We extend the Chernozhukov and Hong (2003) Laplace estimation procedure to Kim and Pollard (1990)–type objective functions (e.g. the maximum score estimator (Manski, 1975). Scaling the nonregular objective functions by a sample–size \( (n)–dependent \) input parameter \( \alpha_n^2 \), we establish that the Laplace estimation procedure, given sufficient smoothness, can improve the \( \sqrt{n} \)–convergence rate of Kim and Pollard (1990) to a rate arbitrarily close to \( \sqrt{n} \).

We further show that the proposed estimator has three different types of limiting distribution, depending on the rate at which \( \alpha_n \) diverges. We provide a simple–to–implement uniform inference method which yields (asymptotically) correct inference irrespective of which of the three types applies. Generally, a slower rate of increase of \( \alpha_n \) translates to a faster convergence rate for our estimator \( \hat{\theta} \), albeit that the convergence rate of \( \hat{\theta} \) is never worse than \( \frac{1}{\sqrt{n}} \) and that to achieve faster than \( n^{2/5} \)–convergence a bias–correction procedure is needed. We provide two such procedures: the bias can be estimated and subtracted or one can use a special ‘prior’ resembling Jeffreys’ (1946) prior. Computation is straightforward and can typically be accomplished using Gibbs sampling (Geman and Geman, 1984). A limited simulation study yields encouraging results.

\footnote{We thank the Human Capital Foundation for their support of CAPCP. We thank Jeremy Fox, Bo Honoré, Joel Horowitz, Roger Koenker, Oliver Linton, Haiqing Xu, and Vicky Zinde–Walsh for helpful suggestions.}

\footnote{sjun@psu.edu (corresponding author), 608 Kern Graduate Building, University Park, PA 16802}

\footnote{joris@psu.edu}

\footnote{yxw162@psu.edu}
1. Motivation

We extend the Chernozhukov and Hong (2003) Laplace estimation procedure to Kim and Pollard (1990)–type objective functions (e.g. the maximum score estimator (Manski, 1975). Scaling the non-regular objective functions by a sample-size ($n$)–dependent input parameter $\alpha n^2$, we establish that the Laplace estimation procedure, given sufficient smoothness, can improve the $\sqrt{n}$–convergence rate of Kim and Pollard (1990) to a rate arbitrarily close to $\sqrt{n}$. We further show that the proposed estimator has three different types of limiting distribution, depending on the rate at which $\alpha n$ diverges. We provide a simple–to–implement uniform inference method which yields (asymptotically) correct inference irrespective of which of the three types applies. Generally, a slower rate of increase of $\alpha n$ translates to a faster convergence rate for our estimator $\hat{\theta}$, albeit that the convergence rate of $\hat{\theta}$ is never worse than $\sqrt{n}$ and that to achieve faster than $n^{2/5}$–convergence a bias–correction procedure is needed. We provide two such procedures: the bias can be estimated and subtracted or one can use a special ‘prior’ resembling Jeffreys’ (1946) prior. Computation is straightforward and can typically be accomplished using Gibbs sampling (Geman and Geman, 1984). A limited simulation study yields encouraging results.

Chernozhukov and Hong (2003) proposed integrating instead of optimizing the (exponential of an) objective function of extremum estimators to obtain an estimator of an unknown parameter vector $\theta_0$; they labelled their estimator a Laplace estimator. The objective in Chernozhukov and Hong (2003) is primarily to facilitate the computation of some difficult–to–compute estimators by using the Markov Chain Monte Carlo (MCMC) method. Under the assumption that the objective function of interest admits a classical quadratic expansion, Chernozhukov and Hong (2003) showed that their Laplace–type estimators are generally $\sqrt{n}$–consistent and have a limiting normal distribution. One version of their estimator can be interpreted as the ‘posterior’ mean based on a ‘prior’ and a ‘pseudo–likelihood,’ where the latter is formed using the extremum estimator objective function of interest. Chernozhukov and Hong (2003) derived conditions under which their estimator is efficient and in which its limiting distribution coincides with the (pseudo) posterior distribution.

The results of Chernozhukov and Hong (2003) do not extend to the class of estimators considered by Kim and Pollard (1990) because the Kim and Pollard objective functions do not admit the quadratic expansion needed. Computation of this class of estimators and their confidence regions

---

1Despite the use of terminology which has a Bayesian ring to it, the Chernozhukov and Hong (2003) procedure — and indeed ours — is entirely classical.
can be cumbersome (see e.g. Manski and Thompson, 1986; Pinkse, 1993; Florios and Skouras, 2008), so a Laplace–type procedure would be valuable.

The extension of the Laplace–procedure to Kim and Pollard (1990)–type objective functions is not merely a generalization of the Chernozhukov and Hong (2003) conditions. Indeed, whereas for the class of estimators admitting the standard quadratic expansion, the estimator is always $\sqrt{n}$–consistent and the limiting distribution is always normal regardless of the scaling of the objective function, in the case studied in this paper both the convergence rate of our estimator and the nature of the limiting distribution depend on the input parameter $\alpha_n^2$ that scales the objective function.

If $\alpha_n$ diverges at a rate no slower than $\sqrt{n}$ then our estimator $\hat{\theta}$ is $\sqrt{n}$–convergent. If $\alpha_n$ diverges faster than $\sqrt{n}$ then $\hat{\theta}$ moreover has the same limiting distribution as the corresponding Kim and Pollard (1990) extremum estimator even though our estimator is always a singleton and the Kim and Pollard (1990) estimator can be set–valued as in e.g. the case of the maximum score estimator. When $\alpha_n$ increases at the rate of $\sqrt{n}$, the limiting distribution of $\hat{\theta}$ is characterized by the ratio of two integrals of a certain Gaussian process. The third type of limiting distribution, namely the normal, arises when $\alpha_n$ increases at a rate slower than $\sqrt{n}$. Indeed, we show that subject to sufficient smoothness the convergence rate of $\hat{\theta}$ can be as good as $\sqrt{n}/\alpha_n$; since $\alpha_n$ must increase to infinity with $n$, a convergence rate of $\sqrt{n}$ is not achievable. If $\alpha_n$ increases more slowly (i.e. no faster than $\sqrt{n}$) then asymptotic bias becomes an issue; the asymptotic bias is discussed further below.

This trichotomy of limiting distributions is interesting, but since in practice one only chooses a value of $\alpha_n$, not a rate, deciding which limit distribution to use is problematic. So, we develop a simple–to–execute simulation–based inference procedure which automatically adapts to the correct limit distribution.\footnote{Directly using the quantiles of the pseudo–posterior to conduct inference does not appear to be feasible in our case, although Chernozhukov and Hong (2003) derived conditions under which this can be done in their, regular, case.} So inference is uniform in the choice of $\alpha_n$; this result is at the heart of this paper.

Uniform inference procedures exist in numerous other environments, including weak identification (Staiger and Stock, 1997), roots near unity (Mikusheva, 2007), subsampling (Andrews and Guggenberger, 2008), kernel estimation (Guerre and Lavergne, 2005), HAC estimation (Kiefer and Vogelsang, 2005), and average derivative estimation (Cattaneo, Crump, and Jansson, 2008). Whereas in Staiger and Stock (1997) uniformity is achieved in a nuisance parameter and in Mikusheva (2007) in the parameter of interest, here it is achieved in a sample–size–dependent input parameter $\alpha_n$. In contrast to e.g. Guerre and Lavergne (2005), in our case uniformity is in the rate of the input parameter instead of a constant multiplying the rate. Like Mikusheva (2007), but unlike e.g. Cattaneo.
Crump, and Jansson (2008), our limiting distributions differ not just in parameter values, but also in type, and here there are three distinct types instead of two as in Mikusheva (2007).

As we mentioned earlier, a slow rate of increase of $\alpha_n$ (i.e. no faster than $\sqrt{n}$) introduces an asymptotic bias issue. We provide two methods for removing the asymptotic bias, namely to subtract an estimate of the bias and to use a special prior. Indeed, we show that using the equivalent of Jeffreys' (1946) prior in the current context, or an estimated version thereof, removes the $n^{-2/5}$ order bias. This is the only instance that we are aware of in which the choice of prior matters in large samples when there is point–identification.\footnote{When the model is not fully identified, the choice of prior generally matters in large samples in Bayesian analysis (e.g. Moon and Schorfheide, 2009).

For consistency, $\alpha_n$ must increase to infinity with the sample size; for consistency the rate at which $\alpha_n$ increases is immaterial. If one were to let $\alpha_n$ decrease to zero, however, then our estimator would converge to the mean of the prior, which would also be true for the regular Chernozhukov and Hong (2003) estimator. A consequence of this is that for small choices of $\alpha_n$, the estimator is biased towards the mean of the prior and that this type of bias cannot be corrected using asymptotic methods. A similar issue arises in kernel regression estimation if one lets the bandwidth go to infinity. We investigate both types of bias in our simulation study.

We illustrate our methodology using the maximum score estimator (Manski, 1975, MSE) as a leading example. We emphasize the binary choice version of the MSE, albeit that the discussion applies equally to the ordered response and multinomial choice cases (Lee, 1992; Manski, 1975).

The MSE is $\sqrt{n}$–consistent for the vector of regression coefficients in a binary choice model using little more than a conditional median restriction on the errors in the latent variable equation. In particular, unlike other parametric estimators such as probit or logit, the error distribution need not be specified. Unlike other semiparametric single–index models (e.g. Klein and Spady, 1993), the error distribution need not be independent of the regressors, and heteroskedasticity of unknown form is permitted. Despite this desirable generality, the MSE has not been especially popular because of its computational difficulty, its slow convergence rate, and its nonstandard limiting distribution; we are aware of only a few empirical uses of the MSE and its generalization, including Bajari, Fox and Ryan (2008), Bajari and Fox (2009), Fox (2007, 2009).

Horowitz (1992) has shown that by replacing the maximum score objective function with a smoothed version thereof, a rate arbitrarily close to $\sqrt{n}$ is attainable with a limiting normal distribution, subject to sufficient smoothness. Subsequently de Jong and Woutersen (2007) and Kotlyaroa and Zinde-Walsh (2009) have extended the Horowitz results. Kordas (2006) has extended the
smoothed maximum score estimator (Horowitz [1992, SMS]) to allow for quantiles other than the median and his procedure was implemented empirically by Belluzo (2004); using quantiles other than the median is also possible with the original MSE and the estimator proposed in this paper. Although the best attainable convergence rate of the SMS is the same as ours under similar conditions, our estimator is entirely different from Horowitz’s, our estimator is not specific to the maximum score case, due to the nonconcavity of the objective function the SMS estimator is more difficult to compute than ours, and no uniform inference procedures are available for it.

One problem with the SMS is that the choice of bandwidth is determined by the unknown degree of smoothness. If the degree of smoothness used to determine the bandwidth differs from the true degree of smoothness then the convergence rate of the SMS is suboptimal. Indeed, Pollard (1993) has shown that \( \sqrt[n]{\cdot} \) is the best rate that can be achieved if only the smoothness conditions for the MSE are satisfied and that the SMS will then have a bias term which vanishes more slowly than \( \sqrt[n]{\cdot} \). The same problem arises for our estimator and the choice of \( \alpha_n \). Kotlyarova and Zinde-Walsh (2006, KZ) proposed an estimation procedure which automatically adapts to the unknown degree of smoothness. The advantage of our estimator over the SMS estimator, when both are combined with a KZ–like procedure, is the uniformity of our inference method across input parameter–values including the \( \sqrt[n]{\cdot} \)–convergence case. For example, with a KZ–like estimation method, our inference procedure (with minor adaptations) could accommodate the possibility that only the smoothness conditions for the MSE are satisfied. It is difficult to see how — absent a uniform inference procedure — one could accomplish this for the SMS.

Horowitz (2002) established that the bootstrap (Efron and Tibshirani, 1997) offers an asymptotic refinement for the SMS estimator. In order for such a refinement to obtain, the rate at which the bandwidth tends to zero is different from the one resulting in the optimal convergence rate of the estimator.\(^4\) We have made a preliminary investigation on the bootstrap in our context, but intend to establish rigorous results for a procedure providing asymptotic refinements in the future. On the basis of our preliminary work, we conclude that the bootstrap is inconsistent if \( \alpha_n \) increases no slower than \( \sqrt[n]{\cdot} \) for much the same reasons that the bootstrap is inconsistent for the regular maximum score estimator (Abrevaya and Huang, 2005).\(^2\) We further conclude that if \( \alpha_n \) increases more slowly than \( \sqrt[n]{\cdot} \), then the bootstrap will be consistent. Moreover, asymptotic refinements will

\(^4\)This is typical for nonparametric estimators, but an unattractive consequence is that the ratio of the width of the ‘regular’ confidence interval to that of the bootstrap confidence interval decreases to zero as the sample size tends to infinity.

\(^5\)Subsampling (Politis, Romano, and Wolf, 1999) has been shown to be consistent for a class of \( \sqrt[n]{\cdot} \)-consistent estimators by Delgado, Rodriguez-Poo, and Wolf (2001), but is less attractive than the bootstrap for reasons of efficiency.
obtain provided that there is sufficient smoothness that is not exploited to optimize the convergence rate of the estimator.

The bootstrap and the uniform inference procedure have different goals: with the bootstrap the actual coverage probability of the confidence interval converges faster to the nominal one if $\alpha_n$ increases sufficiently slowly and there is ‘surplus’ smoothness; the uniform inference procedure provides robustness to the choice of $\alpha_n$. These two goals are mutually exclusive in the current context (and for KP estimators generally) since the bootstrap is inconsistent if $\alpha_n$ increases no slower than $\sqrt{n}$. We focus on the uniform inference procedure here because, in view of the two discontinuities in the limiting distribution as a function of the rate of increase of $\alpha_n$, we believe it is the more serious of the two problems these procedures address.

As noted at the beginning of this section, for the maximum score case our estimator can be computed using Gibbs sampling. It turns out that computation is fast, simple, and accurate. If one chooses a prior for which a closed form solution exists for the integrated prior of each coefficient conditional on the others then computation of our estimator involves nothing more complicated than drawing uniform random numbers, sorting, and averaging. Likewise, irrespective of the type of Kim and Pollard (1990) estimator, uniform inference requires little more than drawing random numbers from a multivariate normal.

We study the properties of our estimator in a limited simulation study. The behavior of the proposed estimator reflects what one would expect on the basis of the theory. First, there is a tradeoff between bias and variance, with higher values of $\alpha_n$ resulting in less bias, but more variance. Further, the asymptotic bias can be corrected using Jeffreys’ prior but if $\alpha_n$ is chosen very small then the estimator is biased towards the mean of the prior. Finally, the uniform inference procedure moves smoothly from the limiting distribution of the maximum score estimator to the normal with the value of $\alpha_n$.

There are issues of potential interest that are not studied in this paper in addition to those, like the bootstrap, that were discussed above. First, one could look at statistics other than the posterior mean, as do Chernozhukov and Hong (2003). Further, using our methodology for the least median of squares (LMS) estimator of Rousseeuw (1984) requires a nuisance parameter problem to be addressed and its breakdown point to be determined; see Zinde-Walsh (2002) for an adaptation of the SMS idea to the LMS case.

The paper is organized as follows. In section 2 we discuss our estimation method and derive convergence results in section 3, which also includes a discussion of both of our bias correction
methods. Our uniform inference procedure is described in section 4. Finally, the simulation study is contained in section 5.

2. Estimation Method

2.1. Set Up. Let $L_n(\theta) = n^{-1} \sum_{i=1}^{n} g_i(\theta)$ be the (renormed) objective function of a $\sqrt{n}$-consistent estimator of an unknown parameter vector $\theta_0 \in \Theta \subset \mathbb{R}^d$. The functions $g_i = g(\xi_i, \cdot)$ are defined such that $g_i(\theta_0) = 0$ a.s.. For the specific case of the maximum score estimator with regressor vector $x_i = [a_i, z_i^\top]^\top$ where the coefficient on $a_i$ is normalized to equal minus one we get for $\xi_i = [y_i, x_i^\top]^\top$,

$$g_i(\theta) = (2y_i - 1) \left( I(a_i \leq z_i^\top \theta) - I(a_i \leq z_i^\top \theta_0) \right),$$

but the results below apply to general $g_i$, provided that our assumptions are satisfied.

We consider Laplace–type estimators of the form

$$\hat{\theta} = \frac{\int \theta \pi(\theta) \exp (\alpha_n^2 L_n(\theta)) d\theta}{\int \pi(\theta) \exp (\alpha_n^2 L_n(\theta)) d\theta},$$

where $\{\alpha_n\}$ is some sequence for which $\alpha_n \to \infty$ as $n \to \infty$ we will call $\pi$ a prior, even though for most of our results we do not require it to be nonnegative. If the prior is nonnegative everywhere, then $\hat{\theta}$ can be interpreted as the mean of a posterior distribution; we will use the term ‘posterior’ regardless of whether the prior is nonnegative. If $L_n$ were the objective function of a $\sqrt{n}$-consistent estimator instead of the one considered here and $\alpha_n = \sqrt{n}$, we would have the Laplace estimator of Chernozhukov and Hong (2003).

Provided that $\theta_0$ is a unique maximizer of $Q(\theta) = \mathbb{E}[g_i(\theta)]$, consistency is straightforward to establish; see e.g. Robert and Casella (2004), corollary 5.11. The purpose of this paper is to study the effect of the choice of $\{\alpha_n\}$ on the asymptotic properties of $\hat{\theta}$. If $\alpha_n$ increases faster than $\sqrt{n}$ then it turns out that $\hat{\theta}$ is asymptotically equivalent to the estimator maximizing $L_n$, for which Kim and Pollard (1990) derived limit results. Although we emphasize the case in which $\alpha_n$ increases no faster than $\sqrt{n}$, we provide results for $\alpha_n$ that increase faster, also. In subsequent sections we show that the best achievable convergence rate (given sufficient smoothness) is $\kappa_n = \max(\sqrt{n}, \sqrt{n/\alpha_n})$, albeit that to achieve a convergence rate better than $n^{2/5}$ requires bias correction; this is in line with the properties derived in Horowitz (1992) for the SMS estimator.

6In Chernozhukov and Hong (2003) $\alpha_n = \sqrt{n}$, but its choice does not affect the limiting distribution provided that $\alpha_n$ diverges to infinity.

7For the case in which $L_n$ is the maximum score objective function, if $\alpha_n$ increases faster than $\sqrt{n}$ (not $\sqrt{n}$), $\hat{\theta}$ is for large $n$ in fact a prior–weighted average over the maximum score estimator, i.e. the (possibly noncontiguous) set of values maximizing the maximum score objective function.
2.2. Intuition for our results. We now provide some intuition for our results when \( \alpha_n \) increases no faster than \( \sqrt[7]{n} \), but please note that our results also cover the case in which \( \alpha_n \) increases faster than that. Let \( S_n(\theta) = n^{-1} \sum_{i=1}^{n} \tilde{g}_i(\theta) \) with \( \tilde{g}_i(\theta) = g_i(\theta) - Q(\theta) \). Then (2) can be rewritten as

\[
\theta = \int \frac{\theta \pi(\theta) \exp(\alpha_n^2 S_n(\theta) + \alpha_n^2 Q(\theta))}{\pi(\theta) \exp(\alpha_n^2 S_n(\theta) + \alpha_n^2 Q(\theta))} d\theta.
\]

Let \( \beta_n = \sqrt{\alpha_n^2 / n} \). By applying the substitution \( t = \alpha_n(\theta - \theta_0) \) to (3) we obtain

\[
\frac{n}{\alpha_n} (\hat{\theta} - \theta_0) = \frac{1}{\beta_n} \int t \pi_n(t) \exp(\beta_n S_n(t) + Q_n(t)) dt,
\]

where \( \pi_n(t) = \pi(\theta_0 + t/\alpha_n) \), \( Q_n(t) = \alpha_n^2 Q(\theta_0 + t/\alpha_n) \) and \( \tilde{S}_n(t) = \sqrt{n\alpha_n} S_n(\theta_0 + t/\alpha_n) \).

For large \( n \), \( Q_n(t) \approx -t^T V t / 2 \) for \( V = -Q''(\theta_0) \) and \( \pi_n(t) \approx \pi_0 \). Replacing \( Q_n, \pi_n \) in the right hand side of (4) with their respective approximations yields

\[
\frac{1}{\beta_n} \int t \exp(\beta_n \tilde{S}_n(t) - t^T V t / 2) dt = \frac{1}{\beta_n} \int t \exp(\beta_n \tilde{S}_n(t)) \phi_V(t) dt,
\]

where \( \phi_V(t) \) is the multivariate mean zero, variance \( V^{-1} \), normal density function.

We establish in an appendix that \( \tilde{S}_n \overset{\text{iid}}{\to} G \), where \( G \) is a tight Gaussian process defined on the entire Euclidean space. So if \( \alpha_n = c_3^2 \sqrt{n} \) then (5) converges in distribution to

\[
\frac{1}{c_3^2} \int t \exp(c_3^2 G(t)) \phi_V(t) dt,
\]

which is indeed the result of theorem (3) below. If \( \alpha_n = o(\sqrt{n}) \) then \( \beta_n = o(1) \) and the right hand side in (5) is approximately

\[
\int t \tilde{S}_n(t) \phi_V(t) dt \]

which has a limiting normal distribution since \( S_n \) is a renormed sample average of a sequence of i.i.d. mean zero variates. This is a result to be established in theorem (4) albeit that the result of theorem (4) contains a bias term \( \mathcal{B}_n \). This bias term arises from the approximation of \( Q_n(t) \) by \(-t^T V t / 2\). It is asymptotically negligible if \( \alpha_n \) diverges at a rate faster than \( \sqrt{n} \). If \( \alpha_n = c_3^2 \sqrt{n} \) then the limiting distribution will be a normal with nonzero mean, as shown in theorem (6) much like in the case of nonparametric kernel estimation of a function of a single argument and indeed like Horowitzi (1992). We derive an expansion for the bias in theorem (5) and show that — subject to additional smoothness conditions — the bias can be removed. A bias–corrected estimator then can have a convergence rate arbitrarily close to \( \sqrt{n} \).

---

\[\text{This approximation uses } \exp(\beta_n \tilde{S}_n(t)) \approx 1 + \beta_n \tilde{S}_n(t) \text{ and the fact that } \int \phi_V(t) dt = 1, \int t \phi_V(t) dt = 0.\]
So, letting \( \alpha_n \) increase at a rate slower than \( \sqrt[3]{n} \) implicitly smooths out the discontinuities in \( L_n \), much like the SMS estimator of [Horowitz (1992)] does explicitly.

3. Convergence Results

We now proceed to state our main results. We first state our main assumptions, followed by a discussion of \( \sqrt[3]{n} \)-consistent estimators, followed by a discussion of estimators that converge faster. We will use the maximum score case as an example to motivate the assumptions.

3.1. Assumptions. The first of our assumptions is standard in the literature and is also found in [Horowitz (1992)].

Assumption A. \( \theta_0 \) is in the interior of some compact set \( \Theta \).

Assumption B is a condition necessary for identification. Indeed, note that \( Q(\theta_0) = 0 \) by construction.

Assumption B. \( \forall \theta \in \Theta : \theta \neq \theta_0 \Rightarrow Q(\theta) < 0 \).

Let \( p(a,z) = P[y = 1|a = a,z = z] \) and \( f(a|z) \) denote the conditional density of \( a \) at \( a \) given \( z = z \). For his maximum score estimator [Manski (1985)] requires that the support of \( x \) is not contained in any proper linear subspace of \( \mathbb{R}^d \), that \( 0 < p(a,z) < 1 \) for almost all \( a,z \), and that for almost all \( z \), \( f(a|z) > 0 \) for all \( a \in \mathbb{R} \), which is sufficient for assumption B. Assumption B is less primitive than the identification conditions of [Manski (1985)], but it applies to all estimators satisfying our conditions.

Assumption C. \( \pi(\theta_0) \neq 0 \), \( \pi \) is bounded in absolute value on \( \Theta \) by \( \bar{\pi} \), zero outside of \( \Theta \), and integrates to one.

Since we can choose the prior, assumption C is innocuous. The same cannot be said for assumption D.

Assumption D. The function \( Q \) is continuous on \( \Theta \). Further, for some \( q \geq 0 \), \( Q \) is \( \Delta = q + 2 \) times continuously differentiable at \( \theta_0 \); \( \pi \) is \( q \) times continuously differentiable at \( \theta_0 \); \( V = -Q''(\theta_0) \) is positive definite.

\(^9\)If \( q = 0 \), \( \pi \) is merely assumed continuous at \( \theta_0 \).
Kim and Pollard (1990) derived the $\sqrt{n}$–limiting distribution of the MSE assuming $\Delta = 2$. The degree of smoothness of $Q$ depends on that of $p(\cdot|z)$ and $f(\cdot|z)$.

For the maximum score case, Horowitz (1992) obtained a $n^{2/3}$–consistent estimator subject to assumptions including that (i) $f(a|z)$ has a uniform upper bound in $a$ and (almost all) $z$, (ii) for almost all $z$, $f'(a|z)$ is continuous in $a$ in a neighborhood of $a = z^T\theta_0$ with a uniform bound over $z$, (iii) for almost all $z$, the second partial derivative of $p$ with respect to $a$, $p_{aa}(a,z)$, is a continuous function of $a$ in a neighborhood of $a = z^T\theta_0$, (iv) $E\|z_i\|^4 < \infty$, (v) $V$ is positive definite. To achieve the same convergence rate we need $\Delta = 3$ (see assumption $\mathcal{E}$ below), which implies the existence of three moments on $z_i$, two partial derivatives of $p$ with respect to $a$ at $z^T\theta_0$, and two derivatives of $f(\cdot|z)$ at $z^T\theta_0$. So the conditions are different from but similar to those in Horowitz (1992); neither set of assumptions implies the other and neither estimator yields a better convergence rate under the conditions of the other. The comparison between the conditions necessary for the two estimators to obtain a certain convergence rate is similar under additional smoothness. Since we accommodate estimation problems other than maximum score, our conditions are less primitive. Finally, lemma $G.1$ shows that for the maximum score case $V = -2E[z_iz_i^T p_{a}(z_i^T\theta_0, z_i)f(z_i^T\theta_0|z_i)]$, which by lemma $G.2$ is positive definite under weak conditions.

**Assumption E.** $E[\sup_{\theta \in \mathbb{R}^d} |g_i(\theta)|] < \infty$ and for some $i, i^* > 0$ with $i + i^* > 1$, some function $v$ for which $\int ||t||^2 + v(t)\phi(t)dt < \infty$, and all $t \in \mathbb{R}^d$, $\lim_{a \rightarrow \infty} a^rv \mathbb{E}|\tilde{g}_i(0 + t/a)||^2 + i \leq v(t)$.

Assumption $\mathcal{E}$ is trivially satisfied for the maximum score case since it only involves indicator functions. The second part of assumption $\mathcal{E}$ is implied by assumption $\mathcal{F}$ below if $g$ is a bounded function for $i = i^* = 1$ since the limit in assumption $\mathcal{E}$ is then bounded by a constant times $H(t,t)$. Assumption $\mathcal{F}$ is also used in Kim and Pollard (1990).

**Assumption F.** For all $t, s$, $H(t,s) = \lim_{a \rightarrow \infty} aE[g_1(t\theta_0 + t/a)g_1(s\theta_0 + s/a)]$ exists and is finite.\(^{11}\)

We show in lemma $G.3$ that for the maximum score case using the normalization adopted earlier, $H(t,s) = E[|M(z_t^Tt, z_s^Ts, 0)|f(z_t^T\theta_0|z_t)|f(z_s^T\theta_0|z_s)]$ with $M$ the median of its arguments. Assumption $G$ is needed to establish weak convergence of the process $\tilde{S}_n$ to $G$.

\(^{10}\)Horowitz’s conditions are phrased differently and use the conditional distribution function $F_{\delta(z)}$ of ‘latent variable equation’ errors $u_i$ given regressors. Then, $p(a, z) = 1 - F_{\delta(z)}(a - z^T\theta_0|a, z)$ so that the degree of smoothness of $p(\cdot|z)$ corresponds to that of $F_{\delta(z)}(\cdot|z)$, given that $F_{\delta(z)}(\cdot|z)$ is $\Delta$–times differentiable of $Q$ corresponds to $(\Delta + 1)$–times differentiability of $f(\cdot|z)$ and $F_{\delta(z)}(\cdot|z)$. See also assumptions 8 and 9 in Horowitz (1992).

\(^{11}\)Note that this implies that $\limsup_{\|t\| \rightarrow \infty} H(t,t)/\|t\| < \infty$.\(^{11}\)
**Assumption G.** Let \( \mathcal{F} = \{g(\cdot; \theta)\}_{\theta \in \mathbb{R}^d} \) and for 1 = o(\( \tilde{\alpha}_n \)), let \( \mathcal{F}_n = \{\sqrt{n}g(\cdot; \theta_0 + t/\tilde{\alpha}_n)\}_{t \in \mathcal{T}} \), where \( \mathcal{T} \) is an arbitrary compact subset of \( \mathbb{R}^d \). Then

(i) For all \( \xi \) in the support \( \Xi \) of \( \xi_i \), \( g(\xi, \theta) \) is right– (or left–) continuous at \( \theta_0 \).

(ii) There exists an envelope function \( F_n \) such that for all \( \xi \in \Xi \) \( \sup_{t \in \mathcal{T}} \sqrt{n}|g(\xi; \theta_0 + t/\tilde{\alpha}_n)| \leq F_n(\xi) \) and \( \mathbb{E}[F_n^2] = O(1) \) where \( F_n = F_n(\xi_i) \).

(iii) For any \( \epsilon > 0 \), \( \mathbb{E}[F_n^2I(F_n > \epsilon \sqrt{n})] = o(1) \).

(iv) For any \( \epsilon_n \downarrow 0 \), \( \sup_{||t-s|| < \epsilon_n} \tilde{\alpha}_n \mathbb{E} [\left( g_t(\theta_0 + t/\tilde{\alpha}_n) - g_t(\theta_0 + s/\tilde{\alpha}_n))^2 \right] = o(1) \).

(v) Let \( N(\epsilon, \mathcal{F}_n, L_2(\mathcal{P})) \) be the \( (L_2) \)–covering number for \( \mathcal{F}_n \) with respect to the probability measure \( \mathcal{P} \). Then for every \( \epsilon_n \downarrow 0 \), \( \sup_{\mathcal{D}} \int_{0}^{\epsilon_n} \sqrt{\log \left( N(\epsilon, \mathcal{F}_n, L_2(\mathcal{P})) \right)} \) \( \text{d} \epsilon = o(1) \).

Assumption G is common, but is not always straightforward to verify. It ensures weak convergence of \( \tilde{\mathcal{S}}_n \) to \( G \) and is satisfied if theorems 1 and 2 show.

**Theorem 1.** Under assumptions A−G \( \tilde{\mathcal{S}}_n \overset{w}{\rightarrow} G \) on \( L^\infty(\mathcal{F}_1, \mathcal{F}_2, \ldots) \) for any increasing sequence of compact sets \( \mathcal{F}_j \) such that \( 0 \in \mathcal{F}_1 \) and whose union is \( \mathbb{R}^d \), where \( L^\infty(\mathcal{F}_1, \mathcal{F}_2, \ldots) \) is the space of functions which are uniformly bounded on each \( \mathcal{F}_j \).

**Proof.** The proofs of all theorems are in appendix H.

**Theorem 2.** For the maximum score case, under assumptions A−E if \( \tilde{\alpha}_n = o(n) \) then assumption G is satisfied if \( \mathbb{E} [\sup_{\mathcal{D}_i} f ||s||_{z_i} ||z_i||] < \infty \).

3.2. **Cube–Root–n–Convergence.** We are now in a position to state our limit results. The first of these deals with the case \( \alpha_n = c_3^\sqrt{n} \). Let \( G \) be a mean zero Gaussian process on \( \mathbb{R}^d \) with covariance kernel \( H \).

**Theorem 3.** (i) If \( \alpha_n = c_3^\sqrt{n} \) for some \( 0 < c_3 < \infty \) and assumptions A−G hold with \( \tilde{\alpha}_n = \alpha_n \) then

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \frac{1}{c_3} \int t \exp(c_3^2G(t)) \phi_V(t) \text{d}t.
\]

(ii) If \( \sqrt{n} = o(\alpha_n) \), \( \varpi = \min_{\theta \in \Theta} \pi(\theta) > 0 \), and assumptions A−G hold with \( \tilde{\alpha}_n = \sqrt{n} \) then for \( \tilde{G}(t) = G(t) - t^T V t / 2 \),

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\max_i \tilde{G}(t).
\]

Theorem 3 establishes that our estimator has the same rate of convergence as the Kim and Pollard (1990) estimator, i.e. the equivalent extremum estimator, if one lets \( \alpha_n \) increase at a rate no slower than \( \sqrt{n} \). In fact, part (ii) of theorem 3 demonstrates that if \( \alpha_n \) increases faster than \( \sqrt{n} \) then the limit
Theorem 4. If $c_n$ parts (i) and (ii) of theorem 3 since if one lets $c_n \to \infty$ after $n \to \infty$, then the limit distribution of (i) converges to that of (ii). To see this, consider a different representation of the same limit distribution that arises if $c_n^2$ is incorporated into $g_i$ instead of into $\alpha_n$\textsuperscript{12}. Then
\[
\sqrt{n}(\theta - \theta_0) \xrightarrow{d} \frac{\int t \exp \left\{ c_n^2 (G(t) - t^T V t/2) \right\} dt}{\int \exp \left\{ c_n^2 (G(t) - t^T V t/2) \right\} dt}.
\]

For large values of $c_n$, the right hand side of (8) is close to the value at which $G(t) - t^T V t/2$ is maximized, whose distribution is exactly the limit distribution of the Kim and Pollard (1990) estimator.

Likewise, it follows from (8) and l'Hôpital's rule that $\sqrt{n}(\hat{\theta} - \theta_0)$ converges to zero as $c_n \downarrow 0$. More interestingly, noting that $\sqrt{n}/\alpha_n = \sqrt{n}/c_n$, (9) suggests that $\sqrt{n}/\alpha_n(\hat{\theta} - \theta_0)$ has an approximate $\int t G(t) \phi_V(t) dt$--distribution, which is indeed the normal distribution of theorem 4 below.

Intuitively, then, the accuracy of $\hat{\theta}$ is decreasing in the value of $c_n$ with the Kim and Pollard (1990) estimator the least efficient possibility. In numerical results reported in section 3.3. we find that the limiting distribution indeed becomes more dispersed as $c_n$ increases. But please note that these are asymptotic results; for small values of $c_n$ the small sample bias can be substantial. Nevertheless, the Kim and Pollard (1990) estimator (or indeed our estimator with $\alpha_n = \infty$), and hence the maximum score estimator, is unlikely to be the optimal choice in the class of estimators studied here, even in samples of finite size.

3.3. Faster convergence. We now proceed with the case in which $\alpha_n$ increases more slowly than $\sqrt{n}$. Let
\[
\begin{align*}
N_n &= \alpha_n^d \sqrt{n}/\alpha_n \int \pi(\theta)(\theta - \theta_0) \exp(\alpha_n^2 S_n(\theta) + \alpha_n^2 Q(\theta)) d\theta, \\
D_n &= \alpha_n^d \int \pi(\theta) \exp(\alpha_n^2 S_n(\theta) + \alpha_n^2 Q(\theta)) d\theta, \\
B_n &= \alpha_n^d \sqrt{n}/\alpha_n \int \pi(\theta)(\theta - \theta_0) \exp(\alpha_n^2 Q(\theta)) d\theta.
\end{align*}
\]

Theorem 4. If $\alpha_n = o(\sqrt{n})$ and assumptions A–C are satisfied with $\bar{\alpha}_n = \alpha_n$, then
\[
\sqrt{n}/\alpha_n(\theta - \theta_0) - \frac{N_n}{D_n} = \frac{N_n - D_n}{D_n} \xrightarrow{d} N(0, \mathcal{V}),
\]
where $\mathcal{V} = \int \int t^2 H(t, s) \phi_V(t) \phi_V(s) dt ds$.

\textsuperscript{12}Alternatively, one can carry out the substitution $\bar{t} = t/c_n^2$ and note that the process $G^*$ with $G^*(\bar{t}) = G(c_n^2 t)/c_n$ has the same statistical properties as $G$. 

In lemma G.4 we show that for the maximum score case in which the regressor vector includes a constant \( c_0 \),

\[
V' = \frac{1}{2\sqrt{n}} V^{-1} E \left[ \frac{z_i z_j^T}{\sqrt{z_i^T V^{-1} z_i}} f(z_i^T \theta_0 | z_i) \right] V^{-1},
\] (11)

which suggests that using observation–specific weights can reduce (some scalar–valued function of) the asymptotic variance. Some scale normalization of the weights should be imposed lest they assume the role of \( \alpha_n \). We do not pursue a weighting procedure in this paper.

The bias term \( \beta_n \) can affect the limiting distribution. As will become apparent below, if \( Q \) is sufficiently smooth then \( \beta_n \) decreases at a rate of \( 1/\alpha_n \beta_n = \sqrt{n}/\alpha_n^3 \), meaning that \( \alpha_n \sim \sqrt{n} \) yields the typical nonparametric convergence rate of \( n^{-2/5} \), also found in Horowitz (1992) (for his \( h = 2 \)).

Assumption H. \( q \geq 1 \).

We can now obtain a simple expansion for the bias, followed by an expression of the asymptotic distribution of \( n^{2/5}(\hat{\theta} - \theta_0) \) when \( \alpha_n = c_\alpha^2 \sqrt{n} \).

Theorem 5. Let assumptions \( A \)–\( H \) be satisfied. For finite weights \( b_{q\tau}^* \) independent of \( n \) and which are zero for even \( \tau \), \( \beta_n = \beta_n^{-1} \sum_{\tau=0}^q b_{q\tau}^* \alpha_n^{-q} \).

An expression for the values of \( b_{q\tau}^* \) is provided in the proof of theorem 5. In particular, \( b_{q1}^* = C_V \int (\pi_0 D_{Q3}(t) + D_{\pi_1}(t)) t \phi_V(t) dt \), where \( C_V = 1/\phi_V(0) \), \( \pi_0 = \pi(\theta_0) \), \( D_{\pi_1}(t) \) is the first term in a Taylor expansion of \( \pi(\theta_0 + t) \) about \( \pi(\theta_0) \), i.e. \( t^T \pi_0(\theta_0) \), and \( D_{Q3}(t) \) is the third term in a Taylor expansion of \( Q(\theta_0 + t) \) about \( Q(\theta_0) \). This then leads to the following result.

Theorem 6. For any \( 0 < c_\alpha \leq \infty \), if \( \alpha_n = c_\alpha^2 \sqrt{n} \) and assumptions \( A \)–\( H \) are satisfied, then

\[
\frac{n^{2/5}(\hat{\theta} - \theta_0)}{\sqrt{n}} \xrightarrow{d} N \left( \frac{\int (\pi_0 D_{Q3}(t) + D_{\pi_1}(t)) t \phi_V(t) dt}{\pi_0 c_\alpha^4} \right),
\] (12)

It is possible to minimize the asymptotic mean square error (AMSE, squared mean plus variance in (12)) by using an estimate of \( c_{a1}^2 = \sqrt{10/4 ||b_{q1}^*||^2 / C_V^2 \alpha_0^2 \text{tr}(V')} \) in lieu of \( c_\alpha \). However, because the asymptotic bias can be removed by the choice of a prior and since the asymptotic distribution using \( c_{a1}^2 \) would have nonzero mean — which is at odds with the uniform inference procedure at the heart of this paper — we do not include this result here; it is available from the authors’ website.

---

13What is needed is in fact weaker: for some \( j = 1, \ldots, d \): \( P[z_j = 0] = 0 \).

14Like in other contexts (Cragg, 1992; Pinkse, 2006), a scalar–valued weight function which optimizes the asymptotic variance of all linear combinations of the coefficients of \( \theta \) does not typically exist.
For the specific case of maximum score, the limit distribution of theorem 6 can be compared to that of the SMS estimator of Horowitz (1992). Generally, which estimator has a smaller mean square error depends on the choice of input parameters (kernel and bandwidth in the case of SMS, prior and \( c_{\alpha} \) here) and the joint distribution of \((y_i, a_i, z_i)\). It can be shown that for the very special case in which \( d = 1 \), \( z_i = 1 \) a.s., a normal kernel is used for the SMS estimator and a flat prior is used for ours, and bandwidth (SMS) and \( c_{\alpha} \) are chosen to minimize the asymptotic mean square error, the asymptotic mean square error of both estimators is identical regardless of the joint distribution of \((y_i, a_i)\). A flat prior is however not a particularly good choice as theorem 7 demonstrates.

3.4. Bias Correction. There are many ways to correct the bias. Besides various resampling schemes, one could choose a prior to remove the bias or estimate the bias directly. We discuss both possibilities below.

From (12) it follows that the bias depends on the choice of prior as well as the values of the second and third partial derivatives of \( Q \). These derivatives are estimable. As noted in theorem 5, \( b_{q\tau} = 0 \) for all even values of \( \tau \). Consequently, provided that an estimator \( \hat{b}_{q1} \) of \( b_{q1} \) converges to \( b_{q1} \), we obtain the result

\[
\frac{n^{2/5}}{c_{\alpha}^2} (\hat{\theta} - \theta_0) - \frac{\hat{b}_{q1}}{c_{\alpha} D_n} \xrightarrow{d} N(0, c_{\alpha}^2 \nu),
\]

(13)

With \( q = 3 \) and \( \alpha_n = c_{\alpha}^2 \sqrt{n} \), (13) can be strengthened to

\[
\frac{n^{4/9}}{c_{\alpha}^2} (\hat{\theta} - \theta_0) - \frac{n^{2/9} \hat{b}_{31}}{c_{\alpha}^2 D_n} - \frac{n^{2/9} \hat{b}_{33}}{c_{\alpha}^2 D_n} \xrightarrow{d} N(0, c_{\alpha}^2 \nu),
\]

assuming \( \hat{b}_{31} \) converges to \( b_{31} \) at a rate faster than \( n^{2/9} \), which, in light of the smoothness condition \( q = 3 \), should not be problematic. This process can be repeated to obtain an estimator of \( \theta_0 \) which converges at a rate arbitrarily close to \( \sqrt{n} \), assuming sufficient smoothness. In particular, using \( \alpha_n = c_{\alpha}^2 n^{1/(2q+3)} \) leads to

\[
\frac{n^{q+1}}{c_{\alpha}^2 D_n} (\hat{\theta} - \theta_0) - \frac{1}{c_{\alpha} D_n} \sum_{\tau=0}^{q} \frac{n^{\frac{q-\tau}{3}} \hat{b}_{q\tau}}{c_{\alpha}^{2\tau+2}} \xrightarrow{d} N(0, c_{\alpha}^2 \nu),
\]

(14)

where \( \hat{b}_{q\tau} = 0 \) for even values of \( \tau \).

It turns out that removing the bias in theorem 6 can also be accomplished by choosing a prior that resembles Jeffreys’ prior (Jeffreys, 1946), here \( \pi(\theta) \propto \sqrt{\det(-Q''(\theta))} \), near \( \theta_0 \). However, since \( Q \) is not necessarily globally concave one should choose a prior which is both everywhere bounded and which resembles Jeffreys’ prior in a neighborhood of \( \theta_0 \). An example is \( \pi_1(\theta) \propto \sqrt{|\det(-Q''(\theta))|} \).

\[\text{15}The\ derivation\ is\ available\ upon\ request.\]
which can be estimated by \( \hat{\pi}_I(\theta) \propto \sqrt{|\det(-\hat{Q}''(\theta))|} \) for some estimator \( \hat{Q}'' \) which satisfies the following assumption. Let vec denote the familiar vec–operator, which stacks the columns of a matrix into a vector and let \( Q'' \) and \( \hat{Q}'' \) be defined as \( \partial \text{vec}^\top Q'' / \partial \theta \) and \( \partial \text{vec}^\top \hat{Q}'' / \partial \theta \), respectively.

**Assumption 1.** For \( r^* \geq 0 \) and any \( \rho_n = o(1/\alpha_n) \), \( \hat{Q}'' \) is bounded on \( \Theta \) and satisfies

\[
\sup_{\|\theta-\theta_0\| \leq \rho_n} (\| \hat{Q}''(\theta) - Q''(\theta) \| + \| \hat{Q}'''(\theta) - Q'''(\theta) \|) = o_p(\alpha_n^{-r^*}).
\]

The conditions specified in assumption \( I \) are generic because the properties of \( Q'' \) depend on the nature of \( g_i \), on which we only imposed generic conditions. In the maximum score case, \( Q'' \) and \( Q''' \) are average derivatives [Powell, Stock, and Stoker, 1989] and hence can typically be estimated at a rate much faster than the \( d + 1 \)-dimensional nonparametric rate. For \( r^* = 0 \), no rate is required. For \( r^* = 2 \), a rate of \( n^{2/5} \) is sufficient, but when we require \( r^* = 2 \) more derivatives of \( Q'' \) will be available. Let \( \hat{\theta}_I \) denote the estimator which is identical to \( \hat{\theta} \) except that it uses the prior \( \hat{\pi}_I \) defined above.

**Theorem 7.** Let assumptions \( I \) and \( II \) be satisfied.

(i) If \( \alpha_n = c^2_\alpha \sqrt{n} \) and assumption \( I \) is satisfied with \( r^* \geq 0 \), then \( n^{2/5}(\hat{\theta}_I - \theta_0) \xrightarrow{d} N(0, c_{\alpha}^2 \gamma) \).

(ii) If \( \sqrt{n} = O(\alpha_n) \), \( q \geq 3 \), and assumption \( I \) is satisfied with \( r^* \geq 2 \), then for some finite \( \mathcal{B}_I, \sqrt{n}/\alpha_n(\hat{\theta}_I - \theta_0) \xrightarrow{d} \mathcal{B}_I / \alpha_n^2 \beta_n \mathcal{B}_n \xrightarrow{d} N(0, c_{\alpha}^2 \gamma) \).

So subject to the conditions of theorem \( 4 \) using (an estimate of) Jeffreys’ prior yields a zero mean limiting normal distribution (part (i) of theorem \( 7 \)), and with extra derivatives the \( n^{4/9} \)-rate obtains.

Further bias reduction can be obtained by choosing a prior which knocks out the higher order bias terms, also. Because the bias only depends on derivatives of \( Q'' \) and \( \pi \) at \( \theta_0 \), such priors are straightforward to construct by using suitable polynomials in \( \theta - \theta_0 \) whose coefficients consist of estimated derivatives of \( Q'' \) at \( \theta_0 \), albeit that this procedure would require an initial plugin estimator to use in lieu of \( \theta_0 \), whereas such an initial plugin estimator is not required for \( \hat{\pi} \) used in theorem \( 7 \).

An example of such a polynomial prior which removes the \( \tau = 1 \) term in the bias expansion of theorem \( 5 \) is \( \hat{\pi}(\theta) = \sqrt{\det(-\hat{Q}''(\theta))} \), with

\[
\hat{Q}''(\theta) = Q''(\theta_0)/2
\]

\[
+ \left( Q'''(\theta_0)(I \otimes (\theta - \theta_0)) + Q''(\theta_0) \right)^T (Q''(\theta_0))^{-1} \times \left( Q'''(\theta_0)(I \otimes (\theta - \theta_0)) + Q''(\theta_0) \right),
\]

(15)
which is guaranteed to be negative definite since it is the sum of a negative definite and negative
semidefinite matrix. Methods for estimating $Q''$ and $Q'''$ for the maximum score case are discussed
in appendix I.3.

An alternative to removing the bias by one’s choice of prior is to estimate the bias directly. The
bias term in theorem 6, for instance, can be estimated using

$$\int \left(\pi_0 D_{Q3}(t) + D_{\pi_1}(t)\right) t\phi(t) dt,$$

In particular, for $d = 1$ and a flat prior, (16) reduces to $\hat{Q}''(\hat{\theta}_p)/2c^4_\alpha(\hat{Q}''(\hat{\theta}_p))^2$, with $\hat{\theta}_p$ some preliminary estimator.

Since the form of an estimator of $Q''$ and its derivatives depends on the nature of $g_i$, establishing
general analytical results for the bias correction is not possible. It would be possible to do so for
the maximum score case, but this would entail not much more than a rehash of results that are
well–known and are available in the literature.

For the maximum score case, nonparametric estimators can be shown to be uniformly consistent
for $Q''$ and $Q'''$ in a neighborhood of $\theta_0$ that decreases more slowly than $\hat{\theta}_p$. So consistency obtains
and the bias up to order $n^{-2/5}$ is removed.

4. Uniform Inference

To conduct inference one can draw random numbers from the Kim and Pollard (1990) limit
distribution if $\alpha_n$ increases faster than $\sqrt[3]{n}$, from the limit distribution of theorem 3 if $\alpha_n$ increases at
the $\sqrt[3]{n}$ rate, or use the normal of theorem 4 if $\alpha_n$ increases more slowly than $\sqrt[3]{n}$. Since these are
only rates for $\alpha_n$, in a sample of finite size it is generally not clear which of these three distributions
should be used. Let $\hat{G}$ be $G$ with $H$ replaced with some estimator $\hat{H}$. As it turns out, provided that
the bias is asymptotically negligible, for $\kappa_n$ defined in section 2.1 and $\hat{V}$ some consistent estimator of
$V$, $\chi_n = \kappa_n(\hat{\theta} - \theta_0)$ and

$$\Psi_n = \frac{\int t \exp\left(\beta_{n1}G(t) - \beta_{n2}t^\top Vt/2\right) dt}{\int \exp\left(\beta_{n1}G(t) - \beta_{n2}t^\top \hat{V}t/2\right) dt},$$

have the same limiting distribution in all three cases.
Theorem 8. If \( \hat{V} - V = o_p(1) \) and \( \hat{H} \) is a positive semidefinite covariance kernel, which is pointwise consistent for \( H \) and for some \( 0 < c_H < 2 \) and all sample paths of \( \hat{H} \) satisfy \( \sup_{\|t\| \to \infty} (\hat{H}(t, t)/\|t\|^2) < \infty \), then \( \Psi_n \) has the limiting distribution derived for \( \chi_n \) in the corresponding portion of theorems 3 and 4 if (i) \( \beta_n = o(1) \), (ii) \( 1/\beta_n + \beta_n = O(1) \), (iii) \( 1/\beta_n = o(1) \).

The requirement that \( \hat{H} \) be a positive semidefinite covariance kernel is both necessary and sufficient for there to exist a Gaussian process \( \hat{G} \) with that covariance kernel (Doob, 1953, theorem 3.1). Drawing random numbers using (17) is simple and is explained in appendix I.2.

So the proposed inference procedure is uniform in the divergence rate of \( \alpha_n \) despite the fact that the limiting distribution of \( \hat{\theta} \) depends on the rate at which \( \alpha_n \) increases.

The limiting distribution of Laplace–type estimators sometimes coincides with the limit of the posterior, i.e. \( \lim_{n \to \infty} \exp (\alpha_n^2 L_n(\theta_0 + t/\alpha_n)) \) up to a proportionality constant. Indeed, Chernozhukov and Hong (2003) derived conditions under which this is the case for objective functions that admit a quadratic expansion. In the case of coincidence of the true distribution and the limit of the posterior, the (Gibbs) draws used to obtain \( \hat{\theta} \) can then be used to construct confidence intervals. Here the limit of the posterior is a random object if \( \alpha_n \) increases at the rate of \( 3\sqrt{n} \) (or faster) and is \( N(0, V^{-1}) \) if \( \alpha_n \) increases more slowly than \( \sqrt{n} \). So in neither case are the posterior and estimator limit distributions the same.

5. Performance

As noted before, the main caveat of our estimator is the need to choose input parameters \( \pi, \alpha_n \). The ability to choose a prior has been shown to be valuable in bias reduction, at least theoretically. Its practical value is examined later on in this section. We begin by analyzing the effect of one’s choice of \( \alpha_n \).

We consider a small number of designs, all in the context of the maximum score estimation problem. In all cases \( y_i = l(z_i^\top \theta_0 - a_i + \zeta((a_i, z_i)u_i \geq 0) \), where \( u_i \) is a standard normal independent of \( a_i, z_i \) which are also standard normal and have mutually independent elements. Unless otherwise stated \( \theta_0 = [0, \bar{1}^\top]^\top \), where \( \bar{1} \in \mathbb{R}^{d-1} \) is a vector of ones and the prior is uniform on a compact support, typically \([-2, 2] \).

For figure 1 we used \( d = 2 \) and \( \zeta^2 = (z_i^2 + 1)/2 \). Figure 1 illustrates how (for a single data set) the normalized posterior varies with \( \alpha_n \); the value at \( \theta_0 \) equals one. Larger values of \( \alpha_n \) lead to a

---

\( \hat{H} \) is a positive semidefinite covariance kernel if for any integer \( 0 < T^* < \infty \) and any values \( t_1, t_2, \ldots, t_{T^*} \in \mathbb{R}^d \), the matrix with \((i,j)\)-element \( \hat{H}(t_i, t_j) \) is positive semidefinite.

Please note that for part (i) of theorem 3, \( \kappa_n \) and \( \sqrt{n} \) can differ by a multiplicative constant.
narrower posterior density, so $\hat{\theta}_2$ is an average over a comparatively small range of $\theta$–values. In this particular instance the posterior density for $\alpha_n = 10$ is close to the truth (one), but for other data sets it may not be. Because it is an average over fewer $\theta$–values whose range varies from one data set to another, $\hat{\theta}$ for $\alpha_n = 10$ has a greater variance than for $\alpha_n = 4$. But the bias is less for greater values of $\alpha_n$. One reason is that the posterior is asymmetric; in figure 1 the posterior for $\alpha_n = 4$ tapers off much faster to the left than it does to the right, which is due to the particular design used.

Figure 1. One draw of $\exp(\alpha_n^2 L_n(\theta_{01}, \theta_2))$; $n = 1,000$.

Figure 2. Density of $\hat{\theta}_2$ for various choices of $\alpha_n$; $d = 2$, $n = 1,000$.
A secondary cause of the bias is that for smaller values of $\alpha_n$, the prior receives relatively more weight, thereby biasing one’s estimates in the direction of the mean of the prior. Inspection of the formula for $b_{n,T}$ in the proof of theorem 5 demonstrates that the asymptotic bias depends on the prior and its derivatives only at $\theta_0$; the value of the prior and its derivatives at other values of $\theta$ are irrelevant for the asymptotic bias. For a flat prior the derivatives at $\theta_0$ are all zero and the prior drops out of the bias formula. Therefore, the secondary bias is not correctible with the methods of section 3.4.

To see that this secondary bias can be substantial for small values of $\alpha_n$, consider figure 2. For very small values of $\alpha_n$ the distribution of $\hat{\theta}_2$ is centered around the mean of the prior (zero). As $\alpha_n$ increases this secondary bias decreases until it becomes small compared to the primary (asymptotic) and correctible bias for $\alpha_n = 10$. Here the primary bias is positive.

This point is reinforced by figure 3 which represents the results of an experiment identical to the one for figure 2 except that the mean of the prior coincides with the value of $\theta_{02}$; both equal one. There is some bias for smaller values of $\alpha_n$, but this bias is all in the same direction; positive.

A heuristic method for choosing $\alpha_n$, then, is to compute the value of $\hat{\theta}$ for two flat priors with different means, say $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}$, and to choose the smallest value of $\alpha_n$ for which the difference between $\hat{\theta}_{(1)}$ and $\hat{\theta}_{(2)}$ is small. Subsequently, the imputed value of $\alpha_n$ can be used with a prior of one’s choosing. Alternatively, the procedure of choosing $\alpha_n$ can be tailored to the nature of the
objective function, e.g. one could construct a method for choosing $\alpha_n$ specific to the maximum score case.

The experiment depicted in figure 2 is repeated in figure 4 with $d = 10$. The results suggest that the value of $\alpha_n$ need not depend on $d$, but that the rewards for a good choice of $\alpha_n$ increase. Indeed, the estimator variance is substantial for $\alpha_n = 32$ and would be still greater for the maximum score estimator. In fact, for a $\sqrt{n}$-consistent estimator to overcome the difference in variances between $\alpha_n = 10$ and $\alpha_n = 32$ over ten times as many observations would be needed.

Figure 5 underlines the value of heteroskedasticity-robust estimators for the binary choice model, such as the (smoothed) maximum score estimator and our Laplace-transform based estimator. Even though there is only a modest amount of heteroskedasticity and the error distribution is normal, the probit estimator has substantial bias. Naturally, under homoskedasticity with a normal error distribution the probit estimator will outperform the robust estimators since, as figure 5 indicates, its variance is considerably smaller than that of the other estimators. It is also clear that here too the maximum score estimator has greater variance than the Laplace-type estimators, which suggests that choosing $\alpha_n = \infty$ is suboptimal even without bias correction.

We now investigate the performance of one of the proposed bias-correction techniques, namely the use of Jeffreys’ prior. It should be pointed out that using Jeffreys’ prior takes more time than the

\[ \alpha_n = 32 \]
\[ \alpha_n = 10 \]
\[ \alpha_n = 5.6 \]
\[ \alpha_n = 4.0 \]

\[ \frac{\text{ratio of variances of } \hat{\theta} \text{-values (for } \alpha_n = 32 \text{ versus } \alpha_n = 10)}{\text{in the simulations}} \approx 5.8; \text{ the number ten comes from the fact that } 5.8^{3/2} = 14 > 10. \]

\[ \text{Computed using a grid search.} \]
flat prior unless the conditional prior of one element of the parameter vector given the others has a closed form solution. For our experiments, the cost of using Jeffreys’ prior was far from prohibitive. If computation time becomes a serious issue, one can instead opt for the more convenient, faster, and asymptotically equivalent form presented in (15), albeit that we expect the finite sample performance of such a procedure to be inferior.

Figure 5 depicts the bias for Jeffreys’ prior and the correctly centered uniform prior for varying values of $\alpha_n$. The support of Jeffreys’ prior was chosen such that the mean of Jeffreys’ prior was also equal to one. The purpose of setting the mean of the prior equal to the truth is to assess the ability of Jeffreys’ prior to correct the asymptotic bias without obfuscating the comparison by the
asymptotically negligible bias caused by the discrepancy between the mean of the prior and the true parameter value.

\[ f(\theta^2_n) = 10 \]

\[ f(\theta^2_n) = 5.6 \]

\[ f(\theta^2_n) = 4 \]

**Figure 7.** Density of \( \hat{\theta}_2 \); Jeffreys (solid), uniform (dashed), estimated Jeffreys (dotted) prior; \( n = 1,000, d = 2 \).

Figure 6 demonstrates that the bias decreases in \( \alpha_n \) for both priors, and that using Jeffreys’ prior removes most of the bias of the uniform kernel, which is encouraging. It is apparent, however, that even Jeffreys’ prior does not remove all asymptotic bias, especially when \( \alpha_n \) is small, which is due to the fact that for small values of \( \alpha_n \) the higher order terms in the bias expansion are of comparatively greater importance.

We repeated the same experiment, but now with the estimated version of Jeffreys’ prior included. The results are depicted in figure 7. \( Q'' \) was estimated using the method described in appendix I.3.

Evidently, the error in estimating \( Q'' \) is small relative to the bias reduction vis–à–vis the uniform prior. We conclude that an asymptotic bias correction using Jeffreys’ prior is valuable unless \( \alpha_n \) is chosen large.

Finally, we evaluate the quality of our uniform inference procedure; the results are in figure 8. As figure 8 demonstrates, the uniform inference procedure of theorem 8 moves gradually from the maximum score limiting distribution to the normal limiting distribution as \( \alpha_n \) (and hence \( \beta_n \)) decreases. It does a good job approximating the finite sample distribution of \( \hat{\theta}_2 \) except that for small values of \( \alpha_n \) there is substantial bias, which is due to the fact that estimates used to generate the finite sample distribution are not higher order bias–corrected.

The reason that the results in figure 8 use a design with \( d = 2 \) only is that the limiting distribution of the maximum score estimator becomes very expensive to simulate for higher dimensions. This

---

\[ ^2 \text{We do use Jeffreys’ prior.} \]
suggests that one should use our uniform inference procedure with a high value of $\beta_n$ to simulate the limiting distribution of the maximum score estimator instead of trying to maximize $\hat{G}(t) - t^T \hat{V} t / 2$ repeatedly. We found that the pattern of the remaining three densities for greater values of $d$ (not shown) is similar to the one depicted in figure 8.

REFERENCES CITED


Figure 8. Density of $\kappa_n(\hat{\theta}_2 - \theta_0)$ for various inference procedures: uniform (solid), normal (dashed), maximum score (dotted), finite sample (densely dashed); $n = 1,000, d = 2, R = 3,000$. 


autocorrelation robust tests,” Econometric Theory, 21(06), 1130–1164.


Econometrica, 61, 387–421.

387–407.


——— (2009): “Robust estimation in binary choice models,” Communications in Statistics (forthcom-
ing).

59–77.

MANSKI, C. (1975): “Maximum score estimation of the stochastic utility model of choice,” Journal of
Econometrics, 3(3), 205–228.


MANSKI, C., AND T. THOMPSON (1986): “Operational characteristics of maximum score estimation,”

1452.

models,” Discussion paper, NBER.

models,” Journal of Econometrics, 58, 185–205.


Econometrica, 57(6), 1403–1430.
A1. Notation. Let \( \gamma_n = a_n^{1/(4q+4)} \), \( \Gamma_n = \{ t \in \mathbb{R}^d : \| t \| \leq \gamma_n \} \), and \( \Gamma^c_n = \mathbb{R}^d - \Gamma_n \). Let further for vector-valued \( t, \hat{t} \) denote 1 if \( j = 0 \) and \( t \) if \( j = 1 \). Define \( C_V = \int \exp(-t^TVt/2)dt = 1/\phi_V(0) = (2\pi)^d/\sqrt{\det(V)} \) and let \( \lambda_\gamma \) be the smallest eigenvalue of \( V \). Finally, let \( R_n(t) = Q_n(t) + t^TVt/2. \)

A2. Weak Convergence. The results in this section presume that assumptions 1 and 2 are satisfied. Let \( \Gamma_1 \subseteq \Gamma_2 \cdots \) be a sequence of compact sets such that \( 0 \in \Gamma_1 \) and \( \mathbb{R}^d = \bigcup_{i=1}^\infty \Gamma_i \). We then have \( \tilde{S}_n \stackrel{w}{\rightarrow} G \) in \( \mathcal{L}^\infty(\mathcal{F}_1, \mathcal{F}_2, \ldots) \), where \( G(\cdot) \) is a Gaussian process with covariance kernel \( H \).

Lemma A.1. Let \( \mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \) be a sequence of compact sets such that \( 0 \in \mathcal{F}_1 \) and \( \mathbb{R}^d = \bigcup_{i=1}^\infty \mathcal{F}_i \). We then have \( \tilde{S}_n \stackrel{w}{\rightarrow} G \) in \( \mathcal{L}^\infty(\mathcal{F}) \) for an arbitrary compact set \( \mathcal{F} \subset \mathbb{R}^d \). Since \( \mathcal{F} \) is dense, assumption 1 ensures that for \( j = 1, 2 \),

\[
\mathcal{F}_{j, \epsilon}^\ell = \{ \tilde{a}_n^{j/2} (g(\cdot; \theta_0 + t/\tilde{a}_n) - g(\cdot; \theta_0 + t/\tilde{a}_n))^j \}_{\| t-s \| < \epsilon}
\]

is a pointwise measurable class, and hence \( \mathcal{F} \)-measurable for every \( \mathcal{F} \); see van der Vaart and Wellner (1996, page 110). Note also that \( E[\tilde{S}_n(t)\tilde{S}_n(s)] = \tilde{a}_n E[g_t(\theta_0 + t/\tilde{a}_n)g_t(\theta_0 + s/\tilde{a}_n)] \rightarrow H(t,s) \) for every \( t, s \in \mathcal{F} \) by assumption 1. Therefore, the result follows from van der Vaart and Wellner (1996, theorem 2.11.22).

Lemma A.2. Let \( \tilde{\gamma} > 0, \tilde{T} = \{ t \in \mathbb{R}^d : \| t \| \leq \tilde{\gamma} \} \). Then (i) for any \( 0 < c < \infty \),

\[
\lim_{T \rightarrow -\infty} \lim_{n \rightarrow \infty} P[\sup_{t \in \Gamma_1^c} |\tilde{S}_n(t)|/\| t \|^2 > c] = 0 \quad \text{and} \quad P[\sup_{t \in \Gamma_1^c} |\tilde{S}_n(t)|/\| t \|^2 > c] = o(1).
\]

Proof. Let \( \tilde{S}_n(t) = \tilde{S}_n(t)/\| t \| \) and \( \tilde{G}(t) = G(t)/\| t \| \). Suppose without loss of generality that \( \alpha_1 = 1 \). Then the equicontinuity of \( \tilde{S}_n \) on \( \Gamma_1^c \) follows from that of \( \tilde{S}_n \). Therefore on \( \Gamma_1^c \), \( \tilde{S}_n \stackrel{w}{\rightarrow} \tilde{G} \), which is a uniformly bounded process in \( \mathcal{L}^\infty(\Gamma_1^c) \). Hence, to establish (ii) (i) is similar, note that \( \sup_{t \in \Gamma_1^c} (|\tilde{S}_n(t)|/\| t \|^2) \leq \sup_{t \in \Gamma_1^c} (|\tilde{S}_n(t)|(\log \| t \|)/\| t \|^2)/\log \gamma_n = O_p(1) + o(1) = o_p(1). \)

\[ \Box \]
Lemma A.3. For any $c > 0$, $\sup_t \{ |\mathbf{S}_n(t)| - ct^3 V_t \} = O_p(1)$.

Proof. Follows immediately from lemma A.1 and the properties of $G$. \qed

A.3. Auxiliary Results.

Lemma A.4. $\lim_{\epsilon \to 0} \sup_{\|\theta - \theta_0\| \leq \epsilon} \|Q''(\theta) + V\| = 0$.

Proof. Follows from the continuity of $Q''$ at $\theta_0$. \qed

Lemma A.5. For some $0 < c_q < \infty$ and all $\theta \in \Theta$, $Q(\theta) \leq -\min(c_q, (\theta - \theta_0)'V(\theta - \theta_0)/4)$ and $Q_n(t) \leq -\min(\tilde{\alpha}_n c_q, t^3 V_t/4)$ for all $t$ for which $\theta_0 + t/\tilde{\alpha}_n \in \Theta$.

Proof. Follows from the fact that (i) $\Theta$ is compact, (ii) $Q$ has a unique maximum of zero at $\theta_0$, and (iii) $Q_{\theta_0}$ is continuous at $\theta_0$ with value $-V$. \qed

Lemma A.6. For any $|c| \leq 1$, any $b$, and any nonnegative integer $j$, $|\exp(c b) - \sum_{s=0}^{j} (cb)^s/s!| \leq |c|^{j+1} \exp(|b|)$.

Proof. We have $|\exp(c b) - \sum_{s=0}^{j} (cb)^s/s!| \leq |\sum_{s=j+1}^{\infty} (cb)^s/s!| \leq |c|^{j+1} \sum_{s=j+1}^{\infty} |b|^s/s! \leq |c|^{j+1} \exp(|b|)$. \qed

A.4. Further Auxiliary Results. The results in this section presume that $0 \leq \lim_{n \to \infty} \beta_n = c_\beta < \infty$ and that $\tilde{\alpha}_n = \alpha_n$.

Lemma A.7. For all nonnegative and finite $c_1, c_2, c_3, c_4$ and any polynomial $P$,

(i) if $c_3 > 0$ then $\int_{\Gamma_n} \|P(t)\| \exp(c_1 \beta_n |\mathbf{S}_n(t)|) - c_3 t^3 V_t dt = o_p(\alpha_n^{-c_4})$,

(ii) if $c_2 + c_3 > 0$ then $\int_{\Gamma_n} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\mathbf{S}_n(t)|) + c_2 Q_n(t) - c_3 t^3 V_t dt = o_p(\alpha_n^{-c_4})$.

Proof. First (i). By lemma A.2, the left hand side of (i) is for any $c$ of smaller order than $\int_{\Gamma_n} \|P(t)\| \exp((cc_1 c_\beta - c_3 \lambda_-) \|t\|^2) dt$, which for sufficiently small $c$ decreases exponentially in $n$. The left hand side in (ii) is by lemma A.5 of order no greater than

$$\int_{\Gamma_n} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\mathbf{S}_n(t)|) - c_2 \min(c_q a_n^2, t^3 V_t/4) - c_3 t^3 V_t dt$$

$$\leq \int_{\Gamma_n} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\mathbf{S}_n(t)|) - c_2 c_q a_n^2 - c_3 t^3 V_t dt$$

$$+ \int_{\Gamma_n} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\mathbf{S}_n(t)|) - (c_2/4 + c_3) t^3 V_t dt. \quad (18)$$

The second right hand side term in (18) was dealt with in (i). If $c_3 > 0$, the first right hand side term is bounded by the same expression with $c_2 = 0$, which was dealt with in (i), also. Finally, the first
right hand side term if \( c_3 = 0 \). Since \( \sup_{\theta} |S_n(\theta)| = o_p(1) \) by lemma A.3, it follows that for some \( c^* > 0 \), 
\[
\int_{\Gamma_n} \| \pi_n(t)P(t) \| \exp(c_1 \beta_n |S_n(t)|) \exp\left(c_2 \beta_n^2 \right) dt \text{ is of order no greater than } \exp\left(-c^* \alpha_n^2 \right).
\]

**Lemma A.8.** For any \( c > 0 \), (i) \( \mathbb{P} [\sup_{t \in \Gamma_n} I(|R_n(t)| > ct^T V t/4) \neq 0] = o(1) \).

**Proof.** Follows from lemma A.4 \( \gamma_n = o(a_n) \), and the fact that \( R_n(t) = t^T (Q'(\theta_0 + t/a_n) + V) t/2. \)

**Lemma A.9.** For all nonnegative and finite \( c_1, c_2, c_3, c_4 \) and any polynomial \( P \),

(i) if \( c_3 > 0 \) then 
\[
\int \| P(t) \| \exp(c_1 \beta_n |S_n(t)| - c_3 t^T V t) dt = O_p(1),
\]

(ii) if \( c_2 + c_3 > 0 \) then 
\[
\int \| \pi_n(t)P(t) \| \exp(c_1 \beta_n |S_n(t)|) + c_2 Q_n(t) - c_3 t^T V t) dt = O_p(1).
\]

**Proof.** By lemma A.7, we only need to show that the integrals over \( t \in \Gamma_n \) are \( O_p(1) \). For (i) the stated result follows from lemma A.1. For (ii), it follows from lemmas A.1 and A.8.

**APPENDIX B.** \( a_n = c_n^2 \sqrt{n} \)

**Lemma B.1.** For \( j = 0, 1 \), 
\[
\int_{\Gamma_n} \pi_n(t) t^j \exp\left(c_3 \tilde{S}_n(t) \right) \left( \exp(R_n(t)) - 1 \right) \phi_V(t) dt = o_p(1).
\]

**Proof.** By lemma A.6 for \( b = R_n(t)/c \), the left hand side is bounded in absolute value by 
\[
c \int_{\Gamma_n} \| \pi_n(t) t^j \| \exp(c_3 \tilde{S}_n(t)) + |R_n(t)/c| - t^T V t/2) dt,
\]

which with probability approaching one is bounded above by 
\[
c \int_{\Gamma_n} \| \pi_n(t) t^j \| \exp(c_3 \tilde{S}_n(t)) - t^T V t/4) dt 
\]

by lemma A.1 for some \( Y \) independent of \( c \). Let \( c \downarrow 0 \).

**Lemma B.2.** For \( j = 0, 1 \), 
\[
\int \pi_n(t) t^j \exp\left(c_3 \tilde{S}_n(t) + Q_n(t) \right) dt = C_V \pi_0 \int t^j \exp(c_3 \tilde{S}_n(t)) \phi_V(t) dt + o_p(1)
\]

**Proof.** We have
\[
\int \pi_n(t) t^j \exp\left(c_3 \tilde{S}_n(t) + Q_n(t) \right) dt = C_V \pi_0 \int t^j \exp(c_3 \tilde{S}_n(t)) \phi_V(t) dt
\]
\[
= \int_{\Gamma_n} \pi_n(t) t^j \exp(c_3 \tilde{S}_n(t) + Q_n(t)) dt + C_V \int_{\Gamma_n} \pi_n(t) t^j \exp(c_3 \tilde{S}_n(t)) \left( \exp(R_n(t)) - 1 \right) \phi_V(t) dt
\]
\[
+ C_V \int_{\Gamma_n} (\pi_n(t) - \pi_0) t^j \exp(c_3 \tilde{S}_n(t)) \phi_V(t) dt - \pi_0 \int_{\Gamma_n} t^j \exp(c_3 \tilde{S}_n(t) - t^T V t/2) dt. \quad (19)
\]

The first and last right hand side terms in (19) are dealt with in lemma A.7. The second right hand side term in (19) is \( o_p(1) \) by lemma B.1. Because \( \pi \) is continuous at \( \theta_0 \), the third right hand side term in (19) is also \( o_p(1) \) by lemma A.3.
APPENDIX C. $\sqrt{n} = o(\alpha_n)$

Lemma C.1. Let $\bar{\gamma}, \bar{\Gamma}$ be as in lemma [A.2]. Let further $L_n(t) = S_n(t) + Q_n(t)$ Then

$$\int_{\Gamma} \pi_n(t) \exp\left(\beta_n^{4/3} L_n(t)\right) dt \leq \sup_{t \in \bar{\Gamma}} \bar{\gamma} \exp\left(\beta_n^{4/3} L_n(t)\right) dt \rightarrow \text{argmax} \bar{G}(t).$$

Proof. Note that $L^2(\bar{\Gamma}, \mathbb{B}, \mu)$ is separable where $\mathbb{B}$ is the Borel sigma algebra. Since $\bar{G}(t) = G(t) - t^\text{TV}/2$, $L_n, \bar{G}$ are both in $L^2(\bar{\Gamma}, \mathbb{B}, \mu)$ and $L_n \xrightarrow{w} \bar{G}$ by lemma [A.1] and the fact that

$$\sup_{t \in \bar{\Gamma}} |Q_n(t) + t^\text{TV}/2| \leq \gamma^2 \sup_{\|\theta - \theta_0\| \leq \gamma/\sqrt{\pi}} \|Q_n(\theta) + V\|/2 = o_p(1),$$

by lemma [A.4]. Hence, by the Skorokhod representation theorem [Billingsley 1999, theorem 6.7], there exist $\bar{L}_n, \bar{G}^*$ with the same distributions as $L_n, \bar{G}$, such that for an arbitrary sample path $\bar{L}_n = \bar{L}_n^*(\cdot; \bar{\omega})$ of $L_n^*$ and corresponding sample path $\bar{G}^* = \bar{G}^*(\cdot; \bar{\omega})$ of $\bar{G}^*$,

$$\int_{\bar{\Gamma}} |\bar{L}_n(t) - \bar{G}^*(t)|^2 dt = o(1).$$

(21)

Let for arbitrary sets $T_1, T_2 \subset \bar{\Gamma}, d^*(T_1, T_2) = \mu(T_1 - T_2) + \mu(T_2 - T_1)$. Let further for arbitrary $c > 0$, $T(\bar{x}, c) = \{t \in \bar{\Gamma} : |\bar{x}(t) - \bar{G}^*| \leq c\}$, where $\bar{G}^* = \max_{t \in \bar{\Gamma}} \bar{G}^*(t)$. We first establish that

$$d^*(T(L_n^*, c), T(\bar{G}^*, c)) = o(1).$$

(22)

Let $T_{1n}(c) = T(\bar{L}_n^*, c) - T(\bar{G}^*, c)$ and $T_{2n}(c) = T(\bar{G}^*, c) - T(\bar{L}_n^*, c)$. We show that $\mu(T_{2n}(c)) = o(1)$ where the same result for $T_{1n}(c)$ follows similarly. Let for arbitrary $c^* > 0$, $T_n(c^*) = \{t \in \bar{\Gamma} : |\bar{L}_n(t) - \bar{G}^*(t)| \leq c^*\}$. For the remainder of this lemma, define complements relative to $\bar{\Gamma}$ (e.g. $T_n^c(c^*) = \bar{\Gamma} - T_n(c^*)$). Because $\mu(T_{2n}^c(c^*) \cap T_{2n}(c)) \leq \mu(T_{2n}^c(c^*)) = o(1)$ by (21), we only need to consider $\mu(T_{2n}(c^*) \cap T_{1n}(c))$.

Note first that $T_{1n}(c^*) \cap T_{2n}(c) \subseteq T_{1n}^*(c, c^*) = \{t \in \bar{\Gamma} : c \leq |\bar{L}_n(t) - \bar{G}^*| \leq c + c^*\}$, such that by (21),

$$\lim_{c^* \to 0} \lim_{n \to \infty} \mu(T_{1n}^*(c, c^*)) = \lim_{c^* \to 0} \mu(\{t \in \bar{\Gamma} : c \leq |\bar{G}^*(t) - \bar{G}^*| \leq c + c^*\})$$

$$= \mu(\{t \in \bar{\Gamma} : |\bar{G}^*(t) - \bar{G}^*| = c\}) = 0,$$

(23)

because $\bar{G}^*$ is continuous and nowhere differentiable. So (22) holds.

Finally, note that for $j = 0, 1$ and some finite constant $C$,

$$\int_{T'(L_n, \bar{c})} \|t\|^j |\pi_n(t)| \exp\left(\beta_n^{4/3} (L_n(t) - \bar{G}^*)\right) dt \leq C \exp(-\beta_n^{4/3} c) \int \|t\|^j |\pi_n(t)| dt = o(1),$$

where $\beta_n = \beta_n(c)$ and $\beta_n^*$ is the corresponding process on $\bar{\Gamma}$.
Lemma D.1. \( \int \pi_n(t) \exp \left( O_n(t) \right) dt = \pi_0 C_V + o(1). \)
Proof. We have by lemma \[A.7\]
\[
\int \pi_n(t) \exp(Q_n(t))dt - \pi_0 C_V = \int_{\Gamma_n} \pi_n(t) \exp(Q_n(t))dt + o(1)
\]
\[
= C_V \int_{\Gamma_n} \pi_n(t) \left( \exp(R_n(t)) - 1 \right) \phi_V(t)dt + C_V \int_{\Gamma_n} (\pi_n(t) - \pi_0) \phi_V(t)dt
\]
\[
- C_V \pi_0 \int_{\Gamma_n} \phi_V(t)dt + o(1). \tag{25}
\]

The third right hand side term in (25) vanishes because \(\gamma_n\) increases to \(\infty\), the first term is \(o(1)\) by lemma \([A.8]\) and the second term is \(o(1)\) by the continuity of \(\pi\) at \(\theta_0\).

\[\Box\]

**Lemma D.2.** Let \(\mu_{in} = \sqrt{\frac{1}{n}} \int t g_i(\theta_0 + t/\alpha_n) \phi_V(t)dt\), such that \(\{\mu_{in}\}\) is an independent mean zero array. Then \(\sum_{i=1}^n E\|\mu_{in}\|^{2+i} = o(1)\) for \(i\) defined in assumption [E].

Proof. We have for \(i^*\) defined in assumption [E]
\[
\sum_{i=1}^n E\|\mu_{in}\|^{2+i} \leq \alpha_n^{1-i-i^*} \left(\frac{\alpha_n}{n}\right)^{i^*} \int \|t\|^{2+i^*} \alpha_n^{i^*} E|g_i(\theta_0 + t/\alpha_n)|^{2+i^*} \phi_V(t)dt = o(1)O(1)O(1) = o(1),
\]
by assumption [E].

\[\Box\]

**Lemma D.3.** Let \(\mathcal{V}_n = \pi_0^2 \left\{ \left| \int t^T H(t,s) \phi_V(t)\phi_V(s)dt\right| \right\} dt \rightarrow N(0,\mathcal{V}_N)\).

Proof. By lemma \([A.1]\) \(\int (\pi_n(t) - \pi_0) \pi_0 t S_n(t) \phi_V(t)dt = o_p(1)\). Further,
\[
\pi_0 \int t \pi_0 t S_n(t) \phi_V(t)dt = n^{-1/2} \sum_{i=1}^n \pi_0 \sqrt{\alpha_n} \int t g_i(\theta_0 + t/\alpha_n) \phi_V(t)dt \overset{d}{\rightarrow} N(0,\mathcal{V}_N),
\]
by Lindeberg’s theorem (see e.g. theorem 23.6 in [Davidson, 1994]), the Lindeberg condition is satisfied by lemma \([D.2]\).

\[\Box\]

**Lemma D.4.** \(\int \pi_n(t) t S_n(t) \left( \exp(R_n(t)) - 1 \right) \phi_V(t)dt = o_p(1)\).

Proof. For \(\int_{\Gamma_n}\), use lemma \([A.7]\) and for \(\int_{\Gamma_n}\) follow the same steps as in the proof of lemma \([B.1]\).

\[\Box\]

**Lemma D.5.** For any \(j = 0, 1\), \(\int \|t\|\pi_n(t) |\exp(\beta_n S_n(t)) + Q_n(t)|dt = o_p(1)\).

Proof. Define \(S_n^* = \sup_{t \in \Gamma_n} |\tilde{S}_n(t) - t^T Vt/4|\). For \(\int_{\Gamma_n}\) the stated result follows from lemma \([A.7]\) So we only need to deal with
\[
\int_{\Gamma_n} \|t\|\pi_n(t) |\exp(\beta_n S_n(t)) + Q_n(t)|dt \leq \exp(\beta_n S_n^*) \int_{\Gamma_n} \|t\|\pi_n(t) |\exp(\beta_n t^T Vt/4 + Q_n(t))|dt.
\]
Since \(\lim_{\|t\| \to 0} \frac{Q_n(t)}{t^T Vt} = -1/2\), it follows that for sufficiently large \(n\), \(Q_n(t) \leq -t^T Vt/4\) for all \(t \in \Gamma_n\). Because \(\beta_n = o(1)\) and \(S_n^* = o_p(1)\) by lemma \([A.3]\) the stated result follows.

\[\Box\]
Lemma D.6. For \( j = 0, 1 \), \( \int \pi_n(t) t^j \{ \exp(\beta_n \mathcal{S}_n(t)) - \sum_{j=0}^1 (\beta_n \mathcal{S}_n(t))^j \} \exp(Q_n(t)) dt = O_p(\beta_n^{j+1}). \)

Proof. Apply lemma A.6 with \( c = \beta_n \) and \( b = \mathcal{S}_n(t) \), followed by lemma D.5.

Lemma D.7. \( \mathcal{N}_n - \mathcal{B}_n \xrightarrow{d} N(0, C_V \mathcal{Y}_n). \)

Proof. Use \( t = \alpha_n(\theta - \theta_0) \) to obtain

\[
\mathcal{N}_n - \mathcal{B}_n = \frac{1}{\beta_n} \int \pi_n(t) t \{ \exp(\beta_n \mathcal{S}_n(t)) - 1 \} \exp(Q_n(t)) dt = \int \pi_n(t) t \mathcal{S}_n(t) \exp(Q_n(t)) dt + \frac{1}{\beta_n} \int \pi_n(t) t \{ \exp(\beta_n \mathcal{S}_n(t)) - 1 - \beta_n \mathcal{S}_n(t) \} \exp(Q_n(t)) dt. \tag{26}
\]

The first right hand side term in (26) converges in distribution to the stated normal by lemmas D.3 and D.4. The last term in (26) is \( O_p(\beta_n) = o_p(1) \) by lemma D.6.

Lemma D.8. \( \mathcal{D}_n = \pi_0 C_V + o_p(1). \)

Proof. Use \( t = \alpha_n(\theta - \theta_0) \) to obtain

\[
\mathcal{D}_n = \int \pi_n(t) \exp(\beta_n \mathcal{S}_n(t) + Q_n(t)) dt = \int \pi_n(t) \exp(Q_n(t)) dt + \int \pi_n(t) \{ \exp(\beta_n \mathcal{S}_n(t)) - 1 \} \exp(Q_n(t)) dt. \tag{27}
\]

The first right hand side term in (27) is \( \pi_0 C_V + o_p(1) \) by lemma D.1. The last term in (27) is \( o_p(1) \) by lemma D.6.

**Appendix E. Bias**

**E.1. Bias Expansion.** In this subsection, the assumptions of theorem 5 are used.

**Lemma E.1.** For some finite \( C_R \), \( \sup_{t \in \Gamma_n} |R_n(t)| \leq C_R \gamma_n^3 / \alpha_n = o(1). \)

Proof. Because \( Q \) is three times continuously differentiable at \( \theta_0 \),

\[
\sup_{t \in \Gamma_n} |R_n(t)| \leq \gamma_n^2 \sup_{\|\theta - \theta_0\| \leq \gamma_n / \alpha_n} \| Q''(\theta) + V \| / 2 \leq C_R \gamma_n^3 / \alpha_n,
\]

which is \( o(1) \) by the definition of \( \gamma_n. \)

**Lemma E.2.**

\[
\int_{\Gamma_n} \pi_n(t) t \left( \exp(R_n(t)) - \sum_{p=0}^q \frac{(R_n(t))^p}{p!} \right) \phi_V(t) dt = O((\gamma_n^3 / \alpha_n)^{q+1}).
\]
Proof. The length of the left hand side is by lemma \[\text{Lemma A.6}\] for \( s_n^* = \sup_{t \in \Gamma_n} |R_n(t)| \) equal to
\[
\left\| \sum_{p=q+1}^{\infty} \int_{\Gamma_n} \pi_n(t) t \left( \frac{R_n(t)}{p!} \right)^p \phi_V(t) dt \right\| \leq (s_n^*)^{q+1} \exp(s_n^*) \int_{\Gamma_n} \| \pi_n(t) t \| \phi_V(t) dt = O((s_n^*)^{q+1}).
\]

Apply lemma \[\text{Lemma E.1}\] \( \Box \)

Let \( \mathcal{M}_{pq} \) be the collection of \( q \)-vectors \( m \) of nonnegative integers for which \( \sum_{\ell=1}^{q} m_\ell = p \). Let further \( \mathcal{M}_{pq}^* = \{ m \in \mathcal{M}_{pq} : \sum_{\ell=1}^{q} \ell m_\ell = r \} \), let \( D_{Q^\delta}(t) \) be the \( \delta \)-th term in a Taylor expansion of \( Q(\theta_0 + t) \) about \( Q(\theta_0) \) and let \( D_{\pi^\delta}(t) \) be likewise for \( \pi \), e.g. \( D_{Q^\delta}(t) = t^T Q''(\theta_0) t/2. \) Finally, let \( D_{\pi^\delta t}(t) \) be \( \alpha_t^\delta \) times the remainder term in a \( \delta \)-th order Taylor expansion of \( Q(\theta_0 + t/\alpha_t) \) about \( Q(\theta_0) \) and let \( D_{\pi^\delta t}(t) \) be likewise for \( \pi \).

**Lemma E.3.** For any vector \( a \in \mathbb{R}^d \), \( (\sum_{\ell=1}^{q} a_\ell)^p / p! = \sum_{m \in \mathcal{M}_{pq}} \Pi_{\ell=1}^{q} a_\ell^{m_\ell} / m_\ell! \).

*Proof.* This is a restatement of the multinomial theorem. \( \Box \)

**E.2. Bias Correction.** In this section the assumptions of theorem \[\text{Theorem 7}\] are used; \( r^\ast, \rho_n \) are as defined in assumption \[\text{Assumption II}\] take \( \pi, \hat{\pi} \) to equal \( \pi_j, \hat{\pi}_j \), let \( \hat{\pi}_0 = \hat{\pi}(\theta_0) \), and \( \hat{\pi}_n(t) = \hat{\pi}(\theta_0 + t/\alpha_t) \). Let further \( v(\theta) = \text{vec}\{(-Q''(\theta))^{-1}\} \) and \( \theta(\theta) = \text{vec}\{(-Q''(\theta))^{-1}\} \).

**Lemma E.4.** (i) \( \hat{\pi}_0 = \pi_0 + o_p(\alpha^{-r^\ast}) \) and (ii) \( \sup_{\|\theta - \theta_0\| \leq \rho_n} \| \hat{\pi}'(\theta) - \pi'(\theta) \| = o_p(\alpha^{-r^\ast}). \)

*Proof.* Part (i) is implied by assumptions \[\text{Assumption II}\] and \[\text{Lemma E.3}\] (\( \det(V) > 0 \)). For (ii), note that
\[
\pi'(\theta) = -\pi(\theta) Q'''(\theta) v(\theta)/2, \tag{28}
\]
in a neighborhood of \( \theta_0 \). Hence in such a neighborhood,
\[
2(\hat{\pi}'(\theta) - \pi'(\theta)) = -(\hat{\pi}_0 - \pi_0) Q'''(\theta) \hat{\phi}(\theta) - \pi_0 \{ Q'''(\theta) \hat{\phi}(\theta) - Q''(\theta) v(\theta) \}.
\]
Apply part (i) and assumption \[\text{Assumption II}\] \( \Box \)

**Lemma E.5.** \( b_{q1}^* = \int (\pi_0 D_{\pi^\delta t}(t) + D_{\pi^\delta t}(t)) t \phi_V(t) dt = 0. \)
Proof. Let $V^{j\ell}$ denote the $(j, \ell)$ element of $V^{-1}$ and let $\mathcal{K}_{p\ell} = \partial_{\theta_p, \theta_\ell} Q(\theta_0)$. Then the $j$-th element of $b_{q1}^*$ is equal to

$$
\frac{\pi_0}{6} \sum_{p,s,\ell=1}^d \mathcal{K}_{p\ell} \int t_p t_s t_j \phi_V(t) dt + \sum_{\ell=1}^d \partial_{\theta_\ell} \pi(\theta_0) \int t_j \phi_V(t) dt
$$

$$
= \frac{\pi_0}{6} \left( \sum_{p,s,\ell=1}^d \mathcal{K}_{p\ell} (V^{p\ell}V^{j\ell} + V^{p\ell}V^{sj} + V^{pj}V^{s\ell}) \right) - 3 \sum_{p,s,\ell=1}^d \mathcal{K}_{p\ell} V^{p\ell}V^{j\ell} = 0,
$$

where the first equality follows from (28).

Lemma E.6. \(\int (\alpha_n + D_{Q3}(t)) \hat{\pi}_n(t) t \phi_V(t) dt = o_p(\alpha_n^{-r})\).

Proof. The left hand side in the lemma statement is by the mean value theorem for some $t^*$ between zero and $t$ equal to

$$
\int t t^T \hat{\pi}_n'(t) \phi_V(t) dt + \hat{\pi}_0 \int D_{Q3}(t) t \phi_V(t) dt + \alpha_n^{-1} \int t t^T \hat{\pi}_n'(t^*) D_{Q3}(t) \phi_V(t) dt,
$$

which by lemma E.3 is equal to

$$
\int t t^T (\hat{\pi}_n'(t) - \pi_0') \phi_V(t) dt + (\hat{\pi}_0 - \pi_0) \int D_{Q3}(t) t \phi_V(t) dt
$$

$$
+ \alpha_n^{-1} \int t t^T (\hat{\pi}_n'(t^*) - \pi_0') D_{Q3}(t) \phi_V(t) dt + \alpha_n^{-1} \int D_{\pi_1}(t) D_{Q3}(t) t \phi_V(t) dt,
$$

whose first three terms are $o_p(\alpha_n^{-r})$ by lemma E.4 dominated convergence, and assumption C and whose last term is zero because of the symmetry of the normal distribution.

APPENDIX F. UNIFORM INFERENCES

The results in appendix F presume that the assumptions of theorem 8 are satisfied.

Lemma F.1. \(\hat{\mathcal{G}} \overset{\text{d}}{=} \mathcal{G}\) in $L^\infty(\mathcal{T}_1, \mathcal{T}_2, \ldots)$.

Proof. For any fixed and distinct $t_1, \ldots, t_j$, we can write

$$
\begin{bmatrix}
\hat{\mathcal{G}}(t_1) \\
\vdots \\
\hat{\mathcal{G}}(t_j)
\end{bmatrix} = \hat{\Omega}^{1/2}(t_1, \ldots, t_j) \eta,
$$

where $\hat{\Omega}$ is the estimated covariance matrix.
with \( \eta \sim N(0, I_j) \) and \( \hat{\Omega}(t_1, \ldots, t_j) \in \mathbb{R}^{l \times j} \) with \((t, p)\)-element \( \hat{H}(t, t_p) \). Let \( \Omega \) be as \( \hat{\Omega} \) with \( \hat{H} \) replaced by \( H \). Since \( \hat{H} \) is a consistent estimator of \( H \),

\[
\hat{\Omega}^{1/2}(t_1, \ldots, t_j) \eta = \Omega^{1/2}(t_1, \ldots, t_j) \eta + o_p(1) \sim N(0, \Omega(t_1, \ldots, t_j)) + o_p(1),
\]

which is the joint distribution of \( G(t_1), \ldots, G(t_j) \). The stated result then follows from the fact that \( \hat{G} \) is tight.

**Lemma F.2.** For any \( c = [c_1, c_2]^T \) with \( 0 < c_1, c_2 < \infty \), let \( \hat{G}_n(t) = c_1 \hat{G}(t) - c_2 t^T \hat{\nu} t/2 \) and \( \hat{G}_n(t) = c_1 G(t) - c_2 t^T \nu t/2 \). Then \( \hat{G}_n(\cdot; \omega) \xrightarrow{d} \hat{G}(\cdot; \omega) \) in \( L^\infty(\mathcal{F}_1, \mathcal{F}_2, \ldots) \).

**Proof.** Follows from lemma F.1 combined with the consistency of \( \hat{\nu} \).

**Lemma F.3.** For \( c \) as in lemma F.2 and any vector–valued function \( \psi \) for which for some \( c_\psi < \infty \), \( \sup_j \| \psi(t) \| / (\| t \|^c \psi + 1) < \infty \),

\[
\int \psi(t) \exp(c_1 \hat{G}(t) - c_2 t^T \hat{\nu} t/2) dt \overset{d}{\xrightarrow{} \int \psi(t) \exp(c_1 G(t) - c_2 t^T \nu t/2) dt.
\]

**Proof.** By lemma F.2 and the Skorokhod representation theorem [Billingsley 1999, theorem 6.7], there exist \( \hat{G}^*, \hat{G}^* \), such that \( \hat{G}^*, \hat{G}^* \) have the same properties as \( \hat{G}, \hat{G} \) and such that for all \( \omega: \hat{G}^*(\cdot; \omega) \to \hat{G}^{**}(\cdot; \omega) \). Since \( \| \psi(t) \| \exp(c_1 \hat{G}(t) - c_2 t^T \hat{\nu} t/2) \) is a.s. integrable by the assumptions on \( \hat{H}, \psi \), the stated result then follows from the dominated convergence theorem [Billingsley 1995, theorem 16.4].

**Lemma F.4.** If \( 1 = o(\beta_n) \) then

\[
\int t \exp(\beta_n^{4/3}(\hat{G}(t) - t^T \hat{\nu} t/2)) dt / \int \exp(\beta_n^{4/3}(\hat{G}(t) - t^T \hat{\nu} t/2)) dt \overset{d}{\xrightarrow{} \frac{d}{d \in \mathbb{R}^d}} \arg\max \hat{G}(t).
\]

**Proof.** Let \( \hat{\psi}_0, \hat{\psi}_1, \hat{\psi}_0, \hat{\psi}_1 \in L(\mathbb{R}^d, B, \mu) \) and \( (\hat{\psi}_0, \hat{\psi}_1) \xrightarrow{w} (\psi_0, \psi_1) \). Repeat the arguments of lemma C.1 following (20).

**Lemma F.5.** If \( \beta_n = o(1) \) then

\[
\frac{1}{\beta_n} \int t^j \exp(\beta_n^{3/2}(\hat{G}(t/\beta_n))) \psi(t) dt = o_p(1) + \begin{cases} 1, & j = 0, \\ \sqrt{\beta_n} \int t \hat{G}(t) \psi(t) dt, & j = 1. \end{cases}
\]
Proof. We show the stated result for \( j = 1 \); the case \( j = 0 \) is similar. By lemma \( \text{A.6} \), we have

\[
\left\| \hat{\beta}_{n}^{-1} \int t \exp \left( \hat{\beta}_{n}^{3/2} \hat{G}(t/\beta_{n}) \right) \Phi_{\beta}(t) dt - \sqrt{\beta_{n}} \int t \hat{G}(t/\beta_{n}) \Phi_{\beta}(t) dt \right\|
\leq \beta_{n}C_{V}^{-1} \int \| t \| \exp \left( \sqrt{\beta_{n}} |\hat{G}(t/\beta_{n})| - t^{T} \hat{V}t/2 \right) dt,
\]

which has the same distribution as \( \beta_{n}C_{V}^{-1} \int \| t \| \exp \left( |\hat{G}(t)| - t^{T} \hat{V}t/2 \right) dt = O_{p}(\beta_{n}) = o_{p}(1). \]

Lemma F.6. If \( \beta_{n} = o(1) \) then
\[
\sqrt{\beta_{n}} \int t \hat{G}(t/\beta_{n}) \Phi_{\beta}(t) dt \xrightarrow{d} N(0, \nu').
\]

Proof. The left hand side in (29) has the same distribution as \( \int t \hat{G}(t) \Phi_{\beta}(t) dt \), which by lemma \( \text{F.1} \), the assumption that \( \hat{V} = V + o_{p}(1) \), and the conditions on \( \hat{H} \), converges in distribution to
\[
\int t G(t) \Phi_{\nu}(t) dt,
\]

Because \( \tilde{S}_{n} \xrightarrow{d} G \), (30) is also the limit of \( \int t \tilde{S}_{n}(t) \Phi_{\nu}(t) dt \), which lemma \( \text{D.3} \) established to have a limiting \( N(0, \nu') \) distribution. \( \square \)

APPENDIX G. RESULTS SPECIFIC TO THE MAXIMUM SCORE CASE

G.1. \( V, H, \nu' \).

Lemma G.1. \( V = -2E[z_{i}z_{i}^{T}pa(z_{i}^{T}0, z_{i})f(z_{i}^{T}0|z_{i})]. \)

Proof. Let \( \partial_{\theta} \) denote a partial derivative with respect to \( \theta \). In view of the definition of \( V \), consider

\[
\partial_{\theta}Eg_{i}(\theta) = \partial_{\theta} \int \left( (2p(a, z_{i}) - 1) I(a_{i} \leq z_{i}^{T}\theta) \right) da = \partial_{\theta} \int \left( z_{i}^{T} \left( 2p(a_{i}, z_{i}) - 1 \right) f(a_{i}|z_{i}) \right) da
\]

\[
= \partial_{\theta} [z_{i}z_{i}^{T} \left( 2p(z_{i}^{T}\theta, z_{i}) - 1 \right) f(z_{i}^{T}\theta|z_{i})]
\]

\[
= E[z_{i}z_{i}^{T} \left\{ 2pa(z_{i}^{T}\theta, z_{i})f(z_{i}^{T}\theta|z_{i}) + (2p(z_{i}^{T}\theta, z_{i}) - 1)f'(z_{i}^{T}\theta|z_{i}) \right\}],
\]

which at \( \theta_{0} \) equals \( 2E[z_{i}z_{i}^{T}pa(z_{i}^{T}\theta_{0}, z_{i})f(z_{i}^{T}\theta_{0}|z_{i})] = -V. \)

Lemma G.2. If \( u_{i} \) is the error term in the latent variable equation of Manski (1975) and Manski’s conditional median assumption is satisfied, then \( V = 2E[z_{i}z_{i}^{T}fu_{i}|z(0, z_{i}^{T}\theta_{0}|z_{i})]. \)

Proof. Note first that \( p(a, z) = 1 - F_{u_{i}}(a - z^{T}\theta_{0}, z) \), whose partial derivative with respect to \( a \) for \( a = z^{T}\theta_{0}, z = z \) is \( f_{u_{i}}(0)z^{T}\theta_{0}, z) \) since \( F_{u_{i}}(0|a, z) = 1/2 \) for all values of \( a, z \) by the conditional median assumption. \( \square \)
Lemma G.3. \( H(t,s) = \mathbb{E}[|M(z_1^T t, z_1^T s, 0)|f(z_1^T \theta_0 | z_i)] \).

Proof. Let

\[
\mathcal{H}(t,s) = \lim_{a \to \infty} \mathbb{E} \left[ \alpha \int_{z_1^T \theta_0}^{z_1^T \theta_0 + \min(z_1^T t, z_1^T s)/\alpha} f(a | z_i) da \right] = \mathbb{E} \left[ \min(z_1^T t, z_1^T s) f(z_1^T \theta_0 | z_i) \right].
\]

by the dominated convergence theorem. Thus, noting that \((2y_i - 1)^2 = 1\), we have

\[
H(t,s) = \mathcal{H}(t,s) - \mathcal{H}(t,0) - \mathcal{H}(0,s) + \mathcal{H}(0,0)
\]

\[
= \mathbb{E} \left[ \{\min(z_1^T t, z_1^T s) - \min(z_1^T t, 0) - \min(z_1^T s, 0)\} f(z_1^T \theta_0 | z_i) \right] = \mathbb{E} \left[ |M(z_1^T t, z_1^T s, 0)| f(z_1^T \theta_0 | z_i) \right].
\]

\[\Box\]

Lemma G.4. If for some \( j = 1, \ldots, d \), \( \mathbb{P}[z_{ij} = 0] = 0 \), then \((11)\) holds.

Proof. We first establish the following five results for a generic variance matrix \( \Sigma \in \mathbb{R}^{d \times d} \), where \( \Sigma_{ij} \) denotes the \( 2 \times 2 \) submatrix of \( \Sigma \) containing \( \sigma_{ii}, \sigma_{ij}, \sigma_{ji}, \sigma_{jj} \) with \( \sigma_{ij} \) the \( (i,j) \) element of \( \Sigma \).

\[
\int_{t_1}^{\infty} \int_{s_1}^{\infty} s_1 \phi_{\Sigma_{ij}}(s_1, s_j) ds_1 ds_j = \sigma_{ij} \phi(t_1 / \sigma_1) / \sigma_1, \tag{31}
\]

\[
\int_{t_1}^{\infty} s_1 \phi(s_1 / \sigma_1) ds_1 / \sigma_1 = \sigma_{ij} \phi(t_1 / \sigma_1), \tag{32}
\]

\[
\int_{t_1}^{\infty} \int_{s_1}^{\infty} t_\ell s_j \phi_{\Sigma_{ij}}(s_1, s_j) \phi_{\Sigma_{\ell \ell}^2/4}(t_1, t_\ell) ds_1 ds_j dt_\ell = \sigma_{ij} \sigma_{i/2} / 4 \sigma_1 \sqrt{\pi}, \tag{33}
\]

\[
\int_{t_1}^{\infty} t_\ell s_j \min(t_1, s_1) \phi_{\Sigma_{\ell \ell}^2/4}(s_j) \phi_{\Sigma_{\ell \ell}^2/4}(t_\ell) ds_1 dt_\ell = \sigma_{ij} \sigma_{i/2} / 2 \sigma_1 \sqrt{\pi}, \tag{34}
\]

\[
\int_{t_1}^{\infty} t s^T \min(t, s_1) \phi_{\Sigma_{\ell \ell}^2/4}(s_j) \phi_{\Sigma_{\ell \ell}^2/4}(t_\ell) ds_1 dt_\ell = \Sigma e_1^T \Sigma / 2 \sqrt{\pi e_1^T \Sigma e_1}, \tag{35}
\]

where \( e_1 \) is the first unit vector. The left hand side in \((31)\) is

\[
\frac{\sigma_{ij}}{\sigma_1} \int_{t_1}^{\infty} s_1 \phi(s_1 / \sigma_1) ds_1 = -\frac{\sigma_{ij}}{\sigma_1} \int_{t_1 / \sigma_1}^{\infty} \phi'(s_1) ds_1 = \frac{\sigma_{ij}}{\sigma_1} \phi(t_1 / \sigma_1).
\]

Equality \((32)\) is similar, but easier to establish, and the left hand side in \((33)\) is by \((31)\) equal to

\[
\frac{\sigma_{ij}}{\sigma_1} \int_{t_1}^{\infty} t_\ell t_\ell \phi(t_1 / \sigma_1) \phi_{\Sigma_{\ell \ell}^2/4}(t_1, t_\ell) dt_\ell dt_1 = \frac{\sigma_{ij} \sigma_{i/2}}{\sigma_1^4} \int_{t_1 / \sigma_1}^{\infty} t_\ell^2 \phi^2(t_1 / \sigma_1) dt_1 = \frac{\sigma_{ij} \sigma_{i/2}}{\sqrt{3} \sigma_1} \int_{t_1 / \sigma_1}^{\sqrt{2} \sigma_1} t_\ell^2 \phi^2(t_1 / \sqrt{2}) dt_1
\]

\[
= \frac{\sigma_{ij} \sigma_{i/2}}{4 \sigma_1 \sqrt{\pi}}.
\]

Equality \((34)\) then follows by applying \((33)\) twice and \((35)\) by repeatedly using \((34)\).
Now suppose without loss of generality that \( j = 1 \) and denote the remaining elements by \( \tilde{z}_i \). Let further

\[
A_i = \begin{bmatrix} z_i & \tilde{z}_i^T \\ 0 & I_{d-1} \end{bmatrix}, \quad V_i = (A_i^*)^{-1}VA_i^{-1}, \quad f_i = f(z_i^T \theta_0 | z_i).
\]

Then by substitution of \( \tilde{t} = A_it \) and \( \tilde{s} = A_is \),

\[
\mathcal{V} = \int \int ts^TH(t,s)\phi_V(t)\phi_V(s)dt ds = \mathbb{E} \left[ f_i \int \int ts^T M(z_i^T t, z_i^T s, 0) | \phi_V(t)\phi_V(s) dtds \right] = \mathbb{E} \left[ f_i A_i^{-1} \int \int \tilde{t}\tilde{s}^T M(\tilde{t}_1, \tilde{s}_1, 0) | \phi_V(t)\phi_V(s) d\tilde{t}d\tilde{s}(A_i^*)^{-1} \right] = \frac{1}{2\sqrt{\pi}} V^{-1} \mathbb{E} \left[ \frac{f_i A_i^{-1}V_i^{-1}e_1^T V_i^{-1}(A_i^*)^{-1}}{e_1^T V_i^{-1} e_1} \right] = \frac{1}{2\sqrt{\pi}} V^{-1} \mathbb{E} \left[ \frac{z_i z_i^T}{\sqrt{z_i^T V_i^{-1} z_i}} f(z_i^T \theta_0 | z_i) \right] V^{-1},
\]

where the penultimate inequality follows from (35) and the last from the fact that \( A_i^{-1}V_i^{-1} = V^{-1}A_i^* \) and that \( A_i^* e_1 = z_i \). \( \Box \)

G.2. Weak Convergence. In this section, we show that assumption \( \mathcal{H} \) is satisfied in the maximum score case. In view of equation (31), part \( \mathcal{H} \) is satisfied. Let \( \mathcal{J}(\mathcal{A}) \) be the Vapnik–Černovenkis (VC) index of a function class \( \mathcal{A} \). Let \( \mathcal{J} \subset \mathbb{R}^d \) be an arbitrary compact set.

**Lemma G.5.** Let \( \vartheta_n(\xi, t) = I(z^T \theta_0 + t/\tilde{a}_n) \geq a \) and \( \tilde{\mathcal{F}}_n = \{ \vartheta_n(\cdot ; t) \}_{t \in \mathcal{J}} \) and every \( n, \mathcal{J}(\tilde{\mathcal{F}}_n) \leq d + 3 \).

**Proof.** Since \( \tilde{\mathcal{F}}_n \) is a collection of indicator functions, its VC index is equal to the VC index of the collection of sets \( \{(z, a) : z^T \theta_0 + t/\tilde{a}_n \geq a \} \) with \( t \) ranging over \( \mathcal{J} \). For every \( n \), this collection is a subcollection of the sets \( \{(z, a) : z^T t + as \geq 0 \} \) with \( (t, s) \) ranging over \( \mathbb{R}^{d+1} \). The VC index of the latter collection of sets is equal to \( d + 3 \) by Kosorok (2008, lemma 9.12). \( \Box \)

**Lemma G.6.** For all \( n, \mathcal{J}(\tilde{\mathcal{F}}_n) \leq 2d + 5 \).

**Proof.** Note that for every element \( v \in \tilde{\mathcal{F}}_n \) there is an element \( \tilde{v}_n \in \tilde{\mathcal{F}}_n \) such that \( v_n(\xi) = \sqrt{\tilde{a}_n}(2y - 1)(\tilde{v}_n(\xi) - I(z^T \theta_0 \geq a)) \). Therefore, \( \mathcal{J}(\tilde{\mathcal{F}}_n) \) is bounded by \( 2\mathcal{J}(\tilde{\mathcal{F}}_n) - 1 \) by Kosorok (2008, lemma 9.9). The conclusion then follows from lemma G.5. \( \Box \)

Let \( t_n(z) = \text{argmin}_{t \in \mathcal{J}} z^T(\theta_0 + t/\tilde{a}_n) \) and \( s_n(z) = \text{argsup}_{t \in \mathcal{J}} z^T(\theta_0 + t/\tilde{a}_n) \) and note that \( F_n \) defined in assumption \( \mathcal{H} \) is \( F_n(\xi) = \sqrt{\tilde{a}_n} I\{z^T(\theta_0 + t_n(z)/\tilde{a}_n) < a \leq z^T(\theta_0 + s_n(z)/\tilde{a}_n) \} \).
Lemma G.7. $E_{ni}^2 = O(1)$.

**Proof.** By the law of iterated expectations,

$$E_{ni}^2 = \tilde{\alpha}_n E \left[ P \left( z_i^T \left( \theta_0 + t_n(z_i)/\tilde{\alpha}_n \right) < a_i \leq z_i^T \left( \theta_0 + s_n(z_i)/\tilde{\alpha}_n \right) \right) \right]$$

$$\leq E \left[ \sup_s f_{a|z}(s|z_i)||z_i||t_n(z_i) + s_n(z_i)|| \right].$$

The stated result then follows from $t_n(z_i), s_n(z_i) \in \mathcal{T}$ and $E[\sup_s f_{a|z}(s|z_i)||z_i||] < \infty.$

**Lemma G.8.** For all $\epsilon > 0$, $E \left[ F_{ni}^2 I \left( F_{ni} > \epsilon \sqrt{n} \right) \right] = o(1)$.

**Proof.** Follows from $n/\tilde{\alpha}_n \to \infty.$

**Lemma G.9.** For every $\epsilon_n \downarrow 0$, $\sup_{\|s-t\| \leq \epsilon_n} E \left[ \tilde{\alpha}_n \left( g_i(\theta_0 + t/\tilde{\alpha}_n) - g_i(\theta_0 + s/\tilde{\alpha}_n) \right)^2 \right] = o(1)$.

**Proof.** Follows from

$$E \left[ \tilde{\alpha}_n \left( g_i(\theta_0 + t/\tilde{\alpha}_n) - g_i(\theta_0 + s/\tilde{\alpha}_n) \right)^2 \right] \leq \tilde{\alpha}_n E \left[ P \left( z_i^T \left( \theta_0 + t/\tilde{\alpha}_n \right) < a_i \leq z_i^T \left( \theta_0 + s/\tilde{\alpha}_n \right) \right) \right]$$

$$+ \tilde{\alpha}_n E \left[ P \left( z_i^T \left( \theta_0 + s/\tilde{\alpha}_n \right) < a_i \leq z_i^T \left( \theta_0 + t/\tilde{\alpha}_n \right) \right) \right] \leq 2E \left[ \sup_s f_{a|z}(s|z_i)||z_i|| \right] ||t - s||.$$

**Lemma G.10.** $\tilde{S}_n \xrightarrow{w} G$ in $L^\infty(\mathcal{T}_1, \mathcal{T}_2, \ldots)$, where $G$ is a Gaussian process with covariance kernel $H$ such that $H(t, t) = O(||t||)$ as $||t|| \to \infty$.

**Proof.** Recall that

$$\tilde{\alpha}_n E \left[ g_i(\theta_0 + t/\tilde{\alpha}_n)g_i(\theta_0 + s/\tilde{\alpha}_n) \right] \to H(s, t) = E \left[ f(z_i^T \theta_0|a_i)| M(z_i^T t, z_i^T s, 0) \right].$$

By lemmas G.8, G.9 and theorem A.1, showing that part V holds will complete the proof. Note however that by lemma G.6 each $\mathcal{F}_n$ is a V$	ilde{\text{C}}$ class of which the V$	ilde{\text{C}}$ index is bounded by $2d + 5$. Therefore, for $0 < \epsilon < 1$,$$
\mathcal{N} \left( \epsilon \|\mathcal{T}\|_2, \mathcal{T}_n, L_2(\mathcal{D}) \right) \leq C\epsilon^{-2(2d + 5)}$$

for some constant $C$ that only depends on $d$; see e.g. van der Vaart and Wellner (1996) theorem 2.6.7. Hence, part V of assumption C is satisfied, and the conclusion follows.

**APPENDIX H. PROOFS OF THEOREMS**

Please note that the proofs use some additional notation that was introduced in appendix A.1

**Proof of Theorem I.** See lemma A.1.
Proof of Theorem 2
See lemma B.10.

Proof of Theorem 3
First (i). Note that by substitution of \( t = \alpha_n(\theta - \theta_0) \) we get

\[
\sqrt{\frac{n}{\alpha_n}} \left( \hat{\theta} - \theta_0 \right) = \sqrt{\frac{n}{\alpha_n}} \int \frac{\pi(\theta)\exp\{\alpha_n^2(S_n(\theta) + Q(\theta))\}d\theta}{\int \pi(\theta)\exp\{\alpha_n^2(S_n(\theta) + Q(\theta))\}d\theta}
\]

\[= \sqrt{\frac{n}{\alpha_n}} \int \pi_n(t)\exp\{c_n^2\tilde{S}_n(t) + Q_n(t)\}dt = \frac{1}{c_n^2} \int \frac{\pi\cdot\exp(c_n^2\tilde{S}_n(t))\phi(t)dt}{\exp(c_n^2\tilde{S}_n(t))\phi(t)dt} + o_p(1),\]

by lemma B.2. Multiply both sides by \( c_n \), then apply lemma A.1 and the continuous mapping theorem.

Now (ii). By substitution of \( t = \tilde{\gamma}n(\theta - \theta_0) \) we get

\[\tilde{\gamma}n(\hat{\theta} - \theta_0) = \int \frac{\pi_n(t)\exp\{c_n^4/3\tilde{S}_n(t) + Q_n(t)\}dt}{\int \pi_n(t)\exp\{c_n^4/3\tilde{S}_n(t) + Q_n(t)\}dt} = \frac{N}{\hat{D}}.
\]

Now, for \( \tilde{\Gamma} = \{ t : \|t\| \leq \tilde{\gamma} \} \) for finite positive \( \tilde{\gamma} \). Let further \( N_F = \int_{\Gamma} \pi_n(t)\exp\{\cdots\}dt, \)

\( N_{\hat{F}} = \int_{\tilde{\Gamma}} \pi_n(t)\exp\{\cdots\}dt, \)

\( D_F = \int_{\Gamma} \pi_n(t)\exp\{\cdots\}dt, \)

\( D_{\hat{F}} = \int_{\tilde{\Gamma}} \pi_n(t)\exp\{\cdots\}dt. \)

Simple manipulations show that

\[\frac{N}{D} = \frac{N_F}{D_F} \left( 1 - \frac{D_F}{D} \right) + \frac{N_{\hat{F}}}{D_{\hat{F}}}.
\]

By lemma C.2 both \( D_{\hat{F}} / D \) and \( N_{\hat{F}} / D \) are \( o_p(1) \) and by lemma C.1, \( N_{\hat{F}} / D_{\hat{F}} \xrightarrow{d} \arg\max_{t \in \hat{G}(t)} G(t) \) for any \( \tilde{\gamma} > 0. \) Let \( \tilde{\gamma} \to \infty. \)

Proof of Theorem 4
Combine lemmas D.7 and D.8.

Proof of Theorem 5
For \( \mathcal{M}_{p,q} \) defined as in appendix E. By lemmas A.7 and E.2, the difference between \( \mathcal{R}_n \) and

\[\mathcal{R}_n = \frac{C \pi}{\hat{D}} \sum_{p=0}^{q} \frac{1}{p!} \int \pi_n(t)\exp(\hat{R}_n(t))^{p}\phi(t)dt, \]

is \( o(\alpha_n^{-q}\beta_n^{-1}). \) We employ the Taylor expansions

\[\hat{R}_n(t) = \sum_{\delta=1}^{q} \frac{D_Q(t)}{\alpha_n^{\delta}} + \frac{D_{Q_n}(t)}{\alpha_n^{q}}, \quad \pi_n(t) = \sum_{\ell=0}^{q} \frac{D^\ell(t)}{\alpha_n^{\ell}} + \frac{D^\ell_{Q_n}(t)}{\alpha_n^{q}}.\]

Note that \( \sup_{t \in \Gamma \setminus \{0\}} \left( |D_{Q_n}(t)| / \|t\|^\Delta + |D^\ell_{Q_n}(t)| / \|t\|^q \right) = o(1) \) by the continuity of the \( \Delta \)-th derivative of \( Q \) and the \( q \)-th derivative of \( \pi \) at \( \theta_0 \), and the fact that \( \gamma_n = o(\alpha_n). \) Hence, by lemmas A.7 and A.9 it follows that any expressions in an expansion of (36) using (37) involving the remainder terms in the Taylor expansions of (37) are \( o(\alpha_n^{-q}\beta_n^{-1}). \)
Omitting the remainder terms in the definition of $\mathcal{B}_n^*$ yields

$$
\mathcal{B}_n^* = \frac{C_V}{\beta_n} \sum_{p=0}^{q} \frac{1}{p!} \int \left( \sum_{\ell=0}^{q} \frac{D_n(t)}{\alpha_n^\ell} \right) \left( \sum_{\delta=1}^{\delta} \frac{D_n^{\delta+2}(t)}{\alpha_n^{\delta}} \right)^p t \phi_V(t) dt.
$$

Using lemma [E.3] we obtain

$$
\mathcal{B}_n^* = \frac{C_V}{\beta_n} \sum_{p=0}^{q} \int \left( \sum_{\ell=0}^{q} \frac{D_n(t)}{\alpha_n^\ell} \right) \sum_{\delta=1}^{\delta} \frac{1}{m_\delta!} \left( \sum_{\delta=1}^{\delta} \frac{D_n^{\delta+2}(t)}{\alpha_n^{\delta}} \right)^p t \phi_V(t) dt
$$

and $b_{q^\ell} = C_V \sum_{\tau=0}^{\tau} b_{q^\tau}$. Finally, $b_{q^\ell} = 0$ for even-valued $\tau$ because $b_{q^\tau} = 0$ whenever $r$ and $\ell$ add up to an even number. Indeed, then there is at least one $t_0$ taken to an odd power in (H), whose integral with respect to the normal density equals zero. \hfill \Box

**Proof of Theorem 6** Given theorems 4 and 5, we only need to show that (i) $b_{q^0} = 0$ and (ii) $b_{q^1} = C_V \int (\pi_0 D_{Q_3}(t) + D_{Q_1}(t)) t \phi_V(t) dt$. Both results follow from the definition of $b_{q^\ell}$ in the proof of theorem 5. \hfill \Box

**Proof of Theorem 7** Let $\hat{N}_n, \mathcal{D}_n, \mathcal{B}_n, \mathcal{B}_n^*$ be defined as $\mathcal{N}_n, \mathcal{D}_n, \mathcal{B}_n, \mathcal{B}_n^*$ with $\hat{\pi} = \hat{\pi}_I$ and $\hat{\pi}_n = \hat{\pi}_I (0 + t/\alpha_n)$ replacing $\pi, \pi_n$. By lemma [E.4] we have $\mathcal{N}_n - \mathcal{B}_n \xrightarrow{d} N(0, C_V^2 \gamma_\mathcal{N})$, $\mathcal{D}_n = \pi_0 C_V + O_p(1)$, following the same steps as in lemmas [D.7] and [D.8] It therefore suffices to consider $\mathcal{B}_n$.

The first few lines of the proof of theorem 5 establish that the difference between $\mathcal{B}_n$ and $\mathcal{B}_n^*$ is $o(\alpha_n^{-q} \beta_n^{-1})$. The argument there and those in lemmas [A.7] and [E.2] used therein do not depend on $\pi$ being nonrandom, so $\mathcal{B}_n - \mathcal{B}_n^* = o_p(\alpha_n^{-q} \beta_n^{-1}).$

Expanding $R_n(t)$ and ignoring the remainder term like in the proof of theorem 5, we get

$$
\mathcal{B}_n^* = \frac{C_V}{\beta_n} \sum_{p=0}^{q} \int \hat{\pi}_n(t) \left( \sum_{\delta=1}^{\delta} \frac{D_n^{\delta+2}(t)}{\alpha_n^{\delta}} \right)^p t \phi_V(t) dt + o_p(\alpha_n^{-q} \beta_n^{-1}). \tag{38}
$$
Part (i) of the theorem then follows by applying lemma F.6 to the first right hand side term in (38).

For (ii) the first right hand side term in (38) equals

\[ \frac{C_V}{a_n^3 \beta_n} \int \hat{\Phi}_n(t)(D_{Q5}(t) + D_{Q3}(t)D_{Q4}(t)/2 + D_{Q3}^3(t)/6)t \phi_V(t)dt \]

\[ + \frac{C_V}{a_n^3 \beta_n} \int \hat{\Phi}_n(t)(D_{Q4}(t) + D_{Q3}^2(t))t \phi_V(t)dt. \] (39)

By lemma E.4 the first term in (39) is

\[ \frac{C_V \pi_0}{a_n^3 \beta_n} \int (D_{Q5}(t) + D_{Q3}(t)D_{Q4}(t)/2 + D_{Q3}^3(t)/6)t \phi_V(t)dt + o_p(a_n^{-3} \beta_n^{-1}) = O_p(a_n^{-3} \beta_n^{-1}). \] (40)

Using the same arguments as in the proof of theorem 6 about the remainder terms in an expansion of the prior, the second term in (39) is

\[ \frac{C_V \pi_0}{a_n^3 \beta_n} \int (D_{Q4}(t) + D_{Q3}^2(t))t \phi_V(t)dt + \frac{C_V}{a_n^3 \beta_n} \int (D_{Q4}(t) + D_{Q3}^2(t)) \phi_V(t)tt^t dt \pi'(\theta_0) \]

\[ + o_p(a_n^{-3} \beta_n). \] (41)

The first term in (41) is zero because the normal distribution is even. The second term in (40) is \( O(a_n^{-3} \beta_n^{-1}) \), so part (ii) of the theorem is satisfied with \( \mathcal{B}_f \) equal to the sum of the first term in (40) and the second term in (41).

**Proof of Theorem**

First part (ii). If \( c_\alpha \geq 1 \), then \( \kappa_n = \sqrt{n} \) and by lemma F.3

\[ \Psi_n = \int t \exp(c_\alpha^4 G(t) - c_\alpha^4 t^T \hat{V} t/2)dt / \int \exp(c_\alpha^4 G(t) - c_\alpha^4 t^T \hat{V} t/2)dt \]

has the limiting distribution given in (6). For \( c_\alpha < 1 \), we have likewise \( \kappa_n = \sqrt{n} / c_\alpha \) and

\[ c_\alpha \Psi_n \overset{d}{\rightarrow} c_\alpha \int \frac{t \exp(c_\alpha^{9/2} G(t) - c_\alpha^{5/2} t^T \hat{V} t/2)dt}{\int \exp(c_\alpha^{9/2} G(t) - c_\alpha^{5/2} t^T \hat{V} t/2)dt} = \frac{1}{c_\alpha^2} \int \frac{t \exp(c_\alpha^{9/2} G(t/c_\alpha^2) - c_\alpha^{5/2} t^T \hat{V} t/2)dt}{\int \exp(c_\alpha^{9/2} G(t/c_\alpha^2) - c_\alpha^{5/2} t^T \hat{V} t/2)dt}, \] (42)

by substitution of \( \tilde{t} = c_\alpha^2 t \) and replacing \( \tilde{t} \) with \( t \). The right hand side in (42) has the same distribution as the limit distribution in (6) since \( c_\alpha^{9/2} G(\cdot/c_\alpha^2) \) has the same properties as \( c_\alpha^{5/2} G(\cdot) \).

Part (iii) is established in lemma F.4

Finally, part (i). By substitution of \( \tilde{t} = \beta_n t \) and replacing \( \tilde{t} \) by \( t \) we obtain

\[ \Psi_n = \frac{1}{\beta_n} \int \frac{t \exp(\beta_n^{3/2} \hat{G}(t/\beta_n) - c_\alpha^{3/2} t^T \hat{V} t/2)dt}{\int \exp(\beta_n^{3/2} G(t/\beta_n) - c_\alpha^{3/2} t^T \hat{V} t/2)dt} \]

Apply lemma F.5 followed by lemma F.6.
APPENDIX I. COMPUTATION

I.1. Maximum Score. We now describe the algorithm used to compute our estimates for the maximum score case. A computer program written in C is available upon request.

Let \( \theta_{-j}, z_{i,-j} \) be respectively \( \theta, z_i \) without its \( j \)-th element. Let \( B_{ij} = B_{ij}(\theta_{-j}) = (a_i - z_{i,-j}^T \theta_{-j})/z_{ij} \) if \( z_{ij} \neq 0 \) and be arbitrarily defined if \( z_{ij} = 0 \). Let further \( S_{nj}(\theta_j) = \sum_{i=1}^{n_j} (2y_i - 1) (I(z_{ij} > 0)I(\theta_j \geq B_{ij}) + I(z_{ij} < 0)I(\theta_j \leq B_{ij}) )/\sqrt{n_j} \). Let \( \mathcal{S}(\theta_j|\theta_{-j}) = \int_{-\infty}^{\theta_j} \pi_j(\theta_j|\theta_{-j}) \exp(S_{nj}(\theta_j)) d\theta_j \).

1. Compute and renormalize probit estimates.
2. Set \( j = 1 \).
3. Compute \( B_{ij} \) for all \( i \) for which \( z_{ij} \neq 0 \) and sort all \( n_j \) of them in ascending order, i.e. \( B_{1j} \leq B_{2j} \leq \ldots \leq B_{nj} \).
4. Compute \( S_{nj}(-\infty) \).
5. Note that for all \( \theta_j \leq B_{1j} \), \( \mathcal{I}_{nj}(\theta_j) = \exp(S_{nj}(-\infty)) \Pi_j(\theta_j|\theta_{-j}) \), where \( \Pi_j(\theta_j|\theta_{-j}) \) is the integrated conditional prior.
6. For all \( \theta_j \in (B_{1j}, B_{2j}] \), \( \mathcal{I}_{nj}(\theta_j) = \mathcal{I}_{nj}(B_{1j}) + \exp\{S_{nj}((B_{1j} + B_{2j})/2)\} \{\Pi_j(\theta_j|\theta_{-j}) - \Pi_j(B_{1j}|\theta_{-j})\} \).
7. Repeat this process until \( \mathcal{I}_{nj}(\theta_j) \) is determined for all values of \( \theta_j \).
8. Set \( \mathcal{I}_{nj}^*(\theta_j) = \mathcal{I}_{nj}(\theta_j) / \mathcal{I}_{nj}(\infty) \), which is a distribution function.
9. Draw a random number \( r \) from a uniform distribution.
10. Compute \( i^* = \max\{i : \mathcal{I}_{nj}^*(B_{ij}) \leq r\} \).
11. Set \( \theta_j = \theta_j(r) = \Pi^{-1}(\Pi_j(B_{i^*j}) + \mathcal{I}_{nj}(\infty)r + \mathcal{I}_{nj}(B_{ij})|\theta_{-j}) \).
12. Increase \( j \). If \( j > d \) set \( j = 1 \). Go to step 2.
13. Repeat for a large number of times to burn in.
14. Then start using them as actual draws.

I.2. Inference. We now describe how one can draw \( J \) random numbers with the same distribution as \( \{ \} \). Suppose that consistent estimates \( \hat{V}, \hat{H} \) are available, where \( \hat{H} \) is moreover a positive semi-definite covariance kernel.\(^2\) One can then draw random numbers \( t_1, \ldots, t_{T^*} \) from the multivariate normal \( (\theta, \beta_n) \) and define

\[
\xi_j = \frac{\sum_{s=1}^{T^*} t_s \exp(\beta_n G_{js})}{\sum_{s'=1}^{T^*} \exp(\beta_n G_{js'})},
\]

where \( G_{1s}, \ldots, G_{T^*s} \) are drawn independently across \( j \) and have for all given \( j = 1, \ldots, J \) a joint normal distribution with \( \text{Cov}[G_{js}, G_{js'}] = \hat{H}(t_s, t_{s'}) \), for all \( s, s' \). In the limit \( (n \text{ and } T^*) \), each \( \xi_j \) has

\(^2\)In practice one only needs to ensure that the \( T^* \times T^* \) matrix with \( (i,j)\)-element \( \hat{H}(t_i, t_j) \) is positive semidefinite.
the same distribution as $\Psi_n$ defined in [17]. Note that the $t_s$ draws are made only once. The draws $\hat{\xi}_1, \ldots, \hat{\xi}_J$ can then be used to construct confidence intervals.

I.3. Derivatives of $Q$. For the maximum score case, $Q(\theta) = \mathbb{E} \left[ (2y_i - 1)(I(a_i \leq z_i^T\theta) - I(a_i \leq z_i^T\theta_0)) \right]$. If derivatives of $Q$ must be estimated, we recommend using the desired derivative of

$$
Q'(\theta) = \frac{1}{nb} \sum_{i=1}^{n} \tilde{k} \left( \frac{z_i^T\theta - a_i}{b} \right) (2y_i - 1)z_i,
$$

where $\tilde{k}$ is a kernel and $b$ a bandwidth.