

Asymptotics for Statistical Treatment Rules

Keisuke Hirano
Department of Economics
University of Arizona
hirano@u.arizona.edu

Jack R. Porter
Department of Economics
University of Wisconsin
jrporter@ssc.wisc.edu

August 10, 2008

1 Introduction

One major goal of treatment evaluation in the social and medical sciences is to provide guidance on how to assign individuals to treatments. For example, a number of studies have examined the problem of “profiling” individuals to identify those likely to benefit from a social program.¹ Manski (2000, 2002, 2004) and Dehejia (2005) suggest placing the problem within a decision-theoretic framework, and specifying a loss function that quantifies the consequences of choosing different treatments under different states of nature. Schlag (2006) and Stoye (2006) derive exact minmax-regret rules for randomized experiments with a discrete covariate and a bounded continuous outcome. Despite these important results, it is difficult to obtain exact optimality results in many empirically relevant settings, in the same way that it is difficult to obtain exactly optimal estimators or hypothesis tests.

In this paper, we develop large sample results to compare treatment rules, and show how to construct approximately optimal procedures from efficient estimates of treatment effect parameters. The data could come from a randomized experiment or an observational data source, and we allow for unrestricted outcome and covariate distributions (including continuously distributed covariates). The key requirements are a local asymptotic normality condition and that a welfare contrast parameter be point-identified. When social welfare contrasts are point-identified, there will typically exist many treatment rules that are consistent, in the sense that they assign the “better” treatment with probability approaching one. Our goal here is to make finer comparisons among rules, and to base these comparisons on risk rather than conventional statistical criteria that are not tightly connected to the underlying decision problem.

We first study regular parametric models, using a local parametrization so that the problem of determining whether to assign the treatment does not become trivial as the sample size increases. Using Le Cam’s limits of experiments framework (see Le Cam (1986)) we show that the treatment

¹See for example Worden (1993), O’Leary, Decker, and Wandner (1998), Berger, Black, and Smith (2001), Black, Smith, Berger, and Noel (2003), and O’Leary, Decker, and Wandner (2005)

assignment problem is asymptotically equivalent to a simpler problem, in which one observes a single draw from a shifted Gaussian distribution, and must decide whether a linear function of the mean vector is greater than zero. We solve the approximate version of the decision problem, and then construct a sequence of decision rules in the original problem that asymptotically matches the solution. The forms of optimal rules will depend on the loss function, and the way in which risk (expected loss) is aggregated over the parameter space—for example we could work with average (Bayes) risk, or minmax risk. We consider some specific loss functions, and show that some simple rules based on efficient parameter estimates are asymptotically optimal under average and minmax risk criteria.

We then extend the results to a semiparametric setting, where the welfare gain of the treatment can be expressed as a regular functional of the unknown distribution. The analysis in this setting mirrors the parametric setting, but involves a Gaussian sequence model instead of a finite-dimensional Gaussian model. We obtain both minmax and average risk optimality results; to define average risk in the semiparametric case, we propose a class of prior weightings on the tangent space. We illustrate our results by showing that Manski’s conditional empirical success rules are asymptotically average and minmax risk optimal under certain symmetric loss functions.

2 Statistical Treatment Assignment Problem

Following Manski (2000, 2002, 2004), we consider a social planner, who assigns individuals to different treatments based on their observed background variables. Suppose that a randomly drawn individual has covariates X with probability distribution F_X on a space \mathcal{X} . The set of possible treatment values is $\mathcal{T} = \{0, 1\}$. The planner observes $X = x$, and assigns the individual to treatment 1 according to a treatment rule

$$\delta(x) = Pr(T = 1|X = x).$$

Let Y_0 and Y_1 denote potential outcomes for the individual, with conditional probability distributions $F_0(\cdot|x)$ and $F_1(\cdot|x)$ on the same space \mathcal{Y} . Given a rule δ , the outcome distribution conditional on $X = x$ is

$$F_\delta(\cdot|x) = \delta(x)F_1(\cdot|x) + (1 - \delta(x))F_0(\cdot|x).$$

For a given outcome distribution F , let the *social welfare* be a functional $W(F)$. We define

$$W_0(x) = W(F_0(\cdot|x)), \quad W_1(x) = W(F_1(\cdot|x)).$$

One optimal rule is then $\delta^*(x) = 1(W_1(x) > W_0(x))$. Of course, this rule is generally infeasible since F_0 and F_1 (and hence W_0 and W_1) are unknown.

We suppose that F_0 and F_1 belong to families of distributions indexed by a parameter $\theta \in \Theta$, where the parameter space could be finite-dimensional or infinite-dimensional. Let $w_0(\theta, x)$ and

$w_1(\theta, x)$ denote the values for $W_0(x)$ and $W_1(x)$ under θ . It will be convenient to work with the welfare contrast

$$g(\theta, x) = w_1(\theta, x) - w_0(\theta, x).$$

We assume that w_0 and g are continuously differentiable in θ for F_X -almost all x .²

Suppose we have some data that are informative about θ , such as data from a randomized experiment or an observational study. For simplicity, we assume that the data $Z^n = (Z_1, \dots, Z_n)$ is i.i.d. with $Z_i \sim P_\theta$ on some space \mathcal{Z} .³ We will consider below a sequence of experiments $\mathcal{E}_n = \{P_\theta^n, \theta \in \Theta\}$ as the sample size grows.

Example 2.1 *Dehejia (2005) uses data from a randomized evaluation comparing the Greater Avenues for Independence (GAIN) program to the standard AFDC program for welfare recipients in Alameda County, California. The outcome of interest is individual earnings in various quarters after the program. Since many welfare recipients had zero earnings, Dehejia used a Tobit model*

$$Y_i = \max\{0, \alpha'_1 X_i + \alpha_2 T_i + \alpha'_3 X_i \cdot T_i + \epsilon_i\},$$

where $T_i = 1$ denotes receipt of the experimental program and $\epsilon_i | X_i, T_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$. Dehejia computed posterior distributions based on observation of the n experimental subjects, and then produced predictive distributions for a hypothetical $(n+1)$ th subject to assess different treatment assignment rules. In our notation, $\theta = (\alpha_1, \alpha_2, \alpha_3, \sigma)$, and $Z^n = \{(T_i, X_i, Y_i) : i = 1, \dots, n\}$.

A randomized statistical treatment rule is a mapping $\delta : \mathcal{Z}^n \times \mathcal{X} \rightarrow [0, 1]$. We interpret it as the probability of assigning a (future) individual with covariate X to treatment, given past data Z^n :

$$\delta(z^n, x) = Pr(T = 1 | Z^n = z^n, X = x).$$

Let $L(\delta, \theta, x)$ be some loss function, which specifies penalties for using the rule δ when the true parameter is θ and the future individual's covariate is $X = x$. We will discuss some specific choices for loss below.

Given a loss L , the risk of a rule $\delta(z^n, x)$ under θ is

$$R(\delta, \theta, x) = E_\theta L(\delta(Z^n, x), \theta, x) = \int L(\delta(z^n, x), \theta, x) dP_\theta^n(z^n).$$

We evaluate risk pointwise in x . In principle, we could integrate the risk over the marginal distribution of X , but this pointwise form fits most naturally with our local asymptotic approximations.

²For a discussion of the relationship between the net social welfare and traditional measures of effects of treatments, such as the average treatment effect, see Dehejia (2003).

³The i.i.d. assumption could be weakened to allow for dependent data satisfying local asymptotic normality, at the cost of complicating the arguments below.

The risk of a decision rule can vary with θ , and typically there will not exist a rule that uniformly dominates all other rules, unless one restricts the class of rules substantially. There are two classic ways of defining an ordering over risk functions: one can average the risk of a rule with respect to some prior measure Π on Θ , obtaining a Bayes risk:

$$\int R(\delta, \theta, x) d\Pi(\theta) = \int \int L(\delta(z^n, x), \theta, x) dP_\theta^n(z^n) d\Pi(\theta).$$

Alternatively, one can focus on worst-case risk:

$$\sup_{\theta \in \Theta} R(\delta, \theta, x) = \sup_{\theta \in \Theta} \int L(\delta(z^n, x), \theta, x) dP_\theta^n(z^n).$$

3 Regular Parametric Models

3.1 Limit Experiment

In this section we consider regular parametric models, where the likelihood is smooth in a finite-dimensional parameter. To develop asymptotic approximations, we adopt a local parametrization, as is standard in the literature on efficiency of estimators and test statistics. The local parametrization is used to derive an asymptotic description of the treatment assignment problem using the limits of experiments framework (Le Cam 1986). Although this framework is typically applied to study point estimation and hypothesis testing, it applies much more broadly, to general statistical decision problems. In regular parametric models, a simple Gaussian shift model provides an approximation to the original decision problem.

We first reparametrize the model in terms of local alternatives.⁴ The idea is to consider values for θ such that $g(\theta, x)$ is “close” to 0, and there is a nontrivial difficulty in distinguishing between the effects of the two treatments as sample size grows. In our setting, for a given value of x , we center the localization around θ_0 such that

$$g(\theta_0, x) = 0, \tag{3.1}$$

and consider parameter sequences of the form $\theta_0 + \frac{h}{\sqrt{n}}$, for $h \in \mathbb{R}^k$. This is the same localization device used in local asymptotic power calculations and efficiency bounds, although here the centering value θ_0 is tied to a particular covariate value x .

Equation (3.1) is not the only case of interest, but for establishing asymptotic optimality, it is the key case to focus on. For combinations of (θ_0, x) such that $g(\theta_0, x) \neq 0$, the treatment that is better at θ_0 will be better for all local alternatives $\theta_0 + h/\sqrt{n}$ asymptotically, and many rules, including the rules we will propose below, will select the better treatment with probability

⁴Alternatively, we could use large-deviations asymptotics, in analogy with Bahadur efficiency of hypothesis tests. Manski (2003) uses finite-sample large-deviations results to bound the risk properties of certain types of treatment assignment rules in a binary-outcome randomized experiment. Puhalskii and Spokoiny (1998) develop a large-deviations version of asymptotic statistical decision theory and apply it to estimation and hypothesis testing.

approaching one. Our localization around a centering value satisfying (3.1) ensures that we are looking at the hardest cases, where it is difficult to determine the best treatment even with large sample sizes.

To simplify the notation, we will suppress the dependence on x in the remainder of the analysis, writing $g(\theta, x)$ as $g(\theta)$ and similarly for other quantities. All results should be interpreted as being stated for a fixed x .⁵

Let Θ be an open subset of \mathbb{R}^k , and suppose $\theta_0 \in \Theta$ satisfies equation (3.1). We assume that the sequence of experiments $\mathcal{E}_n = \{P_\theta^n, \theta \in \Theta\}$ satisfies differentiability in quadratic mean at θ_0 : there exists a function $s : \mathcal{Z} \rightarrow \mathbb{R}^k$ such that

$$\int \left[dP_{\theta_0+h}^{1/2}(z) - dP_{\theta_0}^{1/2}(z) - \frac{1}{2}h's(z)dP_{\theta_0}^{1/2}(z) \right]^2 = o(\|h\|^2), \quad \text{as } h \rightarrow 0. \quad (\text{DQM})$$

The function s is the score function associated with the statistical model \mathcal{E}_1 , and can usually be calculated as the derivative of the log likelihood function. Let $I_0 = E_{\theta_0}[ss']$.

The DQM assumption implies that the log-likelihood ratios of the original model converge weakly to the log-likelihood ratios of a multivariate normal experiment, and is the basis for the following result, which specializes Theorems 7.2 and 15.1 of Van der Vaart (1998).

Proposition 3.1 *Let Θ be an open subset of \mathbb{R}^k , and suppose $\theta_0 \in \Theta$. Let $\mathcal{E}_n = \{P_\theta^n, \theta \in \Theta\}$ satisfy DQM with I_0 nonsingular. Consider a sequence of treatment assignment rules $\delta_n(z^n)$ in the experiments \mathcal{E}_n , and let $\beta_n(h) = E_{\theta_0+h/\sqrt{n}}[\delta_n(Z^n)]$. Suppose $\beta_n(h) \rightarrow \beta(h)$ for every h . Then there exists a function $\delta : \mathbb{R}^k \rightarrow [0, 1]$ such that for every $h \in \mathbb{R}^k$,*

$$\beta(h) = \int \delta(\Delta) dN(\Delta|h, I_0^{-1}),$$

where $N(\Delta|h, I_0^{-1})$ is the multivariate normal distribution with mean h and variance I_0^{-1} .

Proposition 3.1 shows that any converging sequence of treatment rules in the original problem is matched by some treatment rule in a simpler experiment where Δ has a shifted normal distribution with known variance. In this sense, the $N(h, I_0^{-1})$ model is a “limit experiment” for the original problem.

3.2 Loss Functions

Having obtained an asymptotic version of the statistical experiment, we need to complete the specification of the decision problem by choosing a loss function, and then examine the limiting forms of loss, risk, and Bayes risk. Generally, we will need to normalize, or modify somewhat, the loss function in the original problem so that limiting versions of risk and Bayes risk are well-defined

⁵We will reintroduce the dependence on x in the last example of the paper, where we examine a specific treatment assignment rule based on conditional sample averages.

and lead to useful comparisons of treatment rules. A key component of the loss and risk functions that we will consider is the welfare contrast, $g(\theta)$. Let \dot{g} be the vector of partial derivatives of g at θ_0 . Then, since $g(\theta_0) = 0$, we have $\sqrt{n}g(\theta_0 + h/\sqrt{n}) \rightarrow \dot{g}'h$ as $n \rightarrow \infty$.

3.2.1 Asymmetric Losses

We consider two loss functions that penalize differently for two types of errors—assigning to treatment 1 when treatment 0 is better, and vice versa. The first loss has been used in the literature on hypothesis testing, and gives fixed penalties for the two types of errors:

Hypothesis Testing Loss:

$$L^H(\delta, \theta) = \begin{cases} (1 - \delta) & \text{if } g(\theta) > 0 \\ K \cdot \delta & \text{if } g(\theta) \leq 0 \end{cases}$$

where $K > 0$.

For a rule δ , the loss can be written

$$L^H(\delta, \theta_0 + \frac{h}{\sqrt{n}}) = 1(g(\theta_0 + \frac{h}{\sqrt{n}}) > 0) + \delta \left[K \cdot g(\theta_0 + \frac{h}{\sqrt{n}}) \leq 0 - 1(g(\theta_0 + \frac{h}{\sqrt{n}}) > 0) \right].$$

This converges as $n \rightarrow \infty$ for values of h such that $\dot{g}'h \neq 0$. Due to the discontinuity in the loss L^H , the case $\dot{g}'h = 0$ presents a problem for taking limits, but we can define a lower bound limit as

$$L_\infty^H(\delta, h) = 1(\dot{g}'h > 0) + \delta [K \cdot 1(\dot{g}'h < 0) - 1(\dot{g}'h > 0)].$$

For a converging sequence of rules δ_n with $\beta_n(h)$ and $\beta(h)$ as defined in Proposition 3.1, we can define risk and its limiting lower bound as:

$$R_n^H(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) = 1(g(\theta_0 + \frac{h}{\sqrt{n}}) > 0) + \beta_n(h) \left[K \cdot 1(g(\theta_0 + \frac{h}{\sqrt{n}}) \leq 0) - 1(g(\theta_0 + \frac{h}{\sqrt{n}}) > 0) \right];$$

$$R_\infty^H(\delta, h) = 1(\dot{g}'h > 0) + \beta(h) [K \cdot 1(\dot{g}'h < 0) - 1(\dot{g}'h > 0)].$$

For Bayes risk, let Π be a prior on the parameter space Θ , with a Lebesgue density $\pi(\theta)$ that is positive and continuous at θ_0 . Define

$$B_n^H(\delta_n, \Pi) = \int R_n^H(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \pi(\theta_0 + \frac{h}{\sqrt{n}}) dh.$$

Since the prior is smooth around θ_0 and places mass zero on h such that $\dot{g}'h = 0$, its limit is

$$B_\infty^H(\delta) = \pi(\theta_0) \int R_\infty^H(\delta, h) dh.$$

Asymmetric Welfare Regret: Tetenov (2007) proposed a loss that extends loss H by penalizing based on the amount of welfare lost by choosing the worse treatment:

$$L^T(\delta, \theta) = \begin{cases} g(\theta)(1 - \delta) & \text{if } g(\theta) > 0 \\ -K\delta g(\theta) & \text{if } g(\theta) \leq 0 \end{cases}$$

with $K > 0$. The case $K = 1$ corresponds to “welfare regret,” which we will discuss further below.

The loss can be written as

$$L^T(\delta, \theta_0 + \frac{h}{\sqrt{n}}) = g(\theta_0 + \frac{h}{\sqrt{n}})1(g(\theta_0 + \frac{h}{\sqrt{n}}) > 0)(1 - \delta) - K \cdot g(\theta_0 + \frac{h}{\sqrt{n}})1(g(\theta_0 + \frac{h}{\sqrt{n}}) \leq 0)\delta$$

Here, the case $\dot{g}'h = 0$ does not present a problem for taking limits, but we do need to normalize by \sqrt{n} so that the limit is nondegenerate:

$$\sqrt{n}L^T(\delta, \theta_0 + \frac{h}{\sqrt{n}}) \rightarrow L_\infty^T(\delta, h) = (\dot{g}'h)1(\dot{g}'h > 0)(1 - \delta) - K \cdot (\dot{g}'h)1(\dot{g}'h \leq 0)\delta$$

We then define risk and Bayes risk analogously to Loss H.

3.2.2 Welfare Loss

Since $w_0(\theta) + \delta g(\theta)$ is social welfare, it is natural to use its negative as a loss function:

$$L^W(\delta, \theta) = -w_0(\theta) - \delta \cdot g(\theta).$$

In order to keep the second term on the right nondegenerate when sample size increases, we would typically scale the loss by \sqrt{n} . However, this will lead the first term, involving w_0 , to diverge.⁶ To obtain nondegenerate limits, we could recenter welfare loss as follows:

$$L^R(\delta, \theta) = L^W(\delta, \theta) - [-w_0(\theta) - 1(g(\theta) > 0)g(\theta)] = g(\theta)[1(g(\theta) > 0) - \delta]$$

This subtracts off the loss associated with the infeasible optimal rule $\delta^* = 1(g(\theta) > 0)$, and leads to what is called regret loss.⁷ Then

$$\sqrt{n}L^R(\delta, \theta_0 + \frac{h}{\sqrt{n}}) \rightarrow L_\infty^R(\delta, h) = \dot{g}'h [1(\dot{g}'h > 0) - \delta].$$

Note that regret loss is equal to Tetenov’s loss with $K = 1$, so we will not need to treat regret loss separately in the formal results to follow.

The recentering does not affect Bayes rules. Consider a rule δ_n that minimizes

$$B_n^W(\delta_n, \Pi) = \int \int L^W(\delta_n(z^n), \theta_0 + \frac{h}{\sqrt{n}}) dP_{\theta_0+h/\sqrt{n}}^n(z^n) \pi(\theta_0 + \frac{h}{\sqrt{n}}) dh.$$

Clearly, adding any function $f(\theta)$ of the parameter to the loss function does not change the solution. Hence the minimizer of B_n^W will also minimize Bayes risk for regret loss. However, the recentering does affect the minmax solution. In this sense, only certain combinations of loss function and optimality criterion have nontrivial asymptotic approximations.

⁶This is the motivation for Assumption 1 in the Appendix.

⁷Savage (1951) suggested this type of centering in a discussion of the minmax criterion. The label “regret” is standard in the decision theory literature.

3.3 Asymptotic Optimality

We have approximated the sequence of statistical experiments by a Gaussian one, and we have approximated loss functions by certain asymptotic versions. If we can find an optimal rule (according to some criterion) in this limiting version of the decision problem, then it will serve as a benchmark for how well any sequence of decision rules can perform in the original problem. Typically, the solution will also suggest the form of a sequence of decision rules that asymptotically match the optimal rule. We develop results for two standard optimality concepts, average and worst-case risk.

3.3.1 Average Risk Optimality

First, consider the average (Bayes) risk criterion in the limiting Gaussian model. When the limiting loss function is given by $L_\infty(\delta, h)$, we wish to find the rule δ that minimizes

$$B_\infty(\delta) = \int R_\infty(\delta, h)dh = \int \int L_\infty(\delta(\Delta), h)dN(\Delta|h, I_0^{-1})dh.$$

Directly minimizing this expression would involve searching over the space of decision rules. But the problem can be simplified by reversing the order of integration and noting that, for each Δ , the solution will minimize

$$\int L_\infty(\delta(\Delta), h) \exp(-(h - \Delta)'I_0(h - \Delta)/2) dh,$$

which is equivalent to minimizing posterior expected risk, where the posterior has $h|\Delta \sim N(\Delta, I_0^{-1})$.

Let $\sigma_g^2 = \dot{g}'I_0^{-1}\dot{g}$. In the Appendix, we show that the average risk optimal rule for hypothesis testing loss L_∞^H is

$$\delta(\Delta) = 1 \left(\frac{\dot{g}'\Delta}{\sigma_g} > c^{H,B} \right), \quad \text{where } c^{H,B} = \Phi^{-1} \left(\frac{K}{1+K} \right).$$

For asymmetric welfare regret loss, average risk is minimized by the rule

$$\delta(\Delta) = 1 \left(\frac{\dot{g}'\Delta}{\sigma_g} > c^{T,B} \right), \quad \text{where } c^{T,B} \text{ solves } c = \frac{(K-1)\phi(c)}{\Phi(c) + K\Phi(-c)}.$$

For both L_∞^H and L_∞^T , the cutoff is equal to 0 when $K = 1$.

Both optimal rules in the limiting Gaussian model have a simple cutoff form, which suggests how to construct rules in the original problem that are asymptotically equivalent. Let $\hat{\theta}_n$ be an estimator in the original sequence of models that is best regular:

$$\sqrt{n}(\hat{\theta}_n - \theta_0 - h/\sqrt{n}) \overset{h}{\rightsquigarrow} N(0, I_0^{-1}), \quad \forall h \in \mathbb{R}^k, \quad (3.2)$$

where $\overset{h}{\rightsquigarrow}$ denotes convergence in distribution under the sequence of probability measures $P_{\theta_0+h/\sqrt{n}}^n$. Both the MLE and the Bayesian posterior mean would usually satisfy this condition. If we also

have a consistent estimator $\hat{\sigma}_g$ of σ_g , then the feasible decision rule

$$\delta_n^{H,B}(Z^n) = 1 \left(\frac{\sqrt{n}g(\hat{\theta}_n)}{\hat{\sigma}_g} > c^{H,B} \right) \quad (3.3)$$

will have limiting distributions that match $1(j'\Delta/\sigma_g > c^{H,B})$ for every h . For Tetenov's loss, we define $\delta_n^{T,B}$ analogously. These two decision rule are asymptotically optimal for average risk:

Theorem 3.2 *Suppose the conditions of Proposition 3.1 are satisfied, $g(\theta_0) = 0$, $g(\theta)$ is differentiable at θ_0 , and the prior measure Π admits a density π with respect to Lebesgue measure that is continuous and positive at θ_0 . Suppose $\hat{\theta}_n$ is a best regular estimator satisfying Equation (3.2) and $\hat{\sigma}_g \xrightarrow{P} \sigma_g$ under θ_0 . Then*

$$\lim_{n \rightarrow \infty} B_n^H(\delta_n^{H,B}, \Pi) = \inf_{\delta_n \in \mathcal{D}} \liminf_{n \rightarrow \infty} B_n^H(\delta_n, \Pi),$$

and

$$\lim_{n \rightarrow \infty} \sqrt{n}B_n^T(\delta_n^{T,B}, \Pi) = \inf_{\delta_n \in \mathcal{D}} \liminf_{n \rightarrow \infty} \sqrt{n}B_n^T(\delta_n, \Pi),$$

where \mathcal{D} denotes the set of all sequences of decision rules that converge in the sense of Proposition 3.1.

Remark: This result focuses on the case where $g(\theta_0) = 0$. When $g(\theta_0) \neq 0$, one treatment is always preferred under all local parameters, so any rule choosing the appropriate treatment with probability approaching one will be asymptotically optimal. In particular, the rules given in Theorem 3.2 remain optimal when $g(\theta_0) \neq 0$ under suitable rate normalizations for the Bayes risk. Similar remarks apply to the other asymptotic optimality results to follow. \square

Theorem 3.2 shows that a simple rule, which replaces θ by an efficient estimator $\hat{\theta}_n$ and σ_g by a consistent estimator, is approximately Bayes optimal. If the posterior distribution is tractable, we could also solve the finite-sample Bayes problem directly. Not surprisingly, this will also be asymptotically optimal:

Corollary 3.3 *For $j = H, T$, let $\delta_n^{j, Bayes} = \arg \min_{\delta} B_n^j(\delta, \Pi)$. If the argmin does not exist for any n , let $\delta_n^{j, Bayes}$ be any rule such that $B_n^j(\delta_n^{j, Bayes}, \Pi) \leq B_n^j(\delta_n^{j, B}, \Pi)$. Then Theorem 3.2 also holds with $\delta_n^{j, B}$ replaced by $\delta_n^{j, Bayes}$.*

3.4 Minmax Optimality

Next, consider the minmax criterion in the limiting Gaussian model. We wish to solve the functional minimization problem $\inf_{\delta} \sup_h R_{\infty}(\delta, h)$ over the class of all decision rules, a difficult task in general. However, the structure of our problem can be used to simplify the solution. We consider

“slices” of the parameter space constructed in the following way: fix an h_0 such that $\dot{g}'h_0 = 0$, and for any $b \in \mathbb{R}$, define

$$h_1(b, h_0) = h_0 + \frac{b}{\dot{g}'I_0^{-1}\dot{g}}I_0^{-1}\dot{g}.$$

Along each slice, the quantity $b = \dot{g}'h_1$ gives the welfare contrast. It turns out that, for many loss functions of interest, rules of the form $\delta_c = 1(\dot{g}'\Delta > c)$, for $c \in \mathbb{R}$, form an essential complete class on each slice, so that it is sufficient to search among cutoff rules to solve the minmax problem along a slice. Furthermore, when the loss function only depends on $\dot{g}'h$, the same cutoff value c solves the minmax problem along each slice, and leads to a minmax rule over the entire parameter space.

Theorem 3.4 *Suppose that $\Delta \sim N(h, I_0^{-1})$ for $h \in \mathbb{R}^k$, and consider a decision problem with action space $\{0, 1\}$ and loss $L(a, h)$ such that for all h with $\dot{g}'h \neq 0$,*

$$[L(1, h) - L(0, h)](\dot{g}'h) < 0$$

- (i) *For any randomized decision rule $\tilde{\delta}(\Delta)$ and any fixed $h_0 \in \mathbb{R}^k$, there exists a rule of the form $\delta_c(\Delta) = 1(\dot{g}'\Delta > c)$ which is at least as good as $\tilde{\delta}$ on the subspace $\{h_1(b, h_0) : b \in \mathbb{R}\}$.*
- (ii) *Additionally, suppose $L(a, h)$ depends on h only through $\dot{g}'h$.⁸ If a minmax decision rule exists, then $\delta_{c^*}(\Delta)$ is minmax for some c^* . Moreover, the optimal value c^* can be obtained by solving $\inf_c \sup_b E_{h_1(b, 0)}L(\delta_c, h_1(b, 0))$.*

The condition $[L(h, 1) - L(h, 0)](\dot{g}'h) < 0$ requires that the loss impose greater penalties for incorrect assignment, and is satisfied by losses L_∞^H and L_∞^T . Part (i) of this result is a mild extension of the essential complete class theorem of Karlin and Rubin (1956). Part (ii) provides a simple method for constructing a minmax rule.

Using (ii), the minmax rule for loss L_∞^H is derived in the Appendix, and is a cutoff rule $1(\dot{g}'\Delta/\sigma_g > c^{H,M})$ where $c^{H,M} = c^{H,B}$. In the case of loss L_∞^T , Tetenov (2007) provided a solution in the scalar normal case with known variance, which extends to our multivariate setting in light of Theorem 3.4(ii). The minmax rule for loss L_∞^T is a cutoff rule where the cutoff $c^{T,M}$ is the solution to:

$$\sup_{b \leq 0} -K \cdot b \cdot \Phi(b - c^{T,M}) = \sup_{b > 0} b \cdot \Phi(c^{T,M} - b).$$

As with the Bayes criterion, if either loss is symmetric ($K = 1$), then the optimal cutoff is zero. For general K , the minmax criterion and the Bayes risk criterion lead to the same optimal decision rules for loss L_∞^H , but interestingly not for loss L_∞^T . These limit experiment solutions lead to the following asymptotic minmax result:

Theorem 3.5 *Suppose the conditions of Proposition 3.1 are satisfied, $g(\theta_0) = 0$, and $g(\theta)$ is differentiable at θ_0 . Suppose $\hat{\theta}_n$ is a best regular estimator satisfying Equation (3.2) and $\hat{\sigma}_g \xrightarrow{P} \sigma_g$*

⁸There exists a function L_g such that $L(a, h) = L_g(a, \dot{g}'h)$ for all a, h .

under θ_0 . Let $\delta_n^{H,M}$ and $\delta_n^{T,M}$ be defined analogously to Equation (3.3). Then these rules are locally asymptotically minmax:

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} R_n^H(\delta_n^{H,M}, \theta_0 + \frac{h}{\sqrt{n}}) = \inf_{\delta_n \in \mathcal{D}} \sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} R_n^H(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}),$$

and

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} \sqrt{n} R_n^T(\delta_n^{T,M}, \theta_0 + \frac{h}{\sqrt{n}}) = \inf_{\delta_n \in \mathcal{D}} \sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} \sqrt{n} R_n^T(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}),$$

where the outer supremum is over all finite subsets J of \mathbb{R}^k .

Remark: Although the average and minmax asymptotic risk comparisons allow h , and hence $\theta_0 + h/\sqrt{n}$, to take on arbitrary values, this reparametrization has important consequences. The localization reduces the statistical information to have an approximately Gaussian form (a type of asymptotic sufficiency), and it leads to limiting risk functions that depend on the parameter only through $\dot{g}'h$. These two effects lead to the simplification of the general decision problem, with the tradeoff that only certain combinations of loss function and optimality notion lead to nontrivial comparisons of decision procedures. \square

4 Semiparametric Models

4.1 Gaussian Sequence Limit Experiment

Empirical studies of treatment effects often use nonparametric or semiparametric specifications, to allow for more flexibility in the modeling of treatment effects. In this section, we extend the results from the previous section to models with an infinite-dimensional parameter space.

Suppose Z^n consists of an i.i.d. sample of size n drawn from a probability measure $P \in \mathcal{P}$, where \mathcal{P} is the set of probability measures defined by the underlying semiparametric model. In some cases the set \mathcal{P} will include all distributions satisfying certain weak conditions (so that the model is nonparametric); in other cases the form of the semiparametric model may restrict the feasible distributions in \mathcal{P} .

We fix $P_0 \in \mathcal{P}$, and define local alternatives to P_0 in a standard way following Van der Vaart (1991a). Consider subsets of the form $\{P_{t,h} : t \in (0, \eta)\} \subset \mathcal{P}$, where $\eta > 0$, and h is a real-valued measurable function on \mathcal{Z} satisfying

$$\int \left[\frac{1}{t} \left(dP_{t,h}^{1/2} - dP_0^{1/2} \right) - \frac{1}{2} h dP_0^{1/2} \right]^2 \rightarrow 0 \quad \text{as } t \downarrow 0. \quad (4.1)$$

Each subset $\{P_{t,h} : t \in (0, \eta)\}$ is then a smooth one-dimensional submodel (or path) of \mathcal{P} . Given P_0 , the collection of such paths will be denoted $\mathcal{P}(P_0)$. The function h is the score function associated with the submodel, and satisfies $\int h dP_0 = 0$ and $\int h^2 dP_0 < \infty$.

For fixed t and h , $P_{t/\sqrt{n},h}$ is a sequence of measures that approaches P_0 as $n \rightarrow \infty$. It will be enough to consider the sequences $P_{1/\sqrt{n},h}$, so we can view each h as a local parameter, in analogy with the parametric case. Let the tangent set $T(P_0) \subset L_2(P_0)$ be the set of (equivalence classes of) functions h satisfying Equation (4.1).

We will assume that $T(P_0)$ is a separable linear space, so that $\overline{T(P_0)}$ is a separable Hilbert space with the usual inner product and norm for $L_2(P_0)$. Let ϕ_1, ϕ_2, \dots denote any orthonormal basis of $\overline{T(P_0)}$. We identify $\overline{T(P_0)}$ with an l_2 space in the usual way, through the isomorphism $h \mapsto (h_1, h_2, \dots)$ with $h_j = \langle h, \phi_j \rangle$, so that $h(\cdot) = \sum_j h_j \phi_j(\cdot)$.

Again, we use g to denote the difference in social welfare $W_1(x) - W_0(x)$. For a probability measure $P \in \mathcal{P}$, we denote this welfare contrast by $g(P, x)$, or $g(P)$ for short. We assume that there exists a continuous linear map $\dot{g} : T(P_0) \rightarrow \mathbb{R}$ such that

$$\frac{1}{t}(g(P_{t,h}) - g(P_0)) \rightarrow \dot{g}(h) \quad \text{as } t \downarrow 0 \quad (4.2)$$

for every path in $\mathcal{P}(P_0)$.⁹ This implies

$$\sqrt{n}(g(P_{1/\sqrt{n},h}) - g(P_0)) \rightarrow \dot{g}(h).$$

By the Riesz representation theorem, the functional $\dot{g}(\cdot)$ can be associated with an element $\dot{g} \in \overline{T(P_0)}$ such that $\dot{g}(h) = \langle \dot{g}, h \rangle$ for all $h \in T(P_0)$. (This parallels the notation $\dot{g}'h$ in the parametric case.) Assume $\|\dot{g}\|^2 = \langle \dot{g}, \dot{g} \rangle > 0$.

Van der Vaart (1991a) shows that an asymptotic representation theorem similar to the parametric case holds, where the shifted multivariate Gaussian limit experiment is replaced by an infinite (shifted) Gaussian sequence. This leads to the following result for treatment rules:

Proposition 4.1 *Let $\mathcal{E}_n = \{P_{1/\sqrt{n},h}^n : h \in T(P_0)\}$ satisfy Equation (4.1). Consider a sequence of treatment rules $\delta_n(z^n)$ in the experiments \mathcal{E}_n , and let $\beta_n(h) = \int \delta_n dP_{1/\sqrt{n},h}$. Suppose $\beta_n(h) \rightarrow \beta(h)$ for every h . Then there exists a function δ such that $\beta(h) = E_h[\delta(\Delta_1, \Delta_2, \dots)]$, where $(\Delta_1, \Delta_2, \dots)$ is a sequence of independent random variables with $\Delta_j \stackrel{h}{\sim} N(h_j, 1)$.*

4.2 Semiparametric Optimality

We consider loss functions that are analogs of losses H and T in the parametric case, where the parameter θ is replaced by $P \in \mathcal{P}$. We denote these by L^H and L^T as before, in a slight abuse of notation. Then, the limiting versions of the loss functions will have the same form as before, with $\langle \dot{g}, h \rangle$ replacing $\dot{g}'h$.

Defining and working with average risk is more complicated in infinite-dimensional models. In the parametric case, simple conditions ensure that a prior measure behaves locally like Lebesgue

⁹Van der Vaart (1991b) provides a thorough discussion of this differentiability notion, which is related to Hadamard differentiability.

measure. In infinite product spaces, however, there is no natural analog of Lebesgue measure, and the asymptotic properties of Bayes procedures can be quite sensitive to the choice of prior and the specific model at hand (see, for example, Diaconis and Freedman (1986)). Instead of working with some fixed prior on the space \mathcal{P} , we define a prior on the tangent space, and compare procedures by their average risk with respect to this prior.¹⁰

It is useful to choose the orthonormal basis ϕ_1, ϕ_2, \dots so that the welfare contrast $\langle \dot{g}, h \rangle$ is attached to the leading term. Let $\phi_1 = \dot{g}/\|\dot{g}\|$, and let ϕ_2, ϕ_3, \dots be an orthonormal basis for the orthocomplement of the space spanned by ϕ_1 . We can view the Gaussian sequence $\Delta_1, \Delta_2, \dots$ in Proposition 4.1 as being defined relative to this choice of orthonormal basis. In particular, under h , $\Delta_1 \sim N(\langle \dot{g}, h \rangle / \|\dot{g}\|, 1)$, and $\Delta_2, \Delta_3, \dots$ have distributions that do not depend on the value of $\langle \dot{g}, h \rangle$.

Define a prior measure for (h_1, h_2, \dots) by $\Pi = \lambda \times \rho$, where λ is Lebesgue measure on the real line and ρ is some finite or σ -finite measure on l^2 . Since Π is a finite product of σ -finite measures on separable spaces, it is well defined.

Define $\Delta = (\Delta_1, \Delta_2, \dots)$, let R_∞ be the associated risk function and let

$$B_\infty(\delta(\Delta), \Pi) = \int R_\infty(\delta(\Delta), h) d\Pi(h).$$

Suppose that the loss function only depends on h through $\langle \dot{g}, h \rangle$, so we can write (with slight abuse of notation) $L_\infty(\delta(\Delta), \langle \dot{g}, h \rangle)$ (see the supposition of Theorem 3.4(ii)). By interchanging the order of integration, it follows that the Bayes rule can be obtained by minimizing, for each Δ ,

$$\int L_\infty(\delta(\Delta), u) dN(u|\tilde{g}, \sigma_g^2) = \int L_\infty(0, u) dN(u|\tilde{g}, \sigma_g^2) + \delta(\Delta) \int (L_\infty(1, u) - L_\infty(0, u)) dN(u|\tilde{g}, \sigma_g^2),$$

where $\tilde{g} = \|\dot{g}\|\Delta_1$ and $\sigma_g^2 = \|\dot{g}\|^2$. The optimal rule does not depend on $\Delta_2, \Delta_3, \dots$ or on ρ . From here, the analysis is essentially the same as in the parametric case. The optimal Bayes rules in the Gaussian sequence model for L_∞^H and L_∞^T can be written

$$1(\Delta_1 > c^{H,B}), \quad 1(\Delta_1 > c^{T,B}),$$

where $c^{H,B}$ and $c^{T,B}$ are the same constants as in the parametric case.

To construct asymptotic approximation results, define

$$R_n^H(\delta_n, P_{1/\sqrt{n}, h}) = \int L^H(\delta_n(z^n), P_{1/\sqrt{n}, h}) dP_{1/\sqrt{n}, h}^n(z^n),$$

$$B_n^H(\delta_n, \Pi) = \int R_n^H(\delta_n, P_{1/\sqrt{n}, h}) d\Pi(h),$$

and similarly for loss L^T . Then we have

¹⁰In a different setting, Andrews and Ploberger (1994) use priors on local parameters to define local average power optimality of tests.

Theorem 4.2 Suppose the conditions for Proposition 4.1 are satisfied, $g(P_0) = 0$, g satisfies Equation (4.2), $\hat{\sigma}_g \xrightarrow{P} \|\dot{g}\|$ under P_0 , and $\hat{g}_n(Z^n)$ is a best regular estimator for $g(P)$:

$$\sqrt{n} \left(\hat{g}_n(Z^n) - g(P_{1/\sqrt{n},h}) \right) \overset{h}{\rightsquigarrow} N(0, \|\dot{g}\|^2), \quad (4.3)$$

for all $h \in T(P_0)$, where $\overset{h}{\rightsquigarrow}$ denotes convergence in distribution under $P_{1/\sqrt{n},h}$. Let

$$\delta_n^{H,B}(Z^n) = 1 \left(\frac{\sqrt{n}\hat{g}_n}{\hat{\sigma}_g} > c^{H,B} \right), \quad \delta_n^{T,B}(Z^n) = 1 \left(\frac{\sqrt{n}\hat{g}_n}{\hat{\sigma}_g} > c^{T,B} \right).$$

Then

$$\lim_{n \rightarrow \infty} B_n^H(\delta_n^{H,B}, \Pi) = \inf_{\delta_n \in \mathcal{D}} \liminf_{n \rightarrow \infty} B_n^H(\delta_n, \Pi),$$

and

$$\lim_{n \rightarrow \infty} \sqrt{n} B_n^T(\delta_n^{T,B}, \Pi) = \inf_{\delta_n \in \mathcal{D}} \liminf_{n \rightarrow \infty} \sqrt{n} B_n^T(\delta_n, \Pi).$$

where \mathcal{D} denotes the set of all sequences of decision rules that converge in the sense of Proposition 4.1.

Remark: Theorem 4.2 has a different interpretation than the parametric Bayes result in Theorem 3.2. Here, the prior Π is a weighting on the tangent space of P_0 , so its influence does not disappear as the sample size grows large. We use it here to compare different rules which are derived from asymptotic considerations. In general, a sequence of Bayes rules derived from some fixed prior on \mathcal{P} will not be optimal under our criterion. \square

The slicing argument used for the minmax analysis of the parametric case also extends to the infinite-dimensional case. Our choice of basis proves convenient, because $\langle \dot{g}, h \rangle$ depends on h only through its first term h_1 : $\langle \dot{g}, h \rangle = \|\dot{g}\| h_1$.

Theorem 4.3 Let $\Delta = (\Delta_1, \Delta_2, \dots)$ have $\Delta_j \overset{\text{ind}}{\rightsquigarrow} N(h_j, 1)$ under $h = (h_1, h_2, \dots) \in T(P_0)$. Let the action space be $\{0, 1\}$, and for all h such that $\langle \dot{g}, h \rangle \neq 0$, let the loss $L(a, h)$ satisfy:

$$[L(1, h) - L(0, h)] \langle \dot{g}, h \rangle < 0.$$

(i) Then for any randomized decision rule $\tilde{\delta}(\Delta)$, and $(0, h_2, h_3, \dots) \in T(P_0)$, there exists a rule of the form $\delta_c(\Delta) = 1(\Delta_1 > c)$ which is at least as good as $\tilde{\delta}$ on the one-dimensional subspace $\{(b, h_2, h_3, \dots) : b \in \mathbb{R}\}$.

(ii) Additionally, suppose $L(a, (h_1, h_2, \dots))$ only depends on (h_1, h_2, h_3, \dots) through $\langle \dot{g}, h \rangle$.¹¹ If a minmax decision rule exists, then $\delta_{c^*}(\Delta)$ is minmax for some c^* . Moreover, the optimal value c^* can be obtained by solving $\inf_c \sup_b E_{h=(b,0,0,\dots)} L(\delta_c, h)$.

¹¹This supposition is equivalent to assuming that $L(a, (h_1, h_2, \dots))$ does not depend on h_2, h_3, \dots for $a = 0, 1$.

Theorem 4.3 leads to the optimal rules $1(\Delta_1 > c^{H,M})$ and $1(\Delta_1 > c^{T,M})$, where the constants are the same as in the parametric case. If we can match the distribution of Δ_1 asymptotically, we can obtain asymptotic minmax optimality:

Theorem 4.4 *Suppose the conditions for Proposition 4.1 are satisfied, $g(P_0) = 0$, g satisfies Equation (4.2), $\hat{\sigma}_g \xrightarrow{P} \|\dot{g}\|$ under P_0 , and \hat{g}_n is a best regular estimator for $g(P)$ satisfying Equation (4.3).*

Let

$$\delta_n^{H,M}(Z^n) = 1\left(\frac{\sqrt{n}\hat{g}_n}{\hat{\sigma}_g} > c^{H,M}\right), \quad \delta_n^{T,M}(Z^n) = 1\left(\frac{\sqrt{n}\hat{g}_n}{\hat{\sigma}_g} > c^{T,M}\right).$$

Then

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} R_n^H(\delta_n^{H,M}, P_{1/\sqrt{n},h}) = \inf_{\delta_n \in \mathcal{D}} \sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} R_n^H(\delta_n, P_{1/\sqrt{n},h}),$$

and

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} \sqrt{n} R_n^T(\delta_n^{T,M}, P_{1/\sqrt{n},h}) = \inf_{\delta_n \in \mathcal{D}} \sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} \sqrt{n} R_n^T(\delta_n, P_{1/\sqrt{n},h}),$$

where the outer supremum is over finite subsets J of $T(P_0)$.

To close, we illustrate how this result applies to Manski's conditional empirical success rule.

Example 4.5 *Suppose that $W_0(x) = 0$, and that we observe a random sample (X_i, Y_i) , $i = 1, \dots, n$, where X_i has a finitely supported distribution and $Y_i|X_i$ has conditional distribution $F_1(y|x)$. The social welfare contrast is the functional $g(x, F_1) = \int w(y)dF_1(y|x)$. The conditional distribution function F_1 is unknown, and the set of possible CDFs \mathcal{P} is the largest set satisfying*

$$\sup_{F_1 \in \mathcal{P}} E[|w(Y)|^2 | X = x] < \infty.$$

The conditional empirical success rule of Manski (2004) can be expressed as

$$\hat{\delta}_n(x) = 1(\hat{g}_n(x) > 0),$$

where

$$\hat{g}_n(x) := \frac{\sum_{i=1}^n w(Y_i) \cdot 1(X_i = x)}{\sum_{i=1}^n 1(X_i = x)}.$$

The estimator $\hat{g}_n(x)$ is an asymptotically efficient estimator of $g(x, F_1)$ (Bickel, Klaassen, Ritov, and Wellner (1993), pp. 67-68). Therefore, $\hat{\delta}_n$ is asymptotically minmax and asymptotically Bayes optimal for both losses H and T under $K = 1$.

This result extends easily to the case where $W_0(x)$ is not known; then $\hat{g}_n(x)$ would be a difference of conditional mean estimates for outcomes under treatments 1 and 0.

Appendix A: Proofs for Parametric Case

We give some assumptions and lemmas that can be used to extend the results to other loss functions. Let \mathcal{D}_∞ denote the set of all randomized treatment rules in the $N(h, I_0^{-1})$ experiment.

Assumption 1 *Given a loss function L , there exists $L_\infty(\delta, h)$ such that for some sequence r_n ,*

$$\lim_{n \rightarrow \infty} r_n \left[L \left(1, \theta_0 + \frac{h}{\sqrt{n}} \right) - L \left(0, \theta_0 + \frac{h}{\sqrt{n}} \right) \right] = L_\infty(1, h) - L_\infty(0, h)$$

and

$$\lim_{n \rightarrow \infty} r_n L \left(0, \theta_0 + \frac{h}{\sqrt{n}} \right) = L_\infty(0, h)$$

for almost every h (with respect to Lebesgue measure on \mathbb{R}^k).

Assumption 2 *Assume that loss L_∞ in the limit experiment depends on h only through $\dot{g}'h$. That is there exists L_g such that $L_\infty(a, h) = L_g(a, \dot{g}'h)$ for $a \in \{0, 1\}$.*

Lemma 1 *Suppose the conditions of Theorem 3.2 are satisfied with sequence of treatment assignment rules $\delta_n \in \mathcal{D}$. Let loss L satisfy Assumption 1. Then,*

(i)

$$\liminf_{n \rightarrow \infty} r_n B_n(\delta_n, \Pi) \geq \pi(\theta_0) \inf_{\delta \in \mathcal{D}_\infty} B_\infty(\delta).$$

(ii) *Moreover, if $\delta_n^* \in \mathcal{D}$ is matched by $\delta^* \in \mathcal{D}_\infty$ in the sense of Proposition 3.1, and δ^* is the flat-prior Bayes rule in the limit experiment, then δ_n^* is the asymptotically optimal rule for Bayes risk:*

$$\lim_{n \rightarrow \infty} r_n B_n(\delta_n^*, \Pi) = \pi(\theta_0) B_\infty(\delta^*) = \pi(\theta_0) \inf_{\delta \in \mathcal{D}_\infty} B_\infty(\delta).$$

Proof of Lemma 1:

For the sequence $\{\delta_n\}$, let $\bar{\delta}$ be the matching treatment assignment rule in the limit experiment as given by Proposition 3.1. Then

$$\begin{aligned} \liminf_{n \rightarrow \infty} r_n B_n(\delta_n, \pi) &\geq \int \liminf_{n \rightarrow \infty} \left\{ \int r_n L \left(\delta_n(z), \theta_0 + \frac{h}{\sqrt{n}} \right) dP_{\theta_0+h/\sqrt{n}}^n(z) \pi \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right\} dh \\ &= \int \liminf_{n \rightarrow \infty} \left\{ \left(E_{\theta_0+h/\sqrt{n}}[\delta_n(Z_n)] r_n \left[L \left(1, \theta_0 + \frac{h}{\sqrt{n}} \right) - L \left(0, \theta_0 + \frac{h}{\sqrt{n}} \right) \right] \right. \right. \\ &\quad \left. \left. + r_n L \left(0, \theta_0 + \frac{h}{\sqrt{n}} \right) \right) \pi \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right\} dh \\ &= \pi(\theta_0) \int (E_h[\bar{\delta}(\Delta)] [L_\infty(1, h) - L_\infty(0, h)] + L_\infty(0, h)) dh = \pi(\theta_0) B_\infty(\bar{\delta}) \\ &\geq \pi(\theta_0) \inf_{\delta \in \mathcal{D}_\infty} B_\infty(\delta), \end{aligned}$$

where the first inequality follows by Fatou's Lemma and the second equality follows by Proposition 3.1, Assumption 1, and the continuity of π . An analogous argument yields $\limsup_{n \rightarrow \infty} r_n B_n(\delta_n, \Pi) \leq \pi(\theta_0) B_\infty(\bar{\delta})$. Applying these conclusions to δ_n^* and δ^* proves (ii). \square

Lemma 2 *Suppose that loss L_∞ satisfies Assumption 2. Let*

$$\delta^*(\Delta) = \mathbf{1} \{E_s [L_g(1, s)] \leq E_s [L_g(0, s)]\}$$

where $s \sim N(\dot{g}'\Delta, \sigma_g^2)$. Then, δ^* is the flat-prior Bayes rule in the limit experiment,

$$B_\infty(\delta^*) = \inf_{\delta \in \mathcal{D}_\infty} B_\infty(\delta)$$

Proof of Lemma 2:

By Fubini's Theorem, we can rewrite the flat-prior Bayes risk as

$$B_\infty(\delta) = \int \int L_\infty(\delta(\Delta), h) (2\pi)^{-k/2} |I_0|^{-1/2} \exp(-(\Delta - h)' I_0 (\Delta - h)/2) dh d\Delta.$$

As usual, the Bayes optimality problem is equivalent to minimizing posterior expected loss for each observable Δ . The posterior expected loss for the rule δ in the Gaussian limit experiment, at a fixed Δ , is:

$$\int L_\infty(\delta(\Delta), h) dN(\Delta, I_\theta^{-1})(h) = \int L_g(0, s) dN(\dot{g}'\Delta, \sigma_g^2)(s) + \delta(\Delta) \int [L_g(1, s) - L_\infty^*(0, s)] dN(\dot{g}'\Delta, \sigma_g^2)(s).$$

The optimal Bayes rule then is determined by the last term and the statement of the lemma follows. \square

Lemma 3 *Suppose the conditions of Theorem 3.2 are satisfied. Then, for every $h \in \mathbb{R}^k$,*

$$\sqrt{n} \frac{g(x, \hat{\theta}_n)}{\hat{\sigma}_g} \xrightarrow{h} N(\dot{g}'h, 1).$$

Proof of Lemma 3:

By differentiability in quadratic mean, the sequence of experiments is locally asymptotically normal. For all sequences $h_n \rightarrow h$ in \mathbb{R}^k ,

$$\log \frac{dP_{\theta_0 + h_n/\sqrt{n}}^n}{dP_{\theta_0}^n} = h' S_n - \frac{1}{2} h' I_0 h + o_{P_{\theta_0}}(1),$$

where $S_n \xrightarrow{\theta_0} N(0, I_0)$. Since $\hat{\theta}_n$ is best regular, Lemma 8.14 of Van der Vaart (1998) implies $\sqrt{n}(\hat{\theta}_n - \theta_0) = I_0^{-1} S_n + o_{P_{\theta_0}^n}(1)$ under θ_0 . By Slutsky's Theorem and the Delta Method, under θ_0 ,

$$\begin{aligned} \left(\sqrt{n} \frac{g(x, \hat{\theta})}{\hat{\sigma}_g}, \log \frac{dP_{\theta_0 + h/\sqrt{n}}^n}{dP_{\theta_0}^n} \right) &= \left(\frac{1}{\sigma_g} \dot{g}' I_0^{-1} S_n, h' S_n - \frac{1}{2} h' I_0 h \right) + o_{P_{\theta_0}^n}(1) \\ &\rightsquigarrow N \left(\begin{pmatrix} 0 \\ -\frac{1}{2} h' I_0 h \end{pmatrix}, \begin{pmatrix} 1 & \frac{\dot{g}' h}{\sigma_g} \\ \frac{\dot{g}' h}{\sigma_g} & h' I_0 h \end{pmatrix} \right). \end{aligned}$$

Then by Le Cam's Third Lemma, the conclusion of the lemma follows. \square

Proof of Theorem 3.2:

Since losses L^H and L^T satisfy Assumption 1 and L_∞^H and L_∞^T satisfy Assumption 2, Lemma 2 establishes Bayes rules in the limit experiment and Lemma 3 can be used to show that these rules are the matching rules for the sequences of rules in the statement of the theorem. Lemma 1 then states that these sequences of rules are asymptotically Bayes optimal as desired.

Suppose $s \sim N(\dot{g}'z, \sigma_g^2)$. Then, $E[L_g^H(1, s)] = K\Phi(-\dot{g}'z/\sigma_g)$, and $E[L_g^H(0, s)] = \Phi(\dot{g}'\Delta/\sigma_g)$. By Lemma 2, the flat prior Bayes rule in the limit experiment for L_∞^H is $\delta^{H,B}(z) = \mathbf{1}\{\frac{\dot{g}'\Delta}{\sigma_g} \geq c^{H,B}\}$. By Lemma 3, $\lim_{n \rightarrow \infty} E_{\theta_0+h/\sqrt{n}}[\delta_n^{H,B}(Z_n)] = \Pr_h\left(\frac{\dot{g}'\Delta}{\sigma_g} \geq c^{H,B}\right) = E_h[\delta^{H,B}(\Delta)]$.

Similarly, for loss L^T , $E[L_g^T(1, s)] = -K\left[\dot{g}'\Delta\Phi\left(-\frac{\dot{g}'\Delta}{\sigma_g}\right) - \sigma_g\phi\left(\frac{\dot{g}'\Delta}{\sigma_g}\right)\right]$ and $E[L_g^T(0, s)] = \dot{g}'\Delta\Phi\left(\frac{\dot{g}'\Delta}{\sigma_g}\right) + \sigma_g\phi\left(\frac{\dot{g}'\Delta}{\sigma_g}\right)$. Differentiation shows that $E[L_g^T(0, s)] - E[L_g^T(1, s)]$ is monotonically increasing in $(\dot{g}'\Delta)$, and the optimal decision rule will be determined by the cut-off $c^{T,B}$. By Lemmas 2 and 3, $\delta_n^{T,B}$ is matched by the flat prior Bayes rule in the limit experiment. Lemma 1 then yields asymptotic optimality of $\delta_n^{H,B}$ and $\delta_n^{T,B}$. \square

Proof of Corollary 3.3:

From Lemma 1(i), $\liminf_{n \rightarrow \infty} B_n^j(\delta_n^{j,bayes}, \Pi) \geq \pi(\theta_0) \inf_{\delta \in \mathcal{D}_\infty} B_\infty(\delta)$. Also, by definition, $B_n^j(\delta_n^{j,bayes}, \Pi) \leq B_n^j(\delta_n^{j,B}, \Pi)$ for every n , so $\liminf_{n \rightarrow \infty} B_n^j(\delta_n^{j,bayes}, \Pi) \leq \liminf_{n \rightarrow \infty} B_n^j(\delta_n^{j,B}, \Pi) = \pi(\theta_0) \inf_{\delta \in \mathcal{D}_\infty} B_\infty^j(\delta)$. \square

Proof of Theorem 3.4:

Part (i) would follow from a multivariate extension of Karlin and Rubin (1956) Theorem 1. A direct proof follows.

Note that if $\dot{g}'h_0 \neq 0$, then for $\tilde{h}_0 = h_1(-\dot{g}'h_0, h_0)$, $\dot{g}'\tilde{h}_0 = 0$. Since $\{h_1(b, h_0) : b \in \mathbb{R}\} = \{h_1(b, \tilde{h}_0) : b \in \mathbb{R}\}$, we may assume without loss of generality that, in fact, $\dot{g}'h_0 = 0$.

Recall that $\dot{g}'\Delta \sim N(0, \dot{g}'I_0^{-1}\dot{g})$ under h_0 , so $E_{h_0}[\delta_c(\Delta)] = 1 - \Phi\left(c/\sqrt{\dot{g}'I_0^{-1}\dot{g}}\right)$. Let $\tilde{\delta}$ be an arbitrary treatment assignment rule. We can choose c to satisfy $E_{h_0}[\delta_c(\Delta)] = E_{h_0}[\tilde{\delta}(\Delta)]$.

Now, following the method in the proof of Van der Vaart (1998), Proposition 15.2, take some $b > 0$ and consider the test $H_0 : h = h_0$ against $H_1 : h = h_1(b, h_0)$ based on $\Delta \stackrel{h}{\sim} N(h, I_0^{-1})$. Note that $\dot{g}'h_1 = b > 0$. The likelihood ratio is:

$$LR = \frac{dN(h_1, I_0^{-1})}{dN(h_0, I_0^{-1})} = \exp\left(\frac{b}{\dot{g}'I_0^{-1}\dot{g}}\dot{g}'\Delta - \frac{b^2}{2\dot{g}'I_0^{-1}\dot{g}}\right).$$

By the Neyman-Pearson lemma, a most powerful test is based on rejecting for large values of $\dot{g}'\Delta$. Since the test δ_c has been defined to have the same size as $\tilde{\delta}$, $E_{h_1(b, h_0)}[\delta_c(\Delta)] \geq E_{h_1(b, h_0)}[\tilde{\delta}(\Delta)]$. Moreover, this inequality similarly holds for all $b \geq 0$. Similarly, for $b < 0$, $1 - \delta_c = \mathbf{1}\{\dot{g}'\Delta \leq c\}$ is most powerful, leading to $E_{h_1(b, h_0)}[\delta_c(\Delta)] \leq E_{h_1(b, h_0)}[\tilde{\delta}(\Delta)]$ for all $b \leq 0$. Since $R(\tilde{\delta}, h) - R(\delta_c, h) = [L(1, h) - L(0, h)] \{E_h[\delta_1(\Delta)] - E_h[\delta_2(\Delta)]\}$, we can conclude that $R(\tilde{\delta}, h) \geq R(\delta_c, h)$ for all $h \in \{h_1(b, h_0) : b \in \mathbb{R}\}$.

For part (ii), let $R^* = \inf_{\delta \in \mathcal{D}_\infty} \sup_h R(\delta, h)$, and let δ^* be such that $\sup_h R(\delta^*, h) = R^*$. By part (i), there exists c^* such that $R(\delta^*, h_1(b, 0)) \geq R(\delta_{c^*}, h_1(b, 0))$ for all b . Note that $\dot{g}'\Delta \sim N(b, \dot{g}'I_0^{-1}\dot{g})$ under $h = h_1(b, h_0)$. Hence $E_{h_1(b, h_0)}[\delta_{c^*}(\Delta)] = E_{h_1(b, 0)}[\delta_{c^*}(\Delta)]$ for all h_0 with $\dot{g}'h_0 = 0$. Also, loss can be rewritten $L_g(a, \dot{g}'h) = L(a, h)$ for $a \in \{0, 1\}$. Recalling that $b = \dot{g}'h_1(b, h_0) = \dot{g}'h_1(b, 0)$, we have

$$R(\delta_{c^*}, h_1(b, h_0)) = L_g(0, b) + E_{h_1(b, 0)}[\delta_{c^*}(\Delta)][L_g(1, b) - L_g(0, b)] = R(\delta_{c^*}, h_1(b, 0)).$$

Then,

$$R^* \geq \sup_b R(\delta^*, h_1(b, 0)) \geq \sup_b R(\delta_{c^*}, h_1(b, 0)) = \sup_{h_0} \sup_b R(\delta_{c^*}, h_1(b, h_0)) = \sup_h R(\delta_{c^*}, h) \geq R^*.$$

So, δ_{c^*} attains the bound and must be a solution to $\inf_c \sup_b R(\delta_c, h_1(b, 0))$. \square

Lemma 4 *Suppose the conditions of Proposition 3.1 are satisfied and $\delta_n \in \mathcal{D}$ is a sequence of treatment assignment rules with matching rule δ in the limit experiment as given by Proposition 3.1. Let J be a finite subset of \mathbb{R}^k . If*

$$\begin{aligned} \lim_{n \rightarrow \infty} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) &= R_\infty(\delta, h) \\ \left[\liminf_{n \rightarrow \infty} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \geq R_\infty(\delta, h) \right] \end{aligned} \quad (4.4)$$

for $h \in J$, then

$$\liminf_{n \rightarrow \infty} \sup_{h \in J} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) = [\geq] \sup_{h \in J} R_\infty(\delta, h).$$

Furthermore, if (4.4) holds for all $h \in \mathbb{R}^k$, then

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) = [\geq] \sup_h R_\infty(\delta, h) \quad (4.5)$$

where the outer supremum is taken over all finite subsets of \mathbb{R}^k .

Proof of Lemma 4:

Fix a finite subset J . Then,

$$\liminf_{n \rightarrow \infty} \sup_{h \in J} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \geq \sup_{h \in J} \liminf_{n \rightarrow \infty} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) = \sup_{h \in J} \lim_{n \rightarrow \infty} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) = \sup_{h \in J} R_\infty(\delta, h)$$

$$\left[\liminf_{n \rightarrow \infty} \sup_{h \in J} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \geq \sup_{h \in J} \liminf_{n \rightarrow \infty} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \geq \sup_{h \in J} R_\infty(\delta, h) \right].$$

The bracketed inequality in (4.5) follows trivially from the above bracketed expression. Now we show that the equality in (4.5) holds. By the supposition of the lemma, take $\varepsilon > 0$ and any $h \in J$,

there exists N_h such that for $n \geq N_h$, $r_n R(\delta_n, \theta_0 + h/\sqrt{n}) \leq R_\infty(\delta, h) + \varepsilon \leq \sup_{h' \in J} R_\infty(\delta, h') + \varepsilon$. Let $N = \max_{h \in J} N_h$. Then, for $n \geq N$,

$$\sup_{h \in J} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \leq \sup_{h' \in J} R_\infty(\delta, h') + \varepsilon$$

and

$$\liminf_{n \rightarrow \infty} \sup_{h \in J} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \leq \sup_{h \in J} R_\infty(\delta, h) + \varepsilon.$$

Since this holds for any $\varepsilon > 0$, we have $\liminf_{n \rightarrow \infty} \sup_{h \in J} r_n R(\delta_n, \theta_0 + h/\sqrt{n}) = \sup_{h \in J} R_\infty(\delta, h)$. So,

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} r_n R(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) = \sup_J \sup_{h \in J} R_\infty(\delta, h) = \sup_h R_\infty(\delta, h).$$

□

Let $\bar{\delta}_c(\Delta) = \mathbf{1} \left\{ \frac{\dot{g}'\Delta}{\sigma_g} \geq c \right\}$.

Lemma 5 *The solutions to*

$$\inf_c \sup_h R_\infty^j(\bar{\delta}_c, h)$$

for $j = H, T$ are the constants $c^{j,M}$.

Proof of Lemma 5:

Let $b = \dot{g}'h/\sigma_g$. Then, $R_\infty^H(\bar{\delta}_c, h) = R_g^H(\bar{\delta}_c, \sigma_g b) = \Phi(c - b)\mathbf{1}(b > 0) + K(1 - \Phi(c - b))\mathbf{1}(b \leq 0)$. So, $\sup_h R_\infty^H(\bar{\delta}_c, h) = \max\{\Phi(c), K(1 - \Phi(c))\}$. The terms $\Phi(c)$ and $K(1 - \Phi(c))$ are strictly increasing and decreasing in c , and cross at a unique point, which must minimize maximum risk. The crossing point is the solution given in the conclusion of the lemma for loss L^H .

For loss L^T , we can treat $\dot{g}'\Delta$ as the scalar observable and the cut-off value given in the lemma is derived in Tetenov (2007). We note here that the solution is well-behaved. Let $r^+(c, b) = b\Phi(c - b)$, and $r^-(c, b) = -Kb\Phi(b - c)$. Then, $\sup_b R_g^T(\bar{\delta}_c, \sigma_g b) = \sigma_g \cdot \max\{\sup_{b > 0} r^+(c, b), \sup_{b \leq 0} r^-(c, b)\}$. From the first and second derivatives of r^+ in b (for any fixed finite value of c), it is straightforward to show that r^+ is single-peaked with a unique, finite global maximum on $[0, \infty)$. The same conclusion is true of r^- on $(-\infty, 0]$. Also, $\sup_{b > 0} r^+(c, b)$ is strictly increasing in c and $\sup_{b \leq 0} r^-(c, b)$ is strictly decreasing. They cross at the unique value $c^{T,M}$ which must minimize the maximum risk. □

Proof of Theorem 3.5:

From Theorem 3.4(ii), it suffices to look at cut-off rules along a “slice” $h_1(b, 0)$ to obtain the minmax rule in the limit experiment. Note that the class of rules δ_c and $\bar{\delta}_c$ are equivalent and so it suffices to consider $\bar{\delta}_c$. Lemma 5 provides the minmax rules for losses L^H and L^T in the limit experiment.

Given a sequence of rules δ_n and a matching rule $\bar{\delta}$ in the limit experiment, $\lim_{n \rightarrow \infty} r_n R^T(\delta_n, \theta_0 + h/\sqrt{n}) = R_\infty^T(\bar{\delta}, h)$. Also, by Lemma 3, $\delta_n^{T,M}$ is matched in the limit experiment by $\delta^{T,M}$. Lemma 4 states that the risk bound in the limit experiment is the asymptotic risk bound and that it is attained by $\delta_n^{T,M}$.

For loss L^H , for h such that $\dot{g}'h \neq 0$, $\lim_{n \rightarrow \infty} r_n R^H(\delta_n, \theta_0 + h/\sqrt{n}) = R_\infty^H(\bar{\delta}, h)$. For h such that $\dot{g}'h = 0$, $R_\infty^H(\bar{\delta}, h) = 0$, so for all h , $\liminf_{n \rightarrow \infty} r_n R^H(\delta_n, \theta_0 + h/\sqrt{n}) \geq R_\infty^H(\bar{\delta}, h)$. Hence, by Lemma 4,

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} \sqrt{n} R^H(\delta_n, \theta_0 + \frac{h}{\sqrt{n}}) \geq \sup_h R_\infty^H(\bar{\delta}, h) \geq \inf_{\delta \in \mathcal{D}_\infty} \sup_h R_\infty^H(\delta, h) = \sup_h R_\infty^H(\delta^{H,M}, h)$$

where the outer supremum is taken over all finite subsets of \mathbb{R}^k . Also, by Lemma 3, $\delta_n^{H,M}$ is matched in the limit experiment by $\delta^{H,M}$.

Let \tilde{J} be any finite subset such that $\dot{g}'h \neq 0$ for $h \in \tilde{J}$. By Lemma 4,

$$\sup_{\tilde{J}} \liminf_{n \rightarrow \infty} \sup_{h \in \tilde{J}} R^H(\delta_n^{H,M}, \theta_0 + \frac{h}{\sqrt{n}}) = \sup_{\tilde{J}} \sup_{h \in \tilde{J}} R_\infty^H(\delta^{H,M}, h) \leq \sup_h R_\infty^H(\delta^{H,M}, h) = \inf_{\delta \in \mathcal{D}_\infty} \sup_h R_\infty^H(\delta, h). \quad (4.6)$$

Next, take $\varepsilon > 0$, we will show that

$$\sup_J \liminf_{n \rightarrow \infty} \sup_{h \in J} R^H(\delta_n^{H,M}, \theta_0 + \frac{h}{\sqrt{n}}) \leq \sup_{\tilde{J}} \liminf_{n \rightarrow \infty} \sup_{h \in \tilde{J}} R^H(\delta_n^{H,M}, \theta_0 + \frac{h}{\sqrt{n}}) + \varepsilon. \quad (4.7)$$

Take a finite subset $J \subset \mathbb{R}^k$ such that for exactly one element $h_0 \in J$, $\dot{g}'h_0 = 0$. For $\tau > 0$, define $h' = h_0 + \tau\dot{g}$ and $h'' = h_0 - \tau\dot{g}$. Note that $\dot{g}'h' > 0$ and $\dot{g}'h'' < 0$ and by continuity of $E_h(\delta^*)$ in h we can choose τ small enough that $|E_{h_0}(\delta^*) - E_{h'}(\delta^*)| < \varepsilon$ and $|E_{h_0}(\delta^*) - E_{h''}(\delta^*)| < \varepsilon/K$. Take $\tilde{J} = (J \setminus h_0) \cup \{h', h''\}$.

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \sup_{h \in J} R^H(\delta_n^{H,M}, \theta_0 + \frac{h}{\sqrt{n}}) \\ & \leq \liminf_{n \rightarrow \infty} \left[\max \left\{ \sup_{h \in J \setminus h_0} R^H(\delta_n^{H,M}, \theta_0 + \frac{h}{\sqrt{n}}), (1 - E_{h_0}(\delta_n^{H,M})), K E_{h_0}(\delta_n^{H,M}) \right\} \right] \\ & = \max \left\{ \sup_{h \in J \setminus h_0} R_\infty^H(\delta^{H,M}, h), (1 - E_{h_0}(\delta^{H,M})), K E_{h_0}(\delta^{H,M}) \right\} \\ & \leq \max \left\{ \sup_{h \in J \setminus h_0} R_\infty^H(\delta^{H,M}, h), (1 - E_{h'}(\delta^{H,M})), K E_{h''}(\delta^{H,M}) \right\} + \varepsilon \\ & = \sup_{h \in \tilde{J}} R_\infty^H(\delta^{H,M}, h) + \varepsilon = \liminf_{n \rightarrow \infty} \sup_{h \in \tilde{J}} R^H(\delta_n^{H,M}, \theta_0 + \frac{h}{\sqrt{n}}) + \varepsilon \end{aligned}$$

where the first equality follows by the same argument for the proof of the first conclusion of Lemma 4. This argument clearly generalizes to finite subsets J with more than one element h with $\dot{g}'h = 0$. Hence Equation (4.7) holds. Together with Equation (4.6) and the fact that ε was arbitrary, it follows that $\delta_n^{H,M}$ attains the risk bound. \square

Appendix B: Proofs for Semiparametric Case

Most of the proofs follow by obvious modification of the proofs for the parametric results in Appendix A. Nontrivial modifications are noted below.

Proof of Proposition 4.1:

Let $\Delta = (\Delta_1, \Delta_2, \dots)$. The limit experiment and the asymptotic representation theorem (Theorem 3.1) given in Van der Vaart (1991a) yield a randomized statistic $\tilde{\delta}(\Delta, U)$ that matches the limit distribution of δ_n , where U is uniform on $[0, 1]$ independent of Δ . (This is a “doubly randomized” treatment assignment rule.) The desired rule comes from setting $\delta(\Delta) = \int_0^1 \tilde{\delta}(\Delta, u) du$. \square

Proof of Theorem 4.2:

The analog to Lemma 2 follows by the assumed product measure form of the prior Π . An analog to Lemma 3 follows below. If Assumption 1 is modified to require the analogous conditions hold for almost every h with respect to $\lambda \times \rho$, then the analog to Lemma 1 holds with $B_\infty(\cdot, \Pi)$ replacing $\pi(\theta_0)B_\infty(\cdot)$. For loss L^T , the conditions of Assumption 1 hold for all h , so the modification entails no additional complications. For loss L^H , the conditions of Assumption 1 hold for all (h_2, h_3, \dots) and almost every h_1 (with respect to the Lebesgue measure λ). Hence, Lemma 1 also extends to L^H in the semiparametric case, and the conclusions of the theorem follow. \square

Lemma 3' *Suppose the conditions of Theorem 4.2 are satisfied. Then, for every h ,*

$$\frac{\sqrt{n}\hat{g}_n}{\hat{\sigma}_g} \overset{h}{\rightsquigarrow} N(\langle \dot{g}, h \rangle, 1).$$

Proof of Lemma 3':

We revert to treating h and \dot{g} as functions from \mathcal{Z} to \mathbb{R} . Equation (4.1) implies that

$$\ln \prod_{i=1}^n \frac{dP_{1/\sqrt{n}, h}}{dP_0}(Z_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h(Z_i) - \frac{1}{2} \|h\|^2 + o_{P_0}(1),$$

where $\frac{1}{\sqrt{n}} \sum_{i=1}^n h(Z_i) \overset{P_0}{\rightsquigarrow} N(0, \|h\|^2)$. By Van der Vaart (1998), Lemma 25.23, $\sqrt{n}\hat{g}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{g}(Z_i) + o_{P_0}(1)$. By Slutsky's Lemma and the Delta Method,

$$\begin{aligned} \left(\frac{\sqrt{n}\hat{g}_n}{\hat{\sigma}_g}, \ln \prod_{i=1}^n \frac{dP_{1/\sqrt{n}, h}}{dP_0}(Z_i) \right) &= \left(\frac{\sum_{i=1}^n \dot{g}(Z_i)}{\sigma_g \sqrt{n}}, \frac{1}{\sqrt{n}} \sum_{i=1}^n h(Z_i) - \frac{1}{2} \|h\|^2 \right) + o_{P_0}(1) \\ &\overset{P_0}{\rightsquigarrow} N \left(\begin{pmatrix} 0 \\ -\frac{1}{2} \|h\|^2 \end{pmatrix}, \begin{bmatrix} 1 & \frac{\langle \dot{g}, h \rangle}{\sigma_g} \\ \frac{\langle \dot{g}, h \rangle}{\sigma_g} & \|h\|^2 \end{bmatrix} \right). \end{aligned}$$

The conclusion then follows by applying Le Cam's Third Lemma. \square

Proof of Theorem 4.4:

Note that analogs of Lemmas 4 and 5 follow for the semiparametric case with trivial modification of their proofs. The proof of Theorem 4.4 follows from the proof of Theorem 3.5 after letting $h_0 = (0, h_2, h_3, \dots)$, $h' = (\tau, h_2, h_3, \dots)$, and $h'' = (-\tau, h_2, h_3, \dots)$. \square

References

- ANDREWS, D. W. K., AND W. PLOBERGER (1994): “Optimal Tests when a Nuisance Parameter is Present Only Under an Alternative,” *Econometrica*, 62(6), 1383–1414.
- BERGER, M. C., D. BLACK, AND J. SMITH (2001): “Evaluating Profiling as a Means of Allocating Government Services,” in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 59–84. Physica Heidelberg.
- BICKEL, P. J., C. A. KLAASEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.
- BLACK, D., J. SMITH, M. BERGER, AND B. NOEL (2003): “Is the Threat of Training More Effective than Training Itself? Experimental Evidence from the UI System,” *American Economic Review*, 93(4), 1313–1327.
- DEHEJIA, R. (2003): “When is ATE Enough? Risk Aversion and Inequality Aversion in Evaluating Training Programs,” working paper, Columbia University.
- DEHEJIA, R. H. (2005): “Program Evaluation as a Decision Problem,” *Journal of Econometrics*, 125, 141–173.
- DIACONIS, P., AND D. FREEDMAN (1986): “On the Consistency of Bayes Estimates,” *The Annals of Statistics*, 14(1), 1–26.
- KARLIN, S., AND H. RUBIN (1956): “The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio,” *Annals of Mathematical Statistics*, 27, 272–299.
- LE CAM, L. (1986): *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.
- 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, 415–442.
- (2002): “Treatment Choice Under Ambiguity Induced by Inferential Problems,” *Journal of Statistical Planning and Inference*, 105, 67–82.
- (2003): “Statistical Treatment Rules for Heterogeneous Populations,” working paper, Northwestern University.
- (2004): “Statistical Treatment Rules for Heterogeneous Populations,” *Econometrica*, 72(4), 1221–1246.

- O'LEARY, C. J., P. T. DECKER, AND S. A. WANDNER (1998): "Reemployment Bonuses and Profiling," Discussion Paper 98-51, W. E. Upjohn Institute for Employment Research.
- (2005): "Cost-Effectiveness of Targeted Reemployment Bonuses," *The Journal of Human Resources*, 40(1), 270–279.
- PUHALSKII, A., AND V. SPOKOINY (1998): "On Large Deviation Efficiency in Statistical Inference," *Bernoulli*, 4(2), 203–272.
- SAVAGE, L. (1951): "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 46, 55–67.
- SCHLAG, K. (2006): "Eleven – Tests Needed for a Recommendation," EUI Working Paper ECO 2006-2.
- STOYE, J. (2006): "Minimax Regret Treatment Choice with Finite Samples," Working paper.
- TETENOV, A. (2007): "Statistical Treatment Choice Based on Asymmetric Minmax Regret Criteria," Working paper.
- VAN DER VAART, A. W. (1991a): "An Asymptotic Representation Theorem," *International Statistical Review*, 59, 99–121.
- (1991b): "On Differentiable Functionals," *The Annals of Statistics*, 19, 178–204.
- (1998): *Asymptotic Statistics*. Cambridge University Press, New York.
- WORDEN, K. (1993): "Profiling Dislocated Workers for Early Referral to Reemployment Services," Unpublished manuscript, U.S. Department of Labor.