Off-Policy Evaluation with Policy-Dependent Optimization Response

Wenshuo Guo*
Department of EECS
University of California, Berkeley
wguo@cs.berkeley.edu

Michael I. Jordan*
Department of EECS and Department of Statistics
University of California, Berkeley
jordan@cs.berkeley.edu

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

Abstract

The intersection of causal inference and machine learning for decision-making is rapidly expanding, but the default decision criterion remains an *average* of individual causal outcomes across a population. In practice, various operational restrictions ensure that a decision-maker's utility is not realized as an *average* but rather as an *output* of a downstream decision-making problem (such as matching, assignment, network flow, minimizing predictive risk). In this work, we develop a new framework for off-policy evaluation with *policy-dependent* linear optimization responses: causal outcomes introduce stochasticity in objective function coefficients. Under this framework, a decision-maker's utility depends on the policy-dependent optimization, which introduces a fundamental challenge of *optimization* bias even for the case of policy evaluation. We construct unbiased estimators for the policy-dependent estimand by a perturbation method, and discuss asymptotic variance properties for a set of adjusted plug-in estimators. Lastly, attaining unbiased policy evaluation allows for policy optimization: we provide a general algorithm for optimizing causal interventions. We corroborate our theoretical results with numerical simulations.

1 Introduction

The interface of causal inference and machine learning offers to "deliver the right intervention, at the right time, to the right person". An extensive line of research studies off-policy evaluation (OPE) and learning—evaluating the average causal outcomes under alternative personalized treatment assignment policies that differ from the treatment assignment which generated the data (and may have introduced confounding), so that one may optimize over the best such treatment rule [Manski, 2004, Dudík et al., 2011, Zhao et al., 2012, Thomas et al., 2015, Athey et al., 2017, Kitagawa and Tetenov, 2018, Kallus and Zhou, 2018b]. Most of this work is based on the assumption that the appropriate decision criterion is an *average* of individuals across a population. But various operational restrictions or settings imply that a decision-maker's utility is often not realized as an *average* but rather as an *output* of a downstream planning or decision-making problem.

For example, in studying the effects of price incentives in a matching market (e.g., on a ride-share platform), a firm's revenue is not realized until it matches riders to drivers under certain

*Authors listed in alphabetical order.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

constraints [Mejia and Parker, 2021, Ma et al., 2021]. While the marketplace may offer incentives to drive or accept rides and induce causal effects on individuals, the final utility is determined by the *new* matches, taking into account operational constraints and structure.

As another example, although job training (and personalized provision thereof) is commonly touted in causal inference and machine learning papers as a promising example for personalized treatment policy assignment [Athey et al., 2017, Kitagawa and Tetenov, 2018, Knaus et al., 2020], labor economists voice a general concern that "the possible existence of equilibrium effects on the efficiency of the programs seems quite real" [Crépon and Van Den Berg, 2016, p.541]. The equilibrium concern is that personalized provision of job training may not lead to actual beneficial gains at *the population* level due to externalities (substitution effects/congestion in matching) of the labor search process in a finite market. While impressive cluster-randomized trials have been deployed to assess these effects [Crépon et al., 2013], it would be useful if there exists a framework that can model the equilibrium effect and evaluate treatment policies directly based on available data of individual-level causal effects. In some settings, population-level impacts may be well-modeled as a downstream optimization response. The development of such a framework is our focus in the current paper.

We study a new framework for policy evaluation and optimization where there is a *personalized treatment policy* on individual-level outcomes, and a *policy-dependent* optimization response. The key difference between this model and previous work on off-policy evaluation and optimization is that: although treatments realize causal effects on *individuals*, a treatment policy's value depends on a further downstream *policy-dependent* optimization. We study how to evaluate different policies without bias (off-policy evaluation) and how to optimize for the optimal policy under this framework (policy optimization).

Our contributions are as follows: we first introduce the model of policy-dependent optimization response[2], which we formulate as a nonconvex stochastic optimization problem. For off-policy evaluation, we develop a framework of *policy-dependent optimization response*, decompose the bias that arises in this framework ("optimization bias") and show how to control it via the design of estimators for the policy-dependent estimand. Finally, we provide a general algorithm for optimizing causal interventions. We corroborate the theoretical results with experimental comparisons.

1.1 Related work

We highlight the most relevant work from causal inference, off-policy evaluation, and optimization under uncertainty in the main text. We include additional or tangential discussion in Appendix A.

There is an extensive literature on off-policy evaluation and optimization [see, e.g., Manski, 2004, Dudík et al., 2011, Zhao et al., 2012, Swaminathan and Joachims, 2015]. Relative to this line of work, we focus on the introduction of a downstream decision response, arising for example from operational constraints.

The case of constrained policies has been considered in the OPE literature. Our setting is conceptually different but overlaps in some application contexts. Specifically, we decouple the downstream *policy-dependent response*, i.e. over a similar constraint space, from treatment decisions that have causal effects. For example, Bhattacharya [2009] studies the setting of "roommate assignment" with discrete types; i.e., perfect bipartite matching. A crucial difference is that in their setting, the causal treatment of interest *is* the assignment decision to the other individual type; while in our setting the causal treatment only affects certain *parameters* of the assignment decision, such as edge costs. We instantiate an analogous example in our framework to highlight our decoupled causal intervention and prediction decisions. Consider a setting with a causal intervention, such as a diversity information intervention affecting a student's probability of getting along with various types. Here the policy performs treatments on individual's diversity information, and the final assignment decision is policy-dependent response. Other work considers resource-budgeted allocation Kube et al. [2019], which is structured because of reformulation of thresholds Lopez et al. [2020]. Sun [2021] studies sharp asymptotics for the additional challenge of stochasticity in the budget.

[2]For terminology, we use policy-dependent (optimization) response or downstream policy-dependent response to refer to the same concept.

Some works that illustrate the embedding of causal effect estimates in optimization-based decision problems include Rahmattalabi et al. [2022], although their formulation is ultimately a mixed-integer optimization.

In contrast to an extensive line of work on heterogeneous causal effect estimation [Shalit et al., 2017, Wager and Athey, 2018, Künzel et al., 2019], often crucially leveraging simpler structure of the treatment contrast rather than the conditional outcomes, in this work we require estimation of the latter due to the downstream optimization and distributional convergence for the perturbation method. In turn, combining causal outcome estimation with adjustments for optimization bias requires different properties of the estimation strategy, namely plug-in estimation of a modified regression model; we focus on estimators that modify the first-order conditions of a regression model to algebraically achieve an AIPW-type adjustment as discussed in Bang and Robins [2005]. See also Scharfstein et al. [1999], Tran et al. [2019], Shi et al. [2019], Chernozhukov et al. [2021].

This work focuses on the challenge of *optimization bias* for policy evaluation introduced in our setting, for *generic* linear optimization problems. This well-known challenge of in-sample optimization bias ("sample average approximation bias") fundamentally demarcates the statistical regime of optimization under uncertainty from sample mean estimation [Bayraksan and Morton, 2006, Shapiro et al., 2021]. Recent work develops bagging, jackknife, perturbation and variance-corrected perturbation approaches for bias adjustment [Lam and Qian, 2018, Ito et al., 2018, Kannan et al., 2020, Gupta et al., 2021]. We extend a perturbation method of Ito et al. [2018] to the setting of nonlinear predictions.

2 Preliminaries

We first define the setting for off-policy evaluation with policy-dependent responses. We distinguish between the *causal decision policy* π and the *downstream optimization response* x. The causal decision policy π intervenes on individual units, while the policy-dependent responses are solutions to a downstream optimization problem on the causal responses of all the units.

1. Off-policy evaluation. We first describe the single time step off-policy policy evaluation and optimization problem [see Dudík et al., 2014, Hirano and Porter, 2020, for further context]. Let covariates be $W \in \mathcal{W} \subseteq \mathbb{R}^d$, binary treatment be $T \in \{0,1\}$[3], and potential outcomes be $c(T)$. Denote the covariates' distribution as $\mathcal{P}$. Without loss of generality we consider lower is better for c; e.g. we minimize costs. We consider a setting of learning causal responses from a dataset of tuples $\mathcal{D}_1 = \{(W_i, T_i, c_i)\}_{i=1}^n$ where treatment is assigned randomly or in an observational setting; henceforth we call this the *observational / experimental dataset*.

We let $\pi_t \colon \mathcal{W} \mapsto [0,1]$ denote a personalized policy mapping from covariates to a (probability of) treatment t. Later we will focus on parameterized policies, such as $\pi_t(w) = sigmoid(\varphi^\top w)$ or policies that admit global enumeration. The goal of off-policy learning is to optimize the causal interventions (aka policies) by estimating average outcomes induced by any given policy. Throughout we will follow the convention that, a random variable $c(\pi_t)$ denotes $c(\pi_t) = c(t)\mathbb{I}(Z_t = t)$, where $Z_\pi \in \{0,1\}$ is a Bernoulli random variable of policy assignment: $Z_\pi \sim \text{Bern}(\pi_1)$. Then, the (random) outcome for a given covariate with policy π is:

$$c(\pi) = \textstyle\sum_t c(\pi_t) = \sum_t c(t)\mathbb{I}(Z_\pi = t).$$

The average treatment effect (ATE) of a policy $\pi(\cdot)$ is then $\mathbb{E}[c(\pi)]$, where the expectation is taken over the randomness of the covariates $W \sim \mathcal{P}$, assignments induced by π, and c conditional on realized treatment t and covariates.

2. Policy-dependent responses. *Policy-dependent optimization* solves a downstream stochastic linear optimization problem over a decision problem $x \in \mathcal{X} \subseteq \mathbb{R}^m$ on the m units given a causal intervention policy. In particular, m represents the dimension of the downstream decision problem. Relative to the downstream decision problem, causal outcomes may enter *either* as uncertain objective coefficients (in c) *or* constraint capacities (in b).[4]

[3]The extension to non-binary treatments is immediate.

[4]Throughout the text we focus on uncertainty in c for notational clarity; strong duality implies the same results hold for uncertainty in the constraint right-hand-side, b. The decision is made conditionally on context information W but prior to realizations of potential outcomes, aka a policy-dependent response.

Dimensionality of the responses. We consider two different asymptotic regimes: an *out-of-sample, fixed-dimension, fixed-m* regime and an *in-sample, growing-dimension, growing-n* regime. We formalize the former regime, the main focus of the paper, in the following assumption.

Assumption 2.1 (Out-of-sample, fixed-dimension regime)**.** As $n \to \infty$, the dimension of the optimization problem m, given by a new draw of contexts $\mathcal{D}_2 = \{W_i\}_{1:m}$ remains finite. The decision-dependent response on m units is measurable with respect to $\mathcal{D}_2$. Let $c_i(\pi) \stackrel{\text{def}}{=} \mathbb{E}[c(\pi(w)|w = W_i]$, we have that the policy value v_π^* is:

$$v_\pi^* = \mathbb{E}[\min_x\{\textstyle\sum_{i=1}^m c_i(\pi)x_i \colon Ax \leq b\}]. \tag{1}$$

Assumption 2.1 defines our *policy-dependent estimand* in this regime. The expectation is taken over the randomness of the policy π and the randomness of the finite samples $\{w_i\}_{i=1}^m$. The main text focuses on statements in the regime of Assumption 2.1. Evaluating regret with respect to a fixed dimension is standard or implicit in the predictive optimization literature.[5]

Assumption 2.2 (In-sample, growing-dimension regime)**.** As $n \to \infty$, the limit of the objective function is an expectation over contexts.[6] The estimand is:

$$v_\pi^* = \mathbb{E}[\min_x\{\mathbb{E}[c(\pi)x] \colon Ax \leq b\}]. \tag{2}$$

Recall that a policy maps from covariates to a (probability of) treatment. Assumption 2.2 precisely takes an expectation over the two sources of randomness: the outer expectation is taken over the randomness of the policy, and the inner expectation is taken over the randomness of the covariates w.

The limiting object in the growing-dimension regime is a "fluid limit" or asymptotic regime: informally we assume a meaningfully constrained optimization in the limit. We instantiate our framework in the following example.

Example 2.3 (Min-cost bipartite matching)**.** Our framework is precisely motivated by the practical challenges in causal inference tasks, where the problem of "policy dependent" optimizations pops up repeatedly. For instance, for price incentives in a matching market (such as a rideshare platform), the revenue/welfare outcome is not realized until the riders and drivers are matched under constraints. As another example, consider a manager wants to assign agents to different jobs, and assigning an agent to a job is associated with some cost. Our goal is to assign each agent to at most one job such that the overall cost is minimized. To incentivize the workers to complete the jobs, the company might want to provide some bonus to the agents. However, the overall efficiency and total payments are not realized until all the assignments are determined.

The above type of application can be modeled as a min-cost bipartite matching problem, which is well known to have a totally unimodular linear relaxation. Clearly, the agents (or riders) and the jobs (or passenger requests) form the two sides of nodes for the matching. The edge costs in the matching stand for the cost or payment for an agent to complete that job. A treatment ($T = 1$) serves as intervention on the edge costs for that agent, and the covariates W could be any observable features of the agents, such as preferences, demographic information, etc. Given any allocation rule of the bonuses, the manager faces a downstream min-cost bipartite matching:

$$\min_{x\in\{0,1\}^{|\mathcal{E}|}} \left\{\textstyle\sum_{e\in\mathcal{E}} c_e(\pi)x_e \,:\, \sum_{e\in\mathcal{N}(i)} x_e = 1, \forall i \in \mathcal{V}\right\}. \tag{3}$$

Here $\mathcal{N}(i)$ is the set of all edges contains node i, the c_e are the edge costs, and $x = \{x_e\}_{e\in\mathcal{E}}$ represents the matching where $x_e = 1$ means that edge e is selected[7].

In Appendix E we include an additional example of predictive risk optimization, beyond linear optimization, which requires a different estimation strategy.

[5]The predictive optimization literature instead views each dimension of the decision variable as a multivariate outcome; relative to that, our regime can be interpreted as the setting of a scalar-valued contextual response.

[6]Assume the constraint b scales with n in a meaningful problem-dependent way so that constraints are neither all slack nor infeasible in the limit.

[7]In the later analysis we use the linear relaxation with $x_e \in [0, 1]$ (continuous interval). For bipartite matching because of *total unimodularity* the linear relaxation is tight and equivalent to integral formulation.

	Out of sample, fixed m (Assumption 2.1)		**In-sample, growing n** (Assumption 2.2, Appendix D)	
	Evaluation	Policy optimization	Evaluation	Policy optimization
AIPW	N/A		Sample splitting (finite VC-dim x)	Uniform generalization requires problem-dependent structure (finite VC-dim x)
WDM	Perturbation method	Uniform generalization from out-of-sample risk bounds	Perturbation	
GRDR	Perturbation method		Perturbation Doubly-robust estimation	

Table 1: Summary of regimes and estimation properties. The main text provides methods for Assumption 2.1. Additional structural restrictions permit extensions for Assumption 2.2.

3. Policy optimization with policy-dependent responses. Putting together the pieces of the previous subsections, the off-policy optimization over candidate policies $\pi \in \Pi$ is:

$$\min_{\pi\in\Pi} \min_{x\in\mathcal{X}} \left\{\textstyle\sum_{i=1}^m c_i(\pi)x_i \colon Ax \leq b\right\}, \tag{4}$$

where m represents the dimension of the decision problem (e.g., the number of edges in Example 2.3), and x denotes the whole response vector $\{x_i\}_{i\in[m]}$.

We illustrate this framework by revisiting our examples.

Example 2.4 (Policy optimization for Example 2.3, min-cost matching)**.** In the min-cost bipartite matching example, the optimal assignments with a given policy π can be solved via the linear program in Equation (3). Suppose that we want to find the best intervention policy which gives the lowest matching cost. Then, the policy optimization problem is:

$$\min_{\pi\in\Pi} \min_{x\in\mathcal{X}} \left\{\textstyle\sum_{e\in\mathcal{E}} c_e(\pi)x_e : \sum_{e\in\mathcal{N}(i)} x_e = 1, \forall i; x_e \geq 0, \forall e\right\},$$

where Π denotes the set of all policies that are of interest.

3 Problem Description: Optimization Bias

We focus on off-policy evaluation in view of the downstream optimization over the decision variables $x = \{x_i\}_{i\in[m]}$. We first discuss *plug-in* estimation approaches without causal adjustment to introduce the challenge of optimization bias in this regime. We then discuss causal estimation in Section 4.

From estimation bias to optimization bias. Denote $\mu_t(w) = \mathbb{E}[c(t) \mid W = w]$ as the conditional outcome mean of the population with treatment t and covariates w. We consider "predict-then-optimize" approaches which learn some $\hat{\mu}_t(w) = \mathbb{E}[c \mid W = w, T = t]$ and optimize with respect to it, so that our estimator is:

$$\hat{v}_\pi = \min_{x\in\mathcal{X}} \left\{\textstyle\sum_{i=1}^m \sum_{t\in\{0,1\}} \pi_t(w_i)\hat{\mu}_t(w_i)x_i \colon Ax \leq b\right\}.$$

Note that due to the estimation and minimization step, $\hat{v}_\pi$ is not an unbiased estimator for v_π^*. Define the overall error of $\hat{v}_\pi$ with respect to the target estimand of Equation (1) as: $\text{err} = v_\pi^* - \mathbb{E}\left[\hat{v}_\pi\right]$. We decompose the overall error into two parts: the estimation bias of the plug-in estimator, and the optimization bias. Denote $\tilde{v}_\pi$, the best-in-class feasible estimate using the true conditional expectations μ_t^*:

$$\tilde{v}_\pi = \min_{x\in\mathcal{X}} \left\{\textstyle\sum_{i=1}^m \sum_{t\in\{0,1\}} \pi_t(w_i)\mu_t(w_i)x_i \colon Ax \leq b\right\}.$$

Then, the estimation and optimization biases are: (by triangle inequality, $|\text{err}| \leq |\text{bias}_{\text{est}}| + |\text{bias}_{\text{opt}}|$)

$$\text{bias}_{\text{est}} = \mathbb{E}[\hat{v}_\pi] - \mathbb{E}[\tilde{v}_\pi], \qquad \text{bias}_{\text{opt}} = v_\pi^* - \mathbb{E}[\tilde{v}_\pi].$$

In-sample estimation bias due to optimization. It is well known that in-sample estimation of the value of optimization problems is biased; e.g., $\hat{v}$ is a biased estimate for the true objective value v_π^* due to optimization. Ito et al. [2018] studies a bias correction for affine linear objectives with an unbiased estimate of a parameter θ. To understand the source of the bias due to optimization, observe that clearly $\sum_{i=1}^m \mu_t(w_i)x_i \geq \min_x \sum_{i=1}^m \mu_t(w_i)x_i$. The inequality remains valid when evaluating expectations over training datasets so that $\mathbb{E}[\sum_{i=1}^m \mu_t(w_i)x_i] \geq \mathbb{E}[\min_x \sum_{i=1}^m \mu_t(w_i)x_i]$. Noting that the RHS is the true objective v_π^*, we obtain in general the well-known optimistic bias, that $\mathbb{E}[\tilde{v}_\pi] \geq v_\pi^*$. In the policy evaluation setting, our estimates converge to the LHS, $\tilde{v}_\pi$, so that our estimator $\hat{v}_\pi$ is in general a *biased* estimate of the decision-dependent policy value even if we obtain *unbiased* estimates of the cost coefficient.

4 Causal Estimation with Policy-Dependent Responses

In this section we present an estimation approach building upon a perturbation method that adjusts for the aforementioned optimization bias. We summarize tradeoffs among estimation strategies in different regimes in Table 1 and possible extensions and additional structure in Appendix D.

4.1 Estimating causal effects: estimation bias

Assumption 4.1 (Ignorability, overlap, SUTVA)**.** For all t, $c(t) \perp\!\!\!\perp T \mid W$. The evaluation policy is absolutely continuous with respect to treatment probabilities in the training dataset. Assume the stable unit treatment value assumption.

Confounding-adjusted plug-in estimators. In general, plug-in estimation of $\hat{\mu}_t(W)$ does *not* admit unbiased predictions because of selection bias and model misspecification. Existing importance-sampling based estimators, e.g. the inverse propensity weighting (IPW) estimator and the doubly-robust augmented inverse probability weighting (AIPW) uses the propensity score to adjust confounding, under Assumption 4.1. Note importance sampling cannot *directly* be applied in our main regime of interest with out-of-sample evaluation as in Assumption 2.1, see Appendix B for a detailed overview.

We depart from previous work in off-policy evaluation, in view of the optimization bias adjustment (detailed in the next section), and study estimation methods that are *plug-in estimates* for OPE: $\mathbb{E}[c(\pi)] = \sum_t \mathbb{E}[\pi_t(W)\hat{\mu}_t(W)]$, for some outcome model $\hat{\mu}_t$ that is confounding-adjusted.

Note that IPW/AIPW-type estimators cannot be applied in the out-of-sample regime of Assumption 2.1. However, we may obtain out-of-sample risk bounds on the decision regret in this regime by virtue of out-of-sample generalization risk bounds on the generated regressors. We include more detailed discussion in Appendix D.3.

Weighted direct method (WDM). Outcome regression, learning $\hat{\mu}_t(W) = \mathbb{E}[c \mid T = t, W]$ directly from $\mathcal{D}_1$, is sometimes called the *direct method*. However, when $\hat{\mu}$ is a misspecified regression model such a method incurs bias. Nonetheless, re-weighting the estimation $\hat{\mu}$ (maximum likelihood, empirical risk minimization) by the inverse probability weights $1/e$ is known to adjust for the covariate shift; by a similar argument as that of Shimodaira [2000], Cao et al. [2009], Wang et al. [2019]. We call this approach *weighted direct method* (WDM), which solves:

$$\hat{\mu}_t^{\mathrm{WDM}} \in \arg\min_\mu \mathbb{E}\left[\frac{\mathbb{I}(T=t)}{e_t(W)}(c - \mu_t(W))^2\right]. \tag{5}$$

Doubly-robust direct method (GRDR). We also consider an approach that achieves doubly-robust estimation of the treatment-effect due to Bang and Robins [2005]. [See also Scharfstein et al., 1999, Tran et al., 2019]. This approach has been used for CATE estimation Shi et al. [2019], Chernozhukov et al. [2021]. The inverse propensity score reweighted treatment indicator is added as a covariate in the model, inducing coefficients ϵ_0, ϵ_1. Define

$$\hat{\mu}^{\mathrm{GRDR}} = \mu(W) + \epsilon_1(T/e_1(W)) + \epsilon_0((1-T)/e_0(W)).$$

Optimizing over $\hat{\mu}$ by (nonlinear) least-squares yields the following first-order optimality conditions for $\theta^{\mathrm{GRDR}} = [\bar{\theta}, \epsilon_1, \epsilon_0]$:

$$\mathbb{E}[(c - \hat{\mu})\nabla_\theta \hat{\mu}] = 0, \mathbb{E}[(c - \hat{\mu})(T/e_1(W))] = 0, \mathbb{E}[(c - \hat{\mu})((1-T)/e_0(W))] = 0. \tag{6}$$

Bang and Robins [2005] show that the first-order optimality conditions ensure that plug-in estimation of an average treatment effect with the model is equivalent to AIPW, hence doubly-robust. Because it is designed primarily for estimation of the ATE, its use as an outcome predictor is more speculative. Although one can verify that its output is covariate-conditionally equivalent to CATE in expectation, and one can use this fact to again regress upon the pseudooutcomes, this final procedure would require re-verifying asymptotic convergence; we don't outline those arguments here. We include further discussion on the different estimation interpretations of GRDR in the two regimes in Appendix D.3.

4.2 Estimating the decision-dependent estimand

Our procedure is adapted from the perturbation method of Ito et al. [2018] which we describe here for completeness; we extend it from linear to nonlinear predictors. The method of Ito et al. [2018]

Algorithm 1 Perturbation method, Alg. 2 of Ito et al. [2018])

input Estimation strategy $\diamond \in \{\mathrm{WDM}, \mathrm{GRDR}\}$; h: finite different parameter; π: policy.

Estimate $\hat{\xi}_\diamond = [\hat{\theta}_\diamond, \hat{\gamma}_\diamond]$ for $\hat{\mu}^\diamond$ from $\mathcal{D}_1$

$\hat{v}^{(0)} \leftarrow \min_{x \in \mathcal{X}} \sum_{i=1}^m x_i \sum_{t \in \{0,1\}} \pi_t(w_i) \hat{\mu}_t^\diamond(w_i; \hat{\xi}_\diamond)$

Generate $\{\xi_\diamond^{(j)}\}_{j=1}^s$: if by parametric bootstrap, learn $\hat{\xi}_\diamond^{(j)}$ from $\frac{N}{(1+h)^2}$ samples randomly chosen from $\mathcal{D}_1$ with replacement.
Otherwise if using $\hat{\Sigma}$, estimator of asymptotic variance of ξ, approximate the distribution of $\xi^* + (1+h)\delta$. Add $\hat{\xi}$ to $\hat{\theta}$ where $\hat{\delta} \sim N(0, \frac{(1+h)^2-1}{N}\hat{\Sigma})$. Then set $\hat{\xi}_\diamond^{(j)} = \hat{\xi} + \hat{\delta}_j$.

for $j = 1, \ldots, S$: **do**

    $\hat{v}^{(j)} \leftarrow \min_{x \in \mathcal{X}} \sum_{i=1}^m x_i \sum_{t \in \{0,1\}} \pi_t(w_i) \hat{\mu}_t^\diamond(w_i; \hat{\xi}_\diamond^{(j)})$.

end for

Output $\rho_h = \hat{v}_0 - \frac{1}{h}(\hat{v}^{(0)} - \frac{1}{s}\sum_{j=1}^s \hat{v}^{(j)})$.

focuses on one parameter that we denote $\xi = [\theta, \gamma]$, where we assume as outlined in Assumption 4.4 that it encompasses parameters of the outcome and propensity model (respectively). Define the policy-induced outcome model, $\mu_\pi(w) = \sum_t \pi_t(w)\mu_t(w)$, the estimation error $\delta = \hat{\xi} - \xi^*$, and the (parametrized) optimal solution at a given predictive model $x(\xi)$. The perturbation method is motivated by a finite-difference approximation to the optimization bias induced by estimation error δ. Define the auxiliary functions given a scalar ϵ parametrizing the direction of δ:

$$\eta(\epsilon) = \mathbb{E}_\delta \left[\sum_{i=1}^m x(\xi^* + \epsilon\delta)\pi(W; \xi^*)\right], \quad \phi(\epsilon) = \mathbb{E}_\delta \left[\sum_{i=1}^m x(\xi^* + \epsilon\delta)\pi(W; \xi^* + \epsilon\delta)\right].$$

We require regularity conditions for derivatives of these functions to exist:

Assumption 4.2 (Perturbation method assumptions)**.** (i) The optimal solution $x(\xi)$ is unique. (ii) $\hat{\xi}$ is an unbiased estimator of ξ^*.

We generalize Prop. 3 of Ito et al. [2018] for nonlinear models.

Proposition 4.3. *We have* $\eta(\epsilon) = \phi(\epsilon) - \epsilon\phi'(\epsilon) + O(\epsilon^2)$.

The plug-in estimated optimal value $\hat{v}_\pi$ unbiasedly estimates $\phi(1)$. Note $\phi'(1)$ is equivalent to the value of the bias. The perturbation method estimates $\phi'(1)$ by $(\phi(1+h) - \phi(1))/h$ for some small h.

It remains to estimate $\phi(1+h)$. First we obtain s samples of the perturbed parameter $\hat{\xi}_h = \xi^* + (1+h)\delta$, denoted as $\{\hat{\xi}_h^{(j)}\}_{j=1}^s$. Each replicate of $\hat{\xi}^{(j)}$ leads to an optimization estimate $\hat{v}^{(j)} = \min_x \sum_{i=1}^m x_i \hat{\mu}_\pi(w_i, \hat{\xi}_h^{(j)})$. The debiased estimator is:

$$\rho_h = \hat{v}^{(0)} - \frac{1}{h}(\hat{v}^{(0)} - \frac{1}{s}\sum_{j=1}^s \hat{v}^{(j)})$$

Our Proposition 4.3 then implies asymptotic unbiasedness (cf. Prop. 4 of Ito et al. [2018]) so that $\lim_{h \to 0} \mathbb{E}[\rho_h] = \mathbb{E}[\min_x \sum_{i=1}^m x_i \hat{\mu}_\pi(w_i, \xi^*)]$. We summarize the method in Algorithm 1.

Asymptotic variance of estimation methods. We discuss the asymptotic variance of the *weighted direct method* and GRDR via classical asymptotic analysis of *generated regressors* (specifically, stacked estimation equations of GMM) [Newey and McFadden, 1994]. We summarize this framework in Appendix B.1 for completeness and include the main result here that we invoke.[8]

Assumption 4.4 (Estimators via GMM with generated regressors)**.** Suppose the propensity score e and outcome model μ are indexed by true parameters γ^*, θ^* that solve the respective estimating equations $\mathbb{E}[h(W, \gamma^*)] = 0, \quad \mathbb{E}[g(W, \theta^*, \gamma^*)] = 0$. The functions $e_t(w), \mu_t(w)$ are in a Donsker class.

Remark 4.5 (Strength of assumptions). Algorithm 1 requires both unbiased and asymptotically normal predictions—stronger conditions than merely inference on the ATE. The Donsker assumption preserves asymptotic normality with generated regressors. The framework allows for nonparametric estimation via linear sieves (but not some high-dimensional regimes; see Ackerberg et al. [2012]).

[8] Asymptotic normality of these approaches is taken as given in Cao et al. [2009], Bang and Robins [2005] and so we include these statements for completeness. For exposition and context of Donsker-type conditions in semiparametric inference, see Kennedy [2016] or other references.

Algorithm 2 Subgradient method for policy optimization

1: **Input:** step size η, linear objective function f.
2: **for** $j = 1, 2, \cdots$ **do**
3: At φ^k, obtain a subgradient in subdifferential $\mathcal{S}^*(\pi_\varphi^k) = \{x^* : f(x^*; \pi_\varphi^k) = \min_x f(x; \pi_\varphi^k)\}$
4: Compute subgradient $\nabla_\varphi(\min_x f(x; \pi_\varphi^k)) \leftarrow \nabla_\varphi f(x^*; \pi_\varphi)$
5: Update subgradient step: $\varphi^{k+1} \leftarrow \varphi^k - \eta \nabla_\varphi \left(\min_x f(x; \pi_\varphi^k)\right)$
6: **end for**

Theorem 4.6 (Thm. 6.1, eq. 6.12 of Newey and McFadden [1994])**.** *Suppose Assumption 4.4 holds. Let $\hat{G}_\alpha, \hat{G}_\theta, \hat{H}$ denote the Jacobian matrices of partial derivatives of the moment conditions g, h with respect to the respective parameters, i.e. $\hat{G}_\gamma = n^{-1} \sum_{i=1}^n \nabla_\gamma g(w_i, \hat{\theta}, \hat{\gamma})$. Let $\hat{V}_\gamma = (\hat{H}^{-1} \hat{h}_i)(\hat{H}^{-1} \hat{h}_i)^\top$. Then an estimator of the asymptotic variance is:*

$\hat{V}_\theta = \hat{G}_\theta^{-1} \left(n^{-1} \sum_{i=1}^n \hat{g}_i \hat{g}_i^\top\right) (\hat{G}_\theta^{-1})^\top + \hat{G}_\theta^{-1} \hat{G}_\gamma \hat{V}_\gamma \hat{G}_\gamma^\top (\hat{G}_\theta^{-1})^\top.$

Since $\hat{V}_\gamma$ depends only on the specification of the propensity score, to completely specify the asymptotic variance for the above formula we state the mixed terms $\hat{G}_\gamma, \hat{G}_\theta$.

Proposition 4.7 (Asymptotic normality of WDM)**.** *Let $e_t(w), \mu_t(w)$ satisfy Assumption 4.4 with the moment condition $g_t(W, \theta, \gamma) = e_t(W; \gamma)^{-1}(c - \mu_t(W; \theta))^2$ and $g = [g_0, g_1]$. Then*

$$\hat{G}_\gamma = \begin{bmatrix} \mathbb{E}_n[2T(c - \mu(W; \theta)) \frac{\partial}{\partial \theta}(e_1^{-1}(W, \gamma)) \frac{\partial \mu}{\partial \theta}] \\ \mathbb{E}_n[2(1 - T)(c - \mu(W; \theta)) \frac{\partial}{\partial \theta}(e_0^{-1}(W, \gamma)) \frac{\partial \mu}{\partial \theta}] \end{bmatrix}.$$

These formulas are generally computable from standard output of optimization solvers for nonlinear least squares: gradients and Hessians. In practice, using the parametric bootstrap may be simpler at a higher computational cost.

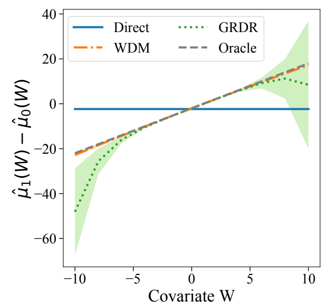
4.3 Optimizing Causal Interventions

Algorithm 1 provides estimation for a fixed policy. We now discuss how to optimize over policies; e.g., implementing the outer optimization over policies $\min_{\pi \in \Pi}$ in Equation (4). We focus on the case where the policy $\pi_t(w)$ is parametrized by and differentiable in a parameter $\varphi \in \Psi$. For example, for the logistic policy parameterization, $\pi_t(w) = sigmoid(\varphi_t^\top w)$. We consider a robust subgradient method, based on Danskin's theorem, detailed in Algorithm 2. Such an approach is a common heuristic used in adversarial machine learning.

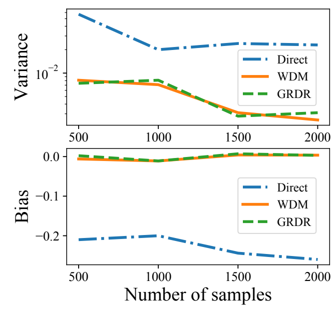
We solve the inner optimization problem to full optimality in line 3 and take (sub)gradient steps for the outer optimization. We evaluate (sub)gradients of the inner optimization solution in line 3 by evaluating the gradient of the objective with respect to φ, fixing the inner optimization variable x^*. Danskin's theorem implies that ∇_φ is a subgradient [Danskin, 1966]. The inner minimization can be solved via a linear optimization oracle for any fixed choice of policy. This use of the linear optimization oracle can be beneficial when special problem structures, such as matching and network flows, may also admit readily-available algorithmic solutions to full optimization.

The perturbation method is compatible with our optimization procedure because the bias-adjusted perturbation estimated from Algorithm 1 is affine in the optimization problems corresponding to each parameter replicate. Hence, run Algorithm 1 with an expanded linear objective over the s-product space $x' \in \mathcal{X}^s$ where $f(\tilde{x}, \pi) = \hat{v}_\pi^{(0)}(\tilde{x}_0) - \frac{1}{h}(\hat{v}_\pi^{(0)}(\tilde{x}_0) - \frac{1}{s} \sum_{j=1}^s \hat{v}_\pi^{(j)}(\tilde{x}_j))$.

So, re-optimize $\tilde{x}_j^* \in \arg\min_{x \in \mathcal{X}} \sum_{i=1}^m x_i \hat{\mu}_\pi^\diamond(w_i; \hat{\xi}_\diamond^{(j)})$ and apply Danskin's theorem to each optimization problem in the sum over $\hat{v}_\pi^{(j)}$ comprising $f(x', \pi)$. In fact, though adversarial machine learning focuses on min-max rather than our min-min optimization problem, this particular approach is simply subgradient descent on a nonconvex function (the solution to the inner optimization).



(a) Effect of W



(b) Effect of dataset size

Figure 1: *(In-sample estimation of $\hat{\mu}_1(W) - \hat{\mu}_0(W)$, with model mis-specification).* Comparison of direct / WDM / GRDR to the oracle. (a) Conditional estimation error averaged over ten random train sets; shaded area indicates std. error. (b) Bias / variance comparison with varying training data size.

5 Experimental Evaluation

Since real data suitable for both policy evaluation and downstream optimization is unavailable, we focus on synthetic data and downstream bipartite matching. We first illustrate estimation properties of the different approaches before showing the improvement obtained via policy optimization. Though we are not aware of prior approaches that are directly comparable for optimizing causal policy with a downstream optimization-dependent response, we include more comparisons to nonparametric estimators (e.g. causal forests [Wager and Athey, 2018]), and full implementation details. All code will be published.

1. Causal effect estimation. First, we investigate and illustrate the properties of different estimators. We generated dataset $\mathcal{D}_1 = \{(W, T, c)\}$ with covariate $W \sim \mathcal{N}(0, 1)$, confounded treatment T, and outcome c. Treatment is drawn with probability $\pi_t^b(W) = (1 + e^{-\varphi_1 W + \varphi_2})^{-1}$, $\varphi_1 = \varphi_2 = 0.5$. The true outcome model is given by a degree-2 polynomial,[9] $c_t(w) = poly_\theta(t, w) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. . In Figure 1a and 1b, we illustrate the (covariate-conditional) estimation error of the three estimators. In the mis-specified setting that induces confounding, the outcome model is a vanilla linear regression over W without the polynomial expansion. The direct method results in more bias under mis-specification, while WDM and GRDR are robust as expected.

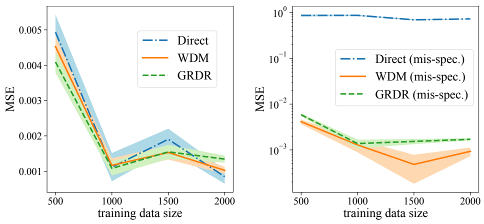
2. Policy evaluation. We compare the perturbation method (Algorithm 1) with three different estimators (direct, WDM, and GRDR). In both the well-specified / mis-specified model setting, we evaluate the mean-squared-error (MSE) of the estimated policy value with the three estimators, where the MSE is computed with regard to the ground-truth outcome model. Training data size n increases from 500 to 2000 samples. We scaled the MSE down by the number of edges (a constant) and computed the MSE in terms of the averaged cost per edge in the matching.

For the policy-dependent optimization, we evaluate a min-cost bipartite matching problem, where the causal policy intervene on the edge costs (as detailed in Example 2.3). Specifically, the bipartite graph contains $m = 500$ left side nodes $W_1, \cdots, W_m$, and $m' = 300$ right side nodes. The policy π_t applies treatments to the left side nodes and the outcome is the edge cost of edges with that node. While we grow the training data size, we fixed m, m' (with $m > m'$) and evaluate over ten random draw of train/test data for each value of n. Figure 2 plots the results. When there is mis-specification, even a large training dataset cannot bring bias correction for the direct method, where both WDM and GRDR enjoy smaller and decreasing MSE.

We also conduct an ablation study for the corresponding performance in the mis-specified setting (i.e., no bootstrapping in Alg. 1). Results indicate that the perturbation method is helpful for MSE reduction for both WDM and GRDR. We further conduct evaluations with different bootstrap replicates' sizes, and the above conclusions remain robust for different replicate sizes (additional results in Appendix F).

3. Policy optimization. Lastly, we integrate the policy evaluation and the sub-gradient method (Alg 2) to conduct policy optimization. At each iteration of Alg 2, the perturbation algorithm (Alg 1) and one of the three different estimation methods are applied to evaluate the policy objective. We consider

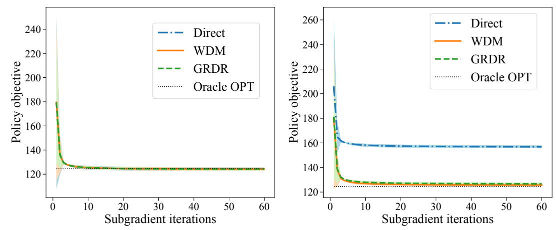
[9] If not stated otherwise we spread the coefficients as $poly_\theta(t, w) = (1, w, t, w^2, wt, t^2) \cdot ([5, 1, -1, 2, 2, -1])^\top$. Additional supporting experiments under other nonlinear data-generating processes are in Appendix F.



(a) well-specified models

(b) with mis-specification

Figure 2: *(Policy evaluation via perturbation method (Algorithm 1)*. Comparison of direct / WDM / GRDR estimators over increasing size of training data (averaged over ten runs).



(a) well-specified models

(b) with mis-specification

Figure 3: *(Policy optimization)*. Subgradient policy optimization with direct / WDM / GRDR estimation methods and a fixed test set. Averaged over ten random training datasets of size=1000.

a logistic policy $\pi_t(W) = sigmoid(\varphi_1 \cdot W + \varphi_2)$. To study the convergence and the effectiveness of the subgradient algorithm for minimization, we fix a test set and perform subgradient descent over 60 iterations for each run. We average the policy values at each iteration over ten runs, where in each run we generate a random set of training data and a random initialization of the starting policy.

We compare to the oracle estimator using the ground truth outcome model (Oracle OPT). Results are presented in Figure 3. Again, WDM and GRDR quickly converge to the oracle estimation, while the large bias of the direct method leads to poor policy optimization. We further evaluate the impact of average random selected initial policies to the performance, and compared Figure 3 with the results using a fixed initial policy. We observe that in this relatively low-dimensional example, the policy value converges to estimation-oracle-optimal after a few iterations (additional results and full training details in Appendix F).

6 Conclusion

We studied a new framework for causal policy optimization with a *policy-dependent* optimization response. We proposed evaluation algorithms and analysis to address the fundamental challenge of an additional optimization bias. Simulations for both the policy evaluation and optimization algorithms demonstrate the effectiveness of this approach. Interesting further directions include studying individual fairness of optimal allocations in applications such as school assignments or job matching, and/or computational improvements to the policy optimization algorithms.

Acknowledgements

We wish to acknowledge support from the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941. WG acknowledges support from a Google PhD fellowship. AZ acknowledges support from the Foundations of Data Science Institute.

References

Daniel Ackerberg, Xiaohong Chen, and Jinyong Hahn. A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, 94(2):481–498, 2012.

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

Susan Athey, Stefan Wager, et al. Efficient policy learning. Technical report, 2017.

Eduardo M Azevedo and Jacob D Leshno. A supply and demand framework for two-sided matching markets. *Journal of Political Economy*, 124(5):1235–1268, 2016.

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

Güzin Bayraksan and David P Morton. Assessing solution quality in stochastic programs. *Mathematical Programming*, 108(2):495–514, 2006.

Patrice Bertail, Stephan Clémençon, Yannick Guyonvarch, and Nathan Noiry. Learning from biased data: A semi-parametric approach. In *International Conference on Machine Learning*, pages 803–812. PMLR, 2021.

Debopam Bhattacharya. Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association*, 104(486):486–500, 2009.

Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, 2009.

Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *arXiv preprint arXiv:2104.12820*, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Victor Chernozhukov, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. *arXiv preprint arXiv:2110.03031*, 2021.

Bruno Crépon and Gerard J Van Den Berg. Active labor market policies. *Annual Review of Economics*, 8:521–546, 2016.

Bruno Crépon, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2):531–580, 2013.

John M Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, 385(3):283–286, 2021.

Vishal Gupta, Michael Huang, and Paat Rusmevichientong. Debiasing in-sample policy performance for small-data, large-scale optimization. *Large-Scale Optimization (June 2, 2021)*, 2021.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016.

Keisuke Hirano and Jack R Porter. Asymptotic analysis of statistical decision rules in econometrics. In *Handbook of Econometrics*, volume 7, pages 283–354. Elsevier, 2020.

Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Shinji Ito, Akihiro Yabe, and Ryohei Fujimaki. Unbiased objective estimation in predictive optimization. In *International Conference on Machine Learning*, pages 2176–2185. PMLR, 2018.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in Neural Information Processing Systems*, 31, 2018a.

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

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

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Data-driven sample average approximation with covariate information. *Optimization Online. URL: http://www. optimization-online. org/DB_HTML/2020/07/7932. html*, 2020.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and their Applications in Public Health Research*, pages 141–167. Springer, 2016.

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

Michael C Knaus, Michael Lechner, and Anthony Strittmatter. Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, pages 0718–9615R1, 2020.

Amanda Kube, Sanmay Das, and Patrick J Fowler. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 622–629, 2019.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Henry Lam and Huajie Qian. Bounding optimality gap in stochastic optimization via bagging: Statistical efficiency and stability. *arXiv preprint arXiv:1810.02905*, 2018.

Liu Leqi and Edward H Kennedy. Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802*, 2021.

Hannah Li, Geng Zhao, Ramesh Johari, and Gabriel Y Weintraub. Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. *arXiv preprint arXiv:2104.12222*, 2021.

Romain Lopez, Chenchen Li, Xiang Yan, Junwu Xiong, Michael Jordan, Yuan Qi, and Le Song. Cost-effective incentive allocation via structured counterfactual inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4997–5004, 2020.

Hongyao Ma, Fei Fang, and David C Parkes. Spatio-temporal pricing for ridesharing platforms. *Operations Research*, 2021.

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

Jorge Mejia and Chris Parker. When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science*, 67(1):166–184, 2021.

Evan Munro, Stefan Wager, and Kuang Xu. Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647*, 2021.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing predictive models using electronic medical records: challenges and pitfalls. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1109. American Medical Informatics Association, 2013.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Zhengling Qi, Jong-Shi Pang, and Yufeng Liu. Estimating individualized decision rules with tail controls. *arXiv preprint arXiv:1903.04367*, 2019.

Aida Rahmattalabi, Phebe Vayanos, Kathryn Dullerud, and Eric Rice. Learning resource allocation policies from observational data with an application to homeless services delivery. *arXiv preprint arXiv:2201.10053*, 2022.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models (rejoinder). *Journal of the American Statistical Association*, 94(448):1135–1146, 1999.

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

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Liyang Sun. Empirical welfare maximization with constraints. *arXiv preprint arXiv:2103.15298*, 2021.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Linh Tran, Constantin Yiannoutsos, Kara Wools-Kaloustian, Abraham Siika, Mark Van Der Laan, and Maya Petersen. Double robust efficient estimators of longitudinal treatment effects: Comparative performance in simulations and a case study. *The international Journal of Biostatistics*, 15(2), 2019.

Sara van de Geer and Leen Stougie. On rates of convergence and asymptotic normality in the multiknapsack problem. *Mathematical Programming*, 51(1):349–358, 1991.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Stefan Wager and Kuang Xu. Experimenting in equilibrium. *Management Science*, 2021.

Lequn Wang, Yiwei Bai, Arjun Bhalla, and Thorsten Joachims. Batch learning from bandit feedback through bias corrected reward imputation. In *ICML Workshop on Real-World Sequential Decision Making*, 2019.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Appendix Section	Referencing Section
Appendix A	Section 1.1, Related Work
Appendix B	Section 4, Causal Estimation with Policy-Dependent Responses
Appendix D	Sections 2 and 4, Preliminaries / Causal Estimation with Policy-Dependent Responses
Appendix E	Section 2, Preliminaries
Appendix F	Section 5, Experimental Evaluation

Table 2: Table of contents: Appendix and referring sections.

A Further Related Work and Comparisons

Other variants of off-policy evaluation and optimization with global dependence on the population. There is also a line of work on off-policy evaluation, for example evaluating policies by non-average functionals over populations (median [Leqi and Kennedy, 2021], quantile [Chandak et al., 2021, Qi et al., 2019]); but estimation crucially depends on specific reformulations based on problem structure. These functionals are typically risk deviations and reformulated in relation to estimation of quantiles rather than generic optimization formulations. In contrast, our debiasing is more general and independent of the functional form of the risk deviation beyond the second-stage problem being a stochastic linear optimization problem. However, risk deviations typically lead to *min-max* problems, while our formulation has aligned objective functions of treatment and response and is inherently a partial optimization of a nonconvex problem, hence a *min-min* problem.

Such min-max formulations also appear in sensitivity analysis and approaches to unobserved confounding via min-max robustness [Kallus and Zhou, 2018a, 2021] and similar algorithms for policy optimization appear there. However, the formulation of this work focuses specifically on *generic optimization problems* in a different conceptual setting. Estimation is quite different due to the requirement of compatibility of estimation and a perturbation approach.

Our focus on the downstream global system response bears a distant conceptual resemblance to interference [Hudgens and Halloran, 2008], but is fundamentally very different: we require the *causal* response satisfies the stable unit treatment value assumption (SUTVA), while the source of interaction across units (analogous to the exposure mapping) *is completely known to* and *under the control of* the decision-maker.

Lastly, recent work develops specialized estimation more closely using the structure of economic systems, such as marketplace interference [Li et al., 2021] or mean-field equilibrium [Wager and Xu, 2021, Munro et al., 2021]. This work is often closely coupled with the application structure. Our use of policy-dependent structure is coarse, considering only a generic linear optimization response and in turn adjusting for the introduced optimization bias.

Decision-dependent classifier shift. Our last example of estimating decision-dependent predictive loss is broadly motivated by decision-dependent distribution shifts but focuses on settings with distinct treatments, rather than other frameworks where a classifier is a treatment studying model-based or utility-model based approaches [Hardt et al., 2016, Perdomo et al., 2020].

Structured prediction. Finally, although there is extensive work on structured prediction in machine learning, our framework is very different: while structured prediction maps contextual inputs to the space of complex outputs (the space of network flows, matchings; the optimization decision vector), our data environment consists of contexts and separable, unit-dependent outcomes.

B Additional Details for Estimation

B.1 Preliminaries

Adjusting for confounding: IPW and AIPW. In general, plug-in estimation of $\hat{\mu}_t(W)$ does *not* admit unbiased predictions because of selection bias and model misspecification. A key object that adjusts for selection bias is the

$$\textit{propensity score: } e_t(W) = \mathbb{P}(T = t \mid W).$$

Although importance sampling cannot *directly* be applied in our main regime of interest with out-of-sample evaluation as in Assumption 2.1, we introduce key properties which can be used to debias outcome models. (See Appendix D for discussion of an alternative in-sample OPE regime).

Inverse propensity weighting (IPW) transforms treatment-conditional expectations to the population expectation—by iterated expectations and Assumption 4.1 (ignorability), we have:

$$\sum_t \mathbb{E}\left[\mathbb{E}\left[c\frac{\mathbb{I}[Z_\pi = t]}{e_t(W)} \mid W\right]\right] = \sum_t \mathbb{E}\left[c\frac{\mathbb{I}[Z_\pi = t]}{e_t(W)}\right] = \sum_t \mathbb{E}[c(\pi_t)] = \mathbb{E}[c(\pi)]. \tag{7}$$

In general, *doubly-robust augmented inverse probability weighting (AIPW)* estimation improves the variance when both models are well-specified and achieves overall unbiasedness under unbiasedness of either outcome or propensity:

$$\textstyle\sum_t \mathbb{E}\left[\pi_t(W)\mathbb{I}[Z_\pi = t](\frac{\mathbb{I}[T=t]}{e_t(W)}(c - \mu_t(W)) + \mu_t(W))\right] = \mathbb{E}[c(\pi)].$$

C Proofs

Asymptotic variance of two-step estimation via GMM asymptotic variance We first recall a general framework for deriving asymptotic variance with generated regressors as discussed in Newey and McFadden [1994]. This section is summarized from the presentation in there to keep derivations self-contained.

We focus on the approach based on the asymptotic variance for GMM, viewing the nuisance estimations as “stacked" moment equations for $\hat{\theta}$ (second-stage estimate) and $\hat{\gamma}$ (the first-stage estimation). Then, applying blockwise inversion to the GMM asymptotic variance obtains the asymptotic variance of the (parameter) vector, $\begin{bmatrix}\theta\\ \gamma\end{bmatrix}$ in terms of the first component (top left block).

Stack the moment equations for $\hat{\theta}, \hat{\gamma}$ to get:

$$\begin{aligned} \mathbb{E}[g(W, \theta_0, \gamma_0)] &= 0, \\ \mathbb{E}[h(W, \gamma_0] &= 0. \end{aligned}$$

Note adaptivity occurs iff $G_\gamma = \nabla_\gamma \mathbb{E}[g(W, \theta_0, \gamma_0)] = 0$

Define the following sample Jacobians of moment conditions with respect to the parameters:

$$\hat{G}_\theta = n^{-1}\sum_{i=1}^n \nabla_\theta g(w_i, \hat{\theta}, \hat{\gamma}), \qquad \hat{G}_\gamma = n^{-1}\sum_{i=1}^n \nabla_\gamma g(w_i, \hat{\theta}, \hat{\gamma}), \qquad \hat{H} = n^{-1}\sum_{i=1}^n \nabla_\gamma h(w_i, \hat{\gamma}).$$

For short introduce notation for the empirical moments $\hat{g}_i, \hat{h}_i$:

$$\hat{g}_i = g(w_i, \hat{\theta}, \hat{\gamma}), \qquad \hat{h}_i = h(w_i, \hat{\gamma}),$$

so that we can define the sample second moment matrix $\hat{\Omega}$ as follows:

$$\hat{\Omega} = n^{-1}\sum_{i=1}^n (\hat{g}_i, \hat{h}_i)^\top (\hat{g}_i, \hat{h}_i).$$

The estimator for asymptotic variance is given by:

$$\begin{aligned}\hat{V} &= \begin{bmatrix}\hat{G}_\theta & \hat{G}_\gamma \\ 0 & \hat{H}\end{bmatrix}^{-1} \Omega^{-1} \begin{bmatrix}\hat{G}_\theta & \hat{G}_\gamma \\ 0 & \hat{H}\end{bmatrix}^{-1} \\ &= \left[\begin{array}{cc}\hat{G}_\theta^{-1} & -\widehat{G}_\theta^{-1}\hat{G}_\gamma \hat{H}^{-1} \\ 0 & \hat{H}^{-1}\end{array}\right] \widehat{\Omega} \left[\begin{array}{cc}\hat{G}_\theta^{-1} & -\widehat{G}_\theta^{-1}\widehat{G}_\gamma \hat{H}^{-1} \\ 0 & \hat{H}^{-1}\end{array}\right].\end{aligned}$$

If the moment functions are uncorrelated then the first-step estimation increases the second-step variance. We may obtain a simplification as follows. Let $\hat{\phi}_i = -\hat{H}^{-1}\hat{h}_i$. Then the upper left block is

$$\hat{V}_\theta = \widehat{G}_\theta^{-1} \left[n^{-1} \sum_{i=1}^{n} \left\{ \hat{g}_i + \hat{G}_\gamma \hat{\psi}_i \right\} \left\{ \hat{g}_i + \widehat{G}_\gamma \hat{\psi}_i \right\}' \right] \widehat{G}_\theta^{-1}.$$

For $\hat{V}_\gamma = n^{-1} \sum_{i=1}^{n} \hat{\phi}_i \hat{\phi}_i'$, an asymptotic variance estimator for $\hat{\theta}$ is

$$\hat{V}_\theta = \hat{G}_\theta^{-1} \left(n^{-1} \sum_{i=1}^{n} \hat{g}_i \hat{g}_i' \right) (\hat{G}_\theta^{-1})' + \hat{G}_\theta^{-1} \hat{G}_\gamma \hat{V}_\gamma (\hat{G}_\gamma)' (\hat{G}_\theta^{-1})'.$$

C.0.1 Outcome regressions with generated regressors

In the appendix we also discuss an approach based on GRDR .

Proposition C.1 (Asymptotic normality of GRDR)**.** *Let $e_t(w), \mu(w)$ satisfy Assumption 4.4 with the moment condition for μ given by Equation* (6)*. Then*

$$\hat{G}_\gamma = \begin{bmatrix} \mathbb{E}_n[2(\epsilon_1(\frac{\partial}{\partial\gamma} e_1^{-1}(W;\gamma))) + \epsilon_0(\frac{\partial}{\partial\gamma} e_0^{-1}(W;\gamma)))\frac{\partial\mu}{\partial\tilde{\theta}} \\ -\mathbb{E}_n[2T(c-\mu(W;\theta))(\frac{\partial}{\partial\gamma} e_1^{-1}(W;\gamma))] \\ -\mathbb{E}_n[2(1-T)(c-\mu(W;\theta))(\frac{\partial}{\partial\gamma} e_0^{-1}(W;\gamma))] \end{bmatrix}.$$

Proof of Proposition 4.7.

$$\mathbb{E}_n[\nabla_\gamma g(W;\theta,\beta)] = \mathbb{E}_n\left[T(\frac{\partial e}{\partial\gamma} e^{-1}) \cdot 2(c-\mu(W;\beta))\frac{\partial\mu}{\partial\tilde{\theta}}\right].$$

□

Proof of Proposition C.1, GRDR*, Bang and Robins [2005].* The stacked estimation equations are as follows, for the parameters $[\gamma, \theta, \epsilon_1, \epsilon_0]$:

$$\begin{aligned} &-\mathbb{E}\left[2(T-e_T(W;\gamma))\frac{\partial e}{\partial\gamma}\right] = 0, \\ &-\mathbb{E}\left[2(c-\tilde{\mu}(W;\theta)\frac{\partial\mu}{\partial\theta}\right] = 0, \\ &-\mathbb{E}[2T(c-\tilde{\mu}_1(W;\theta)e_1^{-1}(W;\gamma)] = 0, \\ &-\mathbb{E}[2(1-T)(c-\tilde{\mu}_0(W;\theta)e_0^{-1}(W;\gamma)] = 0 \end{aligned}$$

and the Jacobian of partial derivatives is:

$$G_\gamma = \begin{bmatrix} \mathbb{E}[2(\epsilon_1(\frac{\partial}{\partial\gamma} e_1^{-1}(W;\gamma))) + \epsilon_0(\frac{\partial}{\partial\gamma} e_0^{-1}(W;\gamma)))\frac{\partial\mu}{\partial\theta} \\ -\mathbb{E}[2(c-\mu(W;\theta))T(\frac{\partial}{\partial\gamma} e_1^{-1}(W;\gamma))] \\ -\mathbb{E}[2(c-\mu(W;\theta))(1-T)(\frac{\partial}{\partial\gamma} e_0^{-1}(W;\gamma)] \end{bmatrix}.$$

□

C.1 Nonlinear Generalization of Perturbation Method

Preliminaries. We include the proof for completeness. It is the same argument of Ito et al. [2018] with the addition of linear expansions of the nonlinear model μ around the parameter ξ.

Let μ denote a generic prediction model which may depend nonlinearly upon its parameter ξ. Define for our context the true-optimal-decision $x(\xi^*)$, and sample-optimal-decision $\hat{x}(\hat{\xi})$:

$$x^* \in \operatorname{argmax} \sum_{i=1}^m \mu^*(W_i, \xi^*) x_i,$$
$$\hat{x} \in \operatorname{argmax} \sum_{i=1}^m \mu(W_i; \hat{\xi}) x_i.$$

Define the auxiliary functions η, ϕ evaluated along paths indexed by ϵ:

$$\eta(\epsilon) = \mathbb{E}_\delta \left[\sum_{i=1}^m x(\xi^* + \epsilon\delta) \mu(W_i; \xi^*) \right],$$
$$\phi(\epsilon) = \mathbb{E}_\delta \left[\sum_{i=1}^m x(\xi^* + \epsilon\delta) \mu(W_i; \xi^* + \epsilon\delta) \right].$$

We focus on the case exclusively where the function f of interest is affine, i.e. so that $f(z^*, \xi^*) = \sum_{i=1}^m z_i^* g(\xi^*; X_i)$.

Proof of Proposition 4.3. Let $\xi^*_\epsilon = \xi^* + \epsilon\delta$. First, we will show

$$\eta(\epsilon) - \phi(\epsilon) = \epsilon \, \mathbb{E}_\delta \left[\sum_{i=1}^m \hat{x}_i (\nabla_\xi \mu \Big|_{\xi_\epsilon} \delta) \right] + O(\epsilon^2). \tag{8}$$

and then that $\epsilon\phi'(\epsilon)$ equals the right-hand-side of the above.

Step 1 (Showing Equation (8)):

Expand the definition of η, ϕ and apply a Taylor expansion of $\mu(W_i; \xi^* + \epsilon\delta)$ from $\mu(W_i; \xi^*)$.

$$\eta(\epsilon) - \phi(\epsilon) = \mathbb{E}_\delta \left[\sum_{i=1}^m \left(\hat{x}_i \mu(W_i; \xi^*) - \hat{x}_i \mu(W_i; \hat{\xi}) \right) \right] = \mathbb{E}_\delta \left[\sum_{i=1}^m \hat{x}_i \left(\mu(W_i; \xi^*) - \mu(W_i; \hat{\xi}) \right) \right]$$
$$= \mathbb{E}_\delta \left[\sum_{i=1}^m \hat{x}_i \left(\epsilon (\nabla_\xi \mu \Big|_{\xi_\epsilon} \delta) + O(\|\epsilon\delta\|_2^2) \right) \right].$$

Step 2: $\epsilon\phi'(\epsilon) =$ RHS of Equation (8).

Let $\xi^*_{\epsilon+h} = \xi^* + (\epsilon + h)\delta$.

By definition,

$$\phi'(\epsilon) = \lim_{h \to 0} \frac{\phi(\epsilon + h) - \phi(\epsilon)}{h}$$
$$= \lim_{h \to 0} \frac{1}{h} \left(\mathbb{E}_\delta \left[\sum_{i=1}^m x_i(\xi^*_{\epsilon+h}) \mu(W_i; \xi^*_{\epsilon+h}) - \sum_{i=1}^m x_i(\xi^*_\epsilon) \mu(W_i; \xi^*_\epsilon) \right] \right).$$

Add / subtract $\hat{x}(\xi_\epsilon)\mu(W_i; \xi_\epsilon)$:

$$\phi'(\epsilon) = \lim_{h \to 0} \left\{ \frac{1}{h} \left(\mathbb{E}_\delta \left[\sum_{i=1}^m \left(x_i(\xi^*_{\epsilon+h}) \mu(W_i; \xi^*_\epsilon) + x_i(\xi^*_{\epsilon+h}) (\mu(W_i; \xi^*_{\epsilon+h}) - \mu(W_i; \xi^*_\epsilon)) \right) \right] \right) - \frac{1}{h} \left(\mathbb{E}_\delta \left[\sum_{i=1}^m x_i(\xi^*_\epsilon) \mu(W_i; \xi^*_\epsilon) \right. \right. \right.$$
$$= \lim_{h \to 0} \left\{ \frac{1}{h} \left(\mathbb{E}_\delta \left[\sum_{i=1}^m (x_i(\xi^*_{\epsilon+h}) - x_i(\xi^*_\epsilon)) \mu(W_i; \xi^*_\epsilon) + \sum_{i=1}^m (x_i(\xi^*_{\epsilon+h}) (\mu(W_i; \xi^*_{\epsilon+h}) - \mu(W_i; \xi^*_\epsilon)) \right] \right) \right\}$$
$$= \lim_{h \to 0} \left\{ \frac{1}{h} \left(\mathbb{E}_\delta \left[\sum_{i=1}^m (x_i(\xi^*_{\epsilon+h}) - x_i(\xi^*_\epsilon)) \mu(W_i; \xi^*_\epsilon) + \sum_{i=1}^m (x_i(\xi^*_{\epsilon+h}) (\mu(W_i; \xi^*_{\epsilon+h}) - \mu(W_i; \xi^*_\epsilon)) \right] \right) \right\}.$$

The last line follows by a Taylor expansion of μ from ξ^*_ϵ to $\xi^*_{\epsilon+h}$ and noting that the first term converges as x does, $\lim_{h\to 0}\frac{1}{h}(\sum_{i=1}^m (x_i(\xi^*_{\epsilon+h}) - x_i(\xi^*_\epsilon))\mu(W_i;\xi^*_\epsilon)) = 0$. under regularity conditions common in perturbation analysis of stochastic programs, such as uniqueness of the solution.

Therefore, interchanging limits and the expectation:

$$\begin{aligned}
\phi'(\epsilon) &= \mathbb{E}_\delta\left[\lim_{h\to 0}\left\{\frac{1}{h}\left(\sum_{i=1}^m x_i(\xi^*_{\epsilon+h})(\nabla_\xi \mu\Big|_{\xi_\epsilon}(h\delta) + O(\|h\epsilon\|_2^2))\right)\right\}\right] \\
&= \mathbb{E}_\delta\left[\lim_{h\to 0}\left\{\sum_{i=1}^m x_i(\xi^*_{\epsilon+h})(\nabla_\xi \mu\Big|_{\xi_\epsilon}\delta + O(h)\right\}\right] \\
&= \mathbb{E}_\delta\left[\sum_{i=1}^m x_i(\xi^*_\epsilon)(\nabla_\xi \mu\Big|_{\xi_\epsilon}\delta)\right].
\end{aligned}$$

□

D Alternative Asymptotic Regime: In-sample, growing-dimension (Assumption 2.2)

In the main text we focused on a fixed-dimension regime. We describe some extensions that may be possible to handle an in-sample, growing dimension, growing-n regime described in Assumption 2.2. We do generally require additional structural information to apply more familiar OPE estimators such as IPW/AIPW and other adaptations of bias adjustment methods.

The strongest such additional structural knowledge is that the optimization is highly structured so as to admit a finite VC dimension; to circumvent issues related to the growing dimension.

Assumption D.1. $x(\pi, W)$ has finite VC dimension.

Such a structural characterization is established in special cases such as multi-knapsack linear programs [van de Geer and Stougie, 1991], or large-market limits of stable matching markets [Azevedo and Leshno, 2016]; but need not hold in general.

D.1 Preliminaries

For completeness, we describe the analogous estimands/estimators for the *in-sample, growing-n* regime as described in the main text for the out-of-sample, fixed-m regime.

Plug-in estimation is evaluated on the training dataset as follows:

$$\hat{v}_\pi = \min_{x \in \mathcal{X}} \left\{ n^{-1} \sum_{i=1}^{n} \sum_{t \in \{0,1\}} \pi_t \hat{\mu}_t(w_i) x_i \colon Ax \le b \right\}. \tag{9}$$

We describe extensions of approaches to handle in-sample bias in this growing-dimension setting, although we generally require more structure on the problem. As described in Assumption 2.2 we typically require a problem-dependent asymptotic scaling; for example that we jointly scale up the problem size as well as the constraints. We provide a concrete example for Example 2.3.

Example D.2 (Fluid limit for Example 2.3)**.** The number of workers is αn and the number of jobs is βn.

D.2 Sample splitting

We first discuss an analogous sample splitting extension of Ito et al. [2018] which combines their sample splitting procedure with standard cross-fitting for doubly robust estimators [Chernozhukov et al., 2018]. However, a naive extension requires four folds and is therefore expected to perform poorly in finite samples.

Let K denote the number of folds, we will exposit the case of two folds and the K-fold generalization is standard. Denote two main folds of the data, $\mathcal{I}_{k_1}, \mathcal{I}_{k_2}$ denoting the index sets for the data, and $\mathcal{I}_{k_{1e}}, \mathcal{I}_{k_{1\mu}}$ be subfolds of $\mathcal{I}_{k_1}$ (respectively subfolds for $\mathcal{I}_{k_2}$). As suggested by the notation, we use distinct subfolds $\mathcal{I}_{k_{1e}}, \mathcal{I}_{k_{1\mu}}$ to learn the nuisance estimates e, μ (from the respective subfold).

The main difference from standard cross-fitting is that in Assumption D.1 we assume the optimization problem is well-parametrized in covariates: the optimization solution is well-described as a function of $x(W)$. We also require that there is a sensible way of sampling datapoints and projecting the feasible set $\mathcal{X}$ onto each subsampled index set. E.g. when subsampling in a matching example, after subsampling nodes the new feasible set in each index set $\mathcal{X}_1, \mathcal{X}_2$ preserves all edges between nodes in the original feasible set $\mathcal{X}$.

Let $\Gamma_t(O_i; e, \mu)$ denote the score associated with observation $O_i = (W_i, T_i, c_i)$ under either IPW or AIPW, with input nuisance functions e, μ. For example, as in Section 4.1, $\Gamma_t^{\text{IPW}}(O; e, \mu) = \frac{\mathbb{I}[T=t]c}{e_t(w)}$ and $\Gamma_t^{\text{AIPW}}(O; e, \mu) = \frac{\mathbb{I}[T=t](c-\mu_t(w)}{e_t(w)} + \mu_t(w)$.

As is standard in cross-fitting we use distinct main folds in order to estimate nuisances (indexed by parameters ξ) for input into $\hat{x}$: that is,

$$\hat{x}_1(\hat{\xi}_2; W) \in \underset{x(W) \in \mathcal{X}_2}{\operatorname{argmin}} \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \sum_{t \in \{0,1\}} \pi_t(W_i) \Gamma_t(O_i; e^{-k(i)}, \mu^{-k(i)}) x_i,$$

and analogously for $\hat{x}_2$.

Then evaluation estimates the value within each fold using the optimal solution from the other fold:

$$\hat{v}_1 = \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \sum_{t \in \{0,1\}} \pi_t(W_i) \Gamma(O_i; e^{-k(i)}, \mu^{-k(i)}) \hat{x}_1(\hat{\xi}_1, W_i),$$

and we return the average over folds, $\frac{1}{2}(\hat{v}_1 + \hat{v}_2)$ or more generally the average of $\{\hat{v}_k\}_{k \in K}$.

When $K > 2$, for each fold k, we will use two subfolds $I_{-k,e}$ and $I_{-k,\mu}$ to estimate e, μ, and then obtain $\hat{x}_{-k}$ from I_{-k}. We evaluate the estimated objective with $\hat{x}_{-k}$, $\hat{e}_k$ and $\hat{\mu}_k$ and average over all folds.

Proposition D.3 (Unbiased estimation by sample splitting.). $\frac{1}{K} \sum_i^K \hat{v}_k = \mathbb{E}[\tilde{v}]$

Proof. Immediate from standard analysis of AIPW and sample-splitting of Ito et al. [2018]. □

D.3 Comparison of estimation properties in the two regimes (Table 1)

Out-of-sample, fixed-dimension. IPW/AIPW-type estimators cannot be applied in the out-of-sample regime of Assumption 2.1, by definition of the regime. However, we may obtain out-of-sample risk bounds on the decision regret in this regime, simply by virtue of out-of-sample generalization risk bounds on the generated regressors. For example, we effectively assume near-parametric regimes for the propensity score so that the conditions of Theorem 1 of Bertail et al. [2021], providing a generalization risk bound for two-stage reweighted empirical risk minimization with estimated weights (as in our Section 4.1), are met. Under assumption of uniformly bounded decision variables, applying the Cauchy-Schwarz inequality directly implies that statistical estimation consistency of our estimation approaches imply decision regret consistency, so that the estimation bias vanishes at a $O_p(n^{-\frac{1}{2}})$ rate. (However the statistical rate of optimization bias adjustment remains unclear).

In-sample, growing-dimension, growing-n. An analogous extension to sample splitting as in Ito et al. [2018] is possible in highly structured situations satisfying Assumption D.1. For a *fixed* optimization solution x, uniform generalization over $\pi \in \Pi$ is a consequence of uniform generalization with a stochastic (bounded) envelope function. However, in this regime, uniform generalization over both $\pi \in \Pi$ and $x \in \mathcal{X}$ is difficult because in the regime of Assumption 2.2, the dimension of the optimization grows as $n \to \infty$. Typical approaches to uniform convergence would require $x^*(\pi, W)$ (the optimal optimization solution at a fixed π) to converge uniformly over the space of policies and W.

Different estimation interpretations of GRDR **in the two regimes.** Note that benefits of GRDR in terms of doubly-robust estimation of the ATE (mixed-bias, rate double-robustness) are only relevant in the *in-sample regime* of Assumption 2.2. Recent work does show this specification obtains empirical benefits for confounded outcome estimation, in appeal to the sufficient balancing properties of the propensity score, that may also apply to the regime of Assumption 2.1.

E Beyond Linearity: Decision-Dependent Classifier Risk

The downstream optimization can also in turn be a prediction risk problem: treatments shift distributions upon which predictive risk models are trained [Paxton et al., 2013]. For example, the medical system simultaneously treats individuals but is also interested in large-scale predictive models from passively collected electronic health records, trained upon the realizations of health outcomes of the entire population, and so may generate distribution shifts in these predictive risk models [Agniel et al., 2018, Finlayson et al., 2021]. Therefore, the post-treatment predictive risk model introduces a downstream causal-policy dependent optimization response.

In the previous sections, we focused on linear optimization because plug-in estimation is consistent when the random variable enters linearly into the optimization problem. The challenge with nonlinearity is that such plug-in-approaches are no longer consistent and can introduce policy-dependent nuisance estimation functions.

Nonetheless, special structure of the problem can admit alternative estimation strategies.

E.1 Problem setup

Example E.1 (Decision-dependent classifier drift.)**.** We shift to notation more typical in statistics/machine learning to emphasize the setting. We model decision-dependent shift of predictive risk models in a repeated measurement setting.[10] Our observation trajectories[11] each comprise of $(L_0, Y_0, T_0, L_1(T_0), Y_1(T_0))$: baseline covariates (L_0, Y_0), time-0 treatment $T_0 \in \{0, 1\}$, and post-treatment covariates and outcome $(L_1(T_0), Y_1(T_0))$.

For example, L could measure patient state and Y a cardiac event within a given time period. Upon observation of (L_0, Y_0) a patient is treated with T_0; for example with more aggressive or wait-and-see treatment depending on $Y_0(T_0)$. While optimal treatment regimes focus on averages of individual-level outcomes $Y_1(T_0)$; in our policy-dependent response setting we model the problem of, for example, continuously monitoring "feedback loops" that may surface in predictive risk models that may generally be trained using large electronic health record databases. Said differently, we could have modeled this abstractly as a policy evaluation with the augmented set of "covariates", jointly (L_0, Y_0), and "outcomes" $(L_1(T_0), Y_1(T_0))$. However, we focus on the ultimate downstream predictive risk model which depends (nonlinearly) on all the outcomes of the population's units. The policy evaluation problem could evaluate the predictive loss of the downstream predictive model $f(L_1(\pi), \beta)$, where there is downstream optimization of the squared loss over β.

$$\min_\beta \mathbb{E}[(Y_1(\pi) - f(L_1(\pi), \beta))^2]. \tag{10}$$

The causal graph of Appendix E.2 describes the two-stage observation of individuals[12], comprising observation trajectories $(L_0, Y_0, T_0, L_1(T_0), Y_1(T_0))$. We consider in the general case a two-stage setting with a treatment affecting covariates and outcomes; upon which a predictive risk model is trained.

Example E.2 (Policy optimization for Example E.1, policy-dependent prediction.)**.** Consider the case of two different treatments with similar (conditional) average treatment effects, but one induces higher variability in outcomes which increases the fundamental noise level in the regression: harming the population prediction model and incurring higher loss. Optimizing between these two treatments, scalarizing population outcomes by the global term would result in choosing the less variable treatment.

Therefore, in this framework we may be interested in the following scalarized policy optimization problem:

$$\min_\pi \min_\beta \lambda\, \mathbb{E}[Y_1(\pi)] + (1-\lambda)\, \mathbb{E}[(Y_1(\pi) - f(L_1(\pi), \beta))^2].$$

[10] That is, observing covariates and outcomes from the same unit, measured at different time periods.

[11] We are therefore modeling "feedback loops" between outcomes Y_0 and the treatments administered to manage them via temporally distinct repeated measurements.

[12] We do not consider for now the causal effects of the prediction model, although this could be a direction for future work.

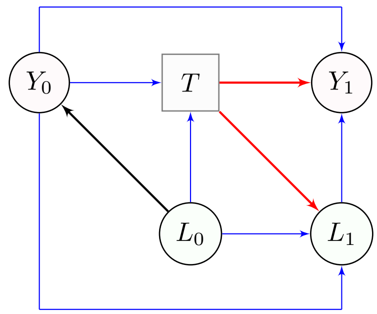


Figure 4: Causal diagram for decision-dependent classifier drift.

E.2 Estimation

Plug-in estimators. Our goal is off-policy evaluation of predictive risk model (parameter) solving the downstream prediction risk optimization after treatment under treatment regime π:

$$\beta^* \in \arg\min_{\beta} \mathbb{E}[(Y_1(\pi) - f(L_1(\pi); \beta)^2].$$

However for the special case that we consider of linear regression response (squared loss, linear parametrization), we may orthogonalize the *first-order optimality* conditions of the policy-dependent optimization, e.g. recognizing that, β^* solves the first order conditions

$$\mathbb{E}[L_1^\top (Y_1 - L_1\beta)] = 0. \tag{11}$$

Hence for the case of least squares and linear regression, we may focus on estimation refinements for β: observe that the estimation requires estimation of certain transformations of (X_1, Y_1) unit's downstream outcome, $x_1 y_1$ and $x_1^\top x_1$, and we may estimate the following matrix-regression or vector-valued regression nuisance estimates:

$$\mathbb{E}[L_1 Y_1 \mid T = t, X_0, Y_0], \qquad \mathbb{E}[L_1 L_1^\top \mid T = t, L_0, Y_0]$$

Hence standard AIPW-type approaches can be applied with the above nuisances and the censored observations $l_1 y_1(t), l_0 y_0(t)$.

This suggests that when β is our parameter of interest (or functions thereof), we can leverage double robustness. And, if we have a small space of policies we can optimize by enumeration. However the same challenges regarding nonlinearity remain if we want to estimate the final squared loss of θ.

Remark E.3 (Restriction to linear models and the challenge for generalization to nonlinear models). Note that the challenge with generalizing to non-least-squares losses or nonlinear predictors is that due to nonlinearity, doubly-robust estimation of outcome X_1 need not provide the same benefits of bias reduction. Although an alternative approach is to instead estimate the *squared loss* as the composite outcome $\mathbb{E}[(Y_1(t) - \theta^\top L_1(t))^2 \mid t, L_0, Y_0]$ because of our *policy-dependent response* optimizing over θ, we would have policy-dependent nuisance functions so this becomes intractable.

F Additional Experiment Details and Results

In this section, we provide more details on the experimental setup as well as further results.

F.1 Causal effect estimation setup

Fr the causal effect estimation we generated the training dataset $\mathcal{D}_1 = \{(W, T, c)\}$ with covariate $W \sim \mathcal{N}(0, 1)$, confounded treatment T, and outcome c. Treatment is drawn with probability $\pi_t^b(W) = \frac{1}{1+e^{-\varphi_1 W + \varphi_2}}, \varphi_1 = \varphi_2 = 0.5$. The true outcome model is given by a degree-2 polynomial:

$$poly_\theta(t, w) = (1, w, t, w^2, wt, t^2) \cdot ([5, 1, -1, 2, 2, -1])^\top.$$

We generate the outcome samples as $c_t(w) = poly_\theta(t, w) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. All random samples are generated using `numpy.random` package. In the mis-specified setting that induces confounding, the outcome model is a vanilla linear regression over W without the polynomial expansion.

In Fig. 1a and Fig. 1b in the main text, we illustrate the (covariate-conditional) estimation over the covariates' landscape for the direct method, the weighted direct method (WDM), and the doubly robust method (GRDR) when there is a model mis-specification. We provide the estimation results without model mis-specification in Fig. 5.

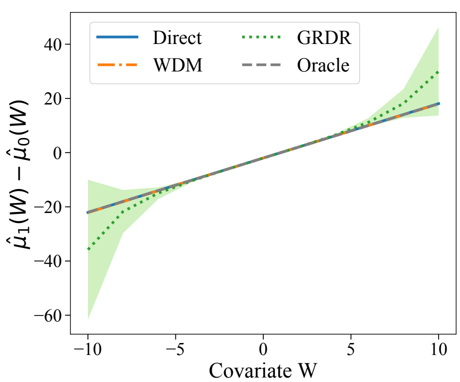


Figure 5: *(In-sample estimation of $\hat{\mu}_1(W) - \hat{\mu}_0(W)$, no model mis-specification).* Comparison of direct / weighted direct (WDM) / doubly robust method (GRDR) to the oracle estimator for estimation of conditional ATE over different covariate values. Results are averaged over ten random training datasets; shading area indicates the standard error.

When there is no model mis-specification that that induces confounding, we observe that both the three estimation methods perform well against the oracle estimation.

F.2 Policy evaluation

For policy evaluation, we compare the perturbation method (Algorithm 1) when being applied with three different estimators (direct, WDM, and GRDR). For consistency, throughout the evaluations, we follow the same true outcome model and covariate distribution as in the previous subsection for causal effect estimation. In both the well-specified model setting and the mis-specified model setting, the mean-squared-error (MSE) of the estimated policy value with the three estimators is computed with regard to the ground truth outcome model (aka oracle).

When evaluating Algorithm 1, we generated $S = 20$ bootstrap replicates. The downstream matching problem is evaluated with $m = 500$ left-hand-side nodes, and $m' = 300$ right-hand-side nodes. The min-cost matching requires each node to be matched to no more than one node on the other side, and was computed by the `linear_sum_assignment` function of the `scipy.optimize` package in Python 3. We evaluated a fixed logistic policy $\pi_t(W) = sigmoid(\phi \cdot W + b)$ with $\phi = 1, b = 0.5$.

Figure 2 shows that when there is mis-specification, even a large training dataset cannot bring bias correction for the direct method, where both WDM and GRDR enjoy smaller and decreasing MSE. As an ablation study, we also compare to the corresponding performance in the mis-specified setting

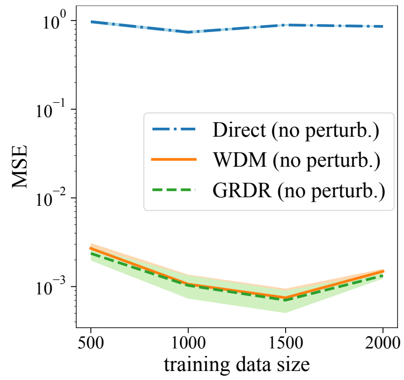


Figure 6: *(In-sample estimation of $\hat{\mu}_1(W) - \hat{\mu}_0(W)$ with model mis-specification, no perturbation applied)*. Comparison of direct / weighted direct (WDM) / doubly robust method (GRDR) over increasing size of training data. Results are averaged over ten random training datasets; shading area indicates the standard error.

Table 3: *(Perturbation method, varying replicate size.)* Performance for different estimator/model combinations. Mean-squared-errors (MSE) are computed with regard to the oracle outcome model.

	Estimation	$h = 1$	$h = 2$	$h = 3$	$h = 4$
Mis-specified model	Direct	0.59±0.05	0.70±0.05	0.76±0.06	0.70±0.06
	WDM	0.00031±0.0001	0.00042±0.0002	0.00041±0.0002	0.00048 ±0.0003
	GRDR	0.00040±0.0002	0.00046±0.0002	0.00031±0.0001	0.00035±0.0001
Well-specified model	Direct	0.00079±0.0004	0.00067±0.0004	0.00062±0.0003	0.00024±0.0002
	WDM	0.00076±0.0004	0.00067±0.0003	0.00080±0.0002	0.00031±0.0001
	GRDR	0.00082±0.0002	0.00067±0.0002	0.00080±0.0002	0.00031±0.0001

when we do not perform the perturbations (i.e. no bootstrapping in Alg. 1). In detail, we directly return $\hat{v}^{(0)}$ without doing the later bootstrap procedure. Figure 6 indicates that the perturbation method is helpful for MSE reduction for both WDM and GRDR.

We further conduct evaluations with different bootstrap replicates' sizes (controlled by variable h in Algorithm 1). We include these results in Table 3. Results in Table 3 show that WDM and GRDR remain more superior and that is robust with different replicate sizes. For the evaluations over different h values, we used training data with 3000 samples. In each iteration, the number of bootstrap replicates is 20.

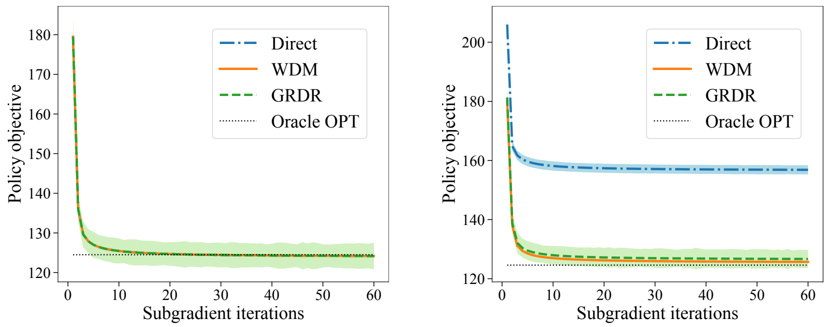
F.3 Policy optimization

For policy optimization, we implemented the subgradient method as in Algorithm 2, and obtained causal effect estimators from Algorithm 1.

In detail, for a given training dataset, we first obtained $S + 1$ outcome estimators (i.e. $\{\hat{\mu}_t^{\diamond}(w_i; \hat{\xi}_{\diamond}), \hat{\mu}_t^{\diamond}(w_i; \hat{\xi}_{\diamond})^{(j)}, j = 1 \cdots S\}$ in Algorithm 1) via bootstrap. Then, at each iteration of running subgradient descent, we evaluate the current policy using the $S + 1$ outcome estimators respectively, and obtain $S + 1$ subgradients of it. We then aggregate these subgradients by the bootstrap aggregation (as in Step 6, Algorithm 1).

We evaluate subgradients of the inner optimization solution in Algorithm 2 (step 4) by evaluating the gradient of the objective with respect to φ, fixing the inner optimization variable x^*. The fact that ∇_{φ} is a subgradient is a consequence of Danskin's theorem [Danskin, 1966]. The inner minimization (the matching problem) is again solved by the `linear_sum_assignment` function of the `scipy.optimize` package in Python 3.

To further study the impact of the random initial policies to begin with the subgradient descent algorithm, in Figure 7 we obtained the corresponding results of Figure 3, but with a fixed initial policy. We observe that again WDM and GRDR quickly converges to the oracle estimation, while the large bias of the direct method leads to poor policy optimization. Moreover, in this relatively low-dimension



(a) Policy optimization with well-specified models

(b) Policy optimization with model mis-specification

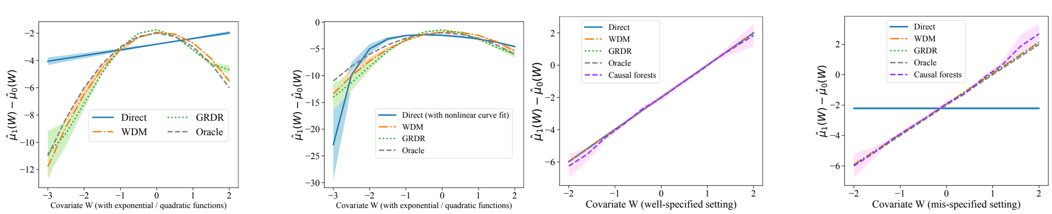
Figure 7: *(Policy optimization (fixed test set))*. Results of subgradient policy optimization with direct / weighted direct (WDM) / doubly robust (GRDR) estimation methods and a fixed test set. Averaged over ten random training datasets of size 1000.

example, although random initialization of the policy leads to a larger variance in earlier iterations, the policy value convergences to oracle policy objective quickly after a few iterations.

For the evaluations of policy optimization, we used training datasets with size 1000, and a downstream min-cost matching with $m = 100, m' = 60$. The learning rate was tuned over $[0.01, 0.1, 1]$. All of our evaluations were run on a 2.3 GHz 8-Core Intel Core i9 CPU. All the differentiation operations were handled by the automatic differentiation library in `JAX`.

F.4 Additinal comparisons and evaluations

Additional evaluations with more complex non-linear outcome model. We conduct further robustness checks with nonlinear data-generating processes: exponential and quadratic. The outcome is $c_t(w) = a_1 + a_2 \exp(b_1 + b_2 w) + c_1 t + c_2 t w^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ is an external noise, and $[a_1, a_2, b_1, b_2, c_1, c_2] = [5, 0.05, 0.5, -2, -2, -1]$. We fit this function with nonlinear least squares (*scipy.optimize.curve_fit*). Indeed, Figure 8 shows that the direct method is sensitive to the model mis-specification bias without using a near-specified nonlinear curve fit. However the weighted direct method (WDM) and the doubly robust estimator (GRDR) remain robust; even if starting with misspecified parametric models.



(a) (b) (c) (d)

Figure 8: (In-sample estimation of $\hat{\mu}_1(W) - \hat{\mu}_0(W)$ for exponential function, (a) without / (b) with curve fit. Comparisons of the CATE estimates with nonparametric estimators. (c) without / (d) with model mis-specification.

Additional comparison to non-parametric estimators. We further compare our estimators against existing CATE estimators (although in general, these estimators tuned to estimate contrasts may improve upon differencing outcome models). We compared the DM/WDM/GRDR to the Causal Random Forests estimator proposed in (Wager and Athey, 2018) [13], following the setup

[13]Based on implementation by Battocchi et al, *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation.*

in Section 5.1. Moreover, we also compared how the CATE estimators affect the policy evaluation task with the perturbation method (section 5.2). In Figure8(c,d) comparing CATE estimates, the non-parametric random forest estimator is indeed unbiased, while the naive direct method has a large bias under model mis-specification.

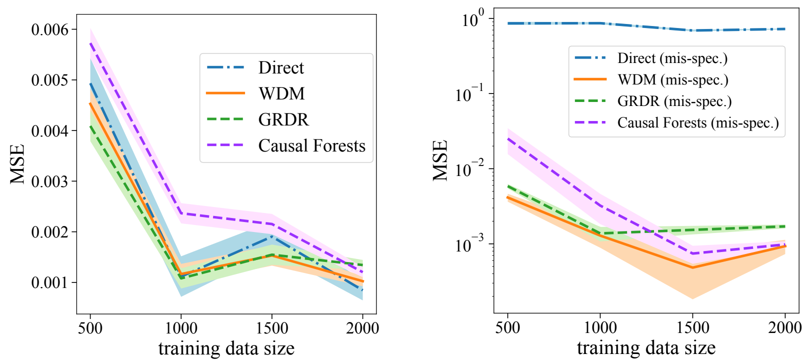


Figure 9: *(Policy evaluation via perturbation method (Algorithm 1)*. Comparison of direct / WDM / GRDR / Causal Forests estimators over increasing size of training data.