sum_{i=1}^{n} R_i^*\psi_i(\pi)}{Z} - \frac{V_{\text{Target}}(\pi)}{Z} + \frac{V_{\text{Target}}(\pi)(1-Z)}{Z}\right| \tag{36}$$

$$\leqslant \frac{1}{Z}\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*\psi_i(\pi) - V(\pi)\right| + \sup_{\pi\in\Pi}\frac{C\left|1-Z\right|}{Z}. \tag{37}$$

We let

$$T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*\psi_i(\pi) - V_{\text{Target}}(\pi)\right|. \tag{38}$$

Since $|Y| \leqslant C$ and, therefore, $|\mu_t(x)| \leqslant C$, $R_i^* \leqslant u$, and $1-\eta \geqslant \pi^b(x) \geqslant \eta$ (for some $\eta > 0$ due to positivity), we have that

1.) for $\psi_i^{\mathrm{DM}}(\pi)$ from (8): $T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*(\pi(X_i)\mu_1(X_i) + (1-\pi(X_i))\mu_0(X_i)) - V(\pi)\right|$ satisfies bounded differences with $\frac{4Cu}{n}$,

2.) for $\psi_i^{\mathrm{NIPW}}(\pi)$ from (9): $T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*(\frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^{n} W_j^{\mathrm{IPW}}}(1-2T_i)(1-T_i-\pi(X_i))Y_i) - V(\pi)\right|$, satisfies bounded differences with $\frac{4Cu}{n}\frac{1-\eta}{\eta}$,

3.) for $\psi_i^{\mathrm{DR}}(\pi)$ from (10): $T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*(\psi_i^{\mathrm{DM}}(\pi) + W_i^{\mathrm{IPW}}(1-2T_i)(1-T_i-\pi(X_i))(Y_i-\mu_{T_i}(X_i))) - V(\pi)\right|$, satisfies bounded differences with $\frac{4Cu}{n}\frac{1+\eta}{\eta}$.

Hence, T satisfies bounded differences with $\frac{4Cu}{n}K_\psi$, where $K_\psi = 1$ for $\psi_i^{\mathrm{DM}}(\pi)$, $K_\psi = \frac{1-\eta}{\eta}$ for $\psi_i^{\mathrm{NIPW}}(\pi)$, and $K_\psi = \frac{1+\eta}{\eta}$ for $\psi_i^{\mathrm{DR}}(\pi)$.

Thus, using McDiarmid's inequality yields

$$\mathbb{P}\left(T - \mathbb{E}[T] \geqslant \epsilon\right) \leqslant \exp(-\frac{n\epsilon^2}{8C^2u^2K_\psi^2}). \tag{39}$$

Therefore, we have that

$$\mathbb{P}\left(T - \mathbb{E}[T] \leqslant \epsilon\right) \geqslant 1 - \exp(-\frac{n\epsilon^2}{8C^2u^2K_\psi^2}). \tag{40}$$

Using $p_1 = \exp(-\frac{n\epsilon^2}{8C^2u^2K_\psi^2})$ and, therefore, $\epsilon = 2CuK_\psi\sqrt{\frac{2\log(1/p_1)}{n}}$, we have that with probability at least $1 - p_1$,

$$T \leqslant \mathbb{E}[T] + 2CuK_\psi\sqrt{\frac{2\log(1/p_1)}{n}}. \tag{41}$$

Since $\mathbb{E}[R_i^*\psi_i(\pi)] = V(\pi)$, a standard symmetrization argument yields

$$\mathbb{E}[T] \leqslant \mathbb{E}\Big[\frac{1}{2^n}\Sigma_{\sigma\in\{-1,+1\}^n}\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i R_i^*\psi_i(\pi)\Big|\Big]. \tag{42}$$

Then, using the Rademacher comparison theorem (Thm 4.12 in Ledoux and Talagrand [2013]), this yields

$$\mathbb{E}[T] \leqslant 2CuK_\psi\mathbb{E}[\mathcal{R}_n(\Pi)], \tag{43}$$

where K_ψ is from above and depends on whether one uses $\psi_i^{\mathrm{DM}}(\pi)$, $\psi_i^{\mathrm{NIPW}}(\pi)$, or $\psi_i^{\mathrm{DR}}(\pi)$. Moreover, $\mathcal{R}_n(\Pi)$ satisfies bounded differences with constants $\frac{2}{n}$ and, hence, we can again use McDiarmid's inequality, which yields

$$\mathbb{P}\left(\mathbb{E}[\mathcal{R}_n(\Pi)] - \mathcal{R}_n(\Pi) \geqslant \epsilon\right) \leqslant \exp(\frac{-\epsilon^2 n}{2}). \tag{44}$$

Therefore, we have that

$$\mathbb{P}\left(\mathbb{E}[\mathcal{R}_n(\Pi)] - \mathcal{R}_n(\Pi) \leqslant \epsilon\right) \geqslant 1 - \exp(\frac{-\epsilon^2 n}{2}). \tag{45}$$

Using $p_2 = \exp(\frac{-\epsilon^2 n}{2})$ and, therefore, $\epsilon = \sqrt{\frac{2\log(1/p_2)}{n}}$, we have that with probability at least $1 - p_2$,

$$\mathbb{E}[\mathcal{R}_n(\Pi)] \leqslant \mathcal{R}_n(\Pi) + \sqrt{\frac{2\log(1/p_2)}{n}}. \tag{46}$$

The second term in (37) can be bounded using $0 \leqslant R_i^* \leqslant u$, $\mathbb{E}\left[R^*\right] = 1$, and Hoeffding's inequality:

$$\mathbb{P}\left(|1 - Z| \geqslant \epsilon\right) \leqslant 2\exp(-2\epsilon^2u^{-2}n). \tag{47}$$

Therefore, we have that

$$\mathbb{P}\left(|1 - Z| \leqslant \epsilon\right) \geqslant 1 - 2\exp(-2\epsilon^2u^{-2}n). \tag{48}$$

Using $p_3 = 2\exp(-2\epsilon^2u^{-2}n)$ and, therefore, $\epsilon = u\sqrt{\frac{\log(2/p_3)}{2n}}$, we have that with probability at least $1 - p_3$,

$$C|1 - Z| \leqslant Cu\sqrt{\frac{\log(2/p_3)}{2n}}. \tag{49}$$

Finally, using that $1/Z \leqslant 1/l$, we get that with probability at least $1 - p_1 - p_2 - p_3$,

$$\sup_{\pi\in\Pi}\left|\hat{V}^*_{\text{Target}}(\pi) - V_{\text{Target}}(\pi)\right| \leqslant 2C\frac{u}{l}K_\psi\mathcal{R}_n(\Pi) + 2C\frac{u}{l}K_\psi\sqrt{\frac{2\log(1/p_2)}{n}} + 2C\frac{u}{l}K_\psi\sqrt{\frac{2\log(1/p_1)}{n}} + C\frac{u}{l}\sqrt{\frac{\log(2/p_3)}{2n}}. \tag{50}$$

Let $p_1, p_2 = \delta/4$ and $p_3 = 2\delta/4$, then, using that $K_\psi \geqslant 1$, the above is bounded by $2C\frac{u}{l}K_\psi \mathcal{R}_n(\Pi) + 2C\frac{u}{l}K_\psi\sqrt{\frac{18\log(4/\delta)}{n}}$. The proof is completed by recognizing that, since the true $R^* \in \mathcal{R}$, we have that $\hat{V}^*_{\text{Target}}(\pi) \leqslant \overline{V}_{\text{Target}}(\pi)$. □

A.3 PROOF OF THEOREM 2

Let (i) denote the ith index of the increasing order statistics, an ordering where $\psi_{(1)}(\theta) \leqslant \ldots \psi_{(n)}(\theta)$. Hence, we address the following optimization problem

$$\max_R \frac{\sum_{i=1}^n R_{(i)}\psi_{(i)}(\theta)}{\sum_{i=1}^n R_{(i)}} \quad \text{s. t.,} \quad l \leqslant R_{(i)} \leqslant u,\ R_{(i)} \geqslant 0, \forall i = 1, \ldots, n. \tag{51}$$

We derive a closed-form solution for any of the $\psi_i(\theta)$ in (8), (9), and (10), which generalizes the solution of Kallus and Zhou [2018] to all standard policy learning methods. Since the constraint on R is linear, the above optimization problem is a linear fractional program. Hence, we can use the Charnes-Cooper transformation [Charnes and Cooper, 1962] with $\tilde{R}_{(i)} = R_{(i)}/\sum_{i=1}^n R_{(i)}$ and $t = 1/\sum_{i=1}^n R_{(i)}$, which yields

$$\max_{\tilde{R}} \sum_{i=1}^n \tilde{R}_{(i)}\psi_{(i)}(\theta) \quad \text{s.t.} \quad t\,l \leqslant \tilde{R}_{(i)} \leqslant t\,u,\ \tilde{R}_{(i)} \geqslant 0 \sum_{i=1}^n \tilde{R}_{(i)} = 1, t \geqslant 0, \forall i = 1, \ldots, n \tag{52}$$

The corresponding dual problem has the dual variables $\lambda \in \mathbb{R}$ for the normalization constraint and $w, v \in \mathbb{R}^n_+$ for the box constraints on the normalized Radon-Nikodým derivative. It is given by

$$\min_{\lambda,v,w} \lambda, \text{ s.t. } \sum_{i=1}^n v_{(i)}\,u + w_{(i)}\,l \geqslant 0,\ \lambda + v_{(i)} - w_{(i)} \geqslant \psi_{(i)}(\theta), \forall i = 1, \ldots, n\, \lambda \in \mathbb{R}, v, w \in \mathbb{R}^n_+. \tag{53}$$

At the optimal solution, only one of the primal weight bound constraints, (for nontrivial bounds $l < u$), $t\,l \leqslant R_{(i)}$ or $R_{(i)} \leqslant t\,u$ will be tight. At the optimal solution, by complementary slackness, either none or one of the nonbinding primal constraints is nonzero, i. e., either $v_{(i)}$, $w_{(i)}$, or none is nonzero. Moreover, $t = 0$ is infeasible, since $t = 0$ would imply $\tilde{R}_{(i)} = 0$ for all i, which contradicts $\sum_{i=1}^n \tilde{R}_{(i)} = 1$. Hence, $t \neq 0$. At the optimal solution, the constraint $\sum_{i=1}^n v_{(i)}\,u + w_{(i)}\,l \geqslant 0$ must be active. Otherwise, we can find a λ which is smaller than the optimal one but still feasible, and hence contradicts the optimality. As a result, at an optimal solution, we have that:

$$\sum_{i=1}^n v_{(i)}\,u + w_{(i)}\,l = 0, \tag{54}$$

$$v_{(i)} - w_{(i)} = \psi_{(i)}(\theta) - \lambda, \forall i = 1, \ldots, n. \tag{55}$$

Since $v, w \geqslant 0$, we see the following by distinction of cases. If $\psi_{(i)}(\theta) \geqslant \lambda$, then $w_{(i)} = 0$ and $v_{(i)} = \psi_{(i)}(\theta) - \lambda$. If $\psi_{(i)}(\theta) < \lambda$, then $v_{(i)} = 0$ and $w_{(i)} = \lambda - \psi_{(i)}(\theta)$.

At optimality, since (i) is the increasing order statistics, there exists some index $k \in \{1, \ldots, n\}$ such that $\psi_{(k)}(\theta) < \lambda \leqslant \psi_{(k+1)}(\theta)$. Hence, we can substitute the solution from (55) in (54) and obtain the following

$$\sum_{i=1}^k l(\lambda - \psi_{(i)}(\theta)) - \sum_{i=k+1}^n u(\psi_{(i)}(\theta) - \lambda) = 0, \tag{56}$$

and, therefore,

$$\lambda(k) = \frac{l\sum_{i=1}^k \psi_{(i)}(\theta) + u\sum_{i=k+1}^n \psi_{(i)}(\theta)}{k\,l + (n-k)\,u}. \tag{57}$$

The optimal k is given by $k^* = \inf\{k : \lambda(k) \leqslant \psi_{(k+1)}(\theta)\}$, which can be seen by the following argument. When $\lambda(k)$ is

maximal, we have that $\lambda(k) \geqslant \lambda(k+1)$. This is equivalent to $\lambda(k) \leqslant \psi_{(k+1)}(\theta)$, since the following steps are equivalent

$$0 \geqslant \lambda(k+1) - \lambda(k) \tag{58}$$
$$0 \geqslant \frac{(lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta)}{l(k+1) + u(n-k-1)} - \lambda(k) \tag{59}$$
$$0 \geqslant (lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta) - \lambda(k)(l(k+1) + u(n-k-1)) \tag{60}$$
$$0 \geqslant (lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta) - \lambda(k)(lk + u(n-k)) - \lambda(k)(l+u) \tag{61}$$
$$0 \geqslant (l-u)\psi_{(k+1)}(\theta) - \lambda(k)(l-u) \tag{62}$$
$$\lambda(k) \leqslant \psi_{(k+1)}(\theta), \tag{63}$$

where the last inequality switches because we divide by $l - u$ which is negative. Next, we show that if $\lambda(k) \geqslant \lambda(k+1)$, then $\lambda(k+1) \geqslant \lambda(k+2)$.

$$\lambda(k+1) \tag{64}$$
$$= \frac{(lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta)}{l(k+1) + u(n-k-1)} \tag{65}$$
$$\leqslant \frac{(lk + u(n-k))\psi_{(k+1)}(\theta) + (l-u)\psi_{(k+1)}(\theta)}{l(k+1) + u(n-k-1)} \tag{66}$$
$$= \psi_{(k+1)}(\theta) \leqslant \psi_{(k+2)}(\theta), \tag{67}$$

and, since we showed above that $\lambda(k) \geqslant \lambda(k+1)$ is equivalent to $\lambda(k) \leqslant \psi_{(k+1)}(\theta)$, we have that $\lambda(k+1) \leqslant \psi_{(k+2)}(\theta)$ is equivalent to $\lambda(k+1) \geqslant \lambda(k+2)$. Thus, $k^* = \inf\{k : \lambda(k) \leqslant \psi_{(k+1)}(\theta)\}$. Hence, the solution of the dual problem is $\tilde{R}_{(i)} = \frac{l\mathbb{1}((i) \leqslant k^*) + u\mathbb{1}((i) > k^*)}{k^*\, l + (n-k^*)\, u}$. Then, the solution of the primal problem can be recovered by $R = \frac{1}{t}\tilde{R}$, where $t = 1/(k^*\, l + (n-k^*)\, u)$. □

A.4 PROOF OF LEMMA 1

For the direct method, we have

$$\psi_i^{\mathrm{DM}}(\theta) = \pi(X_i, \theta)\mu_1(X_i) + (1 - \pi(X_i, \theta))\mu_0(X_i) \tag{68}$$
$$= (\tilde{g}(X_i, \theta) - \tilde{h}(X_i, \theta))\mu_1(X_i) + (1 - \tilde{g}(X_i, \theta) + \tilde{h}(X_i, \theta))\mu_0(X_i). \tag{69}$$

To derive g_i and h_i, we proceed with a case distinction.

Case 1: $\mu_0(X_i) \geqslant 0$ and $\mu_1(X_i) \geqslant 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{g}(X_i, \theta)\mu_1(X_i) + \tilde{h}(X_i, \theta)\mu_0(X_i) + \mu_0(X_i)) \tag{70}$$
$$- (\tilde{h}(X_i, \theta)\mu_1(X_i) + \tilde{g}(X_i, \theta)\mu_0(X_i)), \tag{71}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{g}(X_i, \theta)\mu_1(X_i) + \tilde{h}(X_i, \theta)\mu_0(X_i) + \mu_0(X_i) \tag{72}$$
$$h_i(\theta) = \tilde{h}(X_i, \theta)\mu_1(X_i) + \tilde{g}(X_i, \theta)\mu_0(X_i). \tag{73}$$

Case 2: $\mu_0(X_i) < 0$ and $\mu_1(X_i) \geqslant 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{g}(X_i, \theta)\mu_1(X_i) + \tilde{g}(X_i, \theta)|\mu_0(X_i)| - |\mu_0(X_i)|) \tag{74}$$
$$- (\tilde{h}(X_i, \theta)\mu_1(X_i) + \tilde{h}(X_i, \theta)|\mu_0(X_i)|), \tag{75}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{g}(X_i, \theta)\mu_1(X_i) + \tilde{g}(X_i, \theta)|\mu_0(X_i)| - |\mu_0(X_i)| \tag{76}$$
$$h_i(\theta) = \tilde{h}(X_i, \theta)(\mu_1(X_i) + |\mu_0(X_i)|). \tag{77}$$

Case 3: $\mu_0(X_i) \geqslant 0$ and $\mu_1(X_i) < 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)\mu_0(X_i) + \mu_0(X_i)) \tag{78}$$
$$- (\tilde{g}(X_i,\theta)|\mu_1(X_i)| + \tilde{g}(X_i,\theta)\mu_0(X_i)), \tag{79}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)\mu_0(X_i) + \mu_0(X_i) \tag{80}$$
$$h_i(\theta) = \tilde{g}(X_i,\theta)(|\mu_1(X_i)| + \mu_0(X_i)). \tag{81}$$

Case 4: $\mu_0(X_i) < 0$ and $\mu_1(X_i) < 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{g}(X_i,\theta)|\mu_0(X_i)| - |\mu_0(X_i)|) \tag{82}$$
$$- (\tilde{g}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)|\mu_0(X_i)|), \tag{83}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{g}(X_i,\theta)|\mu_0(X_i)| - |\mu_0(X_i)| \tag{84}$$
$$h_i(\theta) = \tilde{g}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)|\mu_0(X_i)|. \tag{85}$$

For the normalized inverse propensity weights method, we have

$$\psi_i^{\mathrm{NIPW}}(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(1-2T_i)(1-T_i-\pi(X_i,\theta))Y_i. \tag{86}$$

Again, by a case distinction, we yield for $T_i = 1$:

$$\psi_i^{\mathrm{NIPW}}(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\pi(X_i,\theta)Y_i \tag{87}$$
$$= \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{g}(X_i,\theta)Y_i - \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{h}(X_i,\theta)Y_i, \tag{88}$$

and, hence, the claim follows with

$$g_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{g}(X_i,\theta)Y_i \tag{89}$$
$$h_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{h}(X_i,\theta)Y_i. \tag{90}$$

For $T_i = 0$, we derive,

$$\psi_i^{\mathrm{NIPW}}(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(1-\pi(X_i,\theta))Y_i \tag{91}$$
$$= \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(1-\tilde{g}(X_i,\theta)+\tilde{h}(X_i,\theta))Y_i, \tag{92}$$

and, hence, the claim follows with

$$g_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(\tilde{h}(X_i,\theta)+1)Y_i \tag{93}$$
$$h_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{g}(X_i,\theta)Y_i. \tag{94}$$

For the doubly robust method, we can use the decomposition of the direct method. By defining

$$\nu_i = (1 - 2T_i)(Y_i - \mu_{T_i}(X_i)), \tag{95}$$

and rewriting

$$W_i^{\mathrm{IPW}}(1 - 2T_i)(1 - T_i - \pi(X_i, \theta))(Y_i - \mu_{T_i}(X_i)) = W_i^{\mathrm{IPW}}(1 - T_i - \pi(X_i, \theta))\nu_i, \tag{96}$$

we proceed again by a case distinction for the rest. For $\nu_i \geqslant 0$ we have

$$\begin{aligned} W_i^{\mathrm{IPW}}(1 - T_i - \pi(X_i, \theta))\nu_i &= W_i^{\mathrm{IPW}}(1 - T_i)\nu_i + W_i^{\mathrm{IPW}}\nu_i\tilde{h}(X_i, \theta) && (97)\\ &\quad - W_i^{\mathrm{IPW}}\nu_i\tilde{g}(X_i, \theta). && (98) \end{aligned}$$

For $\nu_i < 0$, we have that

$$\begin{aligned} W_i^{\mathrm{IPW}}(1 - T_i - \pi(X_i, \theta))\nu_i &= W_i^{\mathrm{IPW}}(1 - T_i)\nu_i + W_i^{\mathrm{IPW}}|\nu_i|\tilde{g}(X_i, \theta) && (99)\\ &\quad - W_i^{\mathrm{IPW}}|\nu_i|\tilde{h}(X_i, \theta). && (100) \end{aligned}$$

and, hence, the claim follows. □

A.5 PROOF OF THEOREM 3

By Theorem 2, we know that

$$\max_{R \in \mathcal{R}} \frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i} = \max_{R \in \mathcal{S} \subseteq \mathcal{R}} \frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}, \tag{101}$$

with $|\mathcal{S}| = n + 1 < \infty$. Hence, we can write the inner maximum as

$$\max_{j \in \mathcal{J}} \frac{\sum_{i=1}^n R_i^j\psi_i(\theta)}{\sum_{i=1}^n R_i^j}, \tag{102}$$

where R^j for $j \in \mathcal{J} = \{0, \ldots, n\}$ denotes one of the $n + 1$ possible assignments of l and u, i. e., for $j = 0$, it is the vector with all entries equal to l; for $j = 1$, it is the vector with all entries equal to l except for the first one being u and so on. By defining the convex functions

$$G^j(\theta) = \frac{\sum_{i=1}^n R_i^j g_i(\theta)}{\sum_{i=1}^n R_i^j}, \tag{103}$$

$$H^j(\theta) = \frac{\sum_{i=1}^n R_i^j h_i(\theta)}{\sum_{i=1}^n R_i^j}, \tag{104}$$

we have

$$\frac{\sum_{i=1}^n R_i^j\psi_i(\theta)}{\sum_{i=1}^n R_i^j} = G^j(\theta) - H^j(\theta), \tag{105}$$

and, hence,

$$\begin{aligned} \max_{j \in \mathcal{J}} \frac{\sum_{i=1}^n R_i^j\psi_i(\theta)}{\sum_{i=1}^n R_i^j} &= \max_{j \in \mathcal{J}}\{G^j - H^j\} && (106)\\ &= \max_{j \in \mathcal{J}}\{G^j + \sum_{\substack{k=1\\k\neq j}}^n H^k - \sum_{k=1}^n H^k\} && (107)\\ &= \underbrace{\max_{j \in \mathcal{J}}\{G^j + \sum_{\substack{k=1\\k\neq j}}^n H^k\}}_{=:g} - \underbrace{\sum_{k=1}^n H^k}_{=:h}. && (108) \end{aligned}$$

Note that g and h are convex as the sum of convex functions is convex and the maximum of convex functions is convex. Now, g can be rewritten as follows

$$g(\theta) = \max_{j\in\mathcal{J}}\{G^j + \sum_{\substack{k=1\\k\neq j}}^{n} H^k\} = \max_{j\in\mathcal{J}}\{G^j - H^j + \sum_{k=1}^{n} H^k\} \tag{109}$$

$$= \max_{j\in\mathcal{J}}\{G^j - H^j\} + \sum_{k=1}^{n} H^k = \max_{R\in\mathcal{R}}\left\{\frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}\right\} + \sum_{k=1}^{n} H^k \tag{110}$$

$$= \max_{R\in\mathcal{R}}\left\{\frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}\right\} + h. \tag{111}$$

Furthermore, we can use the special structure of the worst case policy solutions to rewrite h as

$$h = \sum_{k=1}^{n} H^k = \sum_{k=1}^{n} \frac{\sum_{i=1}^n R_i^k h_i(\theta)}{\sum_{i=1}^n R_i^k} = \sum_{k=1}^{n}\sum_{i=1}^{n} \frac{R_i^k}{\sum_{i=1}^n R_i^k} h_i(\theta) \tag{112}$$

$$= \sum_{i=1}^{n} h_i(\theta) \underbrace{\sum_{k=1}^{n} \frac{R_i^k}{\sum_{i=1}^n R_i^k}}_{=:c_i} = \sum_{i=1}^{n} h_i(\theta) c_i, \tag{113}$$

where c_i can be calculated as

$$c_i = l\left(\sum_{k=1}^{i} \frac{1}{(n-k+1)l + (k-1)u}\right) + u\left(\sum_{k=i+1}^{n} \frac{1}{(n-k+1)l + (k-1)u}\right), \tag{114}$$

for all i by combinatorial arguments.

A.6 PROOF OF THEOREM 4

The convergence analysis of MMCCP follows from the convergence analysis of the DC-algorithm (DCA) [Tao and An, 1997]. More precisely, DCA for minimizing a function $f = g - h$ reduces to the convex-concave procedure in case that the function h is differentiable [Phan et al., 2018, Sriperumbudur and Lanckriet, 2009]. This is exactly what we have in our case, as by our assumption on $\tilde{g}$ and $\tilde{h}$ we have that each h_i (as a linear combination of differentiable functions) is differentiable and, hence, h is differentiable.

Now, 1. in Theorem 4 directly follows from (i) of Theorem 3 in Tao and An [1997]. For 2. in Theorem 4, we have to proof the following:

1. $\inf_{\theta\in\Theta} \max_{R\in\mathcal{R}} \frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}$ is finite.
2. It holds $\rho(g) + \rho(h) > 0$.
3. $(\theta^k)_{k\in\mathbb{N}}$ is bounded.

Ad Item 1: Since $|Y| \leqslant C$, we have that $|\mu_t(X_i)| \leqslant C$. Also, the rest of the terms involved in each of the three cases for ψ_i are bounded constants, and $l \leqslant R_i \leqslant u$ for all $i \in \{1,\dots,n\}$. Hence, since $\pi(\cdot,\theta) \in [0,1]$, we have that Item 1 holds true.

Ad Item 2: For all $i \in \{1,\dots,n\}$, we have in each of the three cases for ψ_i, that h_i is, up to a constant, a linear combination of $\tilde{g}$ and $\tilde{h}$ with positive weights. By our assumptions, we have that $\rho(\tilde{g}) > 0$ and $\rho(\tilde{h}) > 0$ and, hence, $\rho(h_i) > 0$. By Theorem 3, we have that $h = \sum_{i=1}^n h_i c_i$ with non-negative weights c_i, which yields $\rho(h) > 0$. Item 2 follows by observing that $\rho(g) \geqslant 0$.

Ad Item 3: Follows directly from Assumption 1.

Then, 2. in Theorem 4 follows by (iii) and (iv) of Theorem 3 in Tao and An [1997]. □

B DETAILS ON COVARIATES IN THE ACTG 175 STUDY

The ACTG 175 study assigned four treatments randomly to 2,139 subjects with human immunodeficiency virus (HIV) type 1, whose CD4 counts were 200–500 cells/mm^3. The four treatments that were compared are the zidovudine (ZDV)

monotherapy, the didanosine (ddI) monotherapy, the ZDV combined with ddI, and the ZDV combined with zalcitabine (ZAL).

There are 5 continuous covariates: age (year), weight (kg, coded as wtkg), CD4 count (cells/mm^3) at baseline, Karnofsky score (scale of 0-100, coded as karnof), CD8 count (mm^3) at baseline. They are centered and scaled before further analysis. In addition, there are 7 binary variables: gender (1 = male, 0 = female), homosexual activity (homo, 1 = yes, 0 = no), race (1 = nonwhite, 0 = white), history of intravenous drug use (drug, 1 = yes, 0 = no), symptomatic status (symptom, 1 = symptomatic, 0 = asymptomatic), antiretroviral history (str2, 1 = experienced, 0 = naive) and hemophilia (hemo, 1 = yes, 0 = no).

C ASSUMPTION 1 FOR LINEAR POLICIES

Linear policies are defined by $\pi(X, \theta) = \sigma(\theta^\intercal X)$, where $\sigma(z) = \min(1, \max(z, 0))$. A DC-representation for $\sigma(z)$, with $z = \theta^\intercal X$, is given by

$$\tilde{g}_{\text{Lin}}(z) = \max(z, 0), \tag{115}$$

$$\tilde{h}_{\text{Lin}}(z) = \max(\max(z, 0) - 1, 0). \tag{116}$$

It is straightforward to check that both functions are convex. Again, they can be made strongly convex by adding $\frac{\lambda}{2}z^2$ to both functions. Note however, that $\tilde{g}_{\text{Lin}}$ is not differentiable in 0 and $\tilde{h}_{\text{Lin}}$ is not differentiable in $\{0, 1\}$. As a remedy, one can set $\Theta_i^\epsilon = \{\theta \in \mathbb{R}^d : \epsilon \leqslant \theta^\intercal X_i \leqslant 1 - \epsilon\}$ for an $\epsilon > 0$ and define $\Theta^\epsilon = \bigcap_{i=1}^n \Theta_i^\epsilon$. The intersection has to be nonempty to make this approach work.

D IMPLEMENTATION DETAILS

Our code is available at `github.com/anonymous/GeneralOPL` (link anonymized for peer-review; code for review in the supplements.) For our experiments, we used the policy class of logistic policies as introduced in the main paper. To fulfill Assumption 1, we choose Θ to be a hypercube with large bounds to ensure a large enough search space, i. e., $\Theta = [-10,000; 10,000]^d$. In order to solve the subproblems in MMCCP, we draw upon the L-BFGS-B algorithm implemented in the open-source Python library SciPy. At this point, we note that the subproblems are convex but not necessarily differentiable, as the point-wise maximum of differentiable functions is not necessarily differentiable. However, logistic policies are continuously differentiable and the above choice for Θ is compact. Hence, the functions $\psi_i(\theta)$ are Lipschitz and, thus,

$$\max_{R \in \mathcal{R}} \frac{\sum_{i=1}^n R_i \psi_i(\theta)}{\sum_{i=1}^n R_i} \tag{117}$$

is Lipschitz as the point-wise maximum of Lipschitz functions. By Rademacher's theorem, (117) is therefore almost everywhere differentiable. The points θ where (117) is not differentiable are given by the points in which the maximizing argument R changes. Due to this fact, we find empirically that L-BFGS-B can efficiently solve these subproblems. The rest of the parameters are set as follows. The parameter for the stopping criterion is set to δ_{tol} to 10^{-4}. In order to make $\tilde{g}$ and $\tilde{h}$ strongly convex, λ is set to 10^{-3}. In every run, the starting points are initialized via a normal distribution, i. e., $\theta^0 \sim \mathcal{N}_d(\mathbf{0}_d, 0.1 \cdot \mathbf{I}_d)$. For each method, we ran our algorithm 5 times on the datasets.

We run all of our experiments on a server with two 16 Core Intel Xeon Gold 6242 processors each with 2.8GHz, and 192GB of RAM.

E RESULTS FOR GENDM AND GENNIPW ON ACTG 175 STUDY

We present the results on the ACTG 175 study for our method GenDM, which uses $\psi_i^{\text{DM}}(\pi)$ from (8) and our method GenNIPW, which uses $\psi_i^{\text{NIPW}}(\pi)$ from (9). Analogously to Section 5.2, we study the percentage of patients that are treated (i. e., $\pi(X) > 0.5$) for varying Γ. The results are presented in Figure 3. Similar to the results for GenDR in Section 5.2, we find that compared to the baseline policy, our policy treats fewer patients for increasing Γ. GenNIPW shows little variance across several runs on the dataset. For each run, GenNIPW obtains different, but similar θ. However, the percentage of patients treated remains consistent across different runs.

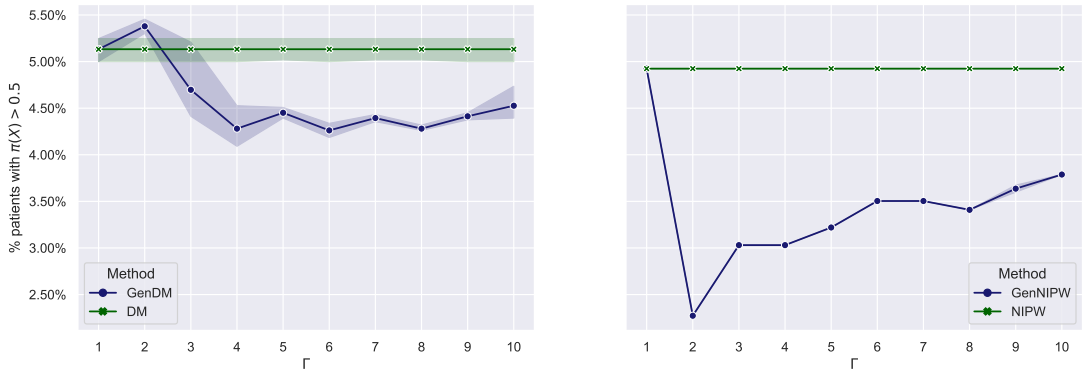

Figure 3: Percentage of patients with $\pi(X) > 0.5$ for our GenDM and GenNIPW policy method. Fewer patients are treated for increasing Γ.