# Optimal Decision Rules when Payoffs are Partially Identified[*]

Timothy Christensen[†]   Hyungsik Roger Moon[‡]   Frank Schorfheide[§]

May 14, 2023

## Abstract

We derive optimal statistical decision rules for discrete choice problems when payoffs depend on a partially-identified parameter $\theta$ and the decision maker can use a point-identified parameter $P$ to deduce restrictions on $\theta$. Leading examples include optimal treatment choice under partial identification and optimal pricing with rich unobserved heterogeneity. Our optimal decision rules minimize the maximum risk or regret over the identified set of payoffs conditional on $P$ and use the data efficiently to learn about $P$. We discuss implementation of optimal decision rules via the bootstrap and Bayesian methods, in both parametric and semiparametric models. We provide detailed applications to treatment choice and optimal pricing. Using a limits of experiments framework, we show that our optimal decision rules can dominate seemingly natural alternatives. Our asymptotic approach is well suited for realistic empirical settings in which the derivation of finite-sample optimal rules is intractable.

**Keywords:** Model uncertainty, statistical decision theory, partial identification, treatment assignment, revealed preference

**JEL codes:** C10, C18, C21, C44, D81

[†]Department of Economics, New York University and Department of Economics, University College London. `t.christensen@ucl.ac.uk`

[‡]Department of Economics, University of Southern California, and Yonsei University. `moonr@usc.edu`

[§]Department of Economics, University of Pennsylvania, CEPR, NBER, and PIER. `schorf@ssc.upenn.edu`

# 1 Introduction

Many important policy decisions involve discrete choices. Examples include whether or not to treat an aggregate population or large sub-population, firm or worker decisions at the extensive margin, and pricing policies when, in practice, prices must be expressed in whole currency units. Suppose a decision maker must choose a policy from a discrete choice set. The decision maker has data that may be used to bound, but not point identify, the payoffs associated with some choices. For instance, the decision maker may be deciding whether or not to assign treatment based on observational data which is sufficient to bound, but not point identify, the average treatment effect. How should they proceed?

In this paper, we propose an approach for making optimal discrete statistical (i.e., data-driven) decisions when the payoffs associated with some choices are only partially identified. In the model described in Section 3, the decision maker observes data which may be used to learn about a vector of parameters $P$. The decision maker then chooses a policy from among a discrete set. The distribution of payoffs associated with the different policies is indexed by a structural parameter $\theta$. A key assumption underlying the analysis in this paper is that $\theta$ is possibly set-identified, but the parameters $P$ may be used to deduce restrictions on $\theta$. The decision maker therefore confronts both ambiguity (the payoff distribution is not point-identified) and statistical uncertainty ($P$ must be estimated from the data).

We propose a theory of optimal statistical decision making in this setting, building on a line of research going back to Manski (2000). We depart from this body of work by adopting a minimax approach to handle the ambiguity that arises from the partial identification of $\theta$ conditional on $P$ and average (or integrated) risk minimization to efficiently estimate $P$. This asymmetric treatment of parameters is in the spirit of the generalized Bayes-minimax principle of Hurwicz (1951). The resulting optimal decision rules are "robust" in the sense that they minimize maximum risk or regret over the identified set for $\theta$ conditional on $P$, and use the data to learn efficiently about features of $P$ germane to the choice problem.

In contrast with existing approaches, our optimal decision rules can be implemented very easily in realistic empirical settings via the bootstrap or (quasi-)Bayesian methods, for both parametric and semiparametric models. To describe the implementation, fix any choice and consider its maximum risk (or regret) over $\theta$ conditional on $P$. The maximum risk (or regret) is averaged across the bootstrap distribution for an efficient estimator $\hat{P}$ of $P$, a posterior distribution for $P$ (in parametric models), or a quasi-posterior based on a limited-information criterion for $P$ (in semiparametric models). Our optimal decision is then to simply to choose whatever choice has smallest average maximum risk (or regret). We show how to implement optimal decisions in the context of two applications, which we use as running examples. The first considers treatment assignment under partial identification of the average treatment effect (ATE). The second considers optimal pricing in an environment with rich unobserved

heterogeneity, where revealed preference arguments may be used to bound, but not point identify, demand responses under counterfactual prices. In both examples, the maximum risk (or regret) of different choices is available in closed form or can be computed by solving a standard optimization problem, e.g., a linear program. Though simple, we provide a formal (frequentist) asymptotic efficiency theory to justify our approach.

To elaborate on practicality a little, consider the two competing paradigms: Bayes and minimax. A Bayes decision requires specifying a prior on the parameter space for $(\theta, P)$, computing the posterior for $(\theta, P)$ having observed the data, then choosing the decision that minimizes posterior risk. A common criticism of Bayes decisions (see, e.g., Manski (2021)) is that they are only justified if one can elicit a credible subjective prior, which can be difficult in practice, and that the resulting decision will depend, to some extent, on the decision maker's choice of prior. This is true under both point- and partial identification. But this problem is more pronounced under partial identification because the decision maker's choice of prior for the partially identified parameter $\theta$ is not updated by the data (e.g., Moon and Schorfheide (2012)). Thus, the resulting decision will depend, even asymptotically, on the decision maker's choice of prior. By contrast, our (quasi-)Bayesian implementations only require specifying a prior for the point-identified parameter $P$ and our decisions are asymptotically independent of the choice of prior. Moreover, our bootstrap implementations sidestep choice of prior altogether.

Minimax decisions are those which maximize risk (or regret) uniformly over $(\theta, P)$. While minimaxity may be desirable, it is often not feasible. That is, to derive minimax decisions one often has to make strong assumptions on the data-generating process (e.g., Gaussian with known variance) and restrict the dependence of payoffs on model parameters to be of a very simple form. Indeed, we are not aware of any work deriving a minimax treatment rule under partial identification even in the simplest case of a binary outcome and binary treatment when randomization is not permitted and bounds on the ATE must be estimated from data. Algorithms such as that of Chamberlain (2000) may be used to compute approximate minimax decisions, but the performance gap between these and the true minimax decision can be difficult to quantify. Adopting a minimax approach with respect to $\theta$ and an average risk minimization approach with respect to $P$ lends a great deal of tractability, allowing us to derive optimal rules for a very broad class of empirically relevant settings where minimax rules are intractable. In particular, our approach does not require strong parametric assumptions on the data-generating process, accommodating semiparametric models, and allows complicated dependence of payoffs on model parameters.

Sections 4 and 5 present optimality results for decisions based on parametric and semiparametric models, respectively. For parametric models, our efficiency criterion extends the asymptotic average risk criterion introduced by Hirano and Porter (2009) for point-identified settings to partially identified settings. We further extend this notion to semiparametric

3

models via a least-favorable parametric submodel. Both of these extensions represent new contributions to the literature on asymptotic efficiency for statistical decision rules. Our main results show formally that the proposed (quasi-)Bayesian implementation is optimal under our asymptotic efficiency criterion. Moreover, any decision rule that is asymptotically equivalent to the Bayes implementation is optimal as well. Most importantly, this includes an implementation that replaces averaging under a (quasi-)posterior by averaging under a bootstrap approximation of the sampling distribution of an efficient estimator $\hat{P}$ of $P$, in cases in which these two distributions are asymptotically equivalent.

Importantly, we show asymptotic equivalence to the proposed Bayesian or bootstrap implementations is in fact *necessary* for optimality: any decision whose asymptotic behavior is different from these is sub-optimal. Hence, it follows from our necessity result that "plug-in rules", which plug an efficient estimator $\hat{P}$ into the oracle decision rule if $P$ were known, can perform sub-optimally. Manski (2021, 2022) refers to plug-in rules as "as-if" optimization, because the estimated parameters $\hat{P}$ are treated as if they are the true parameters. Manski (2021, 2022) shows in numerical experiments that plug-in rules may perform poorly under finite-sample minimax regret criteria. Our necessity result provides a complementary and quite general theoretical explanation for why plug-in rules may perform poorly, albeit under a different (but related) optimality criterion to that of Manski (2021, 2022).[1]

In Section 2 we present an empirical application based on Ishihara and Kitagawa (2021) in which a decision maker decides whether or not to adopt a job-training program based on several RCT estimates and their standard errors from other studies. Intersection bounds are derived by extrapolating multiple studies. Our optimal decision produces different treatment recommendations than the plug-in rule for some sub-populations. In Section 4 we show this difference is explained clearly through our optimality theory and necessity result.

Our application to optimal pricing is presented in Section 7. The problem we study is similar to the problem of robust monopoly pricing studied by Bergemann and Schlag (2011), in which a decision maker is choosing to set a price but faces ambiguity about the true distribution of demand. As in Bergemann and Schlag (2011), we assume the decision maker has a preference for robustness. However, we use revealed preference demand theory to derive bounds on demand, rather than assuming true demand is in a neighborhood of a pre-specified distribution. Moreover, in our setting the decision maker does not know bounds on demand ex ante but must instead estimate bounds from data. This application builds on prior work on revealed-preference demand theory, including Blundell, Browning, and Crawford (2007,

---

[1]For the intuition, consider a treatment assignment problem under partial identification of the ATE. Oracle rules depend on a robust welfare contrast $b(P)$ formed from bounds on the ATE. The bounds are non-smooth functions of $P$ in many empirically relevant settings reviewed in Section 8 and include intersection bounds, nonseparable panel data models, and bounds based on extrapolating IV estimands. This non-smoothness leads to a failure of the $\delta$-method that breaks the asymptotic equivalence between plug-in rules, which depend on $b(\hat{P})$, and optimal rules, which depend on the average of $b(\cdot)$ across a bootstrap or posterior distribution.

2008), Blundell, Kristensen, and Matzkin (2014, 2017), Hoderlein and Stoye (2015), Manski (2007b, 2014), and Kitamura and Stoye (2018, 2019). These works are primarily concerned with testing rationality or deriving bounds on demand. We instead focus on using bounds to solve an optimal pricing problem, and using demand data efficiently in that context.

This paper complements prior work on treatment assignment under partial identification, including Manski (2000, 2007a, 2020, 2021, 2022), Chamberlain (2011), Stoye (2012), Russell (2020), Ishihara and Kitagawa (2021), and Yata (2021). Except for Chamberlain (2011), these works seek decision rules that are optimal under finite-sample minimax regret criteria. We depart from these works in two respects. First, our local asymptotic framework enables us to approximate the finite-sample problem faced by the decision maker without having to explicitly solve the finite-sample problem, which may be intractable in many important applications. This allows us to relax restrictive parametric assumptions on the data-generating process (e.g. Gaussianity) that are typically imposed to derive finite-sample rules. It also allows us to accommodate a much broader class of data-generating processes, including semi-parametric models. Our optimality results apply to settings where the decision maker cannot confidently assert that the data are drawn from a given parametric model, or where bounds on the ATE are estimated using a vector of moments or summary statistics (e.g. regression or IV estimates from observational studies) whose finite-sample distribution is unknown.[2] Second, except for Chamberlain (2011), these works use optimality criteria that (in our notation) are minimax over $(\theta, P)$, whereas our criterion is minimax over the partially-identified parameter $\theta$ and averages over the point-identified parameter $P$, reflecting the asymmetric parameterization of the problem.

Our approach is closely related to the multiple priors framework of Gilboa and Schmeidler (1989) and robust Bayes (or $\Gamma$-minimax) decision making (Robbins, 1951; Berger, 1985) in partially identified models. See Giacomini, Kitagawa, and Read (2021) for a recent review. Our paper makes several novel contributions in connection with $\Gamma$-minimax decisions. First, for parametric models, we provide a large-sample frequentist justification for these types of decision rules, which complements the usual finite-sample robust Bayes justification. Second, we show this optimality carries over to (i) bootstrap-based decisions and (ii) quasi-Bayes decisions in semiparametric models. In these latter cases the decisions we derive are not $\Gamma$-minimax decisions, but rather have a purely frequentist justification.

The remainder of the paper is structured as follows. Section 2 presents the empirical application to treatment assignment based on extrapolating meta-analyses. Section 3 outlines our framework and derives the optimal decisions. Optimality theory is presented in Sections 4 and 5. Section 6 discusses connections with robust Bayes decisions. Further applications to

---

[2]Finite-sample results are sometimes developed for this case assuming Gaussianity of the statistics, by arguing that the studies are sufficiently large that the sampling distribution of the statistics is approximately normal. In these cases, it seems logically consistent to use a large-sample optimality criterion.

optimal pricing and treatment assignment are discussed in Sections 7 and 8. Finally, Section 9 concludes. Technical assumptions and proofs are relegated to the Appendix.

## 2    Motivating Example: Treatment Assignment

**The Treatment Assignment Problem.** Consider a social planner who is choosing whether or not to introduce a treatment in a target population. For concreteness, suppose that the the social planner does not know the target population's average treatment effect (ATE) but, as in Manski (2020) and Ishihara and Kitagawa (2021, hereafter IK), they observe a meta-analysis consisting of estimates $\hat{P}_k$ of the ATE $P_k$ in several related populations $k = 1, \ldots, K$, their standard errors $s_k$, characteristics $x_k$ of the related populations, and characteristics $x_0$ of the target population. The $P_k$ parameters may be used to construct bounds on the ATE of the target population, which we denote by $\theta := \mathbb{E}[Y]$. Here $Y$ is the treatment effect associated with an individual and defined as $Y := Y_1 - Y_0$, where $Y_1$ is the outcome under treatment and $Y_0$ the outcome in the absence of treatment.

As an empirical illustration, we revisit Example 2 of IK. The authors consider a subset of $K = 14$ studies form the database of Card, Kluve, and Weber (2017), each of which is an RCT looking at the impact of job training programs on employment. Studies are implemented in a number of different countries and in groups that differ by characteristics $x_k$ consisting of gender (males, females, or both), age (youths, adults, or both), OECD membership status, GDP growth (standardized) and unemployment (standardized). We consider the hypothetical question of whether to roll out a job-training program in two populations: German male youths in 2010 and German female youths in 2010 (with GDP growth of 3.48% and an unemployment rate of 9.45%).

IK bound the difference between the treatment effect in the target population and in the related population $k$ as

$$|P_k - \theta| \leq C\|x_k - x_0\|, \quad k = 1, \ldots, K, \tag{1}$$

where $C$ is a pre-specified constant. Hence we obtain the bounds

$$b_L(P) := \max_{1 \leq k \leq K}(P_k - C\|x_0 - x_k\|) \leq \theta \leq \min_{1 \leq k \leq K}(P_k + C\|x_0 - x_k\|) =: b_U(P), \tag{2}$$

where $P := (P_1, \ldots, P_K)$. We will refer to $P$ as the identifiable reduced-form parameter. It is assumed to be estimable from the data. The identified set for $\theta$ as a function of $P$ given by (2) is $\Theta_0(P) := [b_L(P), b_U(P)]$. More generally, the following analysis applies in any context where $\theta$ is partially identified and bounds $b_L(P)$ and $b_U(P)$ on $\theta$ are known up to a finite-dimensional, identifiable parameter $P$. How should the social planner use the ATE

bounds $b_L(P)$ and $b_U(P)$, the estimates $\hat{P}_k$, and their standard errors to inform whether or not to treat the target population?

**Oracle Decision.** Following Manski (2000, 2004), it is common to derive treatment rules under a utilitarian social welfare function that is linear in the target population's ATE:

$$W(d, \theta) = d\theta,$$

where $d \in \{0, 1\}$ indicates treatment. The welfare function is maximized by the treatment decision $\mathbb{I}[\theta \geq 0]$, where $\mathbb{I}[\cdot]$ is the indicator function that is equal to one if the condition is satisfied and zero otherwise. Following Manski (2000, 2004), we subsequently use the regret criterion

$$r(d, \theta) = \max_{d'} W(d', \theta) - W(d, \theta) = \big(\mathbb{I}[\theta \geq 0] - d\big)\theta. \tag{3}$$

Failure to treat $(d = 0)$ incurs zero regret when $\theta < 0$, otherwise the regret is $\theta$. Similarly, treating $(d = 1)$ incurs zero regret when $\theta \geq 0$, otherwise the regret is $-\theta$.

If the social planner knew $P$, then they could choose the decision that minimizes maximum regret over $\Theta_0(P)$ as follows. As non-treatment $(d = 0)$ incurs regret $\theta$ when $\theta \geq 0$ and zero otherwise, the maximum regret associated with choice $d = 0$ is

$$R(0, P) := \sup_{\theta \in \Theta_0(P)} r(0, \theta) = (b_U(P))_+ ,$$

where $(a)_+ := \max\{a, 0\}$. Similarly, as treatment $(d = 1)$ incurs regret $-\theta$ when $\theta < 0$ and zero otherwise, the maximum regret associated with choice $d = 1$ is

$$R(1, P) := \sup_{\theta \in \Theta_0(P)} r(1, \theta) = -(b_L(P))_- ,$$

where $(a)_- := \min\{a, 0\}$. This leads to the oracle decision that assigns treatment if the *robust welfare contrast*

$$b(P) := (b_U(P))_+ + (b_L(P))_-$$

is non-negative, and non-treatment otherwise:

$$\delta^o = \mathbb{I}\Big[b(P) \geq 0\Big]. \tag{4}$$

**Efficient Feasible Implementation.** The oracle rule is an infeasible first-best as it requires knowledge of the true $P$. In practice $P$ is unknown and any practical rule must also confront sampling uncertainty about $P$. One option is to simply plug the ATE estimates

$\hat{P} := (\hat{P}_1, \ldots, \hat{P}_K)$ into the oracle decision, which leads to the plug-in rule

$$\delta_n^{plug} = \mathbb{I}\Big[b(\hat{P}) \geq 0\Big]. \tag{5}$$

However, this rule does not account for the precision with which $P$ is estimated. Manski (2022) explores how different estimates of model parameters affects the performance of plug-in rules.

We develop an efficiency theory to guide decision-making in situations such as these where a payoff-relevant parameter $\theta$ is partially identified and its identified set $\Theta_0(P)$ is indexed by a parameter $P$ which can be estimated from data. In our empirical illustration the efficient treatment assignment procedure can be implemented in two steps with the following quasi-Bayesian procedure.

The first step consists of the construction of a limited-information posterior distribution for $P_k$. We assume that sample sizes are sufficiently large, such that the distribution of the ATE estimator $\hat{P}_k$ is approximately $N(P_k, s_k^2)$. The estimates $\hat{P}_k$ are also independent as they come from independent RCTs. Following Doksum and Lo (1990) and Kim (2002), we combine the limited information $N(P, \Sigma)$ quasi-likelihood for $\hat{P}$ with a flat prior for $P$ to obtain a $N(\hat{P}, \Sigma)$ quasi posterior for $P$, where $\Sigma$ denotes a diagonal matrix with $(s_k^2)_{k=1}^K$ down its leading diagonal.

In the second step, we average the robust welfare contrast $b(P)$ across the quasi-posterior for $P$ to compute its quasi-posterior mean $\bar{b}_n$. This is implemented by simply drawing independent $P|\hat{P} \sim N(\hat{P}, \Sigma)$, computing $b(P)$, then averaging across a large number of draws. The efficient decision is then to treat if the quasi-posterior mean $\bar{b}_n$ is non-negative, and not treat otherwise:

$$\delta_n^* = \mathbb{I}\Big[\bar{b}_n \geq 0\Big]. \tag{6}$$

Later we show under the optimality criterion in Section 4, this decision is more efficient than the plug-in rule as it takes into account the sampling uncertainty in the ATE estimates $\hat{P}$.

Using the data set, we compute the optimal decision as described above. For this, we increase the Lipschitz constant in (1) from $C = 0.025$ in IK to $C = 0.25$ so that the bounds on $\theta$ are non-empty. Figures 1a and 1b plot the quasi-posterior distribution of the robust welfare contrast $b(P)$ for males and females, respectively, for $P|\hat{P} \sim N(\hat{P}, \Sigma)$. Both figures display the mean $\bar{b}_n$ whose sign determines the optimal treatment decision $\delta_n^*$ in (6). We also display the plug-in value $b(\hat{P})$ whose sign determines the plug-in rule $\delta_n^{plug}$ in (5).

In Figure 1a (male youths) we see that the distribution has a mode located at zero and a pronounced right-skew. Recall that the lower bound is $\max_{1 \leq k \leq K}(P_k - C\|x_0 - x_k\|)$. When evaluated at the $\hat{P}$, the largest value of $\hat{P}_k - C\|x_0 - x_k\|$ is $-0.3190$ (corresponding to a US study) and the second-largest is $-0.3298$ (corresponding to a Brazilian study). As the maximum is not well separated relative to sampling uncertainty (the average $s_k$ across the $K$
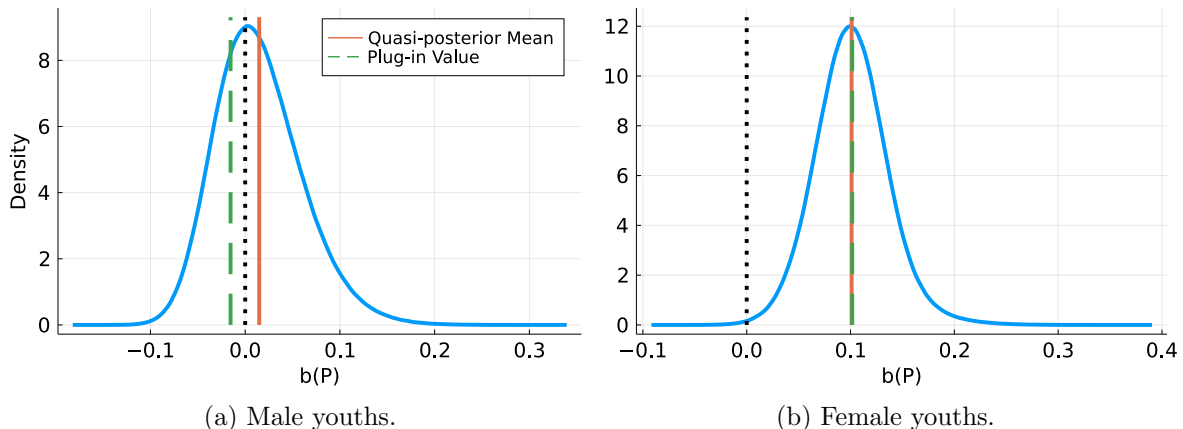
8

(a) Male youths.                    (b) Female youths.

Figure 1: Distribution of the robust welfare contrast $b(P)$ for $P \sim N(\hat{P}, \Sigma)$, its quasi-posterior mean $\bar{b}_n$, and the plug-in value $b(\hat{P})$.

studies is 0.034), the distribution of the lower bound is right-skewed because it behaves like a maximum of two Gaussians. The upper bound is less skewed because the minimum (which corresponds to a Colombian study) is better separated. As a consequence of this asymmetry, the mean $\bar{b}_n$ is positive and the optimal decision is to treat. By contrast, the plug-in value $b(\hat{P})$ is negative so the plug-in decision is to not treat. In Section 4 we show this difference between the two rules is as predicted by our optimality theory.

The situation is different for female youths (Figure 1b). Here the minima and maxima are relatively better separated, so the distribution of $b(P)$ for $P|\hat{P} \sim N(\hat{P}, \Sigma)$ is close to Gaussian. In consequence, the mean $\bar{b}_n$ and the plug-in value $b(\hat{P})$ are almost identical, and the optimal and plug-in decisions are both to treat.

# 3    Optimal Decisions and Their Implementation

We now move beyond this illustration to present our general approach. We begin in Section 3.1 by stating the decision problem. It features a partially-identified payoff-relevant parameter $\theta$ and a point-identified parameter $P$ that determines the identified set for $\theta$. Many empirically relevant examples are covered by our framework: see Sections 7 and 8 for further applications to treatment assignment and optimal pricing and our earlier working paper Christensen et al. (2020) for application to forecasting. We develop our concept of optimal decision rules in Section 3.2. Bayesian and bootstrap implementations are presented in Sections 3.3 and 3.4. Later in Section 6 we interpret our approach in the context of a two-player zero-sum game and relate it to the literature on robust Bayes decision making.

## 3.1 Decision Problem

A decision maker (DM) observes $X^n \sim F_{n,P}$ taking values in a sample space $\mathcal{X}^n$. For instance, $X^n = (X_1, \ldots, X_n)$ may be a random sample of size $n$ with $F_{n,P}$ denoting the $n$-fold product distribution over independent random vectors $X_i \sim F_P$. Alternatively, $X^n$ may be a vector of estimates or summary statistics (e.g., OLS or IV estimates) with $F_{n,P}$ denoting the sampling distribution of $X^n$. The distribution $F_{n,P}$ from which $X^n$ is drawn is indexed by an unknown *reduced-form parameter* $P \in \mathcal{P} \subseteq \mathbb{R}^K$. To simplify exposition, in this section we will be exclusively concerned with parametric models for $X^n$. Our approach extends naturally to semiparametric models, as discussed in Section 5.

After observing $X^n$, the DM chooses an element $d$ from a finite set $\mathcal{D} = \{0, 1, \ldots, D\}$. The DM's *statistical decision rule* $\delta_n : \mathcal{X}^n \to \mathcal{D}$ maps realizations of the data into choices.[3] The DM incurs risk (negative expected utility)

$$r(d, \theta, P) := \mathbb{E}_\theta[-u(d, Y, \theta, P)]$$

from choice $d \in \mathcal{D}$, where the expectation is taken with respect to a random vector

$$Y \sim G_\theta$$

drawn independently of $X^n$. The distribution $G_\theta$ and expectation $\mathbb{E}_\theta[\cdot]$ can be conditional on covariates; we suppress this to simplify notation. The distribution $G_\theta$ is indexed by a *structural parameter* $\theta \in \Theta$, as described in detail below. We may interpret the expectation as a forward-looking expectation over a random outcome, or a cross-sectional average in a social planner's problem. Our notation encompasses welfare regret by setting

$$-u(d, y, \theta, P) = \max_{d' \in \mathcal{D}} \mathbb{E}_\theta[W(d', Y, \theta, P)] - W(d, y, \theta, P)$$

for a welfare function $W(d, y, \theta, P)$. We therefore do not distinguish between "risk" and "regret" in what follows, except when referring to functional forms for $u(\cdot)$.

A key assumption underlying the analysis in this paper is that $\theta$ is possibly set-identified. That is, the most that the DM can infer about $\theta$ is that $\theta \in \Theta_0(P) \subseteq \Theta$, where $\Theta_0(\cdot)$ is a known set-valued mapping from the reduced-form parameter space to the structural parameter space. So while $X^n$ can be used to learn about $P$, it contains no information about the identity of the true $\theta$ within $\Theta_0(P)$.

**Example: Treatment Assignment.** Consider the setup from Section 2. Here $d \in \{0, 1\}$ is a binary treatment indicator. The structural parameter $\theta$ is the ATE. The risk (i.e., regret)

---

[3]We use the $n$ subscript to denote that a decision rule may also depend on sample size $n$.

$r(d, \theta, P) = r(d, \theta)$ is given in display (3). Suppose the DM observes data or summary statistics $X^n \sim F_{n,P}$ that may be used to construct bounds $b_L(P)$ and $b_U(P)$ on $\theta$ as a function of $P$. In the application from Section 2, the statistics were the vector of estimated ATEs $X^n = (\hat{P}_1, \ldots, \hat{P}_k)'$, $F_{n,P}$ is their sampling distribution, the reduced-form parameter is the vector of ATEs $P = (P_1, \ldots, P_k)'$, and $b_L(P)$ and $b_U(P)$ are given in (2). We review other methods for constructing bounds in different empirical settings in Section 8. The identified set is $\Theta_0(P) = [b_L(P), b_U(P)]$. $\qquad\square$

## 3.2 Optimality Criterion

We now define our optimality criterion. Our setup features two types of parameters: a point-identified parameter $P$ and a set-identified parameter $\theta \in \Theta_0(P)$ that affects risk. We adopt a minimax approach to handle the ambiguity that arises from the partial identification of $\theta$ conditional on $P$ and average (or integrated) risk minimization to efficiently estimate $P$. Our asymmetric treatment of parameters is in the spirit of Hurwicz (1951); we discuss the connection in more detail in Section 3.3. We first consider the case in which $P$ is known and then extend the analysis to the case of unknown $P$, which will lead us to the definition of our optimality concept.

**Known $P$.** First suppose that $P$ is known to the DM. To handle the ambiguity about $\theta \in \Theta_0(P)$, the DM evaluates choices $d \in \mathcal{D}$ by their maximum risk

$$R(d, P) := \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[-u(d, Y, \theta, P)]. \tag{7}$$

The minimax decision for known $P$ is

$$\delta^o(P) = \arg\min_{d \in \mathcal{D}} R(d, P)$$

if the argmin is unique, otherwise $\delta^o(P)$ is chosen randomly from $\arg\min_{d \in \mathcal{D}} R(d, P)$. The choice can be interpreted as the equilibrium in a two-player zero-sum game in which an adversarial "nature" chooses $\theta$ in response to the DM's choice of $d$ to minimize $\mathbb{E}[u(d, Y, \theta, P)]$. We refer to $\delta^o(P)$ as the *oracle* decision, as it represents the DM's optimal choice if $P$ was known. The oracle decision is an infeasible first-best, as in any application $P$ will need to be estimated from data. Nevertheless, the oracle decision is a useful benchmark, as its maximum risk, namely $\min_{d \in \mathcal{D}} R(d, P)$, is a lower bound on the minimax risk when $P$ is unknown.

**Unknown $P$.** When the DM does not know $P$, the DM's decision will be data-dependent. For a decision rule $\delta_n : \mathcal{X}^n \to \mathcal{D}$, the DM incurs *frequentist (or average) maximum risk*

$$\mathbb{E}_P[R(\delta_n(X^n), P)], \tag{8}$$

where $\mathbb{E}_P[\cdot]$ denotes expectation with respect to $X^n \sim F_{n,P}$. Our goal is to construct decision rules that use the data efficiently, in the sense that their average maximum risk is as close to that of the oracle over a range of data-generating processes.

Following Hirano and Porter (2009), we use a local asymptotic framework based on per-turbations to the data-generating process of the same order as sampling variation. The idea of this asymptotic framework is that, although asymptotic, it should well approximate the DM's finite-sample problem in which $P$ is unknown.

Formally, fix any limiting $P_0 \in \mathcal{P}$ and let $P_{n,h} = P_0 + h/\sqrt{n}$ denote its local perturbation by $h \in \mathbb{R}^K$ for a sample of size $n$. For the asymptotic analysis it is convenient to express the frequentist maximum risk in excess of the oracle risk (the infeasible first-best) when $P_{n,h}$ is the true reduced-form parameter. We scale the excess risk by $\sqrt{n}$ to ensure the large-sample limit is not degenerate.

$$\mathcal{R}(\{\delta_n\}; P_0, h) := \lim_{n \to \infty} \sqrt{n} \left( \mathbb{E}_{P_{n,h}} \left[ R(\delta_n(X^n), P_{n,h}) \right] - \min_{d \in \mathcal{D}} R(d, P_{n,h}) \right), \tag{9}$$

which is non-negative, and where the limit exists for large class $\mathbb{D}$ of decision rules whose asymptotic behavior is well-defined—see Section 4. This *asymptotic frequentist maximum risk* depends on the perturbation parameter $h$. Integrating over $h$ yields the *asymptotic average maximum risk* criterion:

$$\mathcal{R}(\{\delta_n\}; P_0) := \int \lim_{n \to \infty} \sqrt{n} \left( \mathbb{E}_{P_{n,h}} \left[ R(\delta_n(X^n), P_{n,h}) \right] - \min_{d \in \mathcal{D}} R(d, P_{n,h}) \right) \mathrm{d}h. \tag{10}$$

We define optimality as follows:

**Definition 1** *A sequence of decision rules $\{\delta_n\}$ is optimal if*

$$\mathcal{R}(\{\delta_n\}; P_0) = \inf_{\{\delta'_n\} \in \mathbb{D}} \mathcal{R}(\{\delta'_n\}; P_0) \quad \text{for all } P_0 \in \mathcal{P}.$$

**Remark 3.1** *An alternative is to take the maximum over $h$ rather than integrating over $h$ as in (10), which has the downside that optimal decisions are computationally intractable in most empirically relevant settings. By contrast, to implement optimal decisions under our criterion (10), the DM only needs to be able to compute maximum risk $R(d, P)$ as a function of $P$ for each choice $d$. As we are considering decision problems in which $R(d, P)$ may be non-smooth in $P$, different optimality concepts generally lead to different decisions.*

In the remainder of this section we present a Bayesian implementation and a bootstrap implementation of optimal decisions.

### 3.3 Bayesian Interpretation and Implementation

Let $\pi$ denote a strictly positive, smooth density on $\mathcal{P}$. Integrating the frequentist maximum risk criterion (8) across $P$ using $\pi$ yields the *integrated maximum risk* criterion

$$\int \mathbb{E}_P[R(\delta_n(X^n), P)] \, \mathrm{d}\pi(P). \tag{11}$$

After observing $X^n$, the DM can form a posterior $\pi_n(P|X^n)$ for $P$. Standard arguments (e.g. Wald, 1950, Chapter 5.1) imply that the integrated maximum risk can be minimized by minimizing the posterior maximum risk

$$\bar{R}_n(d|X^n) := \int R(d, P) \, \mathrm{d}\pi_n(P|X^n) \tag{12}$$

with respect to $d \in \mathcal{D}$ for almost every realization of the data. This leads to the Bayes rule

$$\delta_n^*(X^n; \pi) = \arg\min_{d \in \mathcal{D}} \bar{R}_n(d|X^n) \tag{13}$$

if the argmin is unique, otherwise $\delta_n^*(X^n; \pi)$ is chosen randomly from $\arg\min_{d \in \mathcal{D}} \bar{R}_n(d|X^n)$. We include $\pi$ as an argument to indicate that the decision depends on $\pi$ in finite samples.

We formally prove in Section 4 that Bayes decisions $\delta_n^*(X^n; \pi)$ are optimal under criterion (10) for any prior $\pi$ whose density is strictly positive and smooth at $P_0$. To see the connection between (11) and (10), use a change-of-variables to express the prior for $h$ as $\propto \pi(P_0 + h/\sqrt{n})$. This prior becomes uniform as $n \to \infty$, which is the weight function underlying (10).

**Example: Treatment Assignment.** Recall from Section 2 that the maximum risk of $d \in \{0, 1\}$ are positive and negative parts of the bounds on the ATE:

$$R(0, P) = (b_U(P))_+ , \quad R(1, P) = -(b_L(P))_- .$$

Averaging $R(1, P)$ and $R(0, P)$ across the posterior for $P$ yields

$$\bar{R}_n(0|X^n) = \int (b_U(P))_+ \, \mathrm{d}\pi_n(P|X^n), \quad \bar{R}_n(1|X^n) = \int -(b_L(P))_- \, \mathrm{d}\pi_n(P|X^n).$$

We have $\bar{R}_n(1|X^n) \leq \bar{R}_n(0|X^n)$ if and only if the posterior mean $\int b(P) \, \mathrm{d}\pi_n(P|X^n) \geq 0$ for $b(P) = (b_U(P))_+ + (b_L(P))_-$. Hence, an optimal decision is to treat if and only if the posterior mean of the robust welfare contrast $b(P)$ is non-negative:[4]

$$\delta_n^*(X^n; \pi) = \mathbb{I}\left[\int b(P) \, \mathrm{d}\pi_n(P|X^n) \geq 0\right]. \tag{14}$$

---

[4]This rule deterministically chooses treatment when the posterior mean of $b(P)$ is zero. Any (possibly randomized) tie-breaking rule leads to an optimal decision with the same asymptotic average maximum risk.

We discuss different bounds and how to form posteriors in a number of different empirical settings in Section 8. $\hspace{4cm}$ □

## 3.4 Bootstrap Implementation

We now present a bootstrap implementation of optimal decision rules. Let $\hat{P}$ denote an efficient estimator of $P$. Define the bootstrap average maximum risk

$$R_n^*(d) = \mathbb{E}_n^* \left[ R(d, \hat{P}^*) \right],$$

where $\mathbb{E}_n^*$ denotes expectation with respect to the bootstrap version $\hat{P}^*$ of $\hat{P}$ conditional on $X^n$. The bootstrap rule is

$$\delta_n^{**}(X^n) = \arg \min_{d \in \mathcal{D}} R_n^*(d)$$

if the argmin is unique (with a random selection from the argmin otherwise). This rule is asymptotically equivalent to $\delta_n^*$ under mild conditions ensuring asymptotic equivalence of the bootstrap distribution for $\hat{P}^*$ and posterior distribution for $P$. As such, $\delta_n^{**}$ will inherit the asymptotic efficiency properties of the optimal decision rule $\delta_n^*$.

**Example: Treatment Assignment.** Averaging $R(1, P)$ and $R(0, P)$ across the bootstrap distribution of $\hat{P}^*$ yields

$$R_n^*(1|X^n) = \mathbb{E}_n^* \left[ -(b_L(\hat{P}^*))_- \right], \quad R_n^*(0|X^n) = \mathbb{E}_n^* \left[ (b_U(\hat{P}^*))_+ \right].$$

Therefore, the bootstrap rule is to treat if and only if the boostrap average of the robust welfare contrast is non-negative:

$$\delta_n^{**}(X^n) = \mathbb{I} \left[ \mathbb{E}_n^* \left[ b(\hat{P}^*) \right] \geq 0 \right].$$

## 4 Optimality Results

We now present the main optimality results for parametric models. Section 4.1 contains two main results. First, Theorem 1 shows that the Bayes decision rules $\delta_n^*(\,\cdot\,; \pi)$, defined in (13), are optimal. Moreover, any decision that behaves asymptotically like $\delta_n^*(\,\cdot\,; \pi)$ is also optimal. Optimality of the bootstrap-based decision $\delta_n^{**}$ then follows. Second, Theorem 2 shows that any decision whose asymptotic behavior is different from $\delta_n^*(\,\cdot\,; \pi)$ is sub-optimal. Specifically, we show in Section 4.2 that plug-in decisions $\delta_n^{plug}$ are sub-optimal when the maximum risk $R(d, P)$ is a non-smooth function of $P$. As we discuss further in Section 8, this finding has important implications for treatment assignment under partial identification.

14

## 4.1 Optimality of Bayes and Bootstrap Decisions

Appendix A presents Assumptions 1, 2, and 3, which are the main regularity conditions. Assumption 1 imposes continuity and directional differentiability conditions on $R(d, P)$. Assumption 2 states that the model for $X^n$ is locally asymptotically normal at each $P_0 \in \mathcal{P}$. Assumption 3 contains consistency and asymptotic normality assumptions for functionals of the posterior distribution. These conditions implicitly restrict the priors $\pi$ to be positive and smooth on $\mathcal{P}$. Let

$$\mathbb{D} = \left\{ \{\delta_n\} : \delta_n(X^n) \overset{P_{n,h}}{\leadsto} Q_{P_0,h} \text{ for all } h \in \mathbb{R}^K, P_0 \in \mathcal{P} \right\},$$

where $\overset{P_{n,h}}{\leadsto}$ denotes convergence in distribution along a sequence $\{F_{n,P_{n,h}}\}$ with $X^n \sim F_{n,P_{n,h}}$ for each $n$, and $Q_{P_0,h}$ is the (possibly degenerate) limiting probability measure on $\mathcal{D}$. The set $\mathbb{D}$ represents the set of all "well behaved" sequences of decision rules that converge in distribution under each sequence $\{P_{n,h}\}$ for each fixed perturbation direction $h$. In general, $Q_{P_0,h}$ will depend on the sequence $\{\delta_n\}$. We say that two sequences of decisions $\{\delta_n\}$ and $\{\delta_n'\}$ are *asymptotically equivalent* if $\delta_n(X^n)$ and $\delta_n'(X^n)$ have the same asymptotic distribution along $\{P_{n,h}\}$ for all $P_0 \in \mathcal{P}$ and $h \in \mathbb{R}^K$.

**Theorem 1** *Suppose that Assumptions 1 and 2 are satisfied. Let $\Pi$ be the class of priors such that Assumption 3 holds.*

*(i) For any $\pi \in \Pi$, the Bayes decision $\{\delta_n^*(\,\cdot\,;\pi)\}$ is optimal.*

*(ii) For any $\pi, \pi' \in \Pi$, the Bayes decisions $\{\delta_n^*(\,\cdot\,;\pi)\}$ and $\{\delta_n^*(\,\cdot\,;\pi')\}$ are asymptotically equivalent.*

*(iii) If $\{\delta_n\}$ is asymptotically equivalent to $\{\delta_n^*(\,\cdot\,;\pi)\}$ for a $\pi \in \Pi$, then $\{\delta_n\}$ is optimal.*

According to parts (i) and (ii) of Theorem 1, all Bayes decisions based on priors that satisfy some mild regularity conditions are asymptotically equivalent and optimal. Part (iii) implies that a decision that is not a Bayes decisions but is asymptotically equivalent to a Bayes decision under a prior $\pi \in \Pi$ is optimal. Optimality of the bootstrap implementation discussed in Section 3.4 follows from this result under suitable regularity conditions.

Asymptotic equivalence to a Bayes decision is in fact a *necessary* condition for optimality under a side condition ruling out the absence of ties. Say $\{\delta_n\}$ and $\{\delta_n'\}$ fail to be asymptotically equivalent at $P_0$ if $\delta_n(X^n)$ and $\delta_n'(X^n)$ have different asymptotic distributions along $\{P_{n,h_0}\}$ for some $h_0 \in \mathbb{R}^K$. Let $\dot{\rho}_{d,P_0}[\cdot]$ denote the directional derivative of $R(d, P)$ with respect to $P$ at $P_0$ and let $\mathbb{E}^*$ denote expectation with respect to $Z^* \sim N(0, I_0^{-1})$ where $I_0 = I_0(P_0)$ is the (asymptotic) information matrix at $P_0$—see Appendix A for definitions.

**Theorem 2** *Suppose that Assumptions 1 and 2 are satisfied. Let $\Pi$ be the class of priors such that Assumption 3 holds. Suppose that $\{\delta_n\} \in \mathbb{D}$ is not asymptotically equivalent to $\{\delta_n^*(\,\cdot\,;\pi)\}$ for $\pi \in \Pi$. Then at any $P_0$ at which asymptotic equivalence fails,*

$$\mathcal{R}(\{\delta_n\}; P_0) > \mathcal{R}(\{\delta_n^*(\,\cdot\,;\pi)\}; P_0) \quad \text{for all } \pi \in \Pi$$

*provided either of the following hold:*

*(i)* $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ *is a singleton;*

*(ii)* $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ *is not a singleton, and* $\arg\min_{d \in \mathcal{D}} \mathbb{E}^*[\dot{\rho}_{d,P_0}[Z^* + z]]$ *is a singleton for almost every $z$.*

We conclude with a slightly stronger optimality statement than Theorem 1. Define

$$\mathcal{R}_n(\delta_n; P_0) = \int_{\{h : P_{n,h} \in \mathcal{P}\}} \sqrt{n} \left( \mathbb{E}_{P_{n,h}} \left[ R(\delta_n(X^n), P_{n,h}) \right] - \min_{d \in \mathcal{D}} R(d, P_{n,h}) \right) \pi(P_{n,h}) \, \mathrm{d}h,$$

where $\pi$ is the density used to construct $\delta_n^*(X^n; \pi)$. Under mild conditions permitting the exchange of limits and integration (see condition (A.5) in the Appendix), we have

$$\lim_{n \to \infty} \mathcal{R}_n(\delta_n; P_0) = \pi(P_0) \cdot \mathcal{R}(\{\delta_n\}; P_0).$$

As the scale factor $\pi(P_0)$ is independent of the sequence $\{\delta_n\}$, the criterions $\lim_{n \to \infty} \mathcal{R}_n(\delta_n; P_0)$ and $\mathcal{R}(\{\delta_n\}; P_0)$ induce the same rankings over sequences of decision rules.

**Corollary 1** *Suppose that Assumptions 1 and 2 are satisfied. Let $\Pi$ be the class of priors such that Assumption 3 holds. Let $\{\delta_n\} \in \mathbb{D}$ be asymptotically equivalent to $\{\delta_n^*(\,\cdot\,;\pi)\}$ for a $\pi \in \Pi$ and satisfy condition (A.5). Then for all $P_0 \in \mathcal{P}$,*

$$\lim_{n \to \infty} \mathcal{R}_n(\delta_n; P_0) = \liminf_{n \to \infty} \inf_{\delta_n'} \mathcal{R}_n(\delta_n'; P_0),$$

*where the infimum is over all measurable $\delta_n' : \mathcal{X}^n \to \mathcal{D}$.*

## 4.2 Plug-in Rules are Sub-Optimal

A natural alternative to $\delta_n^*(X^n; \pi)$ is to plug an efficient estimator $\hat{P}$ of $P$ into the oracle decision rule, which yields the plug-in rule $\delta_n^{plug}(X^n) = \delta^o(\hat{P})$. Manski (2021, 2022) refers to this approach as "as-if optimization": the estimated parameters $\hat{P}$ are treated "as if" they are the true parameters, with $\delta^o(\hat{P})$ minimizing maximum risk (or regret) over $\Theta_0(\hat{P})$. This approach also has connections with anticipated utility (Kreps, 1998; Cogley and Sargent, 2008). As we show, plug-in rules can differ asymptotically from optimal rules. If so, Theorem 2

implies plug-in rules are sub-optimal. We illustrate this difference within the context of the application to treatment assignment from Section 2.

**Example: Treatment Assignment.** Recall the plug-in rule is $\delta_n^{plug}(X^n) = \mathbb{I}[b(\hat{P}) \geq 0]$ for $b(P) = (b_U(P))_+ + (b_L(P))_-$. The robust welfare contrast $b(P)$ may be only directionally differentiable in $P$ (see Appendix A) rather than fully differentiable because of the outer $(\cdot)_+$ and $(\cdot)_-$ operations and also because the bounds $b_L(P)$ and $b_U(P)$ may themselves depend non-smoothly on $P$. As we show in Section 8, non-smoothness $b_L$ and $b_U$ is a common feature across a wide range of empirical settings. Hence, when $b$ is non-smooth, the plug-in term $b(\hat{P})$ and the posterior mean $\bar{b}_n$ of $b(P)$ may fail to have the same asymptotic distribution (under a suitable centering and scaling). It follows that the plug-in decision $\delta_n^{plug}(X^n) = \mathbb{I}[b(\hat{P}) \geq 0]$ and the optimal decision $\delta_n^*(X^n; \pi) = \mathbb{I}[\bar{b}_n \geq 0]$ may fail to be asymptotically equivalent, in which case $\delta_n^*$ will dominate $\delta_n^{plug}$ under our optimality criterion.

Consider the empirical application from Section 2. The lower bound $b_L(\hat{P})$ on the ATE for male youths behaved approximately like the maximum of two independent Gaussian random variables (corresponding to estimates from US and Brazilian RCTs) while the upper bound on the ATE $b_U(\hat{P})$ behaved approximately like a third independent Gaussian random variable (corresponding to the estimate from a Colombian RCT). The optimal decision for male youths was to treat but the plug-in decision was to not treat.

To understand the difference between the two decisions we use the following stylized example. Let
$$P = (P_1, P_2, P_3)', \quad b_L(P) = (P_1 \vee P_2), \quad b_U(P) = P_3.$$

The robust welfare contrast is $b(P) = (P_1 \vee P_2)_- + (P_3)_+$. The plug-in rule is therefore
$$\delta_n^{plug}(X^n) = \mathbb{I}[(\hat{P}_1 \vee \hat{P}_2)_- + (\hat{P}_3)_+ \geq 0] \equiv \mathbb{I}[(\hat{P}_1 \vee \hat{P}_2) + \hat{P}_3 \geq 0].$$

In the empirical application, the bounds from the three studies are well separated from zero but of roughly the same magnitude (to within sampling uncertainty). To mimic this setting, let $P_0 = (-c, -c, c)$ for $c > 0$. Suppose that $\sqrt{n}(\hat{P} - P_0) \overset{P_{n,h}}{\rightsquigarrow} N(h, I)$. Let $Z \sim N(h, I)$. Then

$$\delta_n^{plug}(X^n) \overset{P_{n,h}}{\rightsquigarrow} \delta_\infty^{plug}(Z; P_0) := \mathbb{I}[(Z_1 \vee Z_2) + Z_3 \geq 0]. \tag{15}$$

For the optimal decision, suppose the posterior for $P$ is $N(\hat{P}, n^{-1}I)$. In Appendix C we derive $\delta_n^*(X^n; \pi)$ in closed form. For any $P_0 = (-c, -c, c)$ with $c > 0$, we show in Appendix C that
$$\delta_n^*(X^n; \pi) \overset{P_{n,h}}{\rightsquigarrow} \delta_\infty^*(Z; P_0) := \mathbb{I}\left[\mathbb{E}^*[(Z_1^* + Z_1) \vee (Z_2^* + Z_2)] + Z_3 \geq 0\right],$$

where $Z \sim N(h, I)$ and $\mathbb{E}^*$ denotes expectation with respect to $Z^* \sim N(0, I)$ given $Z$. By

17

Jensen's inequality,

$$\mathbb{E}^*[(Z_1^* + Z_1) \vee (Z_2^* + Z_2)] > Z_1 \vee Z_2 \quad \text{almost surely.}$$

Comparing the asymptotic representation $\delta_\infty^*(Z; P_0)$ with (15), we see that the optimal decision recommends treatment more aggressively than the plug-in rule does asymptotically. This result agrees with our empirical application from Section 2, in which the optimal rule recommends treatment but the plug-in rule does not.

Evidently, the plug-in and optimal decisions are not asymptotically equivalent. We can show formally that the plug-in decision is sub-optimal by applying Theorem 2. At $P_0 = (-c, -c, c)$, we have that $R(0, P_0) = R(1, P_0) = c$, so $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ is not a singleton. But the directional derivatives of $R(0, P) = (P_3)_+$ and $R(1, P) = -(P_1 \vee P_2)_-$ are $\dot{\rho}_{0,P_0}[h] = h_3$ and $\dot{\rho}_{1,P_0}[h] = -(h_1 \vee h_2)$, respectively. We therefore have

$$\mathbb{E}^*[\dot{\rho}_{0,P_0}[Z^* + z]] = z_3 \,,$$
$$\mathbb{E}^*[\dot{\rho}_{1,P_0}[Z^* + z]] = -z_1 \Phi((z_1 - z_2)/\sqrt{2}) - z_2 \Phi((z_2 - z_1)/\sqrt{2}) - \sqrt{2}\phi((z_1 - z_2)/\sqrt{2}) \,,$$

where the second expression follows, e.g., from Nadarajah and Kotz (2008). Hence, we see that $\arg\min_{d \in \mathcal{D}} \mathbb{E}^*[\dot{\rho}_{d,P_0}[Z^* + z]]$ is a singleton for almost every $z$. Theorem 2 therefore implies that the plug-in rule is sub-optimal: $\mathcal{R}(\{\delta_n^{plug}\}, P_0) > \mathcal{R}(\{\delta_n^*(\,\cdot\,; \pi)\}; P_0)$.

# 5 Semiparametric Models

This section extends our approach to semiparametric models. This extension is relevant for many empirical settings. For instance, the DM may not have grounds for asserting that the data are drawn from a given parametric model. Alternatively, $X^n$ may be a vector of estimators or statistics whose exact finite-sample distribution is unknown, as was the case for the empirical application in Section 2. We first state the model in Section 5.1 then generalize our optimality concept to semiparametric models in Section 5.2. Section 5.3 describes a quasi-Bayesian implementation of optimal decisions in this setting. Section 5.4 presents the main optimality results. In particular, Theorem 3 shows that quasi-Bayes decision rules are optimal. Moreover, Theorem 4 shows that any decision whose asymptotic behavior is different from quasi-Bayes rules are sub-optimal.

## 5.1 Model

Let $X^n \sim F_{n,\beta}$ with $\beta = (P, \eta)$ for $P \in \mathcal{P} \subseteq \mathbb{R}^K$ and $\eta \in \mathcal{H}$, an infinite-dimensional space. For instance, in a GMM model the parameter $\eta$ is the marginal distribution of each observation $X_i$ and $\mathcal{H} = \{\eta : \int g(x, P) \, d\eta(x) = 0 \text{ for some } P \in \mathcal{P}\}$ for a vector of moment functions $g$.

We again assume that given $P$, the structural parameter $\theta$ takes values in a set $\Theta_0(P)$. Therefore, the set of payoff distributions is indexed only by the parametric component $P$. The nonparametric component $\eta$ is a nuisance parameter. For instance, $P$ may be a vector of population moments used to construct bounds on $\theta$. The nuisance parameter $\eta$ represents other features of the distribution of $X^n$ that are irrelevant for the DM's decision problem.

## 5.2 Optimality Criterion

In the parametric case, our asymptotic efficiency criterion (10) integrates the excess maximum risk incurred by $\delta_n(X^n)$ relative to the oracle under $P_{n,h}$ using Lebesgue measure on the local perturbations $h \in \mathbb{R}^K$ of $P_0$. This approach does not extend easily to perturbations of $(P_0, \eta_0)$ in the semiparametric case due to measure-theoretic complications that arise in infinite-dimensional spaces.

We form our asymptotic efficiency criterion by averaging across local perturbations of $P_0$ within an *approximately least-favorable submodel*, in which $X^n$ carries the least information about $P$ of all parametric submodels. The problem of parameter estimation in the least-favorable submodel is asymptotically equivalent to the problem of estimating $P$ in full the semiparametric model. A formal definition of the least-favorable submodel is presented in Appendix A. For now, we let $t \mapsto \beta_{(P_0,\eta_0)}(t)$ denote the mapping from an open neighborhood $\mathcal{P}(P_0, \eta_0)$ of $P_0$ into $\mathcal{P} \times \mathcal{H}$ under the least-favorable submodel at $(P_0, \eta_0)$. Our optimality criterion is analogous to criterion (10), namely

$$\mathcal{R}(\{\delta_n\};(P_0,\eta_0)) := \int \lim_{n\to\infty} \sqrt{n} \left( \mathbb{E}_{\beta_{(P_0,\eta_0)}(P_{n,h})} \left[ R(\delta_n(X^n), P_{n,h}) \right] - \min_{d\in\mathcal{D}} R(d, P_{n,h}) \right) dh, \tag{16}$$

where $\mathbb{E}_{\beta_{(P_0,\eta_0)}(P_{n,h})}$ denotes expectation with respect to $X^n \sim F_{n,\beta_{(P_0,\eta_0)}(P_{n,h})}$.

We say a sequence of decision rules $\{\delta_n\}$ is optimal if it minimizes $\mathcal{R}(\,\cdot\,;(P_0,\eta_0))$ over the class $\mathbb{D}$ of decision rules for which the limit in (16) is well defined, for all $(P_0, \eta_0) \in \mathcal{P} \times \mathcal{H}$.

## 5.3 Quasi-Bayesian Implementation

Optimal decisions are formed similarly to the parametric case, but we replace the posterior with a *quasi-posterior* formed from a limited-information criterion for $P$.

Suppose first that the DM may compute an efficient estimator $\hat{P}$ of $P$ from $X^n$ such that for each $(P_0, \eta_0) \in \mathcal{P} \times \mathcal{H}$, $\sqrt{n}(\hat{P} - P_0) \rightsquigarrow N(0, I_0^{-1})$ with $I_0 = I_0(P_0, \eta_0)$ the semiparametric information bound. Also suppose that the DM may compute a consistent estimator $\hat{I}$ of $I_0$. Following Doksum and Lo (1990) and Kim (2002), the DM could use use a limited information $N(\hat{P}, (n\hat{I})^{-1})$ quasi-likelihood for $P$ which leads to the *quasi-posterior*

$$\pi_n(P|X^n) \propto e^{-\frac{1}{2}(P-\hat{P})'(n\hat{I})(P-\hat{P})}\pi(P),$$

19

for some strictly positive, continuously differentiable density $\pi$. Other limited-information criterions $Q_n(P)$ (e.g. GMM, minimum distance, or simulated method of moments) could be used following Chernozhukov and Hong (2003) (see also Chen, Christensen, and Tamer (2018)), in which case

$$\pi_n(P|X^n) \propto e^{Q_n(P)}\pi(P).$$

While we are deliberately vague about $Q_n$, we require that the quasi-posterior for $\sqrt{n}(P - \hat{P})$ should be well approximated by a $N(0, I_0^{-1})$ distribution in large samples.[5] We therefore implicitly require that the criterion $Q_n$ is "optimal", in the sense that maximizing $Q_n(P)$ leads to a semiparametrically efficient estimator $\hat{P}$ of $P$.

The *quasi-posterior maximum risk* $\bar{R}(d|X^n)$ is calculated by averaging $R(d, P)$ across the quasi-posterior, analogously to display (12). The *quasi-Bayes decision* $\delta_n^*(X^n; \pi)$ is chosen to minimize the quasi-posterior maximum risk $\bar{R}(d|X^n)$, analogously to (13). We again include $\pi$ as an argument of $\delta_n^*(\cdot)$ to indicate that the decision depends on the prior in finite samples. As we will show formally, the decisions $\delta_n^*(X^n; \pi)$ are optimal under the criterion (16) for any prior $\pi$ in the class of priors $\Pi$ for which our regularity conditions hold.

Unlike the parametric case, here the optimal decision cannot be justified on the basis of robust Bayes analysis or maxmin expected utility because $Q_n(P)$ is not the true log-likelihood. A formal Bayesian approach would require specifying a prior on $\mathcal{P} \times \mathcal{H}$ then forming a marginal posterior for $P$.[6] Our approach is computationally simple and avoids the delicate issue of specifying priors in infinite-dimensional parameter spaces.

## 5.4 Optimality of Quasi-Bayes Decisions

Appendix A presents Assumptions 1, 4, and 5, which are the main regularity conditions. Assumption 4 defines the approximately least-favorable model $\{F_{n,\beta_{(P_0,\eta_0)}}(P_{n,h})\}$ for $X^n$ and presents a notion of local asymptotic normality for it. Assumption 5 contains consistency and asymptotic normality assumptions for functionals of the quasi-posterior.

Let $\mathbb{D}$ denote all sequences $\{\delta_n\}$ that converge in distribution along $\{F_{n,\beta_{(P_0,\eta_0)}}(P_{n,h})\}$ for each $(P_0, \eta_0) \in \mathcal{P} \times \mathcal{H}$ and each $h \in \mathbb{R}^K$. Say $\{\delta_n\}$ and $\{\delta_n'\}$ are *asymptotically equivalent* if $\delta_n(X^n)$ and $\delta_n'(X^n)$ have the same asymptotic distribution along $\{F_{n,\beta_{(P_0,\eta_0)}}(P_{n,h})\}$ for all $P_0 \in \mathcal{P}$, $\eta_0 \in \mathcal{H}$, and $h \in \mathbb{R}^K$.

**Theorem 3** *Suppose that Assumptions 1 and 4 are satisfied. Let $\Pi$ be the class of priors such that Assumption 5 holds.*

*(i) For any $\pi \in \Pi$, the quasi-Bayes decision $\{\delta_n^*(\cdot; \pi)\}$ is optimal.*

---

[5]This is an implication of Assumption 5(ii).

[6]See, for instance, the Bayesian exponentially tilted empirical likelihood approach of Schennach (2005) or the Bayesian GMM approach of Shin (2015).

*(ii)* *For any $\pi, \pi' \in \Pi$, the quasi-Bayes decisions $\{\delta_n^*(\,\cdot\,;\pi)\}$ and $\{\delta_n^*(\,\cdot\,;\pi')\}$ are asymptotically equivalent.*

*(iii)* *If $\{\delta_n\}$ is asymptotically equivalent to $\{\delta_n^*(\,\cdot\,;\pi)\}$ for a $\pi \in \Pi$, then $\{\delta_n\}$ is optimal.*

As in the parametric case, asymptotic equivalence to $\{\delta_n^*\}$ is necessary for optimality under a side condition ruling out ties. Say asymptotic equivalence of $\{\delta_n\}$ and $\{\delta_n'\}$ fails at $(P_0, \eta_0)$ if $\delta_n(X^n)$ and $\delta_n'(X^n)$ have different asymptotic distributions along $\{F_{n,\beta_{(P_0,\eta_0)}}(P_{n,h})\}$ for some $h_0 \in \mathbb{R}^K$. Let $\mathbb{E}^*$ denote expectation with respect to $Z^* \sim N(0, I_0^{-1})$ where $I_0 = I_0(P_0, \eta_0)$ is the semiparametric information at $(P_0, \eta_0)$, and let $\dot{\rho}_{d,P_0}[\,\cdot\,]$ denote the directional derivative of $R(d, P)$ with respect to $P$ at $P_0$—see Appendix A.

**Theorem 4** *Suppose that Assumptions 1 and 4 are satisfied. Let $\Pi$ be the class of priors such that Assumption 5 holds. Suppose that $\{\delta_n\} \in \mathbb{D}$ is not asymptotically equivalent to $\{\delta_n^*(\,\cdot\,;\pi)\}$ for $\pi \in \Pi$. Then at any $(P_0, \eta_0)$ at which asymptotic equivalence fails,*

$$\mathcal{R}(\{\delta_n\}; (P_0, \eta_0)) > \mathcal{R}(\{\delta_n^*(\,\cdot\,;\pi)\}; (P_0, \eta_0)) \quad \text{for all } \pi \in \Pi$$

*provided either of the following hold:*

*(i)* $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ *is a singleton;*

*(ii)* $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ *is not a singleton, and* $\arg\min_{d \in \mathcal{D}} \mathbb{E}^*[\dot{\rho}_{d,P_0}[Z^* + z]]$ *is a singleton for almost every $z$.*

# 6  Further Discussion

In this section we discuss the interpretation of our approach in the context of a two-player zero-sum game and relate it to the literature on robust Bayes decision making.

## 6.1  Two-Player Game

Minimax decisions can be interpreted as optimal strategies in a zero-sum game played between the DM and an adversary ("nature"). Recall the case of "Known $P$" discussed in Section 3.2. In a zero-sum game the supremum over $\theta \in \Theta_0(P)$ in the maximum risk definition (7) can be viewed as nature's best response to the DM choosing $d$.

The interpretation of the "Unknown $P$" case is more delicate. Consider the Bayesian construction of the decision rule. Substitute the definition of $R(d, P)$ from (7) into the integrated

maximum risk criterion (11) and manipulate the resulting expression as follows:

$$\min_{\delta_n : \mathcal{X}^n \to \mathcal{D}} \int_{\mathcal{P}} \mathbb{E}_P \left[ \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[-u(\delta_n(X^n), Y, \theta, P)] \right] \mathrm{d}\pi(P)$$

$$= \int_{\mathcal{X}^n} \min_{d \in \mathcal{D}} \left[ \int_{\mathcal{P}} \left( \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[-u(d, Y, \theta, P)] \right) \mathrm{d}\pi_n(P|X^n) \right] \mathrm{d}\pi_n(X^n).$$

On the right-hand side we factorize the joint distribution of $(X^n, P)$ into the posterior $\pi_n(P|X^n)$ and the marginal distribution of the data $\pi_n(X^n)$. The exchange of minimization with respect to $d(X^n)$ and integration over $X^n$ is satisfied under weak regularity conditions because the DM makes a decision conditional on $X^n$. We used this calculation previously to obtain the Bayes decision $\delta_n^*(X^n; \pi)$ in (13).

From a minimax perspective, the non-standard feature of our setup is that the supremum over $\theta$ is taken inside the posterior expectation. This can be justified by assuming that in a zero-sum game nature is allowed to choose $\theta$ conditional on knowing $X^n$ and $P$, after the prior $\pi$ has been set and $(P, X^n)$ have been sampled from their joint distribution.

## 6.2 Relationship to Robust Bayesian Decision Making.

Our setup is closely related to robust Bayes (or $\Gamma$-minimax) decision making (Robbins, 1951; Berger, 1985) and the multiple priors framework of Gilboa and Schmeidler (1989). A key difference is that we distinguish two groups of parameters, $\theta$ and $P$, and apply the minimax reasoning only to $\theta$ because only it is partially identified. As mentioned previously, the approach of treating groups of parameters differently dates back to Hurwicz (1951). He argued that in econometrics and other applications researchers might consider multiple priors rather than a single prior for some of the parameters and referred to it as generalized Bayes-minimax principle. He provided a two-parameter example in which the marginal prior distribution for one of the parameters, $P$ in our notation, is fixed, whereas a family of priors is considered for the conditional distribution of the second parameter, which would be $\theta$, given $P$.

Suppose one combines the unique prior $\pi$ for $P$ with a family $\Lambda$ of conditional priors $\lambda(\theta|P)$ for $\theta$ where, for each $P \in \mathcal{P}$, each $\lambda(\theta|P)$ has support contained in $\Theta_0(P)$. Then one can define $\Gamma = \{\lambda(\theta|P) \cdot \pi(P) : \lambda \in \Lambda\}$. The $\Gamma$-minimax decision rule minimizes the maximum Bayes risk over $\gamma = \lambda \cdot \pi$ on each $X^n$ trajectory:

$$\min_{\delta_n : \mathcal{X}^n \to \mathcal{D}} \sup_{\gamma \in \Gamma} \int_{\mathcal{X}^n} \left( \int_{\mathcal{P}} \left[ \int_\Theta \mathbb{E}_\theta[-u(\delta_n(X^n), Y, \theta, P)] \mathrm{d}\lambda(\theta|P) \right] \mathrm{d}\pi_n(P|X^n) \right) \mathrm{d}\pi_n(X^n)$$

$$= \int_{\mathcal{X}^n} \min_{d \in \mathcal{D}} \sup_{\gamma \in \Gamma} \left( \int_{\mathcal{P}} \left[ \int_\Theta \mathbb{E}_\theta[-u(d, Y, \theta, P) d\theta] \mathrm{d}\lambda(\theta|P) \right] \mathrm{d}\pi_n(P|X^n) \right) \mathrm{d}\pi_n(X^n).$$

22

The key insight is that the marginal distribution of the data, $\pi_n(X^n)$, does not depend on nature's choice of $\gamma$ because conditional on $P$ the distribution of $X^n$ does not depend on $\theta$. This allows us to move the supremum over $\gamma \in \Gamma$ inside of the integral. Moreover, it implies that the posterior of $\theta|(P, X^n)$ is equal to the prior distribution of $\theta|P$.

Let $\theta_*(P) \in \arg\sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[-u(d, Y, \theta, P)]$, assuming that the arg sup is non-empty, and let $\lambda_*(\theta|P)$ be a pointmass at $\theta_*(P)$. Then, by construction,

$$\int_\mathcal{P} \left[ \int_\Theta \mathbb{E}_\theta[-u(d, Y, \theta, P)] \mathrm{d}\lambda_*(\theta|P) \right] \mathrm{d}\pi_n(P|X^n)$$
$$= \int_\mathcal{P} \left[ \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[-u(d, Y, \theta, P)] \right] \mathrm{d}\pi_n(P|X^n). \quad (17)$$

Moreover, for every $P$

$$\sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[-u(d, Y, \theta, P)] \geq \int \mathbb{E}_\theta[-u(d, Y, \theta, P)] \mathrm{d}\lambda(\theta|P). \quad (18)$$

Assuming that $\lambda_*(\theta|P) \cdot \pi(P) \in \Gamma$, we can deduce from (17) and (18) that

$$\int_\mathcal{P} \left[ \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[-u(d, Y, \theta, P)] \right] \mathrm{d}\pi_n(P|X^n)$$
$$= \sup_{\gamma \in \Gamma} \int_\mathcal{P} \left[ \int_\Theta \mathbb{E}_\theta[-u(d, Y, \theta, P)] \lambda(\theta|P) d\theta \right] \mathrm{d}\pi_n(P|X^n).$$

Thus, the optimal decision $\delta_n^*(\,\cdot\,; \pi)$ may be viewed as a $\Gamma$-minimax decision.

In related work, Giacomini and Kitagawa (2021) use a similar multiple priors approach involving a unique prior for a reduced-form parameter and multiple conditional priors for a partially identified structural parameter. They use this approach to study robust Bayesian approaches to estimation and inference on structural parameters. Our approach is different in that it studies a discrete decision problem and we provide a (frequentist) asymptotic efficiency theory for our approach. Further, we consider extensions to semiparametric models using a quasi-Bayesian approach that has a purely frequentist justification.

# 7 Optimal Pricing with Rich Unobserved Heterogeneity

In this section, we show how our methods can be applied to make optimal pricing decisions in models with rich unobserved heterogeneity using revealed-preference demand theory. Section 7.1 first gives the model and empirical setting. Section 7.2 presents techniques for computing sharp bounds on functionals of counterfactual demand using linear programming. Section 7.3 then discusses how to implement our methods.

23

## 7.1 Model

The DM observes repeated cross sections of individual demands $X^n = \left( X_{b,1}, \ldots, X_{b,n} \right)_{b=1}^{B}$, where each $X_{b,i} \in \mathbb{R}^K$ is the demand of individual $i$ for $m$ goods under prices $q_b$:

$$X_{b,i} = \arg \max_{x \in \mathcal{B}_b} v_i(x),$$

where $\mathcal{B}_b = \{x \in \mathbb{R}^m : x' q_b = 1\}$ is the budget set (expenditure is normalized to one). Individuals are heterogeneous in their utility functions $v_i$. We assume that the demand system is rationalized by a random utility model with a probability distribution $\Pi^v$ over utility functions $v$. The demand under $q_b$ of a randomly selected individual may therefore be interpreted as stochastic. The mass $p_b(s)$ of individuals whose demand is in any set $s \subset \mathcal{B}_b$ at price $q_b$ is

$$p_b(s) = \int \mathbb{I} \left\{ \text{argmax}_{x \in \mathcal{B}_b} v(x) \in s \right\} \mathrm{d}\Pi^v(v), \ s \in \mathcal{B}_b, \ b = 1, \ldots, B \tag{19}$$

(see, e.g., Kitamura and Stoye (2018), henceforth KS18).

The DM wishes to choose a new price vector $q_d$ for $d \in \mathcal{D} = \mathcal{O} \cup \mathcal{C}$. The price vectors $\{q_d : d \in \mathcal{O}\}$ are a subset of the observed prices $q_1, \ldots, q_B$, while each $\{q_d : d \in \mathcal{C}\}$ is a set of counterfactual price vectors. In principle, the set $\mathcal{C}$ of new price vectors could be large, representing prices rounded to nearest currency units or tax rates rounded to the nearest percentage. Let $h_d(X_d)$ represent a functional of demand (e.g., revenue or welfare) under prices $q_d$. The DM's goal is to choose $d \in \mathcal{D}$ to maximize the average of $h_d(X_d)$. For $d \in \mathcal{O}$ the average demand $\mathbb{E}[h_d(X_d)]$ is identified from observed choice behavior under $q_d$. However, the observed choice behavior is only sufficient to bound, but not point-identify, $\mathbb{E}[h_d(X_d)]$ for counterfactual prices $q_d$, $d \in \mathcal{C}$.

## 7.2 Bounds on Functionals of Counterfactual Demand

Kitamura and Stoye (2019) present a general approach using linear programming to bound functionals of counterfactual demands. We introduce their approach with an example.

Consider Figure 2. There are two goods, and the DM has observed the demand for two price vectors $q_1$ and $q_2$, which generate the budget sets $\mathcal{B}_1$ and $\mathcal{B}_2$. The counterfactual price vector is $q_0$ which generates the counterfactual budget set $\mathcal{B}_0$.

The budget lines in Figure 2 are divided into segments $s_{jb}$, called "patches" in KS18. Patch $s_{11}$ is the segment of $\mathcal{B}_1$ from the $y$-axis to the intersection of $\mathcal{B}_1$ and $\mathcal{B}_2$, patch $s_{21}$ is the segment of $\mathcal{B}_1$ between $\mathcal{B}_2$ and $\mathcal{B}_0$, and so forth.[7] There are 9 potential combinations of patches consumers may choose from among $\mathcal{B}_1$ and $\mathcal{B}_2$: $(s_{11}, s_{12})$, $(s_{11}, s_{22})$, $(s_{11}, s_{32})$,

---

[7]Like KS18, we suppose for simplicity that the distribution of demand is continuous so we can disregard the "intersection patches" formed at the intersections of budget lines.
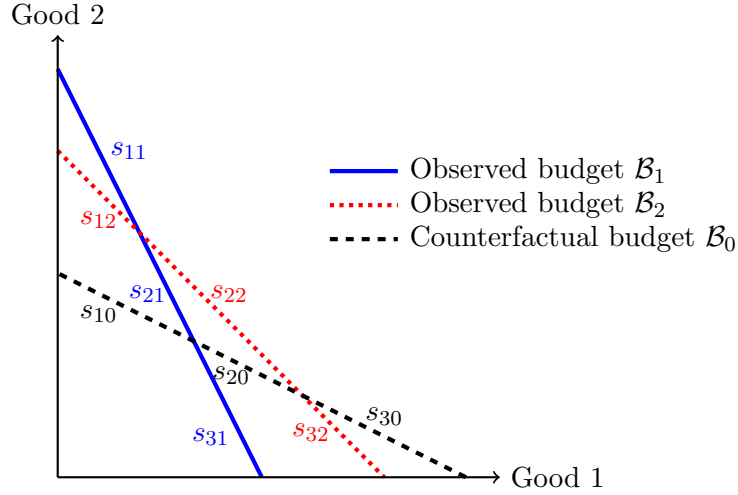
Figure 2: Choices and Budget Lines

$(s_{21}, s_{12})$, $(s_{21}, s_{22})$, $(s_{21}, s_{32})$, $(s_{31}, s_{12})$, $(s_{31}, s_{22})$, $(s_{31}, s_{23})$. Each combination corresponds to a consumer type. By revealed preference we know a consumer will never choose $(s_{21}, s_{12})$ or $(s_{31}, s_{12})$. This leaves a total of 7 rational types of consumer. Let

$$
A = \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1
\end{bmatrix}.
$$

The rows of $A$ correspond to $s_{11}, s_{21}, s_{31}, s_{12}, s_{22}, s_{32}$ and the columns of $A$ correspond to the 7 rational types. Let $p = (p_1(s_{11}), p_1(s_{21}), p_1(s_{31}), p_2(s_{12}), p_2(s_{22}), p_2(s_{32}))'$ collect the corresponding choice probabilities. KS18 showed that the demand system $p$ is rationalizable if and only if

$$
p = A\pi^v,
$$

for some $\pi^v \in \Delta^6$, the unit simplex in $\mathbb{R}^7$. Each entry of $\pi^v$ corresponds to probabilities of the 7 rational types. These probabilities constrain the set of distributions $\Pi^v$ of utilities.

Now consider choice behavior on the counterfactual budget set $\mathcal{B}_0$. Each of the 7 rational types may choose a counterfactual demand in patch $s_{10}$, $s_{20}$, or $s_{30}$, for a total of 21 potential types. By revealed preference, a consumer who chose $s_{31}$ must choose $s_{20}$ or $s_{30}$, and a consumer who chose $s_{32}$ must choose $s_{30}$. This leaves a total of 16 rational types. We may

represent the system as

$$\begin{bmatrix} p \\ p^* \end{bmatrix} = A^* \pi^{v*}, \tag{20}$$

where $p^* = (p_0(s_{10}), p_0(s_{20}), p_0(s_{30}))'$ collects the counterfactual choice probabilities for patches $s_{10}$, $s_{20}$, and $s_{30}$, $\pi^{v*} \in \Delta^{15}$ collects the probabilities of observing each rational type, and

$$A^* = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix},$$

where the rows correspond to choosing $s_{11}, s_{21}, s_{31}, s_{12}, s_{22}, s_{32}, s_{10}, s_{20}, s_{30}$. Partition $A^*$ as

$$A^* = \begin{bmatrix} A^*_{obs} \\ A^*_{unobs} \end{bmatrix},$$

where $A^*_{obs}$ collects the first 6 rows of $A^*$ (corresponding to the observed budget sets) while $A^*_{unobs}$ collects the rows of $A^*$ corresponding to the patches $s_{10}$, $s_{20}$, and $s_{30}$.

Following Kitamura and Stoye (2019), we may deduce sharp bounds on $\mathbb{E}[h_0(X_0)]$ as follows. Let $(\underline{h}_1, \underline{h}_2, \underline{h}_3)$ and $(\overline{h}_1, \overline{h}_2, \overline{h}_3)$ denote row vectors which collect the smallest and largest values of $h_0(x)$ for $x$ in $s_{10}$, $s_{20}$ and $s_{30}$. Then the bounds on $\mathbb{E}[h_0(X_0)]$ are

$$h_L(p) = \min \left\{ (\underline{h}_1, \underline{h}_2, \underline{h}_3) A^*_{unobs} \pi^{v*} : p = A^*_{obs} \pi^{v*}, \pi^{v*} \in \Delta^{15} \right\},$$
$$h_U(p) = \max \left\{ (\overline{h}_1, \overline{h}_2, \overline{h}_3) A^*_{unobs} \pi^{v*} : p = A^*_{obs} \pi^{v*}, \pi^{v*} \in \Delta^{15} \right\}.$$

More generally, the preceding argument applies with a collection of $B$ observed budget sets and multiple goods. In that case, representation (20) holds for observed choice probabilities $p$ of patches on the $B$ observed budget sets and counterfactual choice probabilities $p^*$ of patches on the counterfactual budget set $\mathcal{B}_d$ generated by $q_d$ for $d \in \mathcal{C}$. Suppose there are $J$ patches across $\mathcal{B}_d$ and $T$ rational types. Then letting row vectors $\underline{h}_d$ and $\overline{h}_d$ collect the smallest and largest values of $h_d(x)$ for $x$ in each of the $J$ patches, sharp bounds on $\mathbb{E}[h_d(X_d)]$ are

$$h_{d,L}(p) = \min \left\{ \underline{h}_d A^*_{unobs} \pi^{v*} : p = A^*_{obs} \pi^{v*}, \pi^{v*} \in \Delta^{T-1} \right\},$$
$$h_{d,U}(p) = \max \left\{ \overline{h}_d A^*_{unobs} \pi^{v*} : p = A^*_{obs} \pi^{v*}, \pi^{v*} \in \Delta^{T-1} \right\}, \tag{21}$$

where we have partitioned the $A^*$ matrix analogously to the simple 3 budget example.

## 7.3 Implementation

The DM's problem maps into our framework as follows. For each $d \in \mathcal{O}$, the value $\mathbb{E}[h_d(X^d)]$ is identified from observed choice behavior under $q_d$. The remaining reduced-form parameters are the patch probabilities $p$. Thus, $P = (p, (\mathbb{E}[h(X_d)])_{d \in \mathcal{O}})$.

Here $\theta = \Pi^v$ and $\Theta_0(P)$ is all $\Pi^v$ that rationalize the patch probabilities $p$ consistent with revealed preference. The vector $Y = (h_d(X^d))_{d \in \mathcal{C}}$ collects functionals of demand under the counterfactual prices. Each $\theta \equiv \Pi^v \in \Theta_0(P)$ induces a distribution $G_\theta$ for $Y$. Although we suppressed it in the previous subsections, we now write $\mathbb{E}_\theta[h_d(X^d)]$ for $d \in \mathcal{C}$ to denote that the average counterfactual demand functional depends on the structural parameter $\theta$.

For $d \in \mathcal{O}$, the DM incurs risk

$$r(d, \theta, P) = -\mathbb{E}[h_d(X_d)],$$

which is point identified from $P \equiv (p, (\mathbb{E}[h(X_d)])_{d \in \mathcal{O}})$. For $d \in \mathcal{C}$, the DM incurs risk

$$r(d, \theta, P) = -\mathbb{E}_\theta[h_d(X_d)],$$

which is set-identified and may be bounded as described in the previous subsection. The maximum risk is

$$R(d, P) = \begin{cases} -\mathbb{E}[h_d(X_d)] & d \in \mathcal{O}, \\ -h_{d,L}(p) & d \in \mathcal{C}. \end{cases}$$

This is a semiparametric model and our implementation follows the steps described in Section 5.3. Partition $P = (P_1, \ldots, P_B)$ where $P_b$ collects the parameters corresponding to budget set $\mathcal{B}_b$. As we have assumed that each of the $B$ observed budgets is sampled in a repeated cross section, we can estimate each $P_b$ by just-identified GMM. We can then form a quasi-posterior $\pi_n(P|X^n)$ based on a limited-information Gaussian quasi-likelihood, exponentiating the GMM objective function, or by the Bayesian bootstrap. For $d \in \mathcal{O}$, the expected value $\mathbb{E}[h_d(X_d)]$ is an element of $P$ and so $\bar{R}(d|X^n)$ is simply its posterior mean:

$$\bar{R}(d|X^n) = -\int \mathbb{E}[h_d(X_d)] \, \mathrm{d}\pi(P|X^n).$$

For $d \in \mathcal{C}$, the posterior maximum risk is

$$\bar{R}(d|X^n) = -\int h_{d,L}(p) \, \mathrm{d}\pi(P|X^n).$$

This may be computed by sampling $p$ from the posterior, solving the linear program (21)

27

defining $h_{d,L}(p)$ for each draw of $p$, then taking the average across draws.

The optimal decision minimizes $\bar{R}(d|X^n)$ for $d \in \mathcal{D} = \mathcal{O} \cup \mathcal{C}$. As the value of a linear program is typically directionally differentiable, the posterior mean $\bar{R}(d|X^n)$ for $d \in \mathcal{C}$ may be different, even asymptotically, from the corresponding plug-in values $-h_{d,L}(\hat{p})$. In this case, the optimal decision $\delta_n^*$ may dominate the plug-in rule.

# 8  Treatment Rules: Further Examples and Applications

This section expands on our running example of treatment assignment under partial identification. We first review several empirically relevant approaches for constructing bounds on the ATE. Each of these constructions leads to bounds $b_L(P)$ and $b_U(P)$ that will in general be only directionally differentiable in a vector of reduced-form parameters $P$. We then discuss implementation of the optimal decision rules in these settings. As far as we are aware, ours is the first work to propose optimal decisions for these realistic empirical settings.

**Intersection Bounds.**  This setting generalizes the empirical application from Section 2. Suppose $X^n$ consists of data from $K$ observational studies. In each study $k$ we can consistently estimate lower and upper bounds $b_{L,k}$ and $b_{U,k}$ on the ATE as a function of population moments $P_k$. We then obtain the intersection bounds

$$b_L(P) = \max_{1 \leq k \leq K} b_{L,k}, \qquad b_U(P) = \min_{1 \leq k \leq K} b_{U,k},$$

with $P = (P_k)_{k=1}^K$. While the bounds $b_{L,k}$ and $b_{U,k}$ may themselves be smooth in $P_k$, the presence of the min and max operations makes the intersection bounds $b_L(P)$ and $b_U(P)$ only directionally differentiable in $P$.

**Bounds via IV-like Estimands.**  Mogstad, Santos, and Torgovitsky (2018) present an approach for bounding the ATE and other causal effects using IV-like estimates from observational studies. Suppose treatment is determined by $D = \mathbb{I}[U \leq \nu(Z)]$ where $U \sim \text{Uniform}(0,1)$ and $Z = (X, Z_0)$ collects control variables $X$ and instrumental variables $Z_0$. According to Heckman and Vytlacil (1999, 2005), the ATE may be expressed as a functional $\Gamma_0(m)$, where $m = (m_0, m_1)$ are the marginal treatment response (MTR) functions

$$m_d(u, x) = \mathbb{E}[Y_d | U = u, X = x], \quad d \in \{0, 1\},$$

and

$$\Gamma_0(m) = \mathbb{E}\left[\int_0^1 m_1(u, X) \, \mathrm{d}u - \int_0^1 m_0(u, X) \, \mathrm{d}u\right].$$

Mogstad et al. (2018) show the MTR functions, and hence the identified set for the ATE, can be disciplined if we know the value of certain IV-like estimands. For ease of exposition, consider a single IV estimand

$$\beta_{IV} = \frac{\text{Cov}(Y, Z_0)}{\text{Cov}(D, Z_0)},$$

resulting from using $Z_0$ as an instrument for treatment status dummy $D$ in the observational study. The IV estimand may be expressed as $\beta_{IV} = \Gamma_\beta(m)$ where

$$\Gamma_\beta(m) = \mathbb{E}\left[\int_0^1 m_0(u, X)s(0, Z_0)\mathbb{I}[u > p(z_0)]\,\mathrm{d}u + \int_0^1 m_1(u, X)s(1, Z_0)\mathbb{I}[u \le p(z_0)]\,\mathrm{d}u\right]$$

with $s(d, z) = \frac{z_0 - \mathbb{E}[Z_0]}{\text{Cov}(Z_0, D)}$ and where $p(z_0) = \mathbb{E}[D | Z_0 = z_0]$ is the propensity score. In view of the above, Mogstad et al. (2018) derive (more general versions of) the bounds

$$b_L(P) = \inf_{m \in \mathcal{S}: \Gamma_\beta(m) = \beta_{IV}} \Gamma_0(m), \qquad b_U(P) = \sup_{m \in \mathcal{S}: \Gamma_\beta(m) = \beta_{IV}} \Gamma_0(m),$$

where $\mathcal{S}$ is a class of functions and $P = (\text{Cov}(Y, Z_0), \text{Cov}(D, Z_0), \mathbb{E}[Z_0], p)$, which is finite-dimensional if $Z_0$ has finite support (e.g. binary $Z_0$). They show that $b_L(P)$ and $b_U(P)$ may be expressed as the optimal values of linear programs parameterized by $P$. It is known from Milgrom and Segal (2002) (see also Mills (1956) and Williams (1963) for linear programs) that the value of optimization problems may only be directionally differentiable in parameters.

**Non-separable Panel Data Models.** Suppose the outcome for individual $i$ at date $t$ is of the form $Y_{it} = g(X_{it}, \alpha_i, \varepsilon_{it})$ where $X_{it}$ is a vector of covariates, $\alpha_i$ is a latent individual effect, and $\varepsilon_{it}$ is a vector of disturbances, which are independent across individuals and time. Consider an intervention that changes in covariates from $x^0$ to $x^1$. The ATE associated with the intervention is

$$\int \left(g(x^1, \alpha, \varepsilon) - g(x^0, \alpha, \varepsilon)\right) \mathrm{d}Q(\alpha, \varepsilon),$$

where $Q$ is a distribution over $(\alpha, \varepsilon)$. With discrete regressors $X_{it}$ and discrete outcomes $Y_{it}$, parametric restrictions on the distribution of $\varepsilon$ and functional form restrictions on $g$ are generally insufficient to point identify the ATE without parametric restrictions on the distribution of $\alpha$. A leading example is dynamic panel data models in which $X_{it}$ collects lagged values of a discrete outcome $Y_{it}$—see, e.g., Honoré and Tamer (2006) and Torgovitsky (2019a). Building on Honoré and Tamer (2006), Chernozhukov, Fernández-Val, Hahn, and Newey (2013) and Torgovitsky (2019b) derive bounds on the ATE without parametric assumptions on $Q$. Their bounds may be expressed as the value of optimization problems (linear programs) parameterized by a finite-dimensional vector of choice probabilities $P$. As in the previous example, $b_L(P)$ and $b_U(P)$ may therefore only be directionally differentiable in $P$.

**Implementation.** For the first two examples, $X^n$ may represent the data or the vector of estimates $\hat{P}$ of $P$. The data-generating process for $X^n$ will typically be semiparametric.[8] If the DM has a vector of (efficient) estimates $\hat{P}$ of $P$ and a consistent estimate $\hat{I}^{-1}$ of the asymptotic variance of $\hat{P}$, then the optimal decision can be implemented based on a $N(\hat{P}, (n\hat{I})^{-1})$ quasi-posterior. Alternatively, if the researcher observes the data $X^n$, then a quasi-posterior based on an efficient GMM objective function can be used.

In the nonseparable panel data case with discrete $Y_{it}$ and $X_{it}$, the distribution $F_{n,P}$ of the data $((X_{it})_{t=1}^{T})_{i=1}^{n}$ can be identified with a multinomial distribution parameterized by $P$. In this case our parametric theory applies and the optimal decision can easily be implemented using either the bootstrap or Bayesian methods. For the latter, a Dirichlet prior for $P$ leads to a Dirichlet posterior for $P$, which is trivial to sample from.

## 9 Conclusion

We derived optimal statistical decision rules for discrete choice problems when payoffs depend on a set-identified parameter $\theta$ and the decision maker can use a point-identified parameter $P$ to deduce restrictions on $\theta$. Our optimal decision rules minimize maximum risk or regret over the identified set for $\theta$ conditional on $P$, and use the data efficiently to learn about $P$. In many empirically relevant applications, the maximum risk depends non-smoothly on $P$. In those cases, plug-in rules are suboptimal under our asymptotic efficiency criterion. We provided detailed applications to optimal treatment choice under partial identification and optimal pricing with rich unobserved heterogeneity. Our asymptotic approach is well suited for empirical settings in which the derivation of finite-sample optimal rules is intractable. While our asymptotic optimality theory was developed for discrete decisions, our general approach can be used for continuous decisions as well. We conjecture that similar optimality results also apply in that case.

## References

Bergemann, D. and K. Schlag (2011). Robust monopoly pricing. *Journal of Economic Theory 146*(6), 2527–2543.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York.

Blundell, R., M. Browning, and I. Crawford (2007). Improving revealed preference bounds on demand responses. *International Economic Review 48*(4), 1227–1244.

---

[8]An exception is the intersection bounds example, where in each study we observe a binary treatment indicator and a binary outcome across individuals. That case can be handled by forming a multinomial likelihood parametrized by the reduced-form parameters.

Blundell, R., M. Browning, and I. Crawford (2008). Best nonparametric bounds on demand responses. *Econometrica 76*(6), 1227–1262.

Blundell, R., D. Kristensen, and R. Matzkin (2014). Bounding quantile demand functions using revealed preference inequalities. *Journal of Econometrics 179*(2), 112–127.

Blundell, R. W., D. Kristensen, and R. L. Matzkin (2017). Individual counterfactuals with multidimensional unobserved heterogeneity. Technical report, cemmap working paper.

Card, D., J. Kluve, and A. Weber (2017). What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. *Journal of the European Economic Association 16*(3), 894–931.

Chamberlain, G. (2000). Econometric applications of maxmin expected utility. *Journal of Applied Econometrics 15*(6), 625–644.

Chamberlain, G. (2011). Bayesian aspects of treatment choice. In *The Oxford Handbook of Bayesian Econometrics*. Oxford University Press.

Chen, X., T. M. Christensen, and E. Tamer (2018). Monte carlo confidence sets for identified sets. *Econometrica 86*(6), 1965–2018.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Chernozhukov, V. and H. Hong (2003). An MCMC approach to classical estimation. *Journal of Econometrics 115*(2), 293–346.

Christensen, T., H. R. Moon, and F. Schorfheide (2020). Robust forecasting. *arXiv preprint arXiv:2011.03153*.

Cogley, T. and T. J. Sargent (2008). Anticipated utility and rational expectations as approximations of Bayesian decision making. *International Economic Review 49*(1), 185–221.

Doksum, K. A. and A. Y. Lo (1990). Consistent and Robust Bayes Procedures for Location Based on Partial Information. *The Annals of Statistics 18*(1), 443–453.

Ghosh, J. and R. Ramamoorthi (2003). *Bayesian Nonparametrics*. Springer Verlag, New York.

Giacomini, R. and T. Kitagawa (2021). Robust bayesian inference for set-identified models. *Econometrica 89*(4), 1519–1556.

Giacomini, R., T. Kitagawa, and M. Read (2021). Robust Bayesian analysis for econometrics. *Federal Reserve Bank of Chicago Working Paper 2021-11*.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics 18*(2), 141–153.

Hartigan, J. A. (1983). *Bayes Theory*. Springer Verlag, New York.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy*

*of Sciences 96*(8), 4730–4734.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hoderlein, S. and J. Stoye (2015). Testing stochastic rationality and predicting stochastic demand: the case of two goods. *Economic Theory Bulletin 3*(2), 313–328.

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*(5), 611–629.

Hurwicz, L. (1951). Some specification problems and applications to econometric models. *Econometrica 19*(3), 343–344.

Ishihara, T. and T. Kitagawa (2021). Evidence aggregation for treatment choice. *arXiv preprint arXiv:2108.06473*.

Kim, J.-Y. (2002). Limited information likelihood and bayesian analysis. *Journal of Econometrics 107*(1-2), 175–193.

Kitamura, Y. and J. Stoye (2018). Nonparametric analysis of random utility models. *Econometrica 86*(6), 1883–1909.

Kitamura, Y. and J. Stoye (2019). Nonparametric counterfactuals in random utility models. *arXiv preprint arXiv:1902.08350*.

Kreps, D. M. (1998). Anticipated utility and dynamic choice. In D. P. Jacobs, E. Kalai, M. I. Kamien, and N. L. Schwartz (Eds.), *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures, 1983-1997*, Econometric Society Monographs, pp. 242–274. Cambridge University Press.

Manski, C. F. (2000). Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics 95*(2), 415–442.

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

Manski, C. F. (2007a). Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics 139*(1), 105–115. Endogeneity, instruments and identification.

Manski, C. F. (2007b). Partial identification of counterfactual choice probabilities. *International Economic Review 48*(4), 1393–1410.

Manski, C. F. (2014). Identification of income-leisure preferences and evaluation of income tax policy. *Quantitative Economics 5*(1), 145–174.

Manski, C. F. (2020). Toward credible patient-centered meta-analysis. *Epidemiology 31*(3), 345–352.

Manski, C. F. (2021). Econometrics for decision making: Building foundations sketched by Haavelmo and Wald. *Econometrica 89*(6), 2827–2853.

Manski, C. F. (2022). Probabilistic prediction for binary treatment choice: with focus on personalized medicine. *Journal of Econometrics (Forthcoming)*.

Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601.

Mills, H. D. (1956). Marginal values of matrix games and linear programs. In *Linear Inequalities and Related Systems.*, pp. 183–194. Princeton University Press.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica 86*(5), 1589–1619.

Moon, H. R. and F. Schorfheide (2012). Bayesian and frequentist inference in partially identified models. *Econometrica 80*(2), 755–782.

Murphy, S. A. and A. W. van der Vaart (2000). On profile likelihood. *Journal of the American Statistical Association 95*(450), 449–465.

Nadarajah, S. and S. Kotz (2008). Exact distribution of the max/min of two gaussian random variables. *IEEE Transactions on Very Large Scale Integration Systems 6*(2), 210–212.

Robbins, H. (1951). Asymptocially subminimax solutions of compound decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Volume I. University of California Press, Berkeley and Los Angeles.

Russell, T. M. (2020). Policy transforms and learning optimal policies. *arXiv preprint arXiv:2012.11046*.

Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika 92*(1), 31–46.

Shin, M. (2015). *Bayesian GMM*. Ph. D. thesis.

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics 166*, 138–156.

Torgovitsky, A. (2019a). Nonparametric inference on state dependence in unemployment. *Econometrica 87*(5), 1475–1505.

Torgovitsky, A. (2019b). Partial identification by extending subdistributions. *Quantitative Economics 10*(1), 105–144.

van der Vaart, A. (1991). An asymptotic representation theorem. *International Statistical Review 59*(1), 97–121.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wald, A. (1950). *Statistical Decision Functions*. John Wiley, New York.

Williams, A. C. (1963). Marginal values in linear programming. *Journal of the Society for Industrial and Applied Mathematics 11*(1), 82–94.

Yata, K. (2021). Optimal decision rules under partial identification. *arXiv preprint arXiv:2111.04926*.

## Online Appendix: Optimal Discrete Decisions when Payoffs are Partially Identified

### Timothy Christensen, Hyungsik Roger Moon, and Frank Schorfheide

This Appendix consists of the following sections:

# A    Assumptions

This section presents the main regularity conditions for Theorems 1 and 2. We first place some assumptions on the maximum risk. Say $f : \mathcal{P} \to \mathbb{R}^K$ is *directionally differentiable* at $P_0$ if the limit

$$\lim_{t \downarrow 0} \frac{f(P_0 + th) - f(P_0)}{t} =: \dot{f}_{P_0}[h]$$

exists for every $h \in \mathbb{R}^K$, in which case $\dot{f}_{P_0}[\cdot]$ denotes the directional derivative of $f$ at $P_0$. If it exists, the directional derivative $\dot{f}_{P_0}[\cdot]$ is positively homogeneous of degree one but not necessarily linear. If $\dot{f}_{P_0}[\cdot]$ is linear then $f$ is *fully* differentiable at $P_0$. Define $\rho : \mathcal{P} \to \mathbb{R}^{D+1}$ by $\rho(P) = (R(0, P), R(1, P), \ldots, R(D, P))'$. If $\rho$ is directionally differentiable at $P_0$, we let $\dot{\rho}_{P_0}[\cdot]$ denote its directional derivative.

**Assumption 1**    *(i)  The function $\rho$ is bounded and continuous on $\mathcal{P}$;*

*(ii)  The function $\rho$ is directionally differentiable any $P_0 \in \mathcal{P}$ for which $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ is not a singleton.*

The remaining assumptions are presented separately for parametric and semiparametric models.

### A.1    Parametric Models

We first impose some assumptions on the data-generating process (DGP). Let $\overset{P_{n,h}}{\rightsquigarrow}$ and $\overset{P_{n,h}}{\to}$ denote convergence in distribution and in probability under the sequence of measures $\{F_{n,P_{n,h}}\}$ with $P_{n,h} = P_0 + h/\sqrt{n}$ for $P_0$ fixed and $h$ ranging over $\mathbb{R}^K$. Say the model for $X^n$ is *locally asymptotically normal* at $P_0$ if for each $h_0 \in \mathbb{R}^K$, the likelihood ratio processes indexed by any finite subset $H \subset \mathbb{R}^K$ converge weakly to the likelihood ratio in a shifted normal model:

$$\left( \frac{\mathrm{d}F_{n,P_{n,h}}}{\mathrm{d}F_{n,P_{n,h_0}}} \right)_{h \in H} \overset{P_{n,h_0}}{\rightsquigarrow} \left( \exp\left( (h - h_0)'Z - \frac{1}{2}(h - h_0)'I_0(h - h_0) \right) \right)_{h \in H},$$

with $Z \sim N(h_0, I_0^{-1})$ and $I_0 = I_0(P_0)$ is the (asymptotic) information matrix at $P_0$.

**Assumption 2**    *(i) $\mathcal{P}$ is an open subset of $\mathbb{R}^K$ with $P_0 \in \mathcal{P}$;*
*(ii) The model for $X^n$ is locally asymptotically normal at each $P_0 \in \mathcal{P}$.*

Finally, we impose some assumptions on the large-sample behavior of the posterior. We say that $\pi_n(P|X^n)$ is *consistent* if $\pi_n(P \in N|X^n) \xrightarrow{P_0} 1$ for every neighborhood $N$ containing $P_0$, for each $P_0 \in \mathcal{P}$. Let $\mathbb{P}_h$ denote probability with respect to $Z \sim N(h, I_0^{-1})$ for $I_0 = I_0(P_0)$ and let $\mathbb{E}^*$ denote expectation with respect to $Z^*$ where $Z^* \sim N(0, I_0^{-1})$ independent of $Z$.

**Assumption 3**    *(i) The posterior $\pi_n(P|X^n)$ is consistent;*
*(ii) At any $P_0 \in \mathcal{P}$ with non-singleton $\arg\min_{d \in \mathcal{D}} R(d, P_0)$, for any Borel set $A \subseteq \mathbb{R}^{D+1}$ and $h \in \mathbb{R}^K$, we have*

$$\lim_{n \to \infty} F_{n, P_{n,h}} \left( \int \sqrt{n}(\rho(P) - \rho(P_0)) \, d\pi_n(P|X^n) \in A \right) = \mathbb{P}_h \left( \mathbb{E}^*[\dot{\rho}_{P_0}[Z^* + Z]|Z] \in A \right).$$

Assumption 3(i) is a mild consistency condition that is satisfied under minimal conditions. For the intuition behind Assumption 3(ii), suppose $X^n = (X_1, \ldots, X_n)$ where the $X_i$ are independent $N(P, I_0^{-1})$ random variables under $F_{n,P}$. Under a flat prior for $P$, $\pi_n(P|X^n)$ is $N(\hat{P}, (nI_0)^{-1})$ where $\hat{P} = \bar{X}_n$ is the sample mean, which itself has a $N(P_{n,h}, (nI_0)^{-1})$ distribution under $F_{n,P_{n,h}}$. On a shrinking neighborhood of $P_0$, we then have

$$\sqrt{n}(\rho(P) - \rho(P_0)) \approx \dot{\rho}_{P_0}[\sqrt{n}(P - \hat{P}) + \sqrt{n}(\hat{P} - P_0)],$$

where $\sqrt{n}(P - \hat{P}) \sim N(0, I_0^{-1})$ under $\pi_n(P|X^n)$ and $\sqrt{n}(\hat{P} - P_0) \sim N(h, I_0)$ under $F_{n,P_{n,h}}$. Similar reasoning applies in more general smooth parametric models by the Bernstein–von Mises theorem and standard asymptotic distribution theory for the MLE. Primitive conditions for consistency and the asymptotic normality of posterior distributions can be found, for instance, in the textbook treatments of Hartigan (1983), van der Vaart (1998), or Ghosh and Ramamoorthi (2003).

## A.2   Semiparametric Models

Our conditions for the semiparametric case parallel those for the parametric case. First we place assumptions on the DGP. Similar to Murphy and van der Vaart (2000), we say the model for $X^n$ has an *approximately least-favorable submodel* at $(P_0, \eta_0)$ if (i) there exists a map $t \mapsto \eta_{(P_0, \eta_0)}(t)$ from an open neighborhood $\mathcal{P}(P_0, \eta_0)$ of $P_0$ into $\mathcal{H}$ for which $\eta_{(P_0, \eta_0)}(P_0) = \eta_0$,

and (ii) with $\beta_{(P_0,\eta_0)}(P_{n,h}) = (P_{n,h}, \eta_{(P_0,\eta_0)}(P_{n,h}))$, for each $h_0 \in \mathbb{R}^K$ and any finite subset $H \subset \mathbb{R}^K$, we have

$$\left( \frac{\mathrm{d}F_{n,\beta_{(P_0,\eta_0)}(P_{n,h})}}{\mathrm{d}F_{n,\beta_{(P_0,\eta_0)}(P_{n,h_0})}} \right)_{h \in H} \overset{P_{n,h_0}}{\rightsquigarrow} \left( \exp\left( (h-h_0)'Z - \frac{1}{2}(h-h_0)'I_0(h-h_0) \right) \right)_{h \in H},$$

with $\overset{P_{n,h_0}}{\rightsquigarrow}$ and $\overset{P_{n,h_0}}{\rightarrow}$ denoting convergence in distribution and in probability under the sequence of measures $\{F_{n,\beta_{(P_0,\eta_0)}(P_{n,h})}\}$ with $P_{n,h} = P_0 + h/\sqrt{n}$ for $h \in \mathbb{R}^K$, and where $Z \sim N(h_0, I_0^{-1})$ with $I_0 = I_0(P_0, \eta_0)$ the (nonsingular) semiparametric information bound at $(P_0, \eta_0)$.

**Assumption 4**  *(i) $\mathcal{P}$ is an open subset of $\mathbb{R}^K$ with $P_0 \in \mathcal{P}$;*

*(ii) The model for $X^n$ has an approximately least-favorable submodel at each $(P_0, \eta_0) \in \mathcal{P} \times \mathcal{H}$.*

We then impose some high-level assumptions on the large-sample behavior of the quasi-posterior. We say that $\pi_n(P|X^n)$ is *consistent* if $\pi_n(P \in N|X^n) \overset{P_0}{\rightarrow} 1$ for every neighborhood $N$ containing $P_0$, for each $(P_0, \eta_0) \in \mathcal{P} \times \mathcal{H}$. Let $\mathbb{P}_h$ denote probability with respect to $Z \sim N(h, I_0^{-1})$ where $I_0 = I_0(P_0, \eta_0)$ is the semiparametric information bound and let $\mathbb{E}^*$ denote expectation with respect to $Z^*$ where $Z^* \sim N(0, I_0^{-1})$ independent of $Z$.

**Assumption 5**  *(i) The quasi-posterior $\pi_n(P|X^n)$ is consistent;*

*(ii) At any $P_0 \in \mathcal{P}$ with non-singleton $\arg\min_{d \in \mathcal{D}} R(d, P_0)$, for any Borel set $A \subseteq \mathbb{R}^{D+1}$ and $h \in \mathbb{R}^K$, we have*

$$\lim_{n \to \infty} F_{n,\beta_{(P_0,\eta_0)}(P_{n,h})} \left( \int \sqrt{n}(\rho(P) - \rho(P_0)) \, \mathrm{d}\pi_n(P|X^n) \in A \right)$$
$$= \mathbb{P}_h \left( \mathbb{E}^*[\dot{\rho}_{P_0}[Z^* + Z]|Z] \in A \right)$$

*for all $\eta_0 \in \mathcal{H}$.*

# B  Proofs

To simplify notation, throughout we write $\pi_n(P)$ for the posterior $\pi_n(P|X^n)$. We adopt the convention that $+\infty \times 0 = 0$. We also suppress dependence of $\delta_n^*$ on $\pi$, since any $\pi \in \Pi$ leads to a decision rule with the same asymptotic behavior.

## B.1 Proof of Theorems 1 and 2 and Corollary 1

### B.1.1 Preliminaries

Our proof of Theorem 1 follows the approach of Hirano and Porter (2009), with appropriate modifications to handle non-binary problems. By an asymptotic representation theorem of van der Vaart (1991), Assumption 2 implies that for any $\{\delta_n\} \in \mathbb{D}$ there exists a function $\delta_\infty(z, u; P_0)$ with $\delta_\infty(Z, U; P_0) \sim Q_{P_0, h}$ where $Z \sim N(h, I_0^{-1})$ and $U \sim \text{Uniform}[0, 1]$ with $U$ and $Z$ independent. To summarize the argument, suppose that

$$\lim_{n \to \infty} \sqrt{n} \left( \mathbb{E}_{P_{n,h}} \left[ R(\delta_n(X^n), P_{n,h}) \right] - \min_{d \in \mathcal{D}} R(d, P_{n,h}) \right) = L_\infty(\delta_\infty; P_0, h) \qquad \text{(A.1)}$$

for a suitable function $L_\infty$. We may therefore rewrite the criterion (10) as

$$\mathcal{R}(\{\delta_n\}; P_0) = \int L_\infty(\delta_\infty; P_0, h) \, dh \,.$$

We show that the asymptotic representation $\delta_\infty^*$ of $\{\delta_n^*\}$ is optimal in the limit experiment: for each $P_0 \in \mathcal{P}$,

$$\int L_\infty(\delta_\infty^*; P_0, h) \, dh = \inf_{d^\infty} \int L_\infty(\delta_\infty; P_0, h) \, dh,$$

where the infimum is over all such (possibly randomized) $\mathcal{D}$-valued decisions $\delta_\infty(Z, U; P_0)$ in the limit experiment. It follows that any $\{\delta_n\} \in \mathbb{D}$ which is asymptotically equivalent to $\delta_n^*$ must be asymptotically optimal.

Deriving the asymptotic representation of $\{\delta_n\} \in \mathbb{D}$ requires a tie-breaking rule when the argmin is not a singleton. In what follows, we take the smallest index $d$ among the set of minimizers, though our optimality result does not rely on this and any (possibly randomized) tie-breaking rule is optimal. Because we use a non-randomized rule, asymptotic representations are of the form $\delta_\infty(Z; P_0)$. Randomized rules can easily be handled by letting $\delta_\infty$ also depend on $U$, though this is not necessary as the asymptotic average maximum risk is unaffected. Recall that $\mathbb{P}_h$ denotes probability with respect to $Z \sim N(h, I_0^{-1})$ and $\mathbb{E}^*$ denote expectation with respect to $Z^*$ where $Z^* \sim N(0, I_0^{-1})$ independently of $Z$. Recall $\rho(P) = (R(0, P), R(1, P), \ldots, R(D, P))'$. Let $\rho_d(P) = R(d, P)$ and let $\dot{\rho}_{d, P_0}[\cdot]$ denote its directional derivative at $P_0$.

**Lemma 1** *Suppose that Assumptions 1, 2, and 3 hold. Fix any $P_0 \in \mathcal{P}$. Without loss of generality, reorder the elements of $\mathcal{D}$ so that $R(0, P_0) \leq R(1, P_0) \leq \ldots \leq R(D, P_0)$. Let $k \in \mathcal{D}$ satisfy $R(0, P_0) = \ldots = R(k, P_0)$ with $R(k, P_0) < R(k+1, P_0)$ if $k < D$. Then:*

A.4

*(i) The sequence $\{\delta_n^*\}$ belongs to $\mathbb{D}$, and its asymptotic representation at $P_0$ is*

$$\delta_\infty^*(z; P_0) = \sum_{i=0}^{K} i \times \mathbb{I}\left[\mathbb{E}^*\left[\dot{\rho}_{i,P_0}[Z^* + z]\right] < \min_{j<i} \mathbb{E}^*\left[\dot{\rho}_{j,P_0}[Z^* + z]\right]\right]$$
$$\times \left[\mathbb{E}^*\left[\dot{\rho}_{i,P_0}[Z^* + z]\right] \le \min_{i<j\le k} \mathbb{E}^*\left[\dot{\rho}_{j,P_0}[Z^* + z]\right]\right],$$

*where the minimum over an empty index is $+\infty$;*

*(ii) For any $\{\delta_n\} \in \mathbb{D}$,*

$$L_\infty(\delta_\infty; P_0, h)$$
$$= \sum_{i=0}^{K}(\dot{\rho}_{i,P_0}[h] - \min_{j\le k}\dot{\rho}_{j,P_0}[h]) \times \mathbb{P}_h(\delta_\infty(Z; P_0) = i) + \infty \times \mathbb{P}_h(\delta_\infty(Z; P_0) > k),$$

*where $\delta_\infty(Z; P_0)$ is the asymptotic representation of $\{\delta_n\}$ at $P_0$;*

*(iii) The decision $\delta_\infty^*$ is optimal in the limit experiment.*

**Proof of Lemma 1.** <u>Part (i).</u> As $\delta_n$ is takes values in a finite set $\mathcal{D}$, establishing convergence in distribution under $\{F_{n,P_{n,h}}\}$ is equivalent to showing $\lim_{n\to\infty} F_{n,P_{n,h}}(\delta_n(X^n) = i)$ exists for each $0 \le i \le D$. Suppose $k < D$. Then by Assumption 1(i), choose $\varepsilon > 0$ and a neighborhood $N$ of $P_0$ upon which $\max_{i\le k} R(i, P) \le \min_{j\ge k} R(j, P) - \varepsilon$. By Assumption 3(i),

$$\max_{i\le k} \bar{R}(i|X^n) - \min_{j>k} \bar{R}(j|X^n) \le -\varepsilon\pi_n(P \in N) + 2\|\rho\|_\infty \pi_n(P \notin N) \overset{P_0}{\to} -\varepsilon,$$

where $\|\rho\|_\infty := \max_{d\in\mathcal{D}} \sup_{P\in\mathcal{P}} |R(d, P)| < \infty$ by Assumption 1(i). The convergence in probability in the above display holds along $P_{n,h}$ for any $h$ by by Le Cam's first lemma and Assumption 2. Therefore,

$$\mathbb{I}[\delta_n^*(X^n) > k] \overset{P_{n,h}}{\to} 0.$$

It follows that the asymptotic distribution $Q_{P_0,h}^*$ of $\delta_n^*(X^n)$ along $P_{n,h}$ assigns zero probability to $\{k+1, \ldots, D\}$ for all $h \in \mathbb{R}^K$. We may therefore take $\delta_\infty^*(z; P_0) \in \{0, 1, \ldots, k\}$ for all $z \in \mathbb{R}^K$. In particular, $\delta_\infty^*(z; P_0) \equiv 0$ when $k = 0$. We therefore consider only $k \ge 1$ in what follows.

As $\max_{i\le k} \bar{R}(i|X^n) < \min_{j>k} \bar{R}(j|X^n)$ wpa1 along $P_{n,h}$, in what follows it is without loss of generality to work on the sequence of events with $\max_{i\le k} \bar{R}(i|X^n) < \min_{j>k} \bar{R}(j|X^n)$. Under the above tie-breaking rule, for any $i \le k$ we have

$$\mathbb{I}[\delta_n^*(X^n) = i] = \mathbb{I}\left[\bar{R}(i|X^n) < \min_{j<i} \bar{R}(j|X^n) \text{ and } \bar{R}(i|X^n) \le \min_{i<j\le k} \bar{R}(j|X^n)\right],$$

A.5

where the minimum over an empty set is $+\infty$. As $R(0, P_0) = \ldots = R(k, P_0)$, we may rewrite the previous expression in terms of $\rho_{i0}(P) = R(i, P) - R(i, P_0)$:

$$\mathbb{I}[\delta_n^*(X^n) = i] = \mathbb{I}\left[\int \sqrt{n}\rho_{i0}(P)\,\mathrm{d}\pi_n(P) < \min_{j<i}\int \sqrt{n}\rho_{j0}(P)\,\mathrm{d}\pi_n(P)\right]$$
$$\times \mathbb{I}\left[\int \sqrt{n}\rho_{i0}(P)\,\mathrm{d}\pi_n(P) \leq \min_{i<j\leq k}\int \sqrt{n}\rho_{j0}(P)\,\mathrm{d}\pi_n(P)\right].$$

By Assumption 3(ii) with $A = \{(x_0, x_1, \ldots, x_D) : x_i < \min_{j<i} x_j \text{ and } x_i \leq \min_{i<j\leq k} x_j\}$,

$$\lim_{n\to\infty} F_{n,P_{n,h}}(\delta_n^*(X^n) = i) = \mathbb{P}_h\Big(\Big(\mathbb{E}^*[\dot\rho_{i,P_0}[Z^* + Z]|Z] < \min_{j<i}\mathbb{E}^*[\dot\rho_{j,P_0}[Z^* + Z]|Z]\Big)$$
$$\text{and } \Big(\mathbb{E}^*[\dot\rho_{i,P_0}[Z^* + Z]|Z] \leq \min_{i<j\leq k}\mathbb{E}^*[\dot\rho_{j,P_0}[Z^* + Z]|Z]\Big)\Big).$$

Hence, $\{\delta_n^*\} \in \mathbb{D}$. The asymptotic representation of $\{\delta_n^*\}$ follows from the preceding display.

Part (ii): First write

$$R(d, P_{n,h}) - \min_{d\in\mathcal{D}} R(d, P_{n,h}) = \sum_{i=0}^{D} \mathbb{I}[d = i]\Big(R(i, P_{n,h}) - \min_{d\in\mathcal{D}} R(d, P_{n,h})\Big).$$

By Assumption 1(i), $\min_{d\in\mathcal{D}} R(d, P_{n,h}) = \min_{j\leq k} R(j, P_{n,h})$ for all $n$ sufficiently large, and so

$$L_n(\delta_n; P_0, h) = \sum_{i=0}^{K} \mathbb{E}_{P_{n,h}}\big[\mathbb{I}[\delta_n(X^n) = i]\big]\sqrt{n}\Big(R(i, P_{n,h}) - \min_{j\leq k} R(j, P_{n,h})\Big)$$
$$+ \sum_{i=k+1}^{D} \mathbb{E}_{P_{n,h}}\big[\mathbb{I}[\delta_n(X^n) = i]\big]\sqrt{n}\Big(R(i, P_{n,h}) - \min_{j\leq k} R(j, P_{n,h})\Big),$$

for all $n$ sufficiently large, where the second sum is zero if $k = D$. As $\mathbb{E}_{P_{n,h}}\big[\mathbb{I}[\delta_n(X^n) = i]\big] \to \mathbb{P}_h[\delta_\infty(Z; P_0) = i]$ for all $i \in \mathcal{D}$ and $\liminf_{n\to\infty}(R(i, P_{n,h}) - \min_{j\leq k} R(j, P_{n,h})) > 0$ for $i > k$, we have

$$\sum_{i=k+1}^{D} \mathbb{E}_{P_{n,h}}\big[\mathbb{I}[\delta_n(X^n) = i]\big]\sqrt{n}\Big(R(i, P_{n,h}) - \min_{j\leq k} R(j, P_{n,h})\Big) \to +\infty \times \mathbb{P}_h(\delta_\infty(Z; P_0) > k)$$

whenever $k < D$. Moreover, as $R(0, P_0) = \ldots = R(k, P_0)$, for any $i \leq k$ we have

$$\sqrt{n}\left(R(i, P_{n,h}) - \min_{j\leq k} R(j, P_{n,h})\right) = \left(\sqrt{n}\rho_{i0}(P_{n,h}) - \min_{j\leq k}\sqrt{n}\rho_{j0}(P_{n,h})\right)$$
$$\to \dot\rho_{i,P_0}[h] - \min_{j\leq k}\dot\rho_{j,P_0}[h],$$

A.6

where the final assertion is by Assumption 1(ii) and the fact that if $f_i$, $i = 1, 2$ are directionally differentiable at $P_0$ and $f_1(P_0) = f_2(P_0)$, then $f(P) := \min_{i=1,2} f_i(P)$ has directional derivative $\dot{f}_{P_0}[h] = \min_{i=1,2} \dot{f}_{i,P_0}[h]$.

Part (iii): By part (ii), we see that $\delta_\infty(z; P_0) \in \{0, 1, \ldots, k\}$ for almost every $z$ is necessary for optimality in the limit experiment. For any such $\delta_\infty$, we have by part (ii) that

$$\int L_\infty(\delta_\infty; P_0, h) \, dh = \sum_{i=0}^{K} \int \mathbb{P}_h[\delta_\infty(Z; P_0) = i] \left( \dot{\rho}_{i,P_0}[h] - \min_{j \leq k} \dot{\rho}_{j,P_0}[h] \right) dh$$

$$\propto \sum_{i=0}^{K} \int \int \mathbb{I}[\delta_\infty(z; P_0) = i] \left( \dot{\rho}_{i,P_0}[h] - \min_{j \leq k} \dot{\rho}_{j,P_0}[h] \right) e^{-\frac{1}{2}(z-h)'I_0(z-h)} \, dz \, dh$$

$$= \sum_{i=0}^{K} \int \int \mathbb{I}[\delta_\infty(z; P_0) = i] \left( \dot{\rho}_{i,P_0}[h] - \min_{j \leq k} \dot{\rho}_{j,P_0}[h] \right) e^{-\frac{1}{2}(z-h)'I_0(z-h)} \, dh \, dz.$$

As the integrand is non-negative, changing the order of integration in the final line is justified by Tonelli's theorem which only requires non-negativity. Minimizing pointwise in $z$, we see that if

$$\mathcal{I}(z) = \arg\min_{i \leq k} \int \dot{\rho}_{i,P_0}[h] e^{-\frac{1}{2}(z-h)'I_0(z-h)} \, dh \equiv \arg\min_{i \leq k} \mathbb{E}^*\left[ \dot{\rho}_{i,P_0}[Z + z] \right], \qquad \text{(A.2)}$$

then choosing $\delta_\infty(z; P_0)$ to be any element of $\mathcal{I}(z)$ for each $z$ is optimal in the limit experiment. The tie-breaking rule used in part (i) is a special case with $\delta_\infty^*(z; P_0) = \min \mathcal{I}(z)$. ∎

### B.1.2 Proof of Main Results

**Proof of Theorem 1.** Part (i) is immediate from parts (i) and (iii) of Lemma 1, which together imply

$$\mathcal{R}(\{\delta_n^*\}, P_0) = \int L_\infty(\delta_\infty^*; P_0, h) \, dh = \inf_{\delta_\infty} \int L_\infty(\delta_\infty; P_0, h) \, dh = \inf_{\{\delta_n'\} \in \mathbb{D}} \mathcal{R}(\{\delta_n'\}; P_0). \quad \text{(A.3)}$$

Part (ii) follows from the fact that the asymptotic representation of $\delta_n^*$ derived in part (i) of Lemma 1 does not depend on the prior $\pi \in \Pi$. For part (iii), note if $\{\delta_n\}$ is asymptotically equivalent to $\{\delta_n^*\}$, then $\{\delta_n\}$ must also have asymptotic representation $\delta_\infty^*$. The result now follows from (A.3). ∎

**Proof of Theorem 2.** By Lemma 1(ii) and the proof of Theorem 1, it suffices to show that

$$\int L_\infty(\delta_\infty; P_0, h) \, dh > \int L_\infty(\delta_\infty^*; P_0, h) \, dh, \qquad \text{(A.4)}$$

where $\delta_\infty$ is the asymptotic representation of $\{\delta_n\}$.

First suppose that $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ is a singleton. Without loss of generality let $d = 0$ be the minimizing value. In view of Lemma 1(i), we must have

$$\lim_{n \to \infty} F_{n, P_{n, h_0}}(\delta_n(X^n) = 0) < \lim_{n \to \infty} F_{n, P_{n, h_0}}(\delta_n^*(X^n) = 0) = 1$$

for some $h_0$ in $\mathbb{R}^K$. In terms of the asymptotic representation $\delta_\infty$ of $\{\delta_n\}$, this means

$$\mathbb{P}_{h_0}(\delta_\infty(Z; P_0) = 0) < 1.$$

By continuity of $h \mapsto \mathbb{P}_h(\delta_\infty(Z; P_0) = 0)$, there exists a neighborhood $H \subset \mathbb{R}^K$ of $h_0$ upon which $\mathbb{P}_h(\delta_\infty(Z) = 0) < 1$. In view of Lemma 1(ii), this implies $L_\infty(\delta_\infty; P_0, h) = +\infty$ for all $h \in H$ whereas $L_\infty(\delta_\infty^*; P_0, h) = 0$ for all $h \in \mathbb{R}^K$. Therefore, inequality (A.4) holds.

Now suppose that $\arg\min_{d \in \mathcal{D}} R(d, P_0)$ is not a singleton. As in Lemma 1, it is without loss of generality to reorder elements of $\mathcal{D}$ so that $R(0, P_0) \leq R(1, P_0) \leq \ldots \leq R(D, P_0)$. Let $k \in \{1, \ldots, D\}$ satisfy $R(0, P_0) = \ldots = R(k, P_0)$ with $R(k, P_0) < R(k+1, P_0)$ if $k < D$. If the asymptotic representation $\delta_\infty$ of $\{\delta_n\}$ satisfies

$$\mathbb{P}_{h_0}(\delta_\infty(Z; P_0) \leq k) < 1$$

for any $h_0 \in \mathbb{R}^K$, then by similar arguments to the above we have $L_\infty(\delta_\infty; P_0, h) = +\infty$ for all $h$ in a neighborhood $H$ of $h_0$ and so inequality (A.4) holds. For the remainder of the proof we therefore suppose that

$$\mathbb{P}_h(\delta_\infty(Z; P_0) \leq k) = 1$$

for all $h \in \mathbb{R}^K$. By Lemma 1(ii) it suffices to show

$$\sum_{i=0}^{K} \int (\dot{\rho}_{i, P_0}[h] - \min_{j \leq k} \dot{\rho}_{j, P_0}[h]) \times \mathbb{P}_h(\delta_\infty(Z; P_0) = i)\, \mathrm{d}h$$

$$> \sum_{i=0}^{K} \int (\dot{\rho}_{i, P_0}[h] - \min_{j \leq k} \dot{\rho}_{j, P_0}[h]) \times \mathbb{P}_h(\delta_\infty^*(Z; P_0) = i)\, \mathrm{d}h.$$

Since both integrands are non-negative, by Tonelli's theorem we restate this inequality as

$$\sum_{i=0}^{K} \int \mathbb{I}[\delta_\infty(z; P_0) = i] \left( \int (\dot{\rho}_{i, P_0}[h] - \min_{j \leq k} \dot{\rho}_{j, P_0}[h]) e^{-\frac{1}{2}(z-h)' I_0(z-h)}\, \mathrm{d}h \right) \mathrm{d}z$$

$$> \sum_{i=0}^{K} \int \mathbb{I}[\delta_\infty^*(z; P_0) = i] \left( \int (\dot{\rho}_{i, P_0}[h] - \min_{j \leq k} \dot{\rho}_{j, P_0}[h]) e^{-\frac{1}{2}(z-h)' I_0(z-h)}\, \mathrm{d}h \right) \mathrm{d}z.$$

Recall the set $\mathcal{I}(z)$ from (A.2). The function $\delta_\infty^*(z; P_0)$ takes values in $\mathcal{I}(z)$ for each $z$, so the

A.8

preceding inequality must hold weakly. To establish a strict inequality, note $\delta_\infty(z; P_0)$ and $\delta_\infty^*(z; P_0)$ must disagree on a set of positive Lebesgue measure, say $\mathcal{Z}$ (otherwise $\delta_\infty(Z; P_0)$ and $\delta_\infty^*(Z; P_0)$ would have the same distribution for $Z \sim N(h_0, I_0^{-1})$). For each $z \in \mathcal{Z}$, one of the following must hold:

(a) $\mathcal{I}(z)$ is a singleton and $\delta_\infty(z, I) \notin \mathcal{I}(z)$, or

(b) $\mathcal{I}(z)$ is not a singleton and $\delta_\infty(z, P_0) \neq \delta_\infty^*(z; P_0)$.

Condition (ii) in the statement of the theorem implies that $\mathcal{I}(z)$ is non-singleton on a set of zero Lebesgue measure. Therefore, (a) must hold at almost every $z \in \mathcal{Z}$, and for any such $z$ we have

$$\sum_{i=0}^{K} \mathbb{I}[\delta_\infty(z; P_0) = i] \left( \int (\dot\rho_{i,P_0}[h] - \min_{j \leq k} \dot\rho_{j,P_0}[h]) e^{-\frac{1}{2}(z-h)' I_0(z-h)} \, \mathrm{d}h \right)$$

$$> \sum_{i=0}^{K} \mathbb{I}[\delta_\infty^*(z; P_0) = i] \left( \int (\dot\rho_{i,P_0}[h] - \min_{j \leq k} \dot\rho_{j,P_0}[h]) e^{-\frac{1}{2}(z-h)' I_0(z-h)} \, \mathrm{d}h \right),$$

proving (A.4). ∎

Finally, we prove Corollary 1. For this we require $\{\delta_n\}$ to satisfy a condition permitting the exchange of limits and integration, namely

$$\limsup_{n \to \infty} \mathcal{R}_n(\delta_n; P_0) \leq \int \limsup_{n \to \infty} \sqrt{n} \left( \mathbb{E}_{P_{n,h}} [R(\delta_n(X^n), P_{n,h})] - \min_{d \in \mathcal{D}} R(d, P_{n,h}) \right) w(P_{n,h}) \, \mathrm{d}h \tag{A.5}$$

for all $P_0 \in \mathcal{P}$. By the reverse Fatou lemma, condition (A.5) holds if for each $P_0 \in \mathcal{P}$ there exists an $h$-integrable function $g(h; P_0)$ with

$$\sqrt{n} \left( \mathbb{E}_{P_{n,h}} [R(\delta_n(X^n), P_{n,h})] - \min_{d \in \mathcal{D}} R(d, P_{n,h}) \right) w(P_{n,h}) \leq g(h; P_0)$$

for all $n$.

**Proof of Corollary 1.** The first part of our proof is similar to Lemma 1 of Hirano and Porter (2009). By Fatou's lemma and Lemma 1(ii), for any $\{\delta_n'\} \in \mathbb{D}$ we have

$$\liminf_{n \to \infty} \mathcal{R}_n(\delta_n'; P_0) \geq w(P_0) \cdot \int L_\infty(\delta_\infty'; P_0, h) \, \mathrm{d}h,$$

where $\delta_\infty'$ is the asymptotic representation of $\{\delta_n'\}$. In particular this holds for $\{\delta_n\}$. In view of inequality (A.5) and Lemma 1(ii), we have

$$\liminf_{n \to \infty} \mathcal{R}_n(\delta_n; P_0) = w(P_0) \cdot \int L_\infty(\delta_\infty; P_0, h) \, \mathrm{d}h = w(P_0) \cdot \inf_{\delta_\infty} \int L_\infty(\delta_\infty; P_0, h) \, \mathrm{d}h,$$

A.9

where the final equality is by asymptotic equivalence of $\{\delta_n\}$ and $\{\delta_n^*\}$ and Lemma 1(iii). Therefore,

$$\lim_{n\to\infty} \mathcal{R}_n(\delta_n; P_0) = \lim_{n\to\infty} \mathcal{R}_n(\delta_n^*; P_0) = \inf_{\{\delta_n'\}\in\mathbb{D}} \liminf_{n\to\infty} \mathcal{R}_n(\delta_n'; P_0).$$

Finally, note that

$$\inf_{\{\delta_n'\}\in\mathbb{D}} \liminf_{n\to\infty} \mathcal{R}_n(\delta_n'; P_0) = \liminf_{n\to\infty} \inf_{\delta_n'} \mathcal{R}_n(\delta_n'; P_0).$$

because $\mathcal{R}_n(\delta_n^*; P_0) = \inf_{\delta_n'} \mathcal{R}_n(\delta_n'; P_0)$ for each $n$. $\blacksquare$

### B.2  Proof of Theorems 3 and 4

Note that once we fix $(P_0, \eta_0)$, the approximately least-favorable model $t \mapsto F_{n,\beta_{(P_0,\eta_0)}}(t)$ is a parametric model. In particular, under Assumption 4 any sequence $\{\delta_n\} \in \mathbb{D}$ has an asymptotic representation in the limit experiment associated with $\{F_{n,\beta_{(P_0,\eta_0)}}(P_{n,h}) : h \in \mathbb{R}^K\}$, namely $\delta_\infty(Z, U; (P_0, \eta_0)) \sim Q_{(P_0,\eta_0),h}$ for $Z \sim N(h, I_0^{-1})$ with $I_0 = I_0(P_0, \eta_0) > 0$ denoting the semiparametric information bound and $U \sim \text{Uniform}[0, 1]$ with $U$ and $Z$ independent.[9] Once these modifications are made, the proofs of Theorems 3 and 4 follow identical arguments to the proofs of Theorems 1 and 2.

## C  Derivations for the Example from Section 4.2

By the properties of the truncated Normal distribution, we have

$$\int (b_U(P))_+ \, \mathrm{d}\pi(P|X^n) = \int (P_3)_+ \, \mathrm{d}\pi(P|X^n) = \hat{P}_3 \Phi(\sqrt{n}\hat{P}_3) + n^{-1/2} \phi(\sqrt{n}\hat{P}_3).$$

For the lower bound, note that the posterior distribution of $P_1 \vee P_2$ has PDF

$$f(x) = \sqrt{n}\phi(\sqrt{n}(x - \hat{P}_1))\Phi(\sqrt{n}(x - \hat{P}_2)) + \sqrt{n}\phi(\sqrt{n}(x - \hat{P}_2))\Phi(\sqrt{n}(x - \hat{P}_1));$$

see, e.g., Nadarajah and Kotz (2008). Hence, the posterior mean of $(P_1 \vee P_2)_-$ is

$$\int_{-\infty}^0 \sqrt{n}(x - \hat{P}_1)\phi(\sqrt{n}(x - \hat{P}_1))\Phi(\sqrt{n}(x - \hat{P}_2)) \, \mathrm{d}x$$
$$+ \int_{-\infty}^0 \sqrt{n}(x - \hat{P}_2)\phi(\sqrt{n}(x - \hat{P}_2))\Phi(\sqrt{n}(x - \hat{P}_1)) \, \mathrm{d}x$$
$$+ \sqrt{n}\hat{P}_1 \int_{-\infty}^0 \phi(\sqrt{n}(x - \hat{P}_1))\Phi(\sqrt{n}(x - \hat{P}_2)) \, \mathrm{d}x$$
$$+ \sqrt{n}\hat{P}_2 \int_{-\infty}^0 \phi(\sqrt{n}(x - \hat{P}_2))\Phi(\sqrt{n}(x - \hat{P}_1)) \, \mathrm{d}x =: T_1 + T_2 + T_3 + T_4.$$

---

[9]As before, it suffices to consider asymptotic representations depending only on $Z$.

By integration by parts,

$$T_1 = -n^{-1/2}\phi(\sqrt{n}\hat{P}_1)\Phi(-\sqrt{n}\hat{P}_2) + n^{-1/2}\int_{-\infty}^{-\sqrt{n}\hat{P}_1}\phi(x)\phi(x+\sqrt{n}(\hat{P}_1-\hat{P}_2))\,\mathrm{d}x$$

$$= -n^{-1/2}\phi(\sqrt{n}\hat{P}_1)\Phi(-\sqrt{n}\hat{P}_2) + n^{-1/2}\frac{e^{-\frac{1}{4}n(\hat{P}_1-\hat{P}_2)^2}}{2\pi}\int_{-\infty}^{-\sqrt{n}\hat{P}_1}e^{-(x+\frac{1}{2}\sqrt{n}(\hat{P}_1-\hat{P}_2))^2}\,\mathrm{d}x$$

$$= -n^{-1/2}\phi(\sqrt{n}\hat{P}_1)\Phi(-\sqrt{n}\hat{P}_2) + \frac{1}{\sqrt{2n}}\phi(\sqrt{n}(\hat{P}_1-\hat{P}_2)/\sqrt{2})\Phi(-\sqrt{n}(\hat{P}_1+\hat{P}_2)/\sqrt{2}).$$

So by symmetry,

$$T_1 + T_2 = -n^{-1/2}\phi(\sqrt{n}\hat{P}_1)\Phi(-\sqrt{n}\hat{P}_2) - n^{-1/2}\phi(\sqrt{n}\hat{P}_2)\Phi(-\sqrt{n}\hat{P}_1)$$

$$+ \frac{\sqrt{2}}{\sqrt{n}}\phi(\sqrt{n}(\hat{P}_1-\hat{P}_2)/\sqrt{2})\Phi(-\sqrt{n}(\hat{P}_1+\hat{P}_2)/\sqrt{2}).$$

We also have that

$$T_3 = \hat{P}_1\int_{-\infty}^{\infty}\mathbb{I}[x \leq -\sqrt{n}\hat{P}_1]\Pr(Y \leq \sqrt{n}(\hat{P}_1-\hat{P}_2)+x)\phi(x)\,\mathrm{d}x$$

$$= \hat{P}_1\Pr(Y - \sqrt{n}(\hat{P}_1-\hat{P}_2) \leq X \leq -\sqrt{n}\hat{P}_1),$$

where $Y, X$ are independent $N(0,1)$ random variables and $\Pr$ denotes probabilities with respect to their joint distribution (conditional on $\hat{P}_1$ and $\hat{P}_2$). By symmetry

$$T_4 = \hat{P}_2\Pr(Y - \sqrt{n}(\hat{P}_2-\hat{P}_1) \leq X \leq -\sqrt{n}\hat{P}_2).$$

Hence the optimal decision is

$$\delta_n^*(X^n;\pi) = \mathbb{I}\left[T_1 + T_2 + T_3 + T_4 + \hat{P}_3\Phi(\sqrt{n}\hat{P}_3) + n^{-1/2}\phi(\sqrt{n}\hat{P}_3) \geq 0\right].$$

To derive the asymptotic behavior of the optimal decision, note that along $\{P_{n,h}\}$ with $P_0 = (-c, -c, c)$, we have $\sqrt{n}\hat{P}_1 = -\sqrt{n}c + \sqrt{n}(\hat{P}_1+c)$ where $\sqrt{n}(\hat{P}_1+c)$ is asymptotically tight, and similarly for $\sqrt{n}\hat{P}_2$. Hence by exponential decay of the Gaussian PDF and CDF, we obtain

$$\left|\sqrt{n}(T_1+T_2) - \sqrt{2}\phi(\sqrt{n}(\hat{P}_1-\hat{P}_2)/\sqrt{2})\right| \overset{P_{n,h}}{\to} 0. \tag{A.6}$$

We also have that

$$\sqrt{n}\left(\hat{P}_3\Phi(\sqrt{n}\hat{P}_3) + n^{-1/2}\phi(\sqrt{n}\hat{P}_3)\right) = \sqrt{n}\hat{P}_3 + \sqrt{n}\hat{P}_3\Phi(-\sqrt{n}\hat{P}_3) + \phi(\sqrt{n}\hat{P}_3).$$

Hence, along $\{P_{n,h}\}$ with $P_0 = (-c, -c, c)$, we have $\sqrt{n}\hat{P}_3 = \sqrt{n}c + \sqrt{n}(\hat{P}_3 - c)$ where

A.11

$\sqrt{n}(\hat{P}_3 - c)$ is asymptotically tight. By exponential decay of the Gaussian PDF and CDF, we obtain

$$\sqrt{n}\hat{P}_3\Phi(-\sqrt{n}\hat{P}_3) + \phi(\sqrt{n}\hat{P}_3) \overset{P_{n,h}}{\to} 0.$$

We also have

$$\left|\sqrt{n}(T_3 - \hat{P}_1\Pr(Y - \sqrt{n}(\hat{P}_1 - \hat{P}_2) \leq X)\right| \leq \sqrt{n}\hat{P}_1\Pr(X > -\sqrt{n}\hat{P}_1)$$

$$= \sqrt{n}\hat{P}_1\Phi(\sqrt{n}\hat{P}_1) \overset{P_{n,h}}{\to} 0,$$

because $\sqrt{n}\hat{P}_1 = -\sqrt{n}c + \sqrt{n}(\hat{P}_1 + c)$, where $\sqrt{n}(\hat{P}_1 + c)$ is asymptotically tight. Moreover, because $X$ and $Y$ are independent $N(0,1)$ random vectors, their difference $Y - X$ is $N(0,2)$. Hence,

$$\Pr(Y - \sqrt{n}(\hat{P}_1 - \hat{P}_2) \leq X) = \Pr(Y - X \leq \sqrt{n}(\hat{P}_1 - \hat{P}_2)) = \Phi(\sqrt{n}(\hat{P}_1 - \hat{P}_2)/\sqrt{2}).$$

To relate $T_3$ and $T_4$, also note that

$$\Pr(Y - \sqrt{n}(\hat{P}_2 - \hat{P}_1) \leq X) = 1 - \Pr(Y - \sqrt{n}(\hat{P}_1 - \hat{P}_2) \leq X) = \Phi(\sqrt{n}(\hat{P}_2 - \hat{P}_1)/\sqrt{2}).$$

Hence,

$$\left|\sqrt{n}\left(T_3 + T_4 + \hat{P}_3\Phi(\sqrt{n}\hat{P}_3) + n^{-1/2}\phi(\sqrt{n}\hat{P}_3)\right) - \sqrt{n}(\hat{P}_1 + c)\Phi(\sqrt{n}(\hat{P}_1 - \hat{P}_2)/\sqrt{2})\right.$$
$$\left. - \sqrt{n}(\hat{P}_2 + c)\Phi(\sqrt{n}(\hat{P}_2 - \hat{P}_1)/\sqrt{2}) - \sqrt{n}(\hat{P}_3 - c)\right| \overset{P_{n,h}}{\to} 0. \tag{A.7}$$

Combining (A.6) and (A.7), we see that along $\{P_{n,h}\}$, the optimal decision has the asymptotic representation

$$\delta_\infty^*(Z; P_0) = \mathbb{I}\left[Z_1\Phi((Z_1 - Z_2)/\sqrt{2}) + Z_2\Phi((Z_2 - Z_1)/\sqrt{2}) + Z_3 \geq -2\phi((Z_1 - Z_2)/\sqrt{2})\right],$$

where $Z \sim N(h, I)$. Finally, by the form of the expectation of the maximum of two Gaussian random vectors (see, e.g., Nadarajah and Kotz (2008)), we can write this as

$$\delta_\infty^*(Z; P_0) = \mathbb{I}\left[\mathbb{E}^*[(Z_1^* + Z_1) \vee (Z_2^* + Z_2)] + Z_3 \geq 0\right],$$

where $\mathbb{E}^*$ denotes expectation with respect to $Z^* \sim N(0, I)$ conditional on $Z$.