Policy Learning with Asymmetric Counterfactual Utilities

Eli Ben-Michael\textsuperscript{a}, Kosuke Imai\textsuperscript{b}, and Zhichao Jiang\textsuperscript{c}

\textsuperscript{a}Department of Statistics & Data Science and Heinz College of Information Systems & Public Policy, Carnegie Mellon University, Pittsburgh, PA; \textsuperscript{b}Department of Government and Department of Statistics, Institute for Quantitative Social Science, Harvard University, Cambridge, MA; \textsuperscript{c}School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong, China

ABSTRACT
Data-driven decision making plays an important role even in high stakes settings like medicine and public policy. Learning optimal policies from observed data requires a careful formulation of the utility function whose expected value is maximized across a population. Although researchers typically use utilities that depend on observed outcomes alone, in many settings the decision maker’s utility function is more properly characterized by the joint set of potential outcomes under all actions. For example, the Hippocratic principle to “do no harm” implies that the cost of causing death to a patient who would otherwise survive without treatment is greater than the cost of forgoing life-saving treatment. We consider optimal policy learning with asymmetric counterfactual utility functions of this form that consider the joint set of potential outcomes. We show that asymmetric counterfactual utilities lead to an unidentifiable expected utility function, and so we first partially identify it. Drawing on statistical decision theory, we then derive minimax decision rules by minimizing the maximum expected utility loss relative to different alternative policies. We show that one can learn minimax loss decision rules from observed data by solving intermediate classification problems, and establish that the finite sample excess expected utility loss of this procedure is bounded by the regret of these intermediate classifiers. We apply this conceptual framework and methodology to the decision about whether or not to use right heart catheterization for patients with possible pulmonary hypertension. Supplementary materials for this article are available online.

1. Introduction
The well-known Trolley Problem in ethics goes as follows:

Edward is the driver of a trolley, whose brakes have just failed. On the track ahead of him are five people; the banks are so steep that they will not be able to get off the track in time. The track has a spur leading off to the right, and Edward can turn the trolley onto it. Unfortunately there is one person on the right-hand track. Edward can turn the trolley, killing the one; or he can refrain from turning the trolley, killing the five (Thomson 1976, p. 206).

Should Edward turn the trolley? Is killing someone worse than letting them die? Such ethical dilemmas frequently confront us in moral and legal debates concerning various issues that range from abortion to self-driving cars (e.g., Foot 1967; Lin 2016). Similarly, in the ethics of modern medicine, the Hippocratic principle of “do no harm” remains influential (e.g., Jonsen 1978; Smith 2005; Wiens et al. 2019). In the language of utility theory, a physician may assign a utility loss of greater magnitude to the case where a new drug harms a patient than to the case where not providing the new drug leads to the failure to save a patient (e.g., Bordley 2009).

These examples illustrate the potential applications of asymmetric counterfactual utilities that depend not only on the observed outcome, but also on the counterfactual outcome that could occur under a different action, and treat actions differently depending on their corresponding potential outcomes. Yet, to the best of our knowledge, the existing literature on data-driven decision making and algorithmic policy learning assumes that the decision maker’s utility function only depends on the observed outcome. In this article, we develop the methodological framework for optimal policy learning with asymmetric counterfactual utilities, which includes standard utilities based on marginal outcomes as a special case. We show that in general, asymmetric counterfactual utilities lead to an unidentifiable expected utility function. Therefore, we partially identify the expected utility and propose to minimize the maximum expected utility loss relative to a particular comparison policy. We consider the maximum expected utility loss relative to constant policies such as always-treat and never-treat policies as well as the oracle policy that has complete knowledge of the unidentifiable terms in the expected utility function. We demonstrate that one can learn minimax decision rules from observed data by solving intermediate classification problems. We also establish that the finite sample regret
of this procedure is bounded by the regret of these intermediate classifiers.

We use this framework to reassess the use of Right Heart Catheterization (RHC), an invasive diagnostic tool (Connors et al. 1996). We learn decision rules based on clinical variables as we vary the asymmetry in the costs between failing to prevent a patient’s death and causing it via RHC. These decision rules differ depending on whether we minimize the worst-case expected utility loss relative to a constant policy (always or never using RHC) or the oracle policy that uses RHC optimally. We inspect how the choice of utility function and comparator affect the learned decision rules, finding substantial variability based on these choices. Finally, we translate these findings into directly interpretable patient outcomes, exhibiting a tradeoff between limiting the worst-case proportions of patients that the policy harms or fails to save.

Related Literature. Recent years have seen an increased interest in algorithmic policy learning from randomized control trials or observational data. Many of these approaches follow a similar structure. First, quantify the expected utility of a policy based on the marginal distributions of the potential outcomes. Then, show how to identify the expected utility or regret from observable data and find a policy that optimizes an empirical analog. These approaches typically use inverse propensity score weighting or double-robust methods for the identification and estimation steps (see Zhao et al. 2012; Kitagawa and Tetenov 2018; Athey and Wager 2021, among others). There is also a related literature that focuses on identifying and estimating optimal policies in settings with unmeasured confounding via instrumental variables (see Cui and Tchetgen Tchetgen 2021; Qiu et al. 2021).

More immediately relevant to our discussion here, recent work builds off classical ideas in decision theory and treatment choice (e.g., Manski 2004, 2005, 2011) and considers scenarios where we cannot point identify the expected utility function for possible policies. One strand of work considers choosing between two treatments or fixed decision rules based on a finite sample of data, when treatment effects are partially identified (e.g., Stoye 2012; Ishihara and Kitagawa 2021; Yata 2021). This work typically involves directly solving an empirical minimax regret problem, but does not consider optimization over classes of individualized policies.

In contrast, another line of work considers learning optimal individualized decision rules in situations where treatment effects are only partially identified, using an empirical risk minimization approach. These include settings with unmeasured confounding (e.g., Kallus and Zhou 2021; Pu and Zhang 2021; Han 2023; Cui 2021) or limited overlap between different treatment conditions (e.g., Ben-Michael et al. 2021; Zhang, Ben-Michael, and Imai 2023). D’Adamo (2023) considers a general setup where the conditional expected potential outcomes and treatment effects are partially identified. These approaches take a minimax approach at the population level, deriving the population-level minimum expected utility or maximum regret. They then treat the population-level maximum regret or negative minimum expected utility as a risk, and use empirical risk minimization approaches and propensity score weighting or double-robust methods as above. Our work is in the vein, estimating individualized treatment rules via empirical risk minimization. However, we consider a different setting where treatment effects are point identified, but the expected utility function is partially identified because it is a function of the proportion of units within each principal stratum—an unidentifiable quantity under standard designs.

Finally, Babii et al. (2021) also consider asymmetric utilities, but only using observed outcomes. In contrast, we consider asymmetric counterfactual utilities, which depend on potential outcomes and is a generalization of Babii et al.’s approach (see Appendix D for details).

Paper outline. The article proceeds as follows. Section 2 describes the goal of policy learning with asymmetric counterfactual utilities and reviews the standard symmetric case. Section 3 discusses partial identification of the expected utility function and the minimax population policies relative to different alternatives. Section 4 then shows how to estimate such policies from data. Finally, Section 5 applies this framework to the use of RHC, and Section 6 concludes.

2. Preliminaries

In this section, we introduce the notation and assumptions used throughout this article. We also discuss the nature of asymmetric counterfactual utilities before providing a brief review of policy learning with symmetric utilities, which is a special case of our proposed framework.

2.1. Notation and Assumptions

Suppose that we have a simple random sample of $n$ units from a super population $P$ where each unit $i = 1, \ldots, n$ has a set of characteristics $X_i \in \mathcal{X}$. We consider a binary treatment assignment decision $D_i \in \{0, 1\}$, which can be made by either individual $i$ or a policy maker. We assume that the outcome $Y_i$ is binary with $Y_i = 1$ indicating a desirable outcome (e.g., survival) and $Y_i = 0$ representing an undesirable outcome (e.g., death). Under the assumption that there is only one version of treatment and no interference across units, we have two binary potential outcomes for each unit $i$ where $Y_i(d) \in \{0, 1\}$ represents the potential outcome under the scenarios where the unit receives the decisions $D_i = d$ for $d = 0, 1$.

The setup implies that the observed outcome for unit $i$ can be written as $Y_i = Y_i(1) + (1 - D_i)Y_i(0)$ and the tuple of random variables $(X_i, D_i, Y_i(1), Y_i(0))$ is assumed to be independently and identically distributed. Importantly, under this setting, each unit belongs to one of the four principal strata defined by the values of the two potential outcomes, that is, $(Y_i(1), Y_i(0)) = (y_1, y_0) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ (Frangakis and Rubin 2002). For example, the principal stratum $(y_1, y_0) = (1, 0)$ represents a group of units who would yield the desirable outcome only when they are treated, that is, $D_i = 1$, whereas the principal stratum $(y_1, y_0) = (1, 1)$ indicates a group of units whose outcome is desirable regardless of the treatment decision. Since we never observe the two potential outcomes at the same time for any given unit, it is impossible to know which principal stratum each unit belongs to without additional assumptions.
Throughout this article, for notational simplicity, we will drop the individual $i$ subscript in expressions involving expectations over the distribution. We will also assume the strong ignorability and strict overlap assumptions for observational studies and randomized control trials.

**Assumption 1 (Strong ignorability and strict overlap).** \( \{Y(1), Y(0)\} \perp D \mid X \) and there exists an $\eta > 0$ such that $\eta < d(x) < 1 - \eta$ for all $x \in X$ where $d(x) \equiv \Pr(D = 1 \mid X = x)$ represents the propensity score.

The assumption allows us to identify the expected potential outcome under decision $d$ given covariates $x$, denoted as $m(d, x) \equiv \mathbb{E}[Y(d) \mid X = x]$. However, it is impossible to identify the principal score, or the conditional probability of belonging to a principal stratum given covariates, defined as $e_{Y_1Y_0}(x) \equiv \Pr(Y(1) = y_1, Y(0) = y_0 \mid X = x)$, because we do not observe the two potential outcomes at the same time for a given unit (Ding and Lu 2017; Jiang, Yang, and Ding 2022).

### 2.2. Asymmetric Counterfactual Utilities

We focus on deterministic individualized policies $\pi : X \rightarrow \{0, 1\}$ that assign a binary treatment decision to individual units according to their characteristics $X \in X$. To learn optimal policies from the observed data, we consider a utility function $u(d; y_1, y_0)$ that encodes the utility for taking treatment decision $d$ for a unit in principal stratum $(y_1, y_0)$. Crucially, this utility function depends on the values of both potential outcomes. This contrasts with the standard utility function $u(d; y)$, which only depends on the realized potential outcome $Y_i(d) = y$ under the decision $d$. We measure the overall quality of a policy $\pi$ by its expected utility (also called the value or social welfare),

\[
V(\pi) = \mathbb{E} \left[ \sum_{y_1 = 0}^{1} \sum_{y_0 = 0}^{1} \mathbb{1}\{Y(1) = y_1, Y(0) = y_0\} \left\{ u(0; y_1, y_0)(1 - \pi(X)) + u(1; y_1, y_0)\pi(X) \right\} \right].
\]

\[
= \mathbb{E} \left[ \sum_{y_1 = 0}^{1} \sum_{y_0 = 0}^{1} e_{y_1y_0}(X)\pi(X) \left\{ u(1; y_1, y_0) - u(0; y_1, y_0) \right\} \right] + \mathbb{E} \left[ \sum_{y_1 = 0}^{1} \sum_{y_0 = 0}^{1} e_{y_1y_0}(X)u(0; y_1, y_0) \right].
\]

This setup lets the utility vary across different counterfactual outcomes even when the treatment decision and the realized outcome are the same, allowing for a richer specification of the decision problem. For example, the disutility from assigning treatment to a patient that is harmed by it (i.e., $Y(1) = 0$ and $Y(0) = 1$) can be larger than the disutility from assigning treatment to a patient for whom it is useless (i.e., $Y(1) = Y(0) = 0$), despite the fact that the realized outcome $Y(1)$ is identical in both cases. A standard utility function does not distinguish between these two cases, assigning each a value of $u(1; Y(1))$. This utility function also allows for asymmetry in the utility gain or loss from treating a unit across principal strata. Returning to the Hippocratic oath, we can choose the utility function such that the absolute magnitude of the utility loss for harming a patient through treatment ($\|u(1; 0, 1) - u(0; 0, 1)\|$) is greater than that of the utility gain when the same treatment benefits another patient ($\|u(1; 1, 0) - u(0; 1, 0)\|$).

To encode this, and focus on key ideas, we parameterize the utility function as follows:

1. **Utility Gain:**
   - the utility gain associated with a “useful treatment,” that is, $(y_1, y_0) = (1, 0)$, is $u_g - c_g \equiv u(1; 1, 0) - u(0; 1, 0)$ (e.g., treating with a drug that would benefit the patient)
   - the utility gain associated with a “harmful treatment,” that is, $(y_1, y_0) = (0, 1)$, is $-u_l - c_l \equiv u(1; 0, 1) - u(0; 0, 1)$ (e.g., treating with a drug that would harm the patient)

2. **Utility Loss:**
   - the utility loss associated with a “harmless treatment,” that is, $(y_1, y_0) = (1, 1)$, is $-u_l \equiv u(1; 1, 1) - u(0; 1, 1)$ (e.g., treating with a drug that would not harm the patient)
   - the utility loss associated with a “useless treatment,” that is, $(y_1, y_0) = (0, 0)$, is $-u_g \equiv u(1; 1, 0) - u(0; 0, 0)$ (e.g., treating with a drug that would not benefit the patient)

The values $c_g, c_l$, and $u_g$ denote the cost of administering the treatment $d = 1$ relative to not doing so $d = 0$ in each of the four principal strata. The values $u_g$ and $u_l$ represent the magnitude of the utility gain and loss for administering a useful and harmful treatment, respectively. In this setting, these utility values are known and fixed by the decision maker. Utility functions of this and more general forms have been considered in the literature on decision theory (see, e.g., Stefansson 2015; Bradley and Stefansson 2017). Our focus is, however, on the estimation of individualized decision rules under these asymmetric counterfactual utility functions.

Table 1 summarizes this asymmetric counterfactual utility structure. Fixing the costs to be identical amounts to restricting the utility loss from a harmless and useless treatment to be equal. Without this restriction there will be an additional asymmetry due to the different costs, which would not affect our development, except to make the notation more cumbersome and results less interpretable.

Note that our asymmetric counterfactual utilities include symmetric utilities based on observed outcomes as special cases, thereby generalizing the standard setting considered in the policy learning literature. In Appendix D, we further show that although it is possible to construct asymmetric utilities without using principal strata (Babii et al. 2021), doing so still implies

| $Y(1)$ | $Y(0)$ | $u_g - c_g$ | $u_l - c_l$ | $-u_g$ | $-u_l$ |
|-------|-------|------------|------------|--------|-------|
| 1     | 0     | Harmless   | Useful     | Harmful| Useless|
| 0     | 1     | $-c$       | $u_g - c$  | $-u_l - c$ | $-c$ |

NOTE: Each cell corresponds to the principal stratum defined by the values of the two potential outcomes, $Y(1)$ and $Y(0)$. Each entry represents the utility gain/loss of treatment assignment, relative to no treatment, for a unit that belongs to the corresponding principal stratum $(u(1; y_1, y_0) - u(0; y_1, y_0))$. $c$ is the cost of treatment assignment. We assume that $u_l, u_g > 0$ and $c \geq 0$. Symmetric utilities are a special case with $u_g = u_l$.

### Table 1. Asymmetric counterfactual utility gain/loss for treating each of the four principal strata, relative to not treating.

| $Y(1)$ | $Y(0)$ | $u_g - c_g$ | $u_l - c_l$ | $-u_g$ | $-u_l$ |
|-------|-------|------------|------------|--------|-------|
| 1     | 0     | Harmless   | Useful     | Harmful| Useless|
| 0     | 1     | $-c$       | $u_g - c$  | $-u_l - c$ | $-c$ |
some restrictions on the structure of the resulting counterfactual utilities and hence they are a special case of our framework.

We will primarily be comparing two policies rather than considering one in isolation. We begin by defining the expected utility loss of policy π relative to another policy σ as the difference in values, \( V(\sigma) - V(\pi) \). Using the relations \( m(1,x) = e_{11}(x) + e_{10}(x) \) and \( m(0,x) = e_{11}(x) + e_{01}(x) \), we show in Appendix J that we can write the expected utility loss in a simplified form:

\[
R_{01}(\pi, \sigma) \equiv V(\sigma) - V(\pi) = \mathbb{E} \left[ (\sigma(X) - \pi(X))(u_\pi X + (u_\pi - u_\sigma)e_{01}(X) - c) \right],
\]

where \( r(x) \equiv \mathbb{E}[Y(1) - Y(0) | X = x] = m(1,x) - m(0,x) \) is the conditional average treatment effect (CATE) given the covariates \( X = x \). Comparing two policies to each other allows us to leave the baseline utility for not treating a unit in principal stratum \((y_1, y_0), u(0; y_1, y_0) \) unspecified. In Appendix B we directly consider the expected utility of a single policy, and connect choices of the baseline utility to our discussion below.

Three components contribute to the expected utility loss in (1). First, the difference in the expected treatment effects for those treated under policies \( \sigma \) and \( \pi \), scaled by the utility gain for a useful treatment \( u_\pi \), represents a symmetric component of the utility, where we compare the marginal benefits of the policies. The second component is an asymmetric adjustment term, and relates to the probability of belonging to the principal stratum for whom the treatment is harmful (i.e., \((y_1, y_0) = (0, 1)\)). This can counteract the marginal benefit of treatment and is scaled by the difference between the utility gain for a useful treatment and the loss for a harmful treatment, that is, \( u_\pi - u_\sigma \). The final term \( c \) corresponds to the difference in the overall costs of the two policies.

The first and third components, the difference in effects and costs, are point identifiable under Assumption 1. The second component, however, is only partially identifiable due to the unidentifiability of the principal score \( e_{01}(\cdot) \). We use the \( e_{01} \) subscript for the expected utility loss in (1) to signify this fact. Therefore, we cannot pinpoint whether any policy is superior to any other policy in general. The remainder of this paper focuses on handling this ambiguity.

### 2.3. Policy Learning with Symmetric Utilities: A Review

Before discussing policy learning under asymmetric counterfactual utility functions, we briefly review policy learning with symmetric utilities—a special case of our framework—where the absolute magnitude of the utility gain when the treatment leads to a desirable outcome is equal to that of the expected utility loss when it leads to an undesirable outcome, that is, \( u_\pi = u_\sigma \). This case, a policy can make up for the loss from harming some units by the gain from benefitting other units. This can be seen in the following simplified version of the expected utility loss in (1):

\[
R_{\text{symm}}(\pi, \sigma) = \mathbb{E} \left[ (\sigma(X) - \pi(X))(u_\pi X - c) \right].
\]

The symmetric utility does not involve the principal score \( e_{10}(\cdot) \), and is identifiable under Assumption 1. Thus, under this setting, the oracle optimal policy that minimizes the expected utility loss relative to any other policy is \( \pi_{\text{symm}}^*(x) = \mathbb{1}[u_\pi X \geq c] \). This oracle policy assigns the treatment to all individuals with characteristics \( x \) if their expected utility gain of assigning the treatment relative to not assigning it at least makes up for its cost, that is, \( u_\pi X \geq c \). Note that this is equivalent to maximizing the value \( V(\pi) \) directly.

Under Assumption 1, therefore, we can write the symmetric expected utility loss in (2) in terms of the observed data by using a scoring function \( \Gamma_w(x, d, y) \) such that \( \mathbb{E} \left[ \Gamma_w(X, D, Y) | X = x \right] = m(w, x) \). For example, the Inverse Probability-of-treatment Weighting (IPW) scoring function uses the IP weighting function \( \gamma_w(D, X) \equiv \frac{wD}{\mathbb{E}[w|X]} + (1-w)(1-D) \) to weight the observed outcome by the inverse probability of receiving the decision \( d: \Gamma_w^{\text{ipw}}(X, D, Y) = Y\gamma_w(D, X) \). An alternative is the Doubly Robust (DR) scoring function that combines the observed outcomes and their conditional expectations:

\[
\Gamma_w^{\text{dr}}(X, D, Y) = m(w, X) + (Y-m(w, X))\gamma_w(D, X).
\]

With such a scoring rule, we can then write the symmetric expected utility loss function as

\[
\mathbb{E} \left[ (\sigma(X) - \pi(X))(u_\pi \Gamma_1(1, X, D, Y) - \Gamma_0(1, X, D, Y) - c) \right],
\]

where the observable quantity \( \Gamma_1(1, X, D, Y) - \Gamma_0(1, X, D, Y) \) has replaced the causal quantity \( r(X) \). See Knaus (2020) for a recent review.

In order to empirically find optimal policies from data, recent approaches estimate the propensity score \( \hat{d}(\cdot) \) and/or the conditional expected potential outcome \( \hat{m}(\cdot, \cdot) \) to create estimated scores \( \hat{\Gamma}(X_i, D_i, Y_i) \). For example, we can estimate the IP weights as \( \hat{\gamma}_w(D, X) \equiv \frac{w\hat{D}}{\mathbb{E}[w|X]} + (1-w)(1-D) \), the IPW scoring function as \( \hat{\Gamma}_w^{\text{ipw}}(X, D, Y) := Y\hat{\gamma}_w(D, X) \), and the DR scoring function as \( \hat{\Gamma}_w^{\text{dr}}(X, D, Y) := \hat{m}(w, X) + (Y - \hat{m}(w, X))\hat{\gamma}_w(D, X) \). Then, we solve the sample analog of (2). This leads to finding a policy \( \hat{\pi} \) that solves the following optimization problem:

\[
\min_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \pi(X_i) \left[ u_\pi \left( \hat{\Gamma}_1(X_i, D_i, Y_i) - \Gamma_0(X_i, D_i, Y_i) \right) - c \right],
\]

where \( \Pi \) represents the policy class and restricts the functional form of potential policies. Athey and Wager (2021) establish strong asymptotic guarantees on the regret of the empirical \( \hat{\pi} \) relative to the best-in-class policy when using the DR approach with appropriately cross-fit models (see also Zhao et al. 2012; Kitagawa and Tettenov 2018).

### 3. Policy Learning with Asymmetric Counterfactual Utilities

We now turn to the problem of finding optimal policies in the general asymmetric case where \( u_\pi \neq u_\sigma \). We will first consider the identification problems in the population—that is, with infinite data. We then show how to learn policies empirically from observed data in Section 4. In Appendix A, we consider an alternative formulation as a constrained optimization problem.

#### 3.1. The Oracle Policy with an Asymmetric Counterfactual Utility Function

We begin by considering the oracle policy in the general asymmetric case. By direct computation, the (unconstrained) policy
that has the maximal possible value with an asymmetric counterfactual utility function is given by:

\[ \pi^* \equiv \arg\max_{\pi} V(\pi) = 1 \left\{ \tau(\cdot) \geq \frac{u_l - u_g}{u_g} \epsilon_0(\cdot) + \frac{c}{u_g} \right\}. \tag{3} \]

We refer to this as the oracle policy, because it has access to the unknown (and generally unknowable) principal scores. Unlike in the symmetric case, this policy includes the principal score \( \epsilon_{01} \), which is unidentifiable under Assumption 1. Since \( 0 \leq \epsilon_{01}(x) \leq 1 \) for all \( x \), the asymmetric oracle policy uses a varying threshold for assigning the treatment where the threshold depends on the principal score \( \epsilon_{01}(X) \).

The way in which the oracle policy depends on the principal score is characterized in part by the nature of the symmetry in the utility function. Consider the case where treatment is costless \((c = 0)\). Then \( u_l > u_g \), causing the undesirable outcome by assigning the treatment is considered worse than failing to prevent such an outcome by not providing the treatment. This raises the threshold for assigning the treatment because the expected effect must be larger in order to compensate for the downside risk of causing the undesirable outcome. As a result, this biases the oracle policy toward inaction. Conversely, when \( u_l < u_g \), it is better to cause the undesirable outcome than fail to prevent it by inaction. In this case, the threshold for the treatment assignment is lower, biasing the oracle policy toward action.

Figure 1(a) shows a one-dimensional example of these decision rules where the cost is zero, \( c = 0 \). The principal score \( \epsilon_{01}(x) \) is shown in red whereas the conditional expectations \( m(d, x) = \mathbb{E}[Y(d) \mid X = x] \) are shown in blue \((d = 0)\) and green \((d = 1)\). Figure 1(b) shows the functions that make up the decision rules in this example, centered so that the corresponding policies assign \( d = 1 \) if the function is positive. The symmetric case is shown in green, while the oracle in the asymmetric case \((u_g = 0.8\) and \( u_l = 1)\) is in orange. This plot also shows two other solutions discussed in Section 3.3: the minimal expected utility loss solution relative to always not assigning the treatment (purple), and the minimal regret solution relative to the oracle (black).

In this example, providing the treatment \( d = 1 \) leads to a higher probability of the desirable outcome except near zero. Therefore, with a symmetric utility function, the oracle policy would assign the treatment in most cases (green). However, the asymmetric case is different (orange). There is a region of the covariate space where the principal score \( \epsilon_{01}(x) \) is relatively high, leading to a sufficiently high probability that the treatment causes the undesirable outcome. Therefore, the asymmetric oracle rule has a higher threshold for the treatment assignment, only providing the treatment when the CATE \( \tau(x) \) is large enough and the principal score \( \epsilon_{10}(x) \) is small enough.

### 3.2. Partial Identification and Minimizing Worst-Case Expected Utility Loss

Recall that the unidentifiability of the principal score \( \epsilon_{01}(x) \) for any \( x \in \mathcal{X} \) makes it impossible to identify the expected utility loss in (1) in the general asymmetric case with \( u_g \neq u_l \). However, we can partially identify the principal score by deriving its sharp upper and lower bounds, \( L(x) \) and \( U(x) \). We then take a minimax approach, and find the policy \( \pi^* \) in the policy class \( \Pi \) that minimizes the maximal expected utility loss relative to an alternative policy \( \sigma \):

\[ \pi^* \in \arg\min_{\pi \in \Pi} R_{\sup}(\pi, \sigma), \quad \text{where} \]

\[ R_{\sup}(\pi, \sigma) = \max_{\epsilon_{01}(x) \in [L(x), U(x)]} \epsilon_{01}(\cdot). \tag{4} \]

Note that the maximum expected utility loss \( R_{\sup}(\pi, \sigma) \) is relative to a particular alternative policy \( \sigma \), and is maximal over all possible values for the principal score \( \epsilon_{01}(x) \). As we show below, the choice of this alternative policy will lead to different objectives and optimal solutions (see Cui 2021, for a recent general discussion).

Equation (4) is an example of a treatment choice problem under ambiguity (Manski 2005, 2011). Such minimax formulations of the problem have been widely considered in settings where value functions depend on the marginal distribution of the potential outcomes, but the CATE \( \tau(x) \) is not point identified (e.g., Manski 2007; Stoye 2012; Kallus 2018; Ben-Michael et al. 2021; Ishihara and Kitagawa 2021; Yata 2021; Zhang, Ben-Michael, and Imai 2023; D’Adamo 2023). A key distinction between Eqn (4) and these other problems, however, is that in our setting the value function depends on the principal score \( \epsilon_{10}(x) \), which is not point identified even in randomized control trials.

To derive the sharp lower and upper bounds of the principal score, we first define two classification functions:

\[ \delta_+(x) = \mathbb{I}\{m(0, x) + m(1, x) - 1 \geq 0\} \quad \text{and} \]

\[ \delta_-(x) = \mathbb{I}\{\tau(x) \geq 0\}. \tag{5} \]

Notice that the difference in the probability that both potential outcomes are one and zero is given by \( \epsilon_{11}(x) - \epsilon_{00}(x) = m(0, x) + m(1, x) - 1 \), which is the decision function for classifier \( \delta_+(x) \). In other words, we have \( \delta_+(x) = 1 \) if and only if \( \epsilon_{00}(x) \leq \epsilon_{11}(x) \). Thus, we can view \( \delta_+(x) \) as classifying whether there is a higher probability that both potential outcomes are one rather than zero. In contrast, noting that \( \tau(x) = \epsilon_{10}(x) - \epsilon_{01}(x) \), \( \delta_-(x) \) classifies whether there is a higher probability that the treatment is useful rather than harmful. This corresponds to the symmetric oracle rule with cost \( c = 0 \).

With these classifiers, we can use the Fréchet bounds to find the sharp lower and upper bounds for the principal score, \( \epsilon_{01}(x) \in [L(x), U(x)] \) for all \( x \) (e.g., Heckman, Smith, and Clements 1997; Jiang, Ding, and Geng 2016; Kallus 2018):

\[ L(x) = \max\{0, 1 - m(1, x) + m(0, x) - 1\} \]

\[ = \max\{0, -\tau(x)\} = -\tau(x)\{1 - \delta_+(x)\}, \]

\[ U(x) = \min\{m(0, x), 1 - m(1, x)\} \]

\[ = m(0, x) + \delta_+(x)\{1 - m(0, x) - m(1, x)\}. \]

These lower and upper bounds are sharp (Rüschendorf 1981) and are point-identifiable from observable data. With them we can create a point-identifiable objective.
E. BEN-MICHAEL, K. IMAI, AND Z. JIANG

3.3. Worst Case Expected Utility Loss Relative to Different Alternative Policies

We now inspect the worst-case expected utility loss $R_{\sup}(\pi, \sigma)$ in (4) for different choices of alternative policy $\sigma$. We consider three main alternatives. First, the “never-treat” policy $\pi^O$ that does not treat anyone, that is, $\pi^O(x) = 0$ for all $x$. Second, the “always-treat” policy $\pi^1$ that treats everyone, that is, $\pi^1(x) = 1$ for all $x$. In many cases, the alternative to algorithmic decision making via a data-driven policy is to take the same decision for everyone; thus, these two policies are of interest as they represent the standard of care in the absence of an individualized policy (see Appendix B for connections to maximin policies that maximize the minimum expected utility). We will denote the policies that minimize these worst-case losses as $\pi^O_\sup = \arg\min_\pi R_{\sup}(\pi, \pi^O)$ and $\pi^1_\sup = \arg\min_\pi R_{\sup}(\pi, \pi^1)$.

Finally, we consider the minimax regret policy that minimizes the worst-case expected utility relative to the value of the best-possible policy that has access to the principal scores $\epsilon_{01}(\cdot)$. Formally, the minimax regret policy is defined as $\pi^o_\sup = \arg\min_\pi \max_\sigma \max_{\epsilon_{01}(x)\in[L(x), U(x)]} R_{\epsilon_{01}}(\pi, \pi')$. The definition of the oracle policy $\pi^o$ above implies that this is equivalent to choosing $\pi^o$ as the alternative policy. That is,

$$\pi^o = \arg\min_\pi R_{\sup}(\pi, \pi^o),$$

where

$$R_{\sup}(\pi, \pi^o) = \max_{\epsilon_{01}(x)\in[L(x), U(x)]} \max_{\pi'} R_{\epsilon_{01}}(\pi, \pi').$$

is the regret of policy $\pi$.

Minimax regret policies are often studied in the policy learning literature because alternatives, such as maximin policies, tend to be too conservative (see e.g., Manski 2007, 2011; Stoye 2012; Yata 2021, among many others). Note that when defining the minimax regret policy across a constrained policy class, we compare to the best possible unconstrained policy, that is,

$$\arg\min_\pi \max_\sigma \max_{\epsilon_{01}(x)\in[L(x), U(x)]} R_{\epsilon_{01}}(\pi, \pi') = \arg\min_\pi R_{\sup}(\pi, \pi^o).$$

The resulting policy will be different in general from the policy that minimizes the regret relative to the best-in-class policy, and the unconstrained form of the regret will be larger.

The following theorem shows that the worst-case expected utility loss relative to each of these three policies takes a common form.

**Theorem 3.1 (Worst case expected utility loss).** Let $\pi : \mathcal{X} \rightarrow \{0, 1\}$ be a deterministic policy. For comparison policy $\sigma \in \{\pi^1, \pi^1, \pi^o\}$, the worst-case expected utility loss of $\pi$ relative to $\sigma$ is

$$R_{\sup}(\pi, \sigma) = C - \mathbb{E} \left[ \pi(X) \left\{ c_{\sigma}^0(X) m(1, X) + c_{\sigma}^0(X) m(0, X) + c_{\sigma}^0(X) \right\} \right]$$

$$= C - \mathbb{E} \left[ \pi(X) \left\{ c_{\sigma}^0(X) \Gamma_1(X, D, Y) + c_{\sigma}^0(X) \right\} \right],$$

where $C$ is a constant that does not depend on $\pi$, and $c_{\sigma}^0(\cdot), c_{\sigma}^0(\cdot), c_{\sigma}^o(\cdot) : \mathcal{X} \rightarrow \mathbb{R}$ are functions that depend on $\delta_0(\cdot), \delta_1(\cdot), \pi_{\sigma}^o$, or $\pi_{\sigma}^1$.

The maximum expected utility loss objective in Theorem 3.1 is a weighted average of the expected potential outcomes under treatment and no treatment plus a proxy for the cost. The choice of alternative policy $\sigma$ determines these weights $c_{\sigma}^0(\cdot), c_{\sigma}^0(\cdot)$ and cost $c_{\sigma}^o(\cdot)$, all of which potentially vary with the covariates $X$; we give explicit formulas for these functions in Appendix H. Note that the special case of a symmetric policy (Section 2.3) is also of this form, with $c_{\sigma}^0(X) = -c_{\sigma}^0(X) = u_0$ and $c_{\sigma}^o(X) = c$. Similarly, the two classifiers in (5) have this
form, with \( \delta_r \) corresponding to \( c^{\pi^*}_1(X) = -c^{\pi^*}_0(X) = 1 \) and \( c^{\pi^*}_0(X) = 0 \), and \( \delta_+ \) corresponding to \( c^{\pi^*_+}_1(X) = c^{\pi^*_+}_0(X) = 1 \) and \( c^{\pi^*_+}_0(X) = -1 \).

The second line of (6) shows how to write the worst-case expected utility loss \( R_{\sup}(\pi, \pi') \) in terms of observable data using the scoring functions \( \Gamma_w \) (either IPW or DR) discussed in Section 2.3. So, targeting the worst-case expected utility loss yields an objective function that is identifiable, unlike the true expected utility loss. As shown below, this allows us to construct decision rules based on observable data that control the true expected utility loss by minimizing the worst-case expected utility loss.

Constructing a utility function based on principal strata allows decision makers to define their goals directly in terms of individualized notions of useful and harmful treatments. Nevertheless, Theorem 3.1 shows that the minimax expected utility loss problem reduces to a decision problem that only involves the marginal distribution of the potential outcomes. The principal score \( e_{11}(x) \) will not be involved in the remaining estimation strategies and results, having been replaced with point-identifiable upper and lower bounds.

However, the weighting and cost functions induced by the utility function and choice of alternative policy correspond to a covariate-dependent asymmetry in terms of the marginal potential outcomes. Depending on the values of the nuisance classifiers, (6) places more or less weight on outcomes under treatment versus outcomes under control. Thus, (6) is related to the covariate-dependent loss minimization problem considered by Babii et al. (2021) that depends on marginal outcomes, even though it was derived from placing an asymmetric counterfactual utility on the principal strata. A key distinction is that because (6) involves the unknown nuisance classifiers, we must estimate the corresponding loss function. We analyze the consequences of this in Section 4.2.

Finally, note that here we restrict to deterministic policies to derive the form of the minimax expected utility loss in Theorem 3.1. As Cui (2021) discusses, unlike with the expected utility loss relative to the always treat or never treat policies, allowing for stochastic policies that randomize between actions can lead to lower loss, though this leads to a more complicated form. We leave further understanding the implications for stochastic policies to future work.

Next, we compute and inspect the policy that is the unconstrained minimizer of the maximum expected utility loss in the population, relative to each of the three alternative policies in turn. We will then turn to estimating constrained policies in finite samples in Section 4.

### 3.3.1. Expected Utility Loss Relative to a Constant Decision

We begin by considering the worst-case expected utility loss relative to the never-treat policy.

**Corollary 3.2 (Minimax expected utility loss relative to the never-treat policy).** If \( u_g \geq u_i \), the solution to (4), \( \pi^*_\text{symm} \equiv \arg \min_{\pi} R_{\sup}(\pi, \pi') \) is the symmetric policy,

\[
\pi^*_\text{symm}(x) = \begin{cases} 
1 & \left( \tau(x) \geq \frac{c}{u_g} \right), \\
\pi^*(x) & \text{otherwise.}
\end{cases}
\]

Otherwise, if \( u_g < u_i \), it is given by,

\[
\pi^*_\text{symm}(x) = \begin{cases} 
1 & \left( m(1, x) \geq \frac{u_i}{u_g} m(0, x) + \frac{c}{u_g} \right), \\
0 & \left( m(1, x) < \frac{u_i}{u_g} m(0, x) + \frac{u_i - u_g + c}{u_i} \right).
\end{cases}
\]

**Corollary 3.2** shows that the form of the minimax expected utility loss policy depends on the direction of the asymmetry. To build intuition, consider the case where the treatment is costless \( c = 0 \). If \( u_g > u_i \)—so we would rather cause an undesirable outcome than to fail to prevent it—then the minimax solution relative to the never-treat policy is the same as the optimal rule under a symmetric utility function: assign the treatment when the CATE is positive. In this case, the unit is more likely to be in the \((y_1, y_0) = (1, 0)\) stratum than the \((y_1, y_0) = (0, 1)\) stratum, and since \( u_g > u_i \), it will be better to treat the unit than to not. Conversely, when the CATE is negative it may still be better to treat the unit, but in the worst case it is not. To minimize the worst-case expected utility loss relative to never treating, the minimax loss policy does not treat.

However, the minimax solution is different when \( u_g < u_i \)—that is, when it is worse to cause an undesirable outcome than to fail to prevent it. In this case, the oracle rule depends on the value of the classifier \( \delta_+(x) = 1(e_{00}(x) \leq e_{11}(x)) \). If both potential outcomes are more likely to be zero than one, then the policy only treats if the probability that \( Y(1) \) equals one is higher than the probability that \( Y(0) \) equals one by a factor of \( \frac{u_i}{u_g} > 1 \). Comparing to the decision rule under the symmetric utility, we see that this raises the threshold for assigning the treatment.

In contrast, if both potential outcomes are more likely to be one than zero, the threshold is raised by adding a constant cost \( \frac{u_i - u_g}{u_i} > 0 \), but the multiplicative factor on the probability that \( Y(0) \) equals one is \( \frac{u_g}{u_i} < 1 \). Overall, this has the effect of creating a more cautious policy that provides the treatment less often.

**Figure 1(b)** shows the minimax decision rule relative to \( \pi^*(x) \) (purple) in the one-dimensional example where \( u_g < u_i \) and \( c = 0 \)—that is, it is worse to cause an undesirable outcome than to fail to prevent it. In this case, we see that the decision function is well below the symmetric rule shown in green (i.e., the CATE), leading to a large part of the covariate space being assigned no treatment even though the CATE is positive. In fact, this policy is overly cautious: it does not assign the treatment even in many cases where the oracle rule that knows the principal score would provide the treatment. This is because the alternative policy is to never treat anyone.

Appendix H shows the result for the minimax expected utility loss policy relative to the always-treat policy, which is more aggressive than the symmetric policy. It is the mirror image of the minimax loss policy relative to the never-treat policy, with the relation to \( u_g \) and \( u_i \) reversed.

### 3.3.2. The Minimax Regret Policy

We next consider the policy that minimizes the expected utility loss relative to the oracle \( \pi^* \) in (3), or, equivalently, that minimizes the regret. For simplicity, we assume zero cost, that is, \( c = 0 \); when \( c > 0 \), there will be further terms (see the proof of Theorem 3.1 in Appendix J).

**Corollary 3.3 (Minimax regret policy).** When \( c = 0 \), the minimax regret policy for \( u_g \geq u_i \) is given by,
Corollary 3.3 shows that we can write the worst-case regret and the minimax regret policy in terms of observable data, just as for the constant policies above. But, doing so requires four classifiers rather than one: (i) \( \delta_+ \), which classifies whether \( \epsilon_{\Pi_0}(x) \leq \epsilon_{11}(x) \); (ii) \( \delta_\tau \), which classifies whether the CATE is positive; (iii) the minimax loss solution relative to \( \pi^1 \) and (iv) the minimax loss solution relative to \( \pi^\text{symm} \). Recall from Corollary 3.2 that either \( \pi^\text{symm} \) (when \( u_g \geq u_t \)) or \( \pi^*_o \) (when \( u_g < u_t \)) is the symmetric policy \( \pi^\text{symm} \). Therefore, if the cost \( c = 0 \) as in Corollary 3.3, we only need three classifiers to construct the objective, since \( \pi^\text{symm} = \delta_\tau \) in this case.

Inspecting the minimax solution relative to the oracle policy when \( u_g \geq u_t \), we see that it assigns the treatment if the symmetric rule does, whereas it does not provide the treatment if the minimax solution relative to the always-treat policy does not. In between these two extremes, the decision rule lowers the threshold for the treatment assignment relative to the symmetric rule. The opposite is true when \( u_g < u_t \). If the symmetric rule does not assign treatment, the minimax solution relative to the oracle does not either, but it does provide the treatment whenever the minimax solution relative to the never-treat policy does. In between these two cases, the threshold for treatment assignment is higher than that under the symmetric rule.

Figure 1(b) shows the decision rule (black) in our running one-dimensional example where \( u_g < u_t \). The decision rule is equivalent to \( \pi^*_o \) (purple) when the CATE is negative, and is equal to the CATE decision rule \( \delta_\tau \) when \( \pi^*_o(x) = 1 \). When there is disagreement between the CATE rule \( \delta_\tau \) and \( \pi^*_o \), the minimax oracle rule interpolates between them, leading to a more aggressive policy that treats more individuals than \( \pi^*_o \). Comparing to the oracle rule (orange), we see that this interpolation causes the decision thresholds for the minimax oracle rule to be close to the best possible decision thresholds.

4. Learning a Policy from Data

Having established the behavior and form of the minimax loss policy \( \pi^* \) in (4) in the population for an unconstrained policy class, we now turn to the problem of learning a policy \( \hat{\pi} \) from observed data within a constrained policy class \( \Pi \).

4.1. Estimation Algorithms

To begin, note that in finite samples we know neither the true scoring functions \( \Gamma_w \) nor the true weighting and cost functions \( c^w(\cdot), c^o(\cdot), c^m(\cdot) \)—which depend on the nuisance classifiers—and so they must be estimated from data. As mentioned in Section 2.3, we can obtain estimates of the DR score \( \hat{\Gamma}^m_w \) by plugging in estimates of the nuisance components. Similarly, with estimates of the nuisance classifiers, we can directly obtain estimates of the weighting and cost functions \( \hat{c}^w(\cdot), \hat{c}^o(\cdot), \hat{c}^m(\cdot) \) by plugging in to the formulas in Theorem 3.1.

This leads to the following procedure. First, obtain estimates of the nuisance components \( \hat{m} \) and \( \hat{d} \) and construct the DR scores. Then, estimate the nuisance classifiers and follow Theorem 3.1 to construct estimates of the weighting and cost functions. To find a policy relative to either the always-treat (if \( u_g \geq u_t \)) or never-treat (if \( u_g < u_t \)) policies, we estimate a single nuisance classifier, \( \hat{\delta}_s \). Finding a policy relative to the oracle involves estimating three or four nuisance classifiers: \( \hat{\delta}_s \), \( \hat{\delta}_\tau \), and the minimax loss policies relative to never and always treating, \( \hat{\pi}^\text{symm} \) and \( \hat{\pi}^*_o \). With these in mind, we then find a data-driven policy \( \hat{\pi} \) that solves the following optimization problem (dropping the constant that does not depend on the policy \( \pi \)):

\[
\hat{\pi} \in \arg \min_{\pi \in \Pi} \hat{R}_{\sup}(\pi, \sigma) \quad \text{where}
\]

\[
\hat{R}_{\sup}(\pi, \sigma) = -\frac{1}{n} \sum_{i=1}^{n} \pi(X_i) \left\{ c^w(X_i) \hat{\Gamma}^m_1(X_i, D_i, Y_i) + c^o(X_i) \hat{c}^o_0(X_i) + c^m(X_i) \right\}.
\] (7)

There are two ways to estimate the nuisance classifiers. The first is an empirical risk minimization approach, where we solve (7) with the appropriate weighting and cost functions.1

Appendix C explicitly details this procedure. As shown in Section 4.2, the estimated nuisance classifiers must have low regrets relative to the true ones in order for our learned policy \( \hat{\pi} \) to have low worst-case expected utility loss; therefore, we must choose a flexible policy class. This is in contrast to estimating

---

1For the nuisance classifiers \( \delta_+ \) and \( \delta_\tau \), the weighting and cost functions are known, and so need not be estimated.
our policy of interest \( \hat{\tau} \), whose performance we measure relative to the best possible constrained policy. An alternative is to take a plug-in approach, using our estimates of the conditional expectation function \( \hat{m} \) to directly create estimates of the classifier; for example, \( \hat{\delta}_+(x) = 1\{\hat{m}(1,x) + \hat{m}(0,x) \geq 1\} \) and \( \hat{\delta}_-(x) = 1\{\hat{m}(1,x) - \hat{m}(0,x) \geq 0\} \).

4.2. Excess Worst-Case Expected Utility Loss

To understand the statistical properties of our learned minimax policy \( \hat{\tau} \), we will compare it to the policy \( \pi^* \) that minimizes the worst-case expected utility loss in the population among those in the policy class \( \Pi \) by solving (4). For a given alternative policy \( \sigma_\tau \), we will use the excess worst-case expected utility loss \( R_{\text{sup}}(\hat{\tau}, \sigma_\tau) - R_{\text{sup}}(\pi^*, \sigma_\tau) \) to measure the quality of the learned minimax loss policy \( \hat{\tau} \) since \( R_{\text{sup}}(\pi^*, \sigma_\tau) \) is the best possible expected utility loss in the worst case. We assume that the nuisance components and classifiers have been obtained from a separate sample, and so can be treated as fixed for our finite sample results. However, our results can be extended to solving (7) by cross-fitting nuisance components and classifiers to obtain the estimates of \( \hat{R}_{\text{sup}}^\pi(X_i, D_i, Y_i) \) and \( \hat{c}_{\text{sup}}^\pi(X_i) \) (see Athey and Wager 2021, and Appendix F.1).

To state our results, we define several new quantities. First, we measure the quality of the estimated nuisance classifiers, \( \hat{\delta}_+ \) and \( \hat{\delta}_- \), by their regrets,

\[
R_+(\hat{\delta}_+) \equiv \mathbb{E}\left[1\{\hat{\delta}_+(X) \neq \delta_+(X)\} | m(1,X) + m(0,X) - 1\right],
\]

\[
R_-(\hat{\delta}_-) \equiv \mathbb{E}\left[1\{\hat{\delta}_-(X) \neq \delta_-(X)\} | m(1,X) - m(0,X)\right],
\]

where \( \hat{\delta}_+ \) and \( \hat{\delta}_- \) are treated as fixed and the covariate \( X \) is random. Second, we measure the complexity of the policy class \( \Pi \) by its ability to overfit to noise via the population Rademacher complexity

\[
R_n(\Pi) \equiv \mathbb{E}_X \left[\sup_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \pi(X_i)\right],
\]

where \( \varepsilon_i \) are iid random variables with \( \Pr(\varepsilon_i = 1) = \Pr(\varepsilon_i = -1) = 1/2 \), and the expectation is taken over both \( \varepsilon_i \) and \( X_i \) (Wainwright 2019, sec 4).

We now present two finite sample bounds on the excess worst case expected utility loss, one for learning a minimax loss policy relative to the always or never treat policies (Theorem 4.1), and the other for learning a minimax loss policy relative to the oracle (Theorem 4.2).

Theorems 4.1 and 4.2 reveal three reasons why the data-specific policy \( \hat{\tau} \) can differ from the population policy \( \pi^* \). First, as captured via the Rademacher complexity term, even if the outcome model, propensity score model, and nuisance classifiers were all known, \( \hat{\tau} \) could simply overfit to noisy data. Fortunately, we can choose the complexity of \( \Pi \) and often prefer a relatively simple policy class for its interpretability and transparency. The results above will be relative to the best possible policy in the selected policy class. Thus, we could control this by limiting the complexity of our search space. For example, if the policy class \( \Pi \) has a finite VC dimension \( v \), the Rademacher complexity scales like \( R_n(\Pi) = O\left(\frac{\sqrt{v}}{n}\right) \) (Wainwright 2019, sec 5).

Second, there is error in our estimates of the outcome and propensity score models. However, following Athey and Wager (2021), using the DR scores protects against this error; only the product of the errors enter the bound, which decreases faster than \( 1/\sqrt{n} \) under typical assumptions. These two sources of error occur in symmetric policy learning problems. In the symmetric case when \( u_g = u_l \) and so \( \pi^\perp = \pi^0 = \pi^\circ \), Theorem 4.1 is a special case of the results in Athey and Wager (2021).

Finally, there is error in the nuisance classifiers, which is particular to our setting. For the minimax loss policy relative
to never or always treating, this error is measured by the regret for \( \hat{\delta}_+ \): if it correctly classifies cases that are not very close to the decision boundary (i.e., \( m(0, x) + m(1, x) - 1 \) is not near zero), this component will be small. Similarly for the minimax loss policy relative to the oracle, there are additional terms from the regret of \( \hat{\delta}_r \) and the excess worst case expected utility for the minimax loss policies relative to always and never treating.

If we estimate the nuisance classifiers via empirical risk minimization, results from Kitagawa and Tetlenov (2018) and Athey and Wager (2021) (and Theorem 4.1 for \( \hat{\pi}_1 \) and \( \hat{\pi}_2 \)) imply that the regret will primarily be controlled by the complexity of the policy classes we optimize over for the nuisance classifiers. Unlike for the minimax loss policy class \( \Pi \), unless the nuisance classifier class contains the true function, there will be irreducible approximation error in the misclassification term. Therefore, we might choose more complex classes, in which case the regret of the nuisance classifiers will primarily control the overall excess expected utility loss.

To analyze the plug-in approach, we use a different characterization of the complexity of the learning problem: the proportion of cases that are close to the decision boundary. Focusing on the nuisance classifier \( \delta_+ \), we follow Audibert and Tsybakov (2007) and characterize this via the following margin condition.

**Assumption 2 (Margin condition).** There exists an \( \alpha > 0 \) and a constant \( C \) such that for any \( t \geq 0 \), \( \sup_t \{ m(1, X) + m(0, X) - 1 \} \leq t \leq C t^\alpha \).

The margin parameter \( \alpha \) determines how many cases are allowed to be close to the boundary, with a larger value leading to a stronger assumption that fewer cases are close; for example if \( X \) has a bounded density, then \( \alpha \geq 1 \). Note that the margin condition also leads to faster convergence rates for empirical risk minimization approaches, provided the policy class contains the true classifier. See Audibert and Tsybakov (2007) for further discussion. Under this margin condition, we can further bound the regret of the plug-in nuisance classifier \( R_+ (\hat{\delta}_+) \), leading to the following corollary to Theorem 4.1.

**Corollary 4.3.** Under Assumption 2 and the conditions of Theorem 4.1, using the plug-in nuisance classifier \( \hat{\delta}_+(x) = \mathbb{1}\{ \hat{m}(1, x) + \hat{m}(0, x) \geq 1 \} \), the excess worst-case regret of \( \hat{\pi} \) relative to \( \pi^* \) satisfies

\[
R_{\sup}(\hat{\pi}, \sigma) - R_{\sup}(\pi^*, \sigma) \leq 2U \times \left( \frac{6 + \eta}{\eta} \times \left( 2R_{\eta}(\Pi) + \frac{t}{\sqrt{n}} \right) + \| \hat{\nu} - \gamma \|_2 \sum_{w=0}^{1} \| \hat{m}(w, \cdot) - m(w, \cdot) \|_2 \right) + (u_g - u_\nu) \times \left( 2^{1+\alpha} C \| \hat{m} - m \|_\infty^\alpha + \frac{t}{2\sqrt{n}} \right),
\]

with probability at least \( 1 - 2 \exp \left( \frac{-t^2}{2} \right) \), where \( \| \hat{m} - m \|_\infty \equiv \sup_{w, x} | \hat{m}(w, x) - m(w, x) | \), and \( U \) is a constant depending on the utility values.

With the plug-in nuisance classifier, \( R(\hat{\delta}_+) \) is controlled by the error in the outcome model; however, for \( \alpha > 0 \) this error will be raised to a higher power, leading to a faster rate. In Appendix H, we show an analogous result for the minimax policy relative to the oracle using plug-ins for all nuisance classifiers. See D’Adamo (2023) for an application of these techniques for policy learning in a different partial identification setting, and Kallus (2022) for an application to estimate bounds on \( \Pr(Y(1) < Y(0)) \).

Finally, although the minimax loss policies we consider are designed to minimize the worst-case expected utility loss, in some cases it may be possible that the true, unidentifiable expected utility loss may also be small. In Appendix E, we conduct a brief simulation study to inspect how the misclassification rates and the true expected utility loss behave in finite samples.

5. Application to Right Heart Catheterization

We now apply the proposed methodology to a particular decision problem: whether or not to use Right Heart Catheterization (RHC) in a clinical setting. RHC is a diagnostic tool where a catheter is inserted into the pulmonary artery. In a controversial observational study, Connors et al. (1996) found that RHC led to an increase in mortality on average. RHC, however, can lead to life-saving treatment for some patients. In this section, we will use the data from Connors et al. (1996) to learn policies for using RHC for certain patients, inspecting how asymmetry in the policy maker’s utility function can lead to different data driven decision-making processes.

5.1. Data and Setup

The data from Connors et al. (1996) include \( n = 5735 \) ICU patients, 2184 of whom had RHC applied. We will code the outcome \( Y(d) = 1 \) as survival by thirty days. In this case, the utility value \( u_\nu \) represents the utility gain in saving a patient’s life under RHC who would otherwise die without RHC, and \( u_\gamma \) represents the cost of RHC leading to the death of a patient who would otherwise survive. In this study, RHC use was not experimentally randomized and so we will be relying on Assumption 1, using the same set of socioeconomic and health characteristics as those used by Hirano and Imbens (2001) in their propensity score-based analysis.

Throughout, we will use the estimated doubly robust score \( \hat{d}_{\text{DR}} \). To do so, we need estimates of the conditional expectation function \( m(w, \cdot) \) and the propensity score \( d(\cdot) \). We use a 3-fold cross-fitting procedure to estimate these conditional expectations that we detail in Appendix F.1. With the combined DR scores, we estimate that RHC leads to an overall increase in 30-day mortality by 4.5 percentage points, with an estimated standard error of 1.2 percentage points. This result is consistent with the findings of other existing analyses.

To fit each of these models, we use the full set of socioeconomic and health characteristics in the data. We will consider, however, decision rules that only use a subset of the covariates \( V \subset A \). As we outline in Appendix F.2, Theorem 3.1 implies that the worst-case utility loss will involve nuisance classifiers based on the covariates \( V \), using \( m_V(w, v) \equiv \mathbb{E}[Y(w) \mid V = v] \),...
the expected potential outcome given only the covariates $V$. To adjust for confounding, however, we still require the DR-scores using the full set of covariates. To construct plug-in estimates of the nuisance classifiers, we use a variant of the DR-learner (Kennedy 2022), regressing the DR scores $\hat{\pi}_w$ on the covariates $V$ using gradient boosted decision stumps.

### 5.2. Threshold Decision Rules with Two Variables

We begin by considering decision rules that only use two clinical variables: the estimated probability of surviving two months and whether the patient has a Do-Not-Resuscitate (DNR) order. Throughout we will use threshold decision rules that assign RHC via a cutoff on the estimated probability of surviving two months, using separate thresholds for DNR and non-DNR patients.

First, we consider minimizing the worst-case expected utility loss relative to using RHC for all patients or never using RHC. We estimate the nuisance classifier classifier $\hat{\delta}_w$ via the plug-in approach (Appendix Figure G.1 shows the resulting classifier). We then create estimates of the weighting and cost functions $\hat{\delta}_1(\cdot), \hat{\delta}_1(\cdot), \hat{\delta}_1(\cdot), \hat{\delta}_1(\cdot)$, and solve (7) to estimate the minimax policy $\hat{\pi}_{\delta}$ relative to never using RHC (when $u_q < u_l$) and always using RHC (when $u_q \geq u_l$). We set $u_q = 1$ and vary $u_l \in [0.5, 2]$ so that the utility loss from a harmful treatment moves between half and twice as large as the utility loss from failing to give useful treatment.

Figure 2(a) shows the resulting decision rules for patients with (right) and without (left) a DNR order. We also estimate the CATE conditioned on the estimated probability of survival and DNR status with the DR-learner using kernel smoothing (Kennedy 2022). Note that the estimated CATE is positive for non-DNR patients with less than a 35% or greater than 80% probability of survival. Because we restrict to a single threshold, in the symmetric case, this leads to a decision rule that assigns a threshold of 35% for non-DNR patients, while never assigning RHC to DNR patients. As the utility gain in saving a life becomes greater than the cost of causing death, the estimated threshold increases, leading to a decision rule that uses RHC for non-DNR patients with a higher estimated probability of surviving. Eventually the asymmetry is so large in favor of prioritizing useful treatment that almost all non-DNR patients and most DNR patients would be given RHC, even though the CATE is negative. Conversely, as avoiding harm becomes more important, the decision threshold lowers, assigning RHC for fewer and fewer patients until no patients would receive it.

Next, we consider finding the minimax regret policy relative to the oracle $\hat{\pi}_0$, using plug-in estimates of the classifiers $\hat{\delta}_w$, $\hat{\delta}_I$, $\hat{\tau}$, $\hat{\pi}_I$, and $\hat{\pi}_I$. Figure 2(b) shows the decision functions. Similar to the minimax loss policy relative to always or never using RHC, as $u_l$ decreases the threshold for non-DNR patients increases and as $u_l$ increases the threshold decreases. Even in the extreme case with $u_l = 0.5$, however, DNR patients are not assigned RHC. When $u_l = 2$, DNR patients with a low probability of survival are still assigned RHC. Mirroring our discussion in Section 3.3, measuring regret relative to the best possible policy leads to a less aggressive decision rule than measuring expected utility loss relative to always using RHC.

### 5.3. Decision Trees with Several Clinical Variables

Next we move to decision rules with several clinical variables. Recall that Theorem 3.1 shows how to cast the minimax problem as a weighted policy learning problem; so we can find policies from data by solving (7) using off-the-shelf policy optimization solvers. Here, we focus on learning depth-3 decision trees, using the policytree package (Sverdrup et al. 2020).

Because finding the optimal decision tree scales super-linearly with the number of covariates, we first select variables from the set of clinical covariates by fitting a CATE model given the clinical covariates using the DR-learner with random forest regression. We then measure variable importance as the

---

3Note that we use plug-ins for $\hat{\pi}_I$ and $\hat{\pi}_I$ rather than the simple threshold decision rules above, to try to capture the best possible unconstrained classifiers rather than the best possible constrained ones.

4Of the 66 covariates, 56 are clinical variables while the remaining are socioeconomic variables important to controlling for confounding. See Hirano and Imbens (2001), Table 1, for a full list of covariates.
proportion of times a covariate is split on in the forest, weighted by the node depth, using the grf package, and select the top 10 most important covariates. See Appendix Figure G.2 for the variable importance measures for all clinical covariates. As before, we consider estimating minimax loss policies relative to always or never using RHC as well as the minimax regret policy relative to the oracle, as the utility asymmetry changes with \( u_g = 1 \) and \( u_l \in [0.5, 2] \). We again use plug-in estimates for the nuisance classifiers with the 10 selected covariates.

Figure 3 shows the percent of patients assigned RHC under the different decision rules. As we move away from the symmetric case toward prioritizing using RHC for patients that will benefit from it, the minimax loss policies relative to always using RHC and to the oracle assign more patients to RHC. In the other direction, as we increase \( u_l \) relative to \( u_g \) and so seek to prevent harming patients, the minimax loss policies relative to never using RHC and the oracle assign fewer patients RHC. However, the minimax policy relative to the oracle is less extreme, consistent with the two-covariate case in Figure 2 and our discussion in Section 3.3.2.

We can measure the impact of these policies in terms of directly interpretable patient outcomes rather than the expected utility loss, by inspecting (a) the worst-case proportion of patients that are given a harmful treatment; (b) the worst-case proportion of patients that are failed to be given a useful treatment; and (c) the overall expected mortality. Following the argument in Section 3, we can find upper bounds on the first two proportions (presented in Appendix H), and use the DR scores \( F_{\text{dr}}^w \) and the plug-in classifier \( \hat{\delta}_+ \) to get plug-in estimates of them.

Figure 4 shows the worst-case proportion of patients for whom \( \pi \) fails to give a useful treatment (y-axis) versus the worst-case proportion for whom \( \pi \) gives a harmful treatment (x-axis) as we vary \( u_l \in [0.5, 2] \). We observe the trade-off between these two types of errors, in the worst-case. If \( \pi \) treats almost no one, in the worst-case almost no patients receive a harmful treatment, but \( \pi \) could be failing to help up to 25% of patients. Conversely, if \( \pi \) treats almost everyone, then \( \pi \) necessarily treats almost everyone for whom it is helpful, but could be providing a harmful treatment for up to 30% of patients. Figure 4 shows the intermediate points between these two extreme scenarios, corresponding to a Pareto frontier between the errors in the worst case.

We can also see the impact on overall mortality along the frontier. Appendix Figure G.3 plots the estimated expected mortality under each policy against the worst-case proportion of patients for whom \( \pi \) fails to give a useful treatment or gives a harmful treatment as we vary \( u_l \in [0.5, 2] \). At one extreme,
the policy that uses RHC for a large number of patients has the highest expected mortality, because on average RHC is harmful, but the upper bound guarantees that it fails to give a useful treatment in almost no cases. The symmetric policy is at the other end of this tradeoff, with a lower expected mortality but potentially failing to use RHC when it is useful for over 18% of patients. On the other extreme, the policy that rarely uses RHC has a lower expected mortality than always using RHC, but a higher one than the symmetric policy. While the symmetric policy reduces mortality, up to 6% of patients will receive RHC even though it is harmful for them.

6. Discussion

In this article, we developed a general policy learning framework that allows for asymmetric counterfactual utilities, reflecting common ethical principles including the Hippocratic oath. The asymmetry of utility functions leads to the unidentifiability of expected utilities. We addressed this problem by employing a partial identification approach and then finding a policy that minimizes the maximum expected utility loss relative to a particular policy. We illustrated this framework by reanalyzing the study of Right Heart Catheterization, finding that the minimax optimal policy varies substantially with the asymmetry in the utility and choice of reference policy.

There are several avenues for future research. First, the optimization problem in (4) is a form of distributionally robust optimization (see Bertsimas, Brown, and Caramanis 2011, for a review). Distributionally robust procedures have been used for risk minimization and policy learning, often by assuming that the true, unknown underlying distribution is close to some known reference distribution. (see e.g., Duchi and Namkoong 2021; Kallus and Zhou 2021; Bertsimas, Imai, and Li 2023). In contrast, (4) considers all potential joint distributions between the potential outcomes—that is, all potential principal scores—that agree with the point-identifiable marginal distributions, without making additional assumptions. A direction for future work is to explicitly encode distributional assumptions. For example, we could treat the case where the potential outcomes are independent as a reference distribution, and assume that the true joint distribution is close to it.

Second, we have restricted our attention to binary outcomes and binary treatments. However, many decision problems involve multiple potential actions and categorical or continuous outcomes. This leads to more principal strata and potential asymmetries in the utility function. We briefly discuss extensions to, and difficulties with, the continuous outcome case in Appendix I, and leave a more thorough analysis to future work. Third, we may consider decision problems with utility functions that also depend on other post-treatment variables or mediators, leading to a different principal stratification structure. Finally, we can consider further constraints that encode notions of fairness, such as the concept of principal fairness that is based on the principal strata (Imai and Jiang 2023).

Supplementary Materials

The supplementary materials include additional theoretical and empirical results, a simulation study, and proofs of all theoretical results.

Acknowledgments

We also thank the IQSS’s Alexander and Diviya Magaro Peer Pre-Review Program for feedback.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

We acknowledge the partial support from Cisco Systems, Inc. (CG #2370386), National Science Foundation (SES–2051196), Sloan Foundation (Economics Program; 2020–13946), National Natural Science of China (Grant No. 12371285, 12292984), and Fundamental Research Funds for the Central Universities, Sun Yat-sen University (Grant No. 23hytd010).

ORCID

Eli Ben-Michael http://orcid.org/0000-0002-1175-4129
Kosuke Imai http://orcid.org/0000-0002-2748-1022
Zhichao Jiang http://orcid.org/0000-0002-8571-0217

References

Athey, S., and Wager, S. (2021), “Policy Learning With Observational Data,” Econometrica, 89, 133–161. [2,4,9,10]
Audibert, J. Y., and Tsybakov, A. B. (2007), “Fast Learning Rates for Plug-in Classifiers,” Annals of Statistics, 35, 608–633. [10]
Babii, A., Chen, X., Ghysels, E., and Kumar, R. (2021), “Binary Choice with Asymmetric Loss in a Data-Rich Environment: Theory and an Application to Racial Justice,” arxiv preprint arxiv:2010.08463. [2,3,7]
Ben-Michael, E., Greiner, D. J., Imai, K., and Jiang, Z. (2021), “Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment,” arxiv preprint arxiv:2109.11679. [2,5]
Bertsimas, D., Brown, D. B., and Caramanis, C. (2011), “Theory and Applications of Robust Optimization,” SIAM Review, 53, 464–501. [13]
Bertsimas, D., Imai, K., and Li, M. L. (2023), “Distributionally Robust Causal Inference with Observational Data,” arXiv preprint. [2] Bordley, R. F. (2009), “The Hippocratic Oath, Effect Size, and Utility Theory,” Medical Decision Making, 29, 377–379. [1]
Bradley, R., and Stefánsson, H. O. (2017), “Counterfactual Desirability,” British Journal for the Philosophy of Science, 68, 485–533. [3]
Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J., Vidaliet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. A. (1996), “The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients,” Journal of the American Medical Association, 276, 889–897. [2,10]
Cui, Y. (2021), “Individualized Decision Making under Partial Identification: Three Perspectives, Two Optimality Results, and One Paradox,” Harvard Data Science Review, just accepted. [2,5,7]
Cui, Y., and Tchetgen Tchetgen, E. (2021), “A Semiparametric Instrumental Variable Approach to Optimal Treatment Regimes Under Endogeneity,” Journal of the American Statistical Association, 116, 162–173. [2]
D’Adamo, R. (2023), “Orthogonal Policy Learning Under Ambiguity,” arxiv preprint arxiv:2111.10904. [2,5,9,10]
Ding, P., and Lu, J. (2017), “Principal Stratification Analysis Using Principal Scores,” Journal of the Royal Statistical Society, Series B, 79, 757–777. [3]
Duchi, J. C., and Namkoong, H. (2021), “Learning Models with Uniform Performance via Distributionally Robust Optimization,” Annals of Statistics, 49, 1378–1406. [13]
Foot, P. (1967), “The Problem of Abortion and the Doctrine of the Double Effect,” Oxford Review, 5, 5–15. [1]
Frangakis, C. E., and Rubin, D. B. (2002), “Principal Stratification in Causal Inference,” Biometrics, 58, 21–29. [2]

The supplementary materials include additional theoretical and empirical results, a simulation study, and proofs of all theoretical results.
Han, S. (2023), "Optimal Dynamic Treatment Regimes and Partial Welfare Ordering," *Journal of the American Statistical Association*. DOI:10.1080/01621459.2023.2258941 [2]

Heckman, J. J., Smith, J., and Clements, N. (1997), "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535. [5]

Hirano, K., and Imbens, G. W. (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology*, 2, 259–278. [10,11]

Imai, K., and Jiang, Z. (2023), "Principal Fairness for Human and Algorithmic Decision-Making," *Statistical Science*, 38, 317–328. [13]

Ishihara, T., and Kitagawa, T. (2021), "Evidence Aggregation for Treatment and Surrogate End Point Evaluation by Multiple Trials," *Journal of the Royal Statistical Society*, Series B, 84, 829–848. [5]

Jiang, Z., Yang, S., and Ding, P. (2022), "Multiply Robust Estimation of Causal Effects Under Principal Ignorability," *Journal of the Royal Statistical Society*, Series B, 84, 1423–1445. [3]

Jonsen, A. R. (1978), "Do No Harm," *Annals of Internal Medicine*, 88, 827–832. [1]

Kallus, N. (2018), "Balanced Policy Evaluation and Learning," in *Advances in Neural Information Processing Systems*, pp. 8895–8906. [5]

Kallus, N., and Zhou, A. (2021), "Minimax-Regret Treatment Choice with Missing Outcome Data," *Journal of Econometrics*, 139, 105–115. [5,6]

Pu, H., and Zhang, B. (2021), "Estimating Optimal Treatment Rules with an Instrumental Variable: A Partial Identification Learning Approach," *Journal of the Royal Statistical Society*, Series B, 83, 318–345. [2]

Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., and Luedtke, A. (2021), "Optimal Individualized Decision Rules Using Instrumental Variable Methods," *Journal of the American Statistical Association*, 116, 174–191. [2]

Rüschendorf, L. (1981), "Sharpness of Fréchet-bounds," *Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57, 293–302. [5]

Smith, C. M. (2005), "Origin and Uses of Primum Non Nocere — Above All, Do No Harm!" *Journal of Clinical Pharmacology*, 45, 371–377. [1]

Stefansson, H. O. (2015), "Fair Chance and Modal Consequentialism," *Economics and Philosophy*, 31, 371–395. [3]

Stoye, J. (2012), "Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments," *Journal of Econometrics*, 166, 138–156. [2,5,6]

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., and Wager, S. (2020), "policytree: Policy Learning via Doubly Robust Empirical Welfare Maximization Over Trees," *Journal of Open Source Software*, 5, 2232. [11]

Thomson, J. J. (1976), "Killing, Letting Die, and the Trolley Problem," *The Monist*, 59, 204–217. Philosophical Problems of Death. [1]

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press. [9]

Wiens, F., Jung, K., Heller, K., Krome, D., Saeed, M., Ossorio, P. N., Thadani-Isani, S., and Goldenberg, A. (2019), "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care," *Nature Medicine*, 25, 1337–1340. [1]

Yata, K. (2021), "Optimal Decision Rules Under Partial Identification," arxiv preprint arxiv:2101.04926. [2,5,6]

Zhang, Y., Ben-Michael, E., and Imai, K. (2023), "Safe Policy Learning under Regression Discontinuity Designs with Multiple Cutoffs," arxiv preprint arxiv:2208.13323. [2,5]

Zhao, Y., Zeng, D., Rush, A. I., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [2,4]