Classical Laplace estimation for $\sqrt[3]{n}$–consistent estimators: improved convergence rates and rate–adaptive inference

Sung Jae Jun* Joris Pinkse† Yuanyuan Wan‡
CAPCP, Department of Economics — The Pennsylvania State University
Department of Economics — The University of Toronto

We propose a classical (nonBayesian) Laplace estimator alternative for a large class of $\sqrt[3]{n}$–consistent estimators, including isotonic density and regression estimators, inverse density and regression estimators, the maximum score and mode regression estimators, and interval censoring and monotone hazard rate estimators. The proposed alternative provides a unified method of smoothing that applies to all examples mentioned above; easier computation is a byproduct in the maximum score case. Depending on the choice of input parameter and the degree of smoothness of a population function, the convergence rate of our estimator can be faster than $\sqrt[3]{n}$ and its limit distribution can be normal. With extreme smoothness, a rate close to $\sqrt{n}$ is achievable. We provide a bias reduction method and an inference procedure which automatically adapts to the correct convergence rate and limit distribution.

Keywords and Phrases: Laplace Estimation; $\sqrt[3]{n}$–Consistent Estimators; Rate–Adaptive Inference.
JEL Classification Codes: C13; C14.

*(corresponding author) sjun@psu.edu; Department of Economics, The Pennsylvania State University, 303 Kern Graduate Building, University Park 16802; This paper is based on research supported by NSF grant SES–0922127. We thank the Human Capital Foundation (www.hcfoundation.ru), especially Andrey P. Vavilov, for their support of CAPCP (http://capcp.psu.edu) at the Pennsylvania State University. We thank Don Andrews, Miguel Delgado, Jeremy Fox, Bo Honoré, Joel Horowitz, Roger Koenker, Runze Li, Oliver Linton, Peter Robinson, Neil Wallace, Haiqing Xu, Vicky Zinde–Walsh, and numerous departmental seminar and conference participants for helpful suggestions.
†joris@psu.edu
‡yuanyuan.wan@utoronto.ca

1. Motivation

This paper provides classical (i.e. nonBayesian) Laplace estimator alternatives for a large class of $\sqrt[3]{n}$–consistent estimators. Since the proposed alternatives are based on integration rather than optimization they can have computational advantages. More importantly, they allow for a unified method of smoothing that improves the rate of convergence (under extra smoothness conditions) and a rate–adaptive inference procedure. The extremum estimators for which we provide alternatives can be divided into two classes: θ *class* estimators and η *class* estimators. We describe both classes below.

The θ class of extremum estimators is given by

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{e}} = \operatorname*{argmax}_{\theta \in \Theta} \tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta), \tag{1}$$

with

$$\tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta),$$

where $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = \tilde{g}(\boldsymbol{\xi}_{\boldsymbol{i}}; \theta)$ for i.i.d. $\{\boldsymbol{\xi}_{\boldsymbol{i}}\}$ is such that $\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}$ converges at the $\sqrt[3]{n}$ rate. In all examples that are considered here, the definition of $\tilde{g}$ has the parameter vector θ buried inside an indicator function. Examples of the θ class of estimators are inverse density, regression, and hazard rate estimators, the maximum score estimator (Manski, 1975), a quantile estimator with interval censoring, and the mode regression estimator (Lee, 1989).

For the η class of estimators, $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ depends on two parameters θ, η and hence so do $\tilde{\boldsymbol{L}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}$. The parameter of interest η_0 in the η class of estimators is defined as the solution to $\theta_0(\eta_0) = \theta_0$, where $\theta_0(\eta)$ is the θ class parameter for any given value of η.[1] The estimator $\hat{\boldsymbol{\eta}}_{\boldsymbol{e}}$ is then defined as a value of η for which $\|\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}(\eta) - \theta_0\|$ is minimized for some prespecified value θ_0. Examples of the η class of estimators are the isotonic density and regression estimators (Grenander, 1956; Brunk, 1958), an interval censoring estimator (Ayer, Brunk, Ewing, Reid, and Silverman, 1955), and a monotone hazard rate estimator (Prakasa Rao, 1970). The above–mentioned examples for both the θ class of estimators and the η class of estimators will be explored further in section 6.

Our discussion below focuses on the θ class of estimators, where we briefly return to the η class of estimators at the end of this section.

[1] It is implicit that θ and η have the same dimension.

Asymptotic results for the θ class of extremum estimators are provided by Cavanagh (1987); Kim and Pollard (1990). We instead adapt the classical Laplace estimation technique of Chernozhukov and Hong (CH 2003), whose results do not apply to our case. Our alternative to $\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}$ is

$$\hat{\boldsymbol{\theta}} = \frac{\int \theta \pi(\theta) \exp\{\alpha_n^2 \tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta)\} d\theta}{\int \pi(\theta) \exp\{\alpha_n^2 \tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta)\} d\theta}, \tag{2}$$

where α_n, π are input parameters.[2]

In CH the choice of α_n is largely immaterial for first order asymptotics as long as it diverges sufficiently fast; they use $\alpha_n = \sqrt{n}$. Here, however, both the convergence rate of $\hat{\boldsymbol{\theta}}$ and its limit distributions depend on the divergence rate of α_n and the degree of smoothness of $\tilde{Q}(\theta) = \mathbb{E}\tilde{\boldsymbol{g}}_{\boldsymbol{1}}(\theta)$. The intuition for this is that the more slowly α_n diverges, the more the integrations in (2) smooth out $\tilde{\boldsymbol{L}}_{\boldsymbol{n}}$. We distinguish three cases: *(i)* α_n diverges faster than $\sqrt[3]{n}$ *(ii)* α_n diverges at the $\sqrt[3]{n}$ rate, *(iii)* α_n diverges more slowly than $\sqrt[3]{n}$. Provided that $\tilde{Q}$ is sufficiently smooth for the chosen divergence rate of α_n — more on that in section 2 — these cases correspond to the following convergence rates and limit distributions of $\hat{\boldsymbol{\theta}}$: *(i)* a $\sqrt[3]{n}$ convergence rate and a Chernoff–like (limit) distribution identical to that of $\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}$, *(ii)* a $\sqrt[3]{n}$ rate and a distribution that can be characterized by a ratio of integrals over Gaussian processes, and *(iii)* a $\sqrt{n/\alpha_n}$ rate and a normal distribution.

There are other, optimization–based, techniques that also smooth out the objective function to obtain a better convergence rate under additional conditions, albeit that each method is only relevant for a specific $\sqrt[3]{n}$–consistent estimator. Examples include Barlow and Van Zwet (1969, 1970, 1971); Wright (1982); Friedman and Tibshirani (1984); Mukerjee (1988); Mammen (1991) for either of the isotonic estimators, Horowitz (1992) for the maximum score (MS) estimator, and Zinde-Walsh (2002) for the least median of squares (LMS) estimator; see Jun, Pinkse, and Wan (2011) for a Laplace version of the LMS estimator. That is not to say that there is no way to unify these alternative methods. Indeed, Hong, Mahajan, and Nekipelov (2010) recently developed asymptotic theory for extremum estimators based on numerical derivatives, which allows for a nonsmooth objective function.

A fundamental difference with the aforementioned papers is that our method uses integration rather than optimization. Using the framework of integration–based estimation, we provide a unified theory of smoothing for a large class of estimators and we discuss its computational advantages, e.g. in the maximum score case. Indeed, in the maximum score case Jun, Pinkse, and Wan (2014) show that estimates can be computed using Gibbs sampling. Finally, the integration–based approach provides

[2] We focus on the quasi–posterior mean — CH allow for the use of more general loss functions — to keep the theoretical results tractable. We intend to pursue more general loss functions in future work.

an inference procedure that automatically adapts to the correct limit distribution and convergence rate. Since what we choose in practice is the value of an input parameter rather than its 'rate,' having a rate–adaptive inference procedure is of practical importance. We discuss this issue below.

For case (iii), like with other methods that smooth the objective function, the limiting normal distribution has nonnegligible bias if a convergence rate of $n^{2/5}$ or faster is desired. In our case such bias arises if α_n increases at a rate no faster than $\sqrt[5]{n}$. The bias generally depends on $\tilde{Q}$ and π, and it has a convenient expansion provided that $\tilde{Q}$ is sufficiently smooth. The function π can be chosen to eliminate first order asymptotic bias. For instance, the first order bias can be removed by choosing π proportional to $\sqrt{\det\{-\partial_{\theta\theta^\intercal}\tilde{Q}(\theta)\}}$ in a neighborhood of θ_0;[3] this choice of π resembles the Jeffreys prior (Jeffreys, 1946) but our method is classical, $\tilde{Q}$ is not a loglikelihood, and π is an input parameter we shall, like CH, call a *quasi prior* since it resembles but serves a different purpose from a Bayesian prior.[4] Higher order bias reductions can be achieved by a more complicated choice of π. The function π, then, serves a purpose similar to that of higher order kernels, albeit that its choice need not affect the asymptotic variance and it must satisfy different conditions. Like in the related literature if α_n is chosen based on an assumed degree of smoothness of $\tilde{Q}$ greater than its actual smoothness then the convergence rate will be affected adversely because of excess bias; see Pollard (1993).

The trichotomy of convergence rate and limit distribution raises the practical question of which limit distribution to use since in practice one chooses a value α_n, not a rate. Although our estimator is a Laplace estimator the quasi–posterior does not coincide with the limit distribution, as can be the case in CH. We address the inference issue by providing a single inference procedure which is asymptotically valid in all three cases. This has the advantage that it provides robustness against choosing α_n 'too large,' in much the same way that Cattaneo, Crump, and Jansson (2011) provide robustness against choosing bandwidths that are too small in regular kernel estimation, albeit that there the limit distributions are always normal. To our knowledge, none of the existing literature for any of the estimation problems considered here provides such robustness. Robustness is of particular importance since several authors, including Sen, Banerjee, and Woodroofe (2010); Abrevaya and Huang (2005), have found the standard bootstrap to be inconsistent, although subsampling was shown

[3] $t^\intercal$ denotes the transpose of t.

[4] In regular moment restriction models the CH procedure has a Bayesian interpretation. Indeed, Kim (2002) shows that the exponential of a quadratic form of the sample analog of the moment condition can be interpreted as a nonparametric likelihood in an information–theoretic sense. The function π can then be understood as a Bayesian prior for the Euclidean parameter θ. However, we have failed to uncover a similar interpretation in the nonregular cases studied here.

to work for the MS estimator by Delgado, Rodrıguez-Poo, and Wolf (2001) and for the isotonic density and regression estimators by Sen, Banerjee, and Woodroofe (2010).

Which method to choose will in the end depend in large part on the performance of the estimator and inference procedure in practice. Since ours works for many different estimation problems, a full comparison is beyond the scope of this paper. A detailed simulation comparison to assess the relative merits of each procedure is needed for each application, but the present paper focuses on establishing the theoretical properties of our procedure. We have a separate paper (Jun, Pinkse, and Wan, 2014) — which like the present paper is an offshoot from a single working paper (Jun, Pinkse, and Wan, 2009) —- focusing on several issues, including computation, performance, efficiency, and choice of input parameters, in the specific case of the MS estimator.

We now resume our discussion of the η class of estimators. Please recall that the $\hat{\boldsymbol{\eta}}_{\boldsymbol{e}}$ is a minimizer of $\|\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}(\eta) - \theta_0\|$ where $\theta_0 = \theta_0(\eta_0)$ is a value chosen by us and η_0 is the parameter vector of interest. For instance for the isotonic density estimator of Grenander (1956), $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta, \eta) = \mathbb{1}(\boldsymbol{x}_{\boldsymbol{i}} \leq \theta) - \theta\eta$ whose expectation is for monotonic density $f_{\boldsymbol{x}}$ minimized at $\theta_0 = f_{\boldsymbol{x}}^{-1}(\eta_0)$. Thus, $\theta_0(\eta) = f_{\boldsymbol{x}}^{-1}(\eta)$ (in a neighborhood of η_0) and to obtain a density estimate at a given value of x we simply choose $\theta_0 = x$.

Our analog to the η class extremum estimator is the value $\hat{\boldsymbol{\eta}}$ for which

$$\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}) = \theta_0, \tag{3}$$

where $\hat{\boldsymbol{\theta}}(\eta)$ is defined as in (2) but with $\tilde{\boldsymbol{L}}_{\boldsymbol{n}}$ a function of both θ and η. The proofs for the η class of estimators are more complicated than those for the θ class of estimators, but the intuition is fairly straightforward and described in section 2.

The remainder of this paper is laid out as follows. In section 2 we provide a sketch of our results, which provides some intuition as a bonus. Sections 3 and 4 provide formal statements of the main convergence results in this paper, which can be skipped by the casual reader, who can continue to section 5. In section 3 we obtain results for the θ class of estimators and in section 4 for the η class of estimators. Section 5 describes our inference procedure and section 6 contains a discussion of potential applications.

2. Sketch

2.1. **θ class.** We first provide a sketch of and intuition for our results for the θ class of estimators for the case in which $\tilde{\boldsymbol{L}}_{\boldsymbol{n}}$ does not depend on η. Recall from (2) that

$$\hat{\boldsymbol{\theta}} = \frac{\int \theta\pi(\theta)\exp\left\{\alpha_n^2\tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta)\right\}d\theta}{\int \pi(\theta)\exp\left\{\alpha_n^2\tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta)\right\}d\theta}, \qquad \text{with } \tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta). \tag{4}$$

This corresponds to the θ class of estimators of section 3, albeit that the results of section 3 allow for the presence of η so that the results of section 3 can be used to obtain the extension to the η class of estimators of section 4.

Let $\boldsymbol{g}_{\boldsymbol{i}}(\theta) = \tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) - \tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta_0)$, $Q(\theta) = \mathbb{E}\boldsymbol{g}_{\boldsymbol{1}}(\theta)$, $\boldsymbol{L}_{\boldsymbol{n}}(\theta) = \sum_{i=1}^{n}\boldsymbol{g}_{\boldsymbol{i}}(\theta)/n$, and $\boldsymbol{S}_{\boldsymbol{n}}(\theta) = \boldsymbol{L}_{\boldsymbol{n}}(\theta) - Q(\theta)$. So Q is maximized at θ_0 with the maximum equal to zero and $\boldsymbol{S}_{\boldsymbol{n}}$ is a sample mean of n mean zero random variables, each of which is exactly equal to zero at $\theta = \theta_0$. Then (4) simplifies to

$$\hat{\boldsymbol{\theta}} = \frac{\int \theta\pi(\theta)\exp\left\{\alpha_n^2\boldsymbol{L}_{\boldsymbol{n}}(\theta)\right\}d\theta}{\int \pi(\theta)\exp\left\{\alpha_n^2\boldsymbol{L}_{\boldsymbol{n}}(\theta)\right\}d\theta} = \frac{\int \theta\pi(\theta)\exp\left\{\alpha_n^2\boldsymbol{S}_{\boldsymbol{n}}(\theta) + \alpha_n^2 Q(\theta)\right\}d\theta}{\int \pi(\theta)\exp\left\{\alpha_n^2\boldsymbol{S}_{\boldsymbol{n}}(\theta) + \alpha_n^2 Q(\theta)\right\}d\theta}. \tag{5}$$

We now carry out the substitution $t = \tilde{\alpha}_n(\theta - \theta_0)$ in (5) for a sequence $\{\tilde{\alpha}_n\}$ to be chosen later; $\tilde{\alpha}_n$ is introduced for the purpose of deriving the theoretical properties only. Thus,

$$\hat{\boldsymbol{\theta}} - \theta_0 = \frac{1}{\tilde{\alpha}_n}\frac{\int t\pi_n(t)\exp\left\{\frac{\alpha_n^2}{\sqrt{n\tilde{\alpha}_n}}\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + \frac{\alpha_n^2}{\tilde{\alpha}_n^2}Q_n(t)\right\}dt}{\int \pi_n(t)\exp\left\{\frac{\alpha_n^2}{\sqrt{n\tilde{\alpha}_n}}\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + \frac{\alpha_n^2}{\tilde{\alpha}_n^2}Q_n(t)\right\}dt}, \tag{6}$$

where $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) = \sqrt{n\tilde{\alpha}_n}\boldsymbol{S}_{\boldsymbol{n}}(\theta_0 + t/\tilde{\alpha}_n)$, $Q_n(t) = \tilde{\alpha}_n^2 Q(\theta_0 + t/\tilde{\alpha}_n)$, and $\pi_n(t) = \pi(\theta_0 + t/\tilde{\alpha}_n)$. The normalization in Q_n is due to the fact that $Q(\theta_0)$ and $\partial_\theta Q(\theta_0)$ both equal zero such that the second order derivatives in an expansion dominate. The normalization in the definition of $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ is used for the following reason. Note that

$$\mathrm{Cov}\{\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t), \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(s)\} = \tilde{\alpha}_n \,\mathrm{Cov}\{\boldsymbol{g}_{\boldsymbol{1}}(\theta_0 + t/\tilde{\alpha}_n), \boldsymbol{g}_{\boldsymbol{1}}(\theta_0 + s/\tilde{\alpha}_n)\}. \tag{7}$$

A feature of $\sqrt[3]{n}$–consistent extremum estimators is that the covariance in (7) converges to some function $H(t,s)$ as $\tilde{\alpha}_n \to \infty$, whereas if $\boldsymbol{g}_{\boldsymbol{1}}$ were differentiable then the limit of the covariance in (7) would be zero.[5] The discrepancy between the $\tilde{\alpha}_n$–scaling in (7) and the $\tilde{\alpha}_n^2$–scaling in Q_n is precisely what precludes $\sqrt{n}$–consistency here. Under conditions to be stated later $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ can be shown to behave as a Gaussian process $\mathbb{G}$ with covariance kernel H.

[5] In this case $\boldsymbol{S}_{\boldsymbol{n}}$ would have to be rescaled by $\sqrt{n\tilde{\alpha}_n^2}$.

As noted in section 1, both the convergence rate and limit distribution of $\hat{\boldsymbol{\theta}}$ depend on the choice of α_n and the smoothness of Q. First consider the case in which α_n increases faster than $\sqrt[3]{n}$. Pick $\tilde{\alpha}_n = \sqrt[3]{n}$ in which case (6) can equivalently be written as

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) = \frac{\int t \pi_n(t) \exp\Big[\frac{\alpha_n^2}{n^{2/3}}\big\{\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big\}\Big] dt}{\int \pi_n(t) \exp\Big[\frac{\alpha_n^2}{n^{2/3}}\big\{\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big\}\Big] dt}$$

Now, if Q is twice continuously differentiable at θ_0 then for $V = -\partial_{\theta\theta^\intercal}\tilde{Q}(\theta_0)$, $Q_n(t) \approx -t^\intercal V t/2$, such that

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) \simeq \frac{\int t \pi_n(t) \exp\Big[\frac{\alpha_n^2}{n^{2/3}}\big\{\mathbb{G}(t) - t^\intercal V t/2\big\}\Big] dt}{\int \pi_n(t) \exp\Big[\frac{\alpha_n^2}{n^{2/3}}\big\{\mathbb{G}(t) - t^\intercal V t/2\big\}\Big] dt}, \tag{8}$$

where $\simeq$ means that the left hand side has approximately the same distribution as the right hand side in sufficiently large samples. Since $\alpha_n/\sqrt[3]{n} \to \infty$, (8) suggests that $\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} \operatorname{argmax}_t\big\{\mathbb{G}(t) - t^\intercal V t/2\big\}$, which is exactly the Chernoff distribution of Kim and Pollard (1990). Now consider the case in which $\lim_{n\to\infty}\big(\alpha_n/\sqrt[3]{n}\big) = c_\alpha^2$ for some $0 < c_\alpha^2 < \infty$. Take $\tilde{\alpha}_n = \alpha_n$ in (6) to obtain

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) = \frac{1}{c_\alpha^2}\frac{\int t \pi_n(t) \exp\big\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big\} dt}{\int \pi_n(t) \exp\big\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big\} dt}, \tag{9}$$

where $\beta_n = \sqrt{\alpha_n^3/n}$. Now, since $\lim_{n\to\infty}\beta_n = c_\alpha^3$, (9) suggests that

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} \frac{1}{c_\alpha^2}\frac{\int t \exp\{c_\alpha^3 \mathbb{G}(t) - t^\intercal V t/2\} dt}{\int \exp\{c_\alpha^3 \mathbb{G}(t) - t^\intercal V t/2\} dt}.$$

Finally, suppose that Q is more than twice differentiable at θ_0 and that α_n increases more slowly than $\sqrt[3]{n}$. Indeed, consider the specific example in which Q is three times continuously differentiable at θ_0 and $\alpha_n/\sqrt[5]{n} \to c_\alpha^2$ as $n \to \infty$. Again, take $\tilde{\alpha}_n = \alpha_n$ to obtain

$$\hat{\boldsymbol{\theta}} - \theta_0 = \frac{1}{\alpha_n}\frac{\int t \pi_n(t) \exp\big\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big\} dt}{\int \pi_n(t) \exp\big\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big\} dt}, \tag{10}$$

noting that now $\lim_{n\to\infty}\sqrt[5]{n}\beta_n = c_\alpha^3$. Taking expansions and ignoring remainders in (10) yields

$$\begin{aligned} n^{2/5}(\hat{\boldsymbol{\theta}} - \theta_0) &\simeq \frac{n^{2/5}}{\alpha_n}\frac{\int t\big\{\pi_0 + D_{\pi 1}(t)/\alpha_n\big\} \exp\big\{\beta_n \mathbb{G}(t) - t^\intercal V t/2 + D_{Q3}(t)/6\alpha_n\big\} dt}{\int\big\{\pi_0 + D_{\pi 1}(t)/\alpha_n\big\} \exp\big\{\beta_n \mathbb{G}(t) - t^\intercal V t/2 + D_{Q3}(t)/6\alpha_n\big\} dt} = \\ &\frac{n^{2/5}}{\alpha_n}\frac{\int t\big\{\pi_0 + D_{\pi 1}(t)/\alpha_n\big\} \exp\big\{\beta_n \mathbb{G}(t) + D_{Q3}(t)/6\alpha_n\big\}\phi_V(t) dt}{\int\big\{\pi_0 + D_{\pi 1}(t)/\alpha_n\big\} \exp\big\{\beta_n \mathbb{G}(t) + D_{Q3}(t)/6\alpha_n\big\}\phi_V(t) dt}, \end{aligned} \tag{11}$$

where $D_{Q3}(t)$ is the third order term in a Taylor expansion of $Q(\theta_0 + t)$ around $Q(\theta_0)$, $D_{\pi 1}$ is the first order term in a Taylor expansion of $\pi(\theta_0 + t)$ around $\pi_0 = \pi(\theta_0)$, and ϕ_V is a mean zero, variance V^{-1}, normal density function. Now, noting that $\beta_n \mathbb{G}(t) + D_{Q3}(t)/6\alpha_n = o_p(1)$,[6] expand the right hand side exponential in (11) around one and drop all $o_p(n^{-1/5})$ terms in that expansion to obtain

$$n^{2/5}(\hat{\boldsymbol{\theta}} - \theta_0) \simeq \frac{n^{2/5}}{\alpha_n} \frac{\int t\{\pi_0 + D_{\pi 1}(t)/\alpha_n\}\{1 + \beta_n \mathbb{G}(t) + D_{Q3}(t)/6\alpha_n\}\phi_V(t)dt}{\int\{\pi_0 + D_{\pi 1}(t)/\alpha_n\}\{1 + \beta_n \mathbb{G}(t) + D_{Q3}(t)/6\alpha_n\}\phi_V(t)dt} \simeq$$

$$c_\alpha \int t\mathbb{G}(t)\phi_V(t)dt + \frac{1}{6c_\alpha^4}\int tD_{Q3}(t)\phi_V(t)dt + \frac{1}{\pi_0 c_\alpha^4}\int tD_{\pi 1}(t)\phi_V(t)dt. \quad (12)$$

The first right hand side term in (12) is a mean zero normally distributed random variable and the second and third term combined comprise the mean of the asymptotic distribution, i.e. the asymptotic bias, which is a consequence of approximating $\pi_n(t)$ and $Q_n(t)$ by their expansions around 0.[7] We show in section 3 that π can be chosen to make the asymptotic bias equal to zero by choosing π to behave like the analog of the Jeffreys prior (Jeffreys, 1946).

The case $\alpha_n = c_\alpha^2 \sqrt[5]{n}$ is precisely the case represented in (23) in section 3, but is a special case of the general result that we obtain in section 3. If Q has more than three derivatives at θ_0 then α_n can be allowed to diverge more slowly resulting in a faster convergence rate, provided that the asymptotic bias can be removed successfully, e.g. by the choice of π.

The above–described trichotomy of limit distributions and convergence rates will be formally established in section 3. We note here that the trichotomy is unusual and it makes inference in practice challenging because, as mentioned before, one chooses a value rather than a rate of α_n. So, in section 5 we propose an inference procedure, which automatically adapts to the divergence rate of α_n. The discussion in section 5 can be understood even if one decides to skip the theoretical discussion in sections 3 and 4.

2.2. η **class.** Please recall that for the η class of estimators $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ depends on two parameters θ, η and hence so do $\tilde{\boldsymbol{L}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}$. The parameter of interest is now η_0 that solves $\theta_0(\eta_0) = \theta_0$ for some prespecified value θ_0 and for some function $\theta_0(\eta)$ introduced in section 3. Recall that the estimator $\hat{\boldsymbol{\eta}}$

[6]The formal results of sections 3 and 4, unlike this intuitive description, are rigorous.
[7]This bias issue does not appear if α_n increases faster than $\sqrt[5]{n}$ because then one is effectively undersmoothing.

solves $\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}) = \theta_0$ with $\hat{\boldsymbol{\theta}}(\eta)$ as defined in (2), albeit with an extra η argument, i.e.

$$\hat{\boldsymbol{\theta}}(\eta) = \frac{\int \theta \pi(\theta) \exp\{\alpha_n^2 \tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta, \eta)\} d\theta}{\int \pi(\theta) \exp\{\alpha_n^2 \tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta, \eta)\} d\theta}, \qquad \text{with } \tilde{\boldsymbol{L}}_{\boldsymbol{n}}(\theta, \eta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta, \eta). \tag{13}$$

Other functions that because of the introduction of η now have two arguments instead are all defined as the natural generalization, e.g. $\tilde{Q}(\theta, \eta) = \mathbb{E}\tilde{\boldsymbol{g}}_{\boldsymbol{1}}(\theta, \eta)$. We write

$$\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}) = \hat{\boldsymbol{\theta}}^{\boldsymbol{*}}\{\mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0), \mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0)\},$$

with $\mathfrak{r}_n$ the (at this point unknown) convergence rate of $\hat{\boldsymbol{\eta}} - \eta_0$ and

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*, w) = \frac{\int \theta \pi(\theta) \exp\{\alpha_n^2 \boldsymbol{S}_{\boldsymbol{n}}(\theta, \eta_0 + w^*/\mathfrak{r}_n) + \alpha_n^2 Q(\theta, \eta_0 + w/\mathfrak{r}_n)\} d\theta}{\int \pi(\theta) \exp\{\alpha_n^2 \boldsymbol{S}_{\boldsymbol{n}}(\theta, \eta_0 + w^*/\mathfrak{r}_n) + \alpha_n^2 Q(\theta, \eta_0 + w/\mathfrak{r}_n)\} d\theta}. \tag{14}$$

We will show that for $\theta_{0n}(w) = \theta_0(\eta_0 + w/\mathfrak{r}_n)$ and for wide–ranging $\mathfrak{r}_n$,[8]

$$\zeta_n\{\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*, w) - \theta_{0n}(w)\}$$

converges weakly (in a space to be introduced in section 3) for each w to a limit process whose properties do not depend on w, which is entirely flat in w^* with at each w^* a distribution that corresponds to one of the three limit cases described above for the θ class of estimators under much the same conditions, where ζ_n is the associated convergence rate. There are two local parameters here for a technical reason; for each w we establish weak convergence as a function of w^*, after which we exploit monotonicity in w; see lemma G.3.

Thus, from (3),

$$0 = \zeta_n\{\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}) - \theta_0(\eta_0)\} =$$
$$\zeta_n\Big[\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}\{\mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0), \mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0)\} - \theta_{0n}\{\mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0)\}\Big] + \zeta_n\Big[\theta_{0n}\{\mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0)\} - \theta_{0n}(0)\Big] \tag{15}$$

The first right hand side term in (15) converges, as noted, in distribution to a limit distribution that we have already obtained. The second right hand side term in (15) can be expanded as

$$\frac{\zeta_n}{\mathfrak{r}_n} \partial_{\eta^\intercal} \theta_0(\eta_0) \mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0) + \cdots. \tag{16}$$

Combining (15) and (16) suggests that $\mathfrak{r}_n = \zeta_n$ and that

$$-\partial_{\eta^\intercal} \theta_0(\eta_0) \zeta_n(\hat{\boldsymbol{\eta}} - \eta_0)$$

[8]For this weak convergence result $\mathfrak{r}_n$ does not need to correspond to the convergence rate of $\hat{\boldsymbol{\eta}}$.

has the same limit distribution as $\zeta_n\{\hat{\boldsymbol{\theta}}(\eta_0)-\theta_0\}$. This result will be formally established in section 4.

3. Theta class

Our first set of results covers the case where $\eta_0 \in N \subset \mathbb{R}^d$ is given and $\theta_0 \in \Theta \subset \mathbb{R}^d$ is the unknown maximizer of $\tilde{Q}(\theta,\eta_0)$. The results are however stated more generally in that they cover convergence in a (shrinking) neighborhood of η_0 to prepare for the results for the η class case in section 4, where θ_0 is given and η_0 is estimated.

For any $\eta \in N$, let $\theta_0(\eta)$ be an arbitrary element of $\Theta_0(\eta) = \big\{\theta \in \Theta : \tilde{Q}(\theta,\eta) = \max_{\tilde{\theta}\in\Theta} \tilde{Q}(\tilde{\theta},\eta)\big\}$, and let $\boldsymbol{g_i}(\theta,\eta) = \tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta,\eta) - \tilde{\boldsymbol{g}}_{\boldsymbol{i}}\big\{\theta_0(\eta),\eta\big\}$. Then, by definition, $\Theta_0(\eta_0)$ is a singleton containing θ_0. Further, let Q and $\boldsymbol{L_n}$ be defined by $Q(\theta,\eta) = \mathbb{E}\boldsymbol{g_1}(\theta,\eta)$, $\boldsymbol{L_n}(\theta,\eta) = \sum_{i=1}^n \boldsymbol{g_i}(\theta,\eta)/n$, and let $\boldsymbol{S_n} = \boldsymbol{L_n} - Q$. So Q is a normalized version of $\tilde{Q}$ with the property that its maximum value at each $\eta \in N$ equals zero.

In theorem 1 we derive asymptotic properties of $\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*,w)$ defined in (14) as a process of w^* for fixed values of w, where $\{\alpha_n\}$, $\{\mathfrak{r}_n\}$ are divergent sequences and w^*, w belong to some set $\mathscr{W}$. So for each $w \in \mathscr{W}$, $\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(\cdot,w)$ is a random process with paths in $\mathbb{L}^\infty(\mathscr{W})$, where $\mathbb{L}^\infty(\mathscr{A})$ is the collection of bounded functions from $\mathscr{A}$ to a Euclidean space of implicit dimension. This will suffice for the θ class of estimators since there $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(0,0)$. The results of theorem 1 are used in theorem 3, in which we develop asymptotic results for $\hat{\boldsymbol{\eta}}$. Indeed, $\mathfrak{r}_n$ will turn out to be the (until theorem 3 unknown) convergence rate of $\hat{\boldsymbol{\eta}}$. Consequently, theorem 1 does not need to cover all possible combinations of $(\alpha_n,\mathfrak{r}_n)$ and $\{\mathfrak{r}_n\}$ is entirely irrelevant for $\hat{\boldsymbol{\theta}}(\eta_0) = \hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(0,0)$.

Let $\theta_{0n}(w) = \theta_0(\eta_0 + w/\mathfrak{r}_n)$, such that for any divergent sequence $\{\tilde{\alpha}_n\}$ (whose values we select and which serves only a technical purpose), the substitution $t = \tilde{\alpha}_n\big\{\theta - \theta_{0n}(w)\big\}$ yields

$$\tilde{\alpha}_n\big\{\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*,w) - \theta_{0n}(w)\big\} = \frac{\int t\,\pi_{nw}(t)\exp\Big\{\frac{\alpha_n^2}{\sqrt{n\tilde{\alpha}_n}}\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + \frac{\alpha_n^2}{\tilde{\alpha}_n^2}Q_{nw}(t)\Big\}dt}{\int \pi_{nw}(t)\exp\Big\{\frac{\alpha_n^2}{\sqrt{n\tilde{\alpha}_n}}\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + \frac{\alpha_n^2}{\tilde{\alpha}_n^2}Q_{nw}(t)\Big\}dt}, \tag{17}$$

where $\pi_{nw}(t) = \pi\big\{\theta_{0n}(w)+t/\tilde{\alpha}_n\big\}$, $Q_{nw}(t) = \tilde{\alpha}_n^2 Q\big\{\theta_{0n}(w)+t/\tilde{\alpha}_n, \eta_0+w/\mathfrak{r}_n\big\}$, and $\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) = \sqrt{n\tilde{\alpha}_n}\boldsymbol{S_n}\big\{\theta_{0n}(w)+t/\tilde{\alpha}_n, \eta_0 + w^*/\mathfrak{r}_n\big\}$. As part of the proof of theorem 1, we develop conditions under which the random process $\tilde{\boldsymbol{S}}^{\boldsymbol{*}}_{\boldsymbol{nw}}$ given by $\tilde{\boldsymbol{S}}^{\boldsymbol{*}}_{\boldsymbol{nw}}(t,w^*) = \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)/c_t$, with $c_t = \|t\|+1$, converges weakly in $\mathbb{L}^\infty(\mathbb{R}^d \times \mathscr{W})$ to a limit process $\mathbb{G}^{\boldsymbol{*}}$ which is flat in w^*.[9]

[9]Regarding the weak convergence of $\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}$, see the discussions following assumption G.

We now develop formal results. We start with some assumptions, which will be verified for a number of applications in section 6. Please note that for the θ case $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ does not depend on η and any conditions pertaining only to η (or w, w^*) can be safely ignored.

Let $g^*(\xi; t, w^*, w; \alpha, \mathfrak{r}) = \tilde{g}\{\xi; \theta_0(\eta_0 + w/\mathfrak{r}) + t/\alpha, \eta_0 + w^*/\mathfrak{r}\} - \tilde{g}\{\xi; \theta_0(\eta_0 + w/\mathfrak{r}), \eta_0 + w^*/\mathfrak{r}\}$, and let $\boldsymbol{g}_{\boldsymbol{i}}^*$ be g^* evaluated at $\xi = \boldsymbol{\xi}_{\boldsymbol{i}}$,

Assumption. *For some neighborhood $\mathfrak{N}_0 \subset N$ of η_0,*

A. θ_0 is in the interior of a compact set Θ;

B. for all $\theta \neq \theta_0$, $\tilde{Q}(\theta, \eta_0) < \tilde{Q}(\theta_0, \eta_0)$;

C. $\tilde{Q}$ is continuous on $\Theta \times \mathfrak{N}_0$, for some $q \geq 0$, $\tilde{Q}$ is $\Delta = q + 2$ times continuously differentiable in θ at (θ_0, η_0), $\partial_\eta \tilde{Q}$ is $q + 2$ times continuously differentiable in θ at (θ_0, η_0), and the minimum and maximum eigenvalues of $-\partial_{\theta\theta^\intercal} \tilde{Q}(\theta_0, \eta_0) = V > 0$ satisfy $0 < \lambda^- \leq \lambda^+ < \infty$;

D. π is q times continuously differentiable in θ at θ_0, $\pi_0 = \pi(\theta_0) > 0$, and $\pi(\theta) = 0$ for all $\theta \notin \Theta$;

E. for any $w, w^*, \tilde{w}^* \in \mathscr{W}$, $H(t,s) = \lim_{\alpha\to\infty} \alpha \mathbb{E}\{\boldsymbol{g}_{\boldsymbol{1}}^*(t; w^*, w; \alpha, \alpha) \boldsymbol{g}_{\boldsymbol{1}}^*(s; \tilde{w}^*, w; \alpha, \alpha)\}$ does not depend on $w, w^*, \tilde{w}^*$;

F. for all $\xi \in \Xi$, there is some neighborhood of (θ_0, η_0) such that for every (θ, η) in the neighborhood there exist some sequence $\{\theta_m\}$ and $\{\eta_m\}$ with $\theta_m \neq \theta$ and $\eta_m \neq \eta$ such that $\tilde{g}(\xi; \theta_m, \eta_m) \to \tilde{g}(\xi, \theta, \eta)$, where Ξ is the support of $\boldsymbol{\xi}$. □

Assumptions A to D are standard except for the presence of the "prior" π. Note that by assumptions A to C and an implicit function theorem argument $\mathfrak{N}_0$ can be taken small enough to ensure that $\Theta_0(\eta)$ consists only of a singleton $\theta_0(\eta)$ for all $\eta \in \mathfrak{N}_0$, which we shall do from hereon. Assumption E is the cause of the $\sqrt[3]{n}$ convergence rate of $\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}$ (see Kim and Pollard, 1990); assumption F allows for the presence of discontinuous functions such as indicator functions.

Assumption E implies that H is a positive definite covariance kernel and that

$$\forall t, s \in \mathbb{R}^d : \begin{cases} \forall c > 0 : H(ct, cs) = cH(t,s), \\ H(t,t) + H(s,s) - 2H(t,s) = H(t-s, t-s), \end{cases} \tag{18}$$

where the second implication requires some simple but tedious manipulations. Assumption G below is relatively high level to maintain a desirable degree of generality, but will be shown to be satisfied for a range of applications in section 6.

Let $\{\tilde{\alpha}_n\}$ and $\{\mathfrak{r}_n\}$ be (positive) sequences with $0 \leq \tilde{\alpha}_n \leq \mathfrak{r}_n$ and $1/\tilde{\alpha}_n = o(1)$. For any $w \in \mathscr{W}$ let further $g^*_{nw}(\xi; t, w^*) = g^*(\xi; t, w^*, w; \tilde{\alpha}_n, \mathfrak{r}_n)$ and $g^\circ_{nw}(\xi; t, w^*) = \sqrt{\tilde{\alpha}_n} g^*_{nw}(\xi; t, w^*)/c_t$, where $\xi \in \Xi$.

Assumption G. *Let $\mathscr{W} \subset \mathbb{R}^d$ be compact and let $\mathscr{F}_{nw} = \{g^\circ_{nw}(\cdot; t, w^*) : (t, w^*) \in \mathbb{R}^d \times \mathscr{W}\}$. Then*

(i) there exists an envelope function $\mathfrak{F}_{nw}$ such that for all $\xi \in \Xi$, $\sup_{(t,w^*)\in\mathbb{R}^d\times\mathscr{W}} \left|g^\circ_{nw}(\xi; t, w^*)\right| \leq \mathfrak{F}_{nw}(\xi)$;

(ii) for $\boldsymbol{\mathfrak{F}_{nwi}} = \mathfrak{F}_{nw}(\boldsymbol{\xi_i}) : \mathbb{E}\boldsymbol{\mathfrak{F}^2_{nw1}} = O(1)$;

(iii) for any $\epsilon > 0$, $\mathbb{E}\{\boldsymbol{\mathfrak{F}^2_{nw1}} \mathbb{1}(\boldsymbol{\mathfrak{F}_{nw1}} > \epsilon\sqrt{n})\} = o(1)$;

(iv) for any $0 < \epsilon_n = o(1)$,

$$\sup_{\substack{\|(t,w^*)-(s,\tilde{w}^*)\|\leq\epsilon_n \\ (t,s,w^*,\tilde{w}^*)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathscr{W}\times\mathscr{W}}} \mathbb{E}\left\{\boldsymbol{g^\circ_{nw1}}(t, w^*) - \boldsymbol{g^\circ_{nw1}}(s, \tilde{w}^*)\right\}^2 = o(1);$$

(v) let $\mathscr{N}\{\epsilon, \mathscr{F}_{nw}, \mathbb{L}^2(\mathscr{P})\}$ be the ($\mathbb{L}^2$–) covering number for $\mathscr{F}_{nw}$ with respect to the probability measure $\mathscr{P}$. Then for every $0 < \epsilon_n = o(1)$,

$$\sup_{\mathscr{P}^*} \int_0^{\epsilon_n} \sqrt{\log\left[\mathscr{N}\{\epsilon\|\boldsymbol{\mathfrak{F}_{nwi}}\|_{\mathscr{P}^*,2}, \mathscr{F}_{nw}, \mathbb{L}^2(\mathscr{P}^*)\}\right]} d\epsilon = o(1),$$

where $\sup_{\mathscr{P}^*}$ is the supremum taken over all finitely discrete probability measures $\mathscr{P}^*$ with $\|\boldsymbol{\mathfrak{F}_{nwi}}\|_{\mathscr{P}^*,2} > 0$. □

Please note that the index t runs over $\mathbb{R}^d$ as opposed to a compact set, as in e.g. Kim and Pollard (1990). Because g° involves a division by c_t, g° is bounded on the entire $\mathbb{R}^d$. This construction makes it more convenient to show that large values of t are immaterial asymptotically. See also appendix A.

We make a few notational comments before we proceed. For a fixed value $w \in \mathscr{W}$, we will write $\mathbb{Y}_{\boldsymbol{n}}(w^*, w) \rightsquigarrow \mathbb{Y}(w^*, w)$ in $\mathbb{L}^\infty(\mathscr{W})$ when the process $\mathbb{Y}_{\boldsymbol{n}}(\cdot, w)$ converges weakly to $\mathbb{Y}(\cdot, w)$, where $\mathbb{Y}_{\boldsymbol{n}}(\cdot, w)$ and $\mathbb{Y}(\cdot, w)$ have paths in $\mathbb{L}^\infty(\mathscr{W})$. When the limit process is flat in w^*, we will write $\mathbb{Y}_{\boldsymbol{n}}(w^*, w) \rightsquigarrow \mathbb{Y}(w)$ in $\mathbb{L}^\infty(\mathscr{W})$. Similarly, we will use shorthand notation like $\mathbb{Y}_{\boldsymbol{n}}(t, w^*, w) \rightsquigarrow \mathbb{Y}(t)$ in $\mathbb{L}^\infty(\mathbb{R}^d \times \mathscr{W})$ when the process $\mathbb{Y}_{\boldsymbol{n}}(\cdot, \cdot, w)$ converges weakly to $\mathbb{Y}(\cdot, \cdot, w)$, where $\mathbb{Y}_{\boldsymbol{n}}(\cdot, \cdot, w)$ and $\mathbb{Y}(\cdot, \cdot, w)$ have paths in $\mathbb{L}^\infty(\mathbb{R}^d \times \mathscr{W})$ and the limit process $\mathbb{Y}(\cdot, \cdot, w)$ depends neither on w^* nor on w.

We are now ready to state our first theorem. Let $C_V = \int \exp\big(-t^\intercal V t/2\big) dt$ and let ϕ_V be a normal density function with variance V^{-1}.

Theorem 1. *Suppose that assumptions A through F are satisfied. Then*

(i) if for some $c_\alpha > 0$, $\alpha_n/\sqrt[3]{n} - c_\alpha^2 = o(1)$ and assumption G is satisfied for $\mathfrak{r}_n \geq \tilde{\alpha}_n = \alpha_n$ then for all $w \in \mathscr{W}$,

$$\sqrt[3]{n}\big\{\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*, w) - \theta_{0n}(w)\big\} \rightsquigarrow \frac{1}{c_\alpha^2}\frac{\int t \exp\{c_\alpha^3 \mathbb{G}(t)\}\phi_V(t)dt}{\int \exp\{c_\alpha^3 \mathbb{G}(t)\}\phi_V(t)dt}, \tag{19}$$

in $\mathbb{L}^\infty(\mathscr{W})$, where $\mathbb{G}$ is a tight Gaussian process with covariance kernel H;

(ii) if $\sqrt[3]{n}/\alpha_n = o(1)$ and assumption G is satisfied for $\mathfrak{r}_n \geq \tilde{\alpha}_n = \sqrt[3]{n}$ then for all $w \in \mathscr{W}$,

$$\sqrt[3]{n}\big\{\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*, w) - \theta_{0n}(w)\big\} \rightsquigarrow \operatorname*{argmax}_{t\in\mathbb{R}^d} \mathbb{C}(t),$$

in $\mathbb{L}^\infty(\mathscr{W})$, where $\mathbb{C}(t) = \mathbb{G}(t) - t^\intercal V t/2$;

(iii) if α_n diverges at a polynomial rate, $\alpha_n = o(\sqrt[3]{n})$, and assumption G is satisfied for $\mathfrak{r}_n \geq \tilde{\alpha}_n = \alpha_n$ then for all $w \in \mathscr{W}$,

$$\sqrt{n/\alpha_n}\big\{\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*, w) - \theta_{0n}(w)\big\} - \frac{\mathbb{B}_{nw}}{C_V\pi_0 + o_p(1)} \rightsquigarrow N(0, \mathscr{V}),$$

in $\mathbb{L}^\infty(\mathscr{W})$, where $\mathscr{V} = \iint ts^\intercal H(t,s)\phi_V(t)\phi_V(s)dtds$ and $\mathbb{B}_{nw}$ (defined in (42)) has an expansion provided in part (iv);

(iv) In part (iii), if moreover $q \geq 1$ and $\mathfrak{r}_n/\alpha_n^q = o(1)$ then for all $w \in \mathscr{W}$,

$$\mathbb{B}_{nw} = \frac{C_V}{\beta_n}\sum_{\tau=0}^{q}\frac{b_{q\tau}^*}{\alpha_n^\tau} + o\bigg(\frac{1}{\alpha_n^q\beta_n}\bigg), \tag{20}$$

where $b_{q\tau}^* = \sum_{j=0}^{\tau} b_{qj,\tau-j}$ and

$$b_{qjs} = \sum_{p=0}^{q}\sum_{m\in\mathscr{M}_{pqj}^*}\int D_{\pi s}(t)\bigg\{\prod_{\delta=1}^{q}\frac{D_{Q;\delta+2}^{m_\delta}(t)}{m_\delta!}\bigg\}t\phi_V(t)dt, \tag{21}$$

with $\mathscr{M}_{pqj}^*$ the collection of q–dimensional vectors $m = (m_1, \cdots, m_q)$ consisting of nonnegative integers for which $\sum_{\delta=1}^{q} m_\delta = p$ and $\sum_{\delta=1}^{q}\delta m_\delta = j$ and where $D_{\pi s}$ and D_{Qs} are the term of order s in a Taylor expansion of $\pi(\theta_0 + t)$ and $Q(\theta_0 + t, \eta_0)$ around θ_0, respectively.

Proof. The proof of parts (i), (ii), (iii) and (iv) can be found in appendices B, C, D, and E, respectively.

□

The condition $\mathfrak{r}_n \geq \tilde{\alpha}_n$ ensures that $\mathbb{G}$ depends only on t under assumption E. Theorem 1 implies that, as promised, for the θ class our Laplace estimator can have three different limit distributions, depending on the amount of smoothing, i.e. depending on the rate of the input parameter α_n. First, if little smoothing is applied (α_n increases fast) then the limit distributions of $\hat{\boldsymbol{\theta}}(\eta_0) = \hat{\boldsymbol{\theta}}^*(0,0)$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{e}}$ coincide, but even for $\alpha_n = \infty$ (for fixed n) the estimators generally have different values; indeed, the Manski (1975) estimator is set–valued. The normal limit distribution and improved convergence rate arise only if sufficient smoothing is applied (α_n diverges more slowly than $\sqrt[3]{n}$) which, as noted before, requires extra smoothness of $\tilde{Q}$. Equations (20) and (21) plus some elbow grease reveal that $b^*_{q\tau} = 0$ for even values of τ and that for $q = 1$,

$$\mathbb{B}_{nw} = C_V \sqrt{n/\alpha_n^5} \int \big\{D_{\pi 1}(t)t + \pi_0 D_{Q3}(t)\big\}\phi_V(t)dt + o\Big(\sqrt{n/\alpha_n^5}\Big). \tag{22}$$

Hence for $q = 1$, if $\sqrt[5]{n} = o(\alpha_n)$ then the bias is negligible and if $\alpha_n = o(\sqrt[5]{n})$ then the bias dominates. If $\alpha_n = c_\alpha^2 \sqrt[5]{n}$ then theorem 1 implies that for $q = 1$ in the θ case,

$$n^{2/5}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} N\left\{\frac{\int \big\{D_{\pi 1}(t) + \pi_0 D_{Q3}(t)\big\} t\phi_V(t)dt}{c_\alpha^4 \pi_0}, c_\alpha^2 \mathscr{V}\right\}. \tag{23}$$

The rate in (23) is the familiar nonparametric rate, which is no coincidence. Having chosen a convergence rate, the choice of input parameter is reduced to a choice of c_α, which can be made to minimize the asymptotic mean square error. However, we do not pursue this possibility in the current paper. Instead we focus on rate–adaptive inference for a given value of the input parameter since in practice one chooses a value rather than a rate. See e.g. Cattaneo, Crump, and Jansson (2011).

One way of removing the bias in (23) is to choose a prior for which the mean in (23) equals zero, for which it is required that

$$\partial_\theta \log \pi(\theta_0) = -\left\{\int tt^\intercal \phi_V(t)dt\right\}^{-1} \int D_{Q3}(t)t\phi_V(t)dt. \tag{24}$$

One way of satisfying (24) is for any 'mother prior' π^M with $\pi^M(\theta_0) \neq 0$ and $\partial_\theta \pi^M(\theta_0) \neq 0$ to pick a matrix $\mathbb{A}$ such that (24) is satisfied for $\pi(\theta) \propto \pi^M\{\mathbb{A}(\theta - \theta_0) + \theta_0\}$. Alternatively, one can choose $\pi \propto \sqrt{\det(-\partial_{\theta\theta^\intercal} Q)}$ in a neighborhood of θ_0, which resembles the Jeffreys prior; see lemma F.1. Since $\partial_{\theta\theta^\intercal} Q$ is unknown, it must be estimated. Theorem 2 addresses the issue of estimated priors for the case $q = 1$.

Theorem 2. *Consider the θ–only case, i.e. $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ does not have an η argument. Suppose that assumptions A through F are satisfied for $q = 1$, that assumption G is satisfied for $\tilde{\alpha}_n = \min(\alpha_n, \sqrt[3]{n})$, and* *that* $1/\alpha_n = O(n^{-1/5})$. *Suppose moreover that* $\hat{\hat{\theta}}$ *is identical to* $\hat{\theta}$ *except that it uses a data–dependent prior* $\hat{\pi}$ *in lieu of* π *and (i)* $\forall \theta \notin \Theta : \hat{\pi}(\theta) = 0$*; (ii)* $\partial_\theta \hat{\pi}$ *is continuous on* Θ *with probability approaching one; (iii) for some* $0 < \underline{\pi} < \bar{\pi} < \infty$, $\mathbb{P}\{\inf_{\theta\in\Theta} \hat{\pi}(\theta) < \underline{\pi}\} = o(1)$ *and* $\mathbb{P}\{\sup_{\theta\in\Theta} \hat{\pi}(\theta) > \bar{\pi}\} = o(1)$*; (iv) for some* $\bar{\pi}_1 < \infty$, $\mathbb{P}\{\sup_{\theta\in\Theta} \|\partial_\theta \hat{\pi}(\theta)\| > \bar{\pi}_1\} = o(1)$*; (v)* $\hat{\pi}(\theta_0) - \pi(\theta_0) = o_p(1)$*; (vi) for any* $c > 0$, $\lim_{t\downarrow 0} \lim_{n\to\infty} \mathbb{P}\big\{\sup_{\|\theta-\theta_0\|\le t} \|\partial_\theta \hat{\pi}(\theta) - \partial_\theta \pi(\theta)\| > c\big\} = 0$. *Then*

$$\max(\sqrt{n/\alpha_n}, \sqrt[3]{n})(\hat{\hat{\theta}} - \hat{\theta}) = o_p(1).$$

Proof. See appendix F. □

If α_n diverges faster than $\sqrt[5]{n}$ then there is no asymptotic bias for any prior satisfying the conditions of theorem 1. So the most interesting case in theorem 2 arises when $\alpha_n = c_\alpha^2 \sqrt[5]{n}$. Indeed, then theorem 2 implies that $\hat{\hat{\theta}} - \hat{\theta} = o_p(n^{-2/5})$ such that if π is such that the asymptotic mean in (23) equals zero then

$$n^{2/5}(\hat{\hat{\theta}} - \theta_0) \xrightarrow{d} N(0, c_\alpha^2 \mathscr{V}).$$

An attractive feature of theorem 2 is that the bias correction procedure does not change the asymptotic variance.

The conditions on $\hat{\pi}$ in theorem 2 are mild and since $\hat{\pi}$ can be chosen they are straightforward to satisfy with the possible exception of the uniform convergence of $\partial_\theta \hat{\pi}$ in a neighborhood of θ_0. We chose not to provide the analogous result for the η class of estimators in theorem 2 to conserve space.

Theorem 2 only provides results for the case $q = 1$. Analogous results can be obtained for $q > 1$. Indeed, on our website http://joris.econ.psu.edu/research we provide code that automatically produces explicit expressions for all terms in the bias expansion (20). For $q = 2$ the bias expansion is identical to that for $q = 1$, for $q = 4$ to that for $q = 3$, and so on. It can be seen from these expressions that the Jeffreys prior does not work for $q \ge 3$. However, π can always be chosen as a polynomial in θ around θ_0 whose polynomial coefficients are determined by derivatives of Q at θ_0. Since we, like e.g. Horowitz (1992), assume *continuous* differentiability, the bias can be successfully eliminated by using an estimated prior or indeed be estimated and subtracted by estimating $\mathbb{B}_{nw}$ in (20). Thus, the optimal convergence rate is $n^{(q+1)/(2q+3)}$ for $\alpha_n \sim n^{1/(2q+3)}$.[10]

[10]See also Horowitz (1993).

4. Eta class

We now turn our attention to the η case, in which θ_0 is chosen and η_0 satisfying $\theta_0(\eta_0) = \theta_0$ is the parameter of interest. The convergence results for $\hat{\boldsymbol{\eta}}$ are similar to those for $\hat{\boldsymbol{\theta}}$, but they do require additional assumptions.

Assumption.

H. *the absolute values of the eigenvalues of $\partial_{\theta\theta^\intercal} \tilde{Q}(\theta, \eta)$ are no greater than $\lambda^+(\theta, \eta)$ which is uniformly bounded on $\Theta \times N$;*

I. *for all $\eta \in N$, $\tilde{Q}$ is quasi–concave in θ, i.e. its sublevel sets are convex;*

J. *the matrix $\mathscr{C} = V^{-1} \partial_{\theta\eta^\intercal} \tilde{Q}(\theta_0, \eta_0)$ is nonsingular;*

K. *for all $\eta \in N$, $\partial_\eta \tilde{Q}(\cdot, \eta)$ is bounded and nondecreasing in every element of θ.*

L. *for all $\eta \neq \eta_0$ in N, $\theta_0 \notin \Theta_0(\eta)$;*

M. *$\tilde{g}(\xi; \theta, \eta)$ is continuous in η.* □

Assumption H is implied by continuity of $\partial_{\theta\theta^\intercal} Q$ and assumption I is weaker than concavity. Assumptions J and L are imposed to ensure identification of η_0. Assumption K removes the need to achieve uniformity in w in theorem 1. When θ and η are scalar–valued, assumption L is implied by assumptions J and K. See section 6 for a justification and examples.

Theorem 3. *Let assumptions A through J be satisfied. Then*

(i) $\hat{\boldsymbol{\eta}} = \eta_0 + o_p(1)$;

(ii) $\hat{\boldsymbol{\eta}} = \eta_0 + O_p\big(n^{-1/3}\big)$;

(iii) if the conditions of part (i) of theorem 1 are satisfied then

$$\sqrt[3]{n}(\hat{\boldsymbol{\eta}} - \eta_0) \xrightarrow{d} \frac{1}{c_\alpha^2} \mathscr{C}^{-1} \frac{\int t \exp\{c_\alpha^3 \mathbb{G}(t)\} \phi_V(t) dt}{\int \exp\{c_\alpha^3 \mathbb{G}(t)\} \phi_V(t) dt};$$

(iv) if the conditions of part (ii) of theorem 1 are satisfied then

$$\sqrt[3]{n}(\hat{\boldsymbol{\eta}} - \eta_0) \xrightarrow{d} \mathscr{C}^{-1} \operatorname*{argmax}_{t \in \mathbb{R}^d} \mathbb{C}(t);$$

(v) if the conditions of parts (iii) and (iv) of theorem 1 are satisfied and if $\tilde{\boldsymbol{\eta}}$ solves

$$\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\eta}}) - \theta_0 - \sqrt{\frac{\alpha_n}{n}} \frac{C_V \sum_{\tau=0}^{q} \big(b_{q\tau}^{*}/\alpha_n^{\tau}\big) + o_p(\alpha_n^{-q})}{\beta_n \alpha_n^{d} \int \pi(\theta) \exp\{\alpha_n^2 \boldsymbol{L_n}(\theta, \tilde{\boldsymbol{\eta}})\} d\theta} = 0, \tag{25}$$

then

$$\sqrt{n/\alpha_n}(\tilde{\boldsymbol{\eta}} - \eta_0) \xrightarrow{d} N\big\{0, \mathscr{C}^{-1}\mathscr{V}(\mathscr{C}^{-1})^{\intercal}\big\}.$$

Proof. See appendix G. □

So the results for $\hat{\boldsymbol{\eta}}$ (and its bias–reduced version $\tilde{\boldsymbol{\eta}}$) are thus largely the same as for $\hat{\boldsymbol{\theta}}$, albeit subject to additional assumptions. The extra $\mathscr{C}^{-1}$ matrix in theorem 3 is due to the fact that we are essentially estimating an inverse function here.

Our previous comments on the bias reduction of $\hat{\boldsymbol{\theta}}$ also apply to $\tilde{\boldsymbol{\eta}}$. For instance, for $q = 1$, if $\alpha_n = c_\alpha \sqrt[5]{n}$ then we only need to adjust the first order bias. If the Jeffreys prior is used, then the first order bias becomes zero, so $\tilde{\boldsymbol{\eta}}$ will simply solve $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\eta}}) = \theta_0$ as before.

5. Inference

We now discuss our rate–adaptive inference procedure. We focus on the θ case with no (or fixed) η parameter, because it best highlights the idea with the least number of restrictions. We start by assuming the availability of consistent estimators for H and V.

Assumption.

N. $\hat{\boldsymbol{V}} = V + o_p(1)$;

O. $\forall t, s \in \mathbb{R}^d$, $\hat{\boldsymbol{H}}(t,s) = H(t,s) + o_p(1)$;

P. $\forall t, s \in \mathbb{R}^d$ *and* $c > 0$, $\hat{\boldsymbol{H}}(t-s, t-s) = \hat{\boldsymbol{H}}(t,t) + \hat{\boldsymbol{H}}(s,s) - 2\hat{\boldsymbol{H}}(t,s)$ *and* $\hat{\boldsymbol{H}}(ct, cs) = c\hat{\boldsymbol{H}}(t,s)$;

Q. $\mathbb{E}\big\{\sup_{\|t\|=1} \hat{\boldsymbol{H}}^{\boldsymbol{2}}(t,t)\big\} = O(1)$.

Assumptions N to Q are high–level, but they are satisfied by common estimators; assumption P is the sample counterpart of (18). Letting $\{\hat{\mathbb{G}}\}$ be a (sample–size–dependent) sequence of Gaussian processes with finite marginal distributions characterized by $\hat{\boldsymbol{H}}$, assumption O ensures that $\hat{\mathbb{G}}$ has the same marginal distributions as $\mathbb{G}$ in the limit and assumptions P and Q ensure that $\exp\big\{\hat{\mathbb{G}}(\cdot)\big\}$ is integrable with respect to a normal density with probability one. It is innocuous in that covariance kernel estimates satisfying assumption P are straightforward to construct. Assumption Q is needed for the weak convergence of $\hat{\mathbb{G}}$ to $\mathbb{G}$ in an $\mathbb{L}^\infty$ space on an arbitrary compact set. Note that we do not require uniform convergence of $\hat{\boldsymbol{H}}$. Instead we impose restrictions on how to construct $\hat{\boldsymbol{H}}(t,s)$ such that it suffices to work with $\hat{\boldsymbol{H}}(t,t)$, which then ensures that assumption P is sufficient for stochastic equicontinuity in $\mathbb{L}^\infty$ on compacta.

Let β_n be as defined right after (9) and

$$\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}} = \frac{1}{\min\left(\sqrt[3]{\beta_n},1\right)} \frac{\int t \exp\left[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{V} t/2\}\right] dt}{\int \exp\left[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{V} t/2\}\right] dt},$$

which we can simulate with a given value of α_n and data. We now have the following theorem.

Theorem 4. *Under assumptions N to Q, $\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}} \xrightarrow{d} \mathbb{J}_{\boldsymbol{\alpha}}$, where the distribution of $\mathbb{J}_{\boldsymbol{\alpha}}$ is equal to the limit distribution given in parts (i), (ii), or (iii) of theorem 1, albeit that the norming there is replaced with* $\max\left(\sqrt{n/\alpha_n}, \sqrt[3]{n}\right)$.

Theorem 4 shows that the limiting distribution of $\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}}$ automatically adapts to the rate of α_n. So, for all $w \in \mathscr{W}$, $\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}}$ approximates the distribution given in theorem 1 uniformly in w^*, up to a "bias" term in the normality case.[11] Therefore, using the limiting distributions of theorem 1 for inference does not require to translate a chosen value of α_n to a specific rate condition on it, which removes rate–related uncertainty encountered in practice.

6. Applications

We now provide an incomplete list of applications to which our results can be applied; further applications can be found in Groeneboom and Wellner (2001) and Kim and Pollard (1990), albeit that for LMS Jun, Pinkse, and Wan (2011) already provides a $\sqrt{n}$–consistent estimator. Statistical properties of the extremum estimators of the following examples are nonstandard but well–studied: e.g. the theory of isotonic density estimation is in van der Vaart (1998).

We first list the applications with the formulas for $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}, H, V$ in each case before we verify our assumptions for a few representative cases. Please note that in a few instances, $\tilde{Q}$ is decreasing rather than increasing in η, but only the monotonicity in η is germane; η can always be replaced with $-\eta$.

Let $\text{Med}(t,s,0)$ denote the median of $t,s,0$.

[a] Inverse density: The object of interest is the inverse density function value $\theta_0 = f^{-1}(\eta_0)$ for specified η_0 where f is a (weakly) decreasing density function for which f' is continuous and negative at θ_0. Then one can use $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = \mathbb{1}(\boldsymbol{x}_{\boldsymbol{i}} \leq \theta) - \theta\eta_0$, which produces $H(t,s) = \left|\text{Med}(t,s,0)\right| f(\theta_0)$ and $V = -f'(\theta_0)$.

[11] The bias can be removed without affecting inference, as per the comments in the last paragraph of section 3.

[b] Inverse regression: The object of interest is the inverse regression function value $\theta_0 = \mathfrak{m}^{-1}(\eta_0)$ for specified η_0 where $\mathfrak{m}$ with $\mathfrak{m}(x) = \mathbb{E}(\boldsymbol{y_1}|\boldsymbol{x_1} = x)$ is a (weakly) decreasing regression function of a continuously distributed regressor $\boldsymbol{x_i}$ for which $\mathfrak{m}'$ is negative and continuous at θ_0. One can use $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = \mathbb{1}(\boldsymbol{x_i} \le \theta)(\boldsymbol{y_i} - \eta_0)$, which results in $H(t,s) = \big|\mathrm{Med}(t,s,0)\big| f(\theta_0)\mathbb{V}(\boldsymbol{y_1}|\boldsymbol{x_1} = \theta_0)$ and $V = -\mathfrak{m}'(\theta_0)f(\theta_0)$.

[c] Inverse monotone hazard: Observed are durations $\boldsymbol{x_i}$ drawn from a distribution F for which the hazard rate $\mathscr{H} = f/(1-F)$ is (weakly) decreasing and strictly decreasing at the value of interest $\theta_0 = \mathscr{H}^{-1}(\eta_0)$. Take $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = \mathbb{1}(\boldsymbol{x_i} \le \theta) - \eta_0 \min(\boldsymbol{x_i}, \theta)$, which yields $H(t,s) = |\mathrm{Med}(t,s,0)| f(\theta_0)$ and $V = -f'(\theta_0) - f^2(\theta_0)/\{1-F(\theta_0)\}$.

[d] Quantiles with interval censoring: Observed are (continuously distributed) times $\boldsymbol{y_i}$ with distribution function $F_{\boldsymbol{y}}$ and the indicator $\mathbb{1}\{\boldsymbol{y_i} \ge \boldsymbol{x_i}\}$; $\boldsymbol{x_i}$ is continuous and independent of $\boldsymbol{y_i}$, but its value is not observed. The object of interest is the quantile $\theta_0 = F_{\boldsymbol{x}}^{-1}(\eta_0)$ of $\boldsymbol{x_1}$ at some specified value of η_0. Take $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = \mathbb{1}(\boldsymbol{y_i} \le \theta)\eta_0 - \mathbb{1}\big\{\mathbb{1}(\boldsymbol{x_i} \le \boldsymbol{y_i})\boldsymbol{y_i} \le \theta\big\}$, which results in $H(t,s) = |\mathrm{Med}(t,s,0)| F_{\boldsymbol{x}}(\theta_0)\{1 - F_{\boldsymbol{x}}(\theta_0)\} f_{\boldsymbol{y}}(\theta_0)$ and $V = f_{\boldsymbol{y}}(\theta_0) f_{\boldsymbol{x}}(\theta_0)$, where $f_{\boldsymbol{x}}, f_{\boldsymbol{y}}$ are the density functions corresponding to $F_{\boldsymbol{y}}, F_{\boldsymbol{x}}$, respectively.

[e] Mode regression (Lee, 1989): Observations $\boldsymbol{y_i} = \theta_0^\intercal \boldsymbol{x_i} + \boldsymbol{u_i}$ are available only when $\boldsymbol{y_i} \ge 0$, at least one of the regressors is continuous with nonzero coefficient, and the conditional distribution of $\boldsymbol{u_i}$ given $\boldsymbol{x_i}$ is assumed continuous and even, with the conditional density strictly decreasing at zero and weakly decreasing elsewhere. For given input parameter $\nu > 0$, let $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = \mathbb{1}\big\{|\boldsymbol{y_i} - \max(\theta^\intercal \boldsymbol{x_i}, \nu)| \le \nu\big\}$, resulting in $H(t,s) = 2\mathbb{E}\big\{\mathbb{1}(\theta_0^\intercal \boldsymbol{x_1} \ge \nu)\big|\mathrm{Med}(t^\intercal \boldsymbol{x_1}, s^\intercal \boldsymbol{x_1}, 0)\big| f_{\boldsymbol{u}|\boldsymbol{x}}(\nu|\boldsymbol{x_1})\big\}$ and $V = -2\mathbb{E}\big\{\mathbb{1}(\theta_0^\intercal \boldsymbol{x_1} \ge \nu)\boldsymbol{x_1}\boldsymbol{x_1}^\intercal f'_{\boldsymbol{u}|\boldsymbol{x}}(\nu|\boldsymbol{x_1})\big\}$. As a special case one can estimate the mode of a distribution by replacing $\boldsymbol{x_i}$ with a constant.

[f] Maximum score (Manski, 1975): The estimator is intended for the regression model with binary regressand $\boldsymbol{y_i}$ and regressors $\boldsymbol{x_i}$ under the assumption that $\mathrm{Med}(\boldsymbol{y_1}|\boldsymbol{x_1}) = \mathbb{1}(\tilde{\theta}_0^\intercal \boldsymbol{x_1} \ge 0)$ a.s., and where the regressor vector can be partitioned as $\boldsymbol{x_i} = [\boldsymbol{z_i}^\intercal, \boldsymbol{a_i}]^\intercal$ for a continuous regressor $\boldsymbol{a_i}$, whose regression coefficient can be normalized to (plus or) minus one. Let $\tilde{\theta}_0 = [\theta_0^\intercal, -1]^\intercal$ and take $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = (2\boldsymbol{y_i} - 1)\mathbb{1}(\theta^\intercal \boldsymbol{z_i} - \boldsymbol{a_i} \ge 0)$, resulting in $H(t,s) = \mathbb{E}\big\{\big|\mathrm{Med}(t^\intercal \boldsymbol{z_1}, s^\intercal \boldsymbol{z_1}, 0)\big| f(\theta_0^\intercal \boldsymbol{z_1}|\boldsymbol{z_1})\big\}$ and $V = -2\mathbb{E}\big\{\boldsymbol{z_1}\boldsymbol{z_1}^\intercal \partial_a \mathbb{P}(\boldsymbol{y_1} = 1|\boldsymbol{a_1} = \theta_0^\intercal \boldsymbol{z_1}, \boldsymbol{z_1}) f(\theta_0^\intercal \boldsymbol{z_1}|\boldsymbol{z_1})\big\}$, where f is the conditional density of $\boldsymbol{a_1}$ given $\boldsymbol{z_1}$. The MS estimator is related to the *perceptron* (Rosenblatt, 1957) used in artificial neural networks. Other variations and extensions of the MS estimator include a multinomial version (in Manski, 1975) and panel data model with "fixed effects" (Manski, 1987).

[g] Isotonic density (Grenander, 1956): The object of interest is the density function value $\eta_0 = f(\theta_0)$ for chosen $\theta_0 > 0$ where f is a (weakly) decreasing density function for which $f'(\theta_0) < 0$. Then one can use $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta, \eta) = \mathbb{1}(\boldsymbol{x}_{\boldsymbol{i}} \leq \theta) - \theta\eta$. H and V are as in [a]. The formulation for $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ here is inspired by the characterization of the Grenander nonparametric maximum likelihood estimator (NPMLE) in Groeneboom (1985).

[h] Isotonic regression (Brunk, 1958, 1970): To be estimated is $\eta_0 = \mathfrak{m}(\theta_0)$ for chosen $\theta_0 > 0$ where m is as in [b]. One can use $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta, \eta) = \mathbb{1}(\boldsymbol{x}_{\boldsymbol{i}} \leq \theta)(\boldsymbol{y}_{\boldsymbol{i}} - \eta)$. H, V are as in [b]. There is again a direct analogy to the corresponding NPMLE.

[i] Monotone hazard rate (Prakasa Rao, 1970): The objective is to estimate $\eta_0 = \mathscr{H}(\theta_0)$ where $\mathscr{H}, \boldsymbol{x}_{\boldsymbol{i}}$ are as in [c]. Take $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta, \eta) = \mathbb{1}(\boldsymbol{x}_{\boldsymbol{i}} \leq \theta) - \eta \min(\boldsymbol{x}_{\boldsymbol{i}}, \theta)$; H, V are is in [c].

[j] Interval censoring (Ayer, Brunk, Ewing, Reid, and Silverman, 1955): Using the notation of [d], the object of interest is $\eta_0 = F_{\boldsymbol{x}}(\theta_0)$ at some specified value θ_0. Take $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta, \eta) = \mathbb{1}(\boldsymbol{y}_{\boldsymbol{i}} \leq \theta)\eta - \mathbb{1}\big\{\mathbb{1}(\boldsymbol{x}_{\boldsymbol{i}} \leq \boldsymbol{y}_{\boldsymbol{i}})\boldsymbol{y}_{\boldsymbol{i}} \leq \theta)\big\}$; H, V are as in [d].

To illustrate our assumptions consider example [g]. Example [g] is similar to [h] and indeed [i] and [j] and will yield intuition for [a] to [d] as a byproduct. Example [f] is discussed in depth in Jun, Pinkse, and Wan (2014).

For [g], $\tilde{Q}(\theta, \eta) = F(\theta) - \theta\eta$. In a neighborhood $\mathfrak{N}_0$ of η_0, $\theta_0(\eta) = f^{-1}(\eta)$ and since f' is assumed continuous and negative at θ_0, assumption B is satisfied. In assumption C, which is also satisfied, $q \geq 0$ corresponds to the number of derivatives f' posseses at θ_0. Further, for assumption G, note that $\mathscr{F}_{nw}$ consists of $g^{\circ}_{nw}(x; t, w^*) = \sqrt{\tilde{\alpha}_n}\big\{\mathbb{1}(\theta_n < x \leq \theta_n + t/\tilde{\alpha}_n) - (\eta_0 + w^*/\mathfrak{r}_n)t/\tilde{\alpha}_n\big\}/c_t$, where $t \in \mathbb{R}$, $w^* \in \mathscr{W}$, and θ_n converges to θ_0. Therefore, $\mathscr{F}_{nw}$ is contained in a convex hull of the two classes $\mathscr{F}^1_{nw} = \big\{\sqrt{\tilde{\alpha}_n}\mathbb{1}\{\theta_n < x \leq \theta_n + t/\tilde{\alpha}_n\}/c_t : t \in \mathbb{R}\big\}$ and $\mathscr{F}^2_{nw} = \big\{(\eta_0 + w^*/\mathfrak{r}_n)(t/c_t)/\sqrt{\tilde{\alpha}_n} : t \in \mathbb{R}, w^* \in \mathscr{W}\big\}$. Both $\mathscr{F}^1_{nw}$ and $\mathscr{F}^2_{nw}$ are polynomial classes with polynomial indices independent of n with well–behaved envelope functions $\sqrt{\tilde{\alpha}_n}/\big(1+\tilde{\alpha}_n|x-\theta_n|\big)$ and $(|\eta_0| + \bar{w}/\mathfrak{r}_n)/\sqrt{\tilde{\alpha}_n}$, respectively, where $\bar{w} = \sup \mathscr{W}$. In sum, the assumptions for theorem 1 are satisfied.

Assumption H requires the integration region to be restricted to an area in which f' is bounded; assumption I is implied by the assumed concavity of F (f' is decreasing). $\mathscr{C} = -1/f'(\theta_0)$, which is nonzero, $\partial_\eta Q(\theta) = \theta$ is bounded and nondecreasing, and $\tilde{g}$ is continuous in η, so assumptions J, K, and M hold. Assumption L follows from assumptions J and K, because θ and η are scalars.

If $q \geq 0$ then part (iii) of theorem 3 implies that if $\sqrt[3]{n}/\alpha_n = o(1)$ then our estimator produces the same limiting distribution as the Grenander estimator under the same conditions (Groeneboom, 1985) and if $\alpha_n = c_\alpha^2 \sqrt[3]{n}$ then it produces a $\sqrt[3]{n}$–rate but a different limiting distribution.

If $q \geq 1$ and $\alpha_n = c_\alpha^2 \sqrt[5]{n}$ then by (23) and theorem 3,

$$n^{2/5}\{\hat{\boldsymbol{f}}(\theta_0) - f(\theta_0)\} \xrightarrow{d} N\left[\frac{\pi'(\theta_0) - \pi(\theta_0) f''(\theta_0)/2f'(\theta_0)}{c_\alpha^4 \pi(\theta_0)}, \frac{c_\alpha^2 f(\theta_0)\sqrt{-f'(\theta_0)}}{2\sqrt{\varpi}}\right], \tag{26}$$

where $\varpi = 3.14159....$ For flat π the assumptions needed for and asymptotic bias and variance of the proposed estimator are the same as those of a kernel density estimator using a normal kernel and bandwidth equal to $1/\sqrt[5]{n}\sqrt{-c_\alpha^4 f'(\theta_0)}$. For the Jeffreys–like prior $\pi(\theta) \propto \sqrt{-f'(\theta)}$ the asymptotic bias of our isotonic density estimator is zero.

If α_n is chosen 'too large' then the limit distribution in (26) will not be a good approximation for $\hat{\boldsymbol{f}}$. Since it is hard to know what 'too large' means in this context, the inference procedure of theorem 4 is generally preferable.

References

Abrevaya, J., and J. Huang (2005): "On the bootstrap of the maximum score estimator," *Econometrica*, 73(4), 1175–1204.

Ayer, M., H. Brunk, G. Ewing, W. Reid, and E. Silverman (1955): "An empirical distribution function for sampling with incomplete information," *Annals of Mathematical Statistics*, pp. 641–647.

Barlow, R. E., and W. R. Van Zwet (1969): "Asymptotic properties of isotonic estimators for the generalized failure rate function. Part II: Asymptotic distributions," *Operations Research Center Report ORC 69-10, Univ. California*.

——— (1970): "Asymptotic properties of isotonic estimators for the generalized failure rate function. Part I: Asymptotic distributions," in *Nonparametric Techniques in Statistical Inference*, ed. by M. L. Puri, chap. 2, pp. 159–173. Wiley, New York.

——— (1971): "Comparison of several nonparametric estimators of the failure rate function," in *Operations Research and Reliability*, chap. 2, pp. 375–399. Gordon and Breach, New York.

Berge, C. (1963): *Topological spaces*. Courier Dover.

Brunk, H. (1958): "On the estimation of parameters restricted by inequalities," *Annals of Mathematical Statistics*, pp. 437–454.

Brunk, H. D. (1970): "Estimation of isotonic regression," in *Nonparametric techniques in statistical inference*, pp. 177–195. Cambridge University Press.

Cattaneo, M., R. Crump, and M. Jansson (2011): "Small bandwidth asymptotics for density-weighted average derivatives," *Econometric Theory*, Forthcoming.

Cavanagh, C. (1987): "The limiting behavior of estimators defined by optimization," Discussion paper, Harvard University.

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

Cremers, H., and D. Kadelka (1986): "On weak convergence of integral functionals of stochastic processes with applications to processes taking paths in L_p^E," *Stochastic Processes and their Applications*, 21, 305–317.

Delgado, M., J. Rodriguez-Poo, and M. Wolf (2001): "Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator," *Economics Letters*, 73(2), 241–250.

Friedman, J., and R. Tibshirani (1984): "The monotone smoothing of scatterplots," *Technometrics*, 26(3), 243–250.

Grenander, U. (1956): "On the theory of mortality measurement, Part II," *Skandinavisk Aktuarietidskrift*, 39, 125–153.

Groeneboom, P. (1985): "Estimating a monotone density," in *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, II.*, ed. by L. M. L. Cam, and R. A. Olshen, pp. 535–555. Springer.

Groeneboom, P., and J. A. Wellner (2001): "Computing Chernoff's distribution," *Journal of Computational and Graphical Statistics*, 10(2), 388–400.

Hong, H., A. Mahajan, and D. Nekipelov (2010): "Extremum estimation and numerical derivatives," Discussion paper, The University of California at Berkeley.

Horowitz, J. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, 60(3), 505–31.

Horowitz, J. L. (1993): "Optimal rates of convergence of parameter estimators in the binary response model with weak distributional assumptions," *Econometric Theory*, 9(1), 1–18.

Jeffreys, H. (1946): "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, pp. 453–461.

Jun, S., J. Pinkse, and Y. Wan (2011): "$\sqrt{n}$–consistent robust integration-based estimation," *Journal of Multivariate Analysis*, 102, 828–846.

——— (2014): "Integrated score estimation," Discussion paper, Penn State.

Jun, S. J., J. Pinkse, and Y. Wan (2009): "Cube–root–consistent Laplace–type estimators," Discussion paper, The Pennsylvania State University.

Kim, J., and D. Pollard (1990): "Cube root asymptotics," *Annals of Statistics*, 18(1), 191–219.

Kim, J.-Y. (2002): "Limited information likelihood and Bayesian analysis," *Journal of Econometrics*, 107, 175–193.

Lee, M. (1989): "Mode regression," *Journal of Econometrics*, 42(3), 337–349.

Mammen, E. (1991): "Estimating a smooth monotone regression function," *Annals of Statistics*, 19(2), 724–740.

Manski, C. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3(3), 205–228.

——— (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, pp. 357–362.

Mukerjee, H. (1988): "Monotone nonparametric regression," *Annals of Statistics*, 16(2), 741–750.

Pollard, D. (1993): "The asymptotics of a binary choice model," Discussion paper, Yale University.

Prakasa Rao, B. (1970): "Estimation for distributions with monotone failure rate," *Annals of Mathematical Statistics*, pp. 507–519.

Rosenblatt, F. (1957): "The perceptron, a perceiving and recognizing automaton," Discussion paper, Cornell Aeronautical Laboratory.

Sen, B., M. Banerjee, and M. Woodroofe (2010): "Inconsistency of bootstrap: the Grenander estimator," *Annals of Statistics*, 38(4), 1953–1977.

Steele, J. (2004): *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge Univ Pr.

van der Vaart, A. (1998): *Asymptotic statistics*. Cambridge University Press.

van der Vaart, A., and J. Wellner (1996): *Weak convergence and empirical processes*. Springer.

Wright, F. T. (1982): "Monotone regression estimates for grouped observations," *Annals of Statistics*, 10(1), 278–286.

Zinde-Walsh, V. (2002): "Asymptotic theory for some high breakdown point estimators," *Econometric theory*, 18(05), 1172–1196.

Appendix A. Notation and Comments

Throughout $\{\gamma_n\}$ is a sequence which diverges at a polynomial rate such that $\gamma_n = o\{\tilde{\alpha}_n^{1/3(q+1)}\}$. Further, $\mathfrak{T}_n = \{t \in \mathbb{R}^d : \theta_0 + t/\tilde{\alpha}_n \in \Theta\}$, $\Gamma_n = \{t \in \mathbb{R}^d : \|t\| \leq \gamma_n\}$, $\Gamma_n^c = \mathbb{R}^d - \Gamma_n$, and $\Gamma_n^{c*} = \mathfrak{T}_n - \Gamma_n$. Further, $R_{nw}(t) = Q_{nw}(t) + t^\intercal V t/2$. Additional notation will be introduced in the lemmas and proofs in the appendices below.

Before we proceed, we make some comments on the role of the division by c_t in the definition of $\mathscr{F}_{nw}$. Instead of proving the weak convergence of $\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}$ on compacta and showing that large values of t do not matter for our purpose, we work with the weak convergence of $\tilde{\boldsymbol{S}}^*_{\boldsymbol{nw}}$ (defined in B.2) with the index t running over the entire Euclidean space. We take this approach, because all we need for integrals like $\int \exp\{\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)\}\phi_V(t)dt$ to be well–defined is that $\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)$ does not diverge faster than a quadratic of t as $\|t\| \to \infty$. B.5 shows that the tails of $\tilde{\boldsymbol{S}}^*_{\boldsymbol{nw}}$ are negligible, from which B.6 shows that the tails of $\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}$ diverge no faster than $\|t\|$ with probability approaching to one.

Appendix B. Part (i) of theorem 1

Lemma B.1. *For* $\mathscr{C} = V^{-1}\partial_{\theta\eta^\intercal} Q(\theta_0, \eta_0)$, $\mathfrak{r}_n\{\theta_{0n}(w) - \theta_0\} = \mathscr{C}w + o(1)$.

Proof. Since Q is concave in θ at (θ_0, η_0) and continuous in both θ, η, it follows that $\theta_{0n}(w) - \theta_0 = o(1)$. Thus, by multiple use of the mean value theorem and using assumption C it follows that

$$0 = \partial_\theta Q\{\theta_{0n}(w), \eta_0 + w/\mathfrak{r}_n\} = \partial_{\theta\eta^\intercal} Q(\theta_0, \eta_0)w/\mathfrak{r}_n - V\{\theta_{0n}(w) - \theta_0\} + o\{\|\theta_{0n}(w) - \theta_0\| + 1/\mathfrak{r}_n\}. \quad \square$$

Lemma B.2. $\tilde{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t, w^*) = \sum_{i=1}^n \{\boldsymbol{g}^\circ_{\boldsymbol{nwi}}(t, w^*) - \mathbb{E}\boldsymbol{g}^\circ_{\boldsymbol{nw1}}(t, w^*)\}/\sqrt{n} \rightsquigarrow \mathbb{G}^*(t)$ *in* $\mathbb{L}^\infty(\Re^d \times \mathscr{W})$, *where* $\mathbb{G}^*$ *is a Gaussian process with the covariance kernel* $H^*(t, s) = H(t, s)/c_t c_s$. *Consequently,* $\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*) = \tilde{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t, w^*)c_t \rightsquigarrow \mathbb{G}(t)$ *in* $\mathbb{L}^\infty_L(\Re^d \times \mathscr{W})$, *where* $\mathbb{L}^\infty_L$ *is a space of locally bounded functions on compacta.*

Proof. Since $\Re^d \times \mathscr{W}$ is dense, assumption F ensures that for $j = 1, 2$,

$$\big\{\{g^\circ_{nw}(\cdot; t, w^*) - g^\circ(\cdot; s, \tilde{w}^*)\}^j\big\}_{\|(t,w^*)-(s,\tilde{w}^*)\|<\epsilon}$$

is a pointwise measurable class, and hence $\mathscr{P}$–measurable for every $\mathscr{P}$; see van der Vaart and Wellner (1996, p.110). Note further that $\lim_{n\to\infty} \mathbb{E}\{\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(s, \tilde{w}^*)\} = H^*(t, s)$ for every $t, s \in \Re^d$ and every $w^*, \tilde{w}^* \in \mathscr{W}$ by assumption E. Therefore, the result follows from assumption G and van der Vaart and Wellner (1996, theorem 2.11.22). $\square$

Lemma B.3. *For any* $|c| \leq 1$*, any* b*, and any integer* $j \geq 0$*,* $\left|\exp(cb) - \sum_{s=0}^{j}(cb)^s/s!\right| \leq |c|^{j+1}\exp(|b|)$.

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

Lemma B.4. *For some* $0 < c_q < \infty$*, all sufficiently large* n*, all* t *for which* $\theta_{0n}(w) + t/\tilde{\alpha}_n \in \Theta$*, and all* $w \in \mathscr{W}$*,* $Q_{nw}(t) \leq -\min(c_q\tilde{\alpha}_n^2, t^\intercal Vt/4)$.

Proof. Choose neighborhoods $\mathfrak{N}_\eta, \mathfrak{N}_\theta$ of η_0, θ_0, such that for all $\eta \in \mathfrak{N}_\eta$ and some $0 < c_q < \infty$, *(i)* $\theta \in \mathfrak{N}_\theta \Rightarrow Q(\theta, \eta) \leq -\{\theta - \theta_0(\eta)\}^\intercal V\{\theta - \theta_0(\eta)\}/4$, *(ii)* $\theta \notin \mathfrak{N}_\theta \Rightarrow Q(\theta, \eta) \leq -c_q$. Such neighborhoods exist because of assumptions B and C. Take n large enough to ensure that $\eta_0 + w/\mathfrak{r}_n \in \mathfrak{N}_\eta$. Then, for all $\theta \in \Theta$, $Q(\theta, \eta_0 + w/\mathfrak{r}_n) \leq -\min\left[\{\theta - \theta_{0n}(w)\}^\intercal V\{\theta - \theta_{0n}(w)\}/4, c_q\right]$, which implies the stated result. □

Lemma B.5. *For* $\bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t) = \sup_{w^* \in \mathscr{W}}|\tilde{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t, w^*)|$*, (i)* $\sup_{t \in \Gamma_n^c} \bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t) = o_p(1)$ *and (ii) for all* $\epsilon > 0$*,* $\lim_{\tilde{\gamma}\to\infty}\lim_{n\to\infty}\mathbb{P}\left\{\sup_{t \in \tilde{\Gamma}^c}\bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t) > \epsilon\right\} = 0$*, where* $\tilde{\Gamma} = \left\{t \in \mathbb{R}^d : \|t\| \leq \tilde{\gamma}\right\}$.

Proof. We show (i); (ii) is similar. For any $\epsilon > 0$ and any nonnegative integer j by B.2, the continuous mapping theorem, using the fact that $\mathbb{G}(st)$ and $\sqrt{s}\mathbb{G}(t)$ have the same distribution by assumption E, and by the Markov inequality, we have

$$
\begin{aligned}
\lim_{n\to\infty}\mathbb{P}\left\{\left|\sup_{\|t\|\geq\gamma_n}\bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t)\right| > \epsilon\right\} &\leq \lim_{n\to\infty}\mathbb{P}\left\{\sup_{\|t\|\geq j}|\bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t)| > \epsilon\right\} = \mathbb{P}\left\{\sup_{\|t\|\geq j}|\mathbb{G}^*(t)| > \epsilon\right\} \\
&= \sum_{s=j+1}^{\infty}\mathbb{P}\left\{\sup_{s-1\leq\|t\|<s}|\mathbb{G}^*(t)| > \epsilon\right\} \leq \sum_{s=j+1}^{\infty}\mathbb{P}\left\{\sup_{s-1\leq\|t\|<s}|\mathbb{G}(t)| > \epsilon s\right\} \\
&\leq \sum_{s=j+1}^{\infty}\mathbb{P}\left\{\sup_{\|t\|\leq 1}|\mathbb{G}(t)| > \epsilon\sqrt{s}\right\} \leq \frac{\mathbb{E}\sup_{\|t\|\leq 1}\mathbb{G}^4(t)}{\epsilon^4}\sum_{s=j+1}^{\infty}\frac{1}{s^2}. \qquad (27)
\end{aligned}
$$

By van der Vaart and Wellner (1996, proposition A.2.4), we know that there exists some $C < \infty$ such that $\mathbb{E}\sup_{\|t\|\leq 1}\mathbb{G}^4(t) \leq C\left\{\mathbb{E}\sup_{\|t\|\leq 1}|\mathbb{G}(t)|\right\}^4$, which is finite by van der Vaart and Wellner (1996, corollary 2.2.8), such that the right hand side in (27) converges to zero as $j \to \infty$. □

Lemma B.6. *For any* $c > 0$ *and* $\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) = \sup_{w^* \in \mathscr{W}}|\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)|$*,* $\sup_{t \in \mathbb{R}^d}\left\{\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) - c\|t\|\right\} = O_p(1)$.

Proof. For any $1 \le t^* < \infty$, using B.2 and the fact that $\inf_{\|t\|>t^*} \|t\|/c_t = 1/2$,

$$\limsup_{n\to\infty} \mathbb{P}\Big[\sup_{t\in\mathbb{R}^d} \{\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) - c\|t\|\} > C\Big] \le$$
$$\limsup_{n\to\infty} \mathbb{P}\Big\{\sup_{\|t\|\le t^*} \bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) > C\Big\} + \lim_{n\to\infty} \mathbb{P}\Big\{\sup_{\|t\|>t^*} \bar{\boldsymbol{S}}^{\boldsymbol{*}}_{\boldsymbol{nw}}(t) > \frac{c}{2}\Big\}.$$

The first right hand side term converges to zero as $C \to \infty$ by B.2. The second right hand side term converges to zero as $t^* \to \infty$ by B.5. □

Lemma B.7. *For any $c > 0$ and all sufficiently large n, $\sup_{t\in\Gamma_n} \mathbb{1}\{|R_{nw}(t)| - ct^\top Vt/4 > 0\} = 0$.*

Proof. Note that for some $c^* = c^*(t, n) \in [0, 1]$ and using the short hand $Q''_{nw}(c^*t) = \partial_{\theta\theta^\top} Q(\theta_{0n} + c^*t/\tilde{\alpha}_n, \eta_0 + w/\mathfrak{r}_n)$,

$$\sup_{t\in\Gamma_n} \{|R_{nw}(t)| - ct^\top Vt/4\} \le \frac{1}{2} \sup_{t\in\Gamma_n} \max\Big[t^\top\{Q''_{nw}(c^*t) + (1-c/2)V\}t, t^\top\{-Q''_{nw}(c^*t) - (1+c/2)V\}t\Big].$$

The stated result then follows from the fact that $c^* \in [0, 1]$ and that all eigenvalues of $Q''_{nw}(c^*t) + (1 - c/2)V$ and $-Q''_{nw}(c^*t) - (1 + c/2)V$ are nonpositive for sufficiently large n, uniformly in Γ_n, by the continuity of $\partial_{\theta\theta^\top} Q$ in both θ and η; see assumption C. □

Lemma B.8. *For any polynomial function P and any $0 \le c^* < \infty$,*

$$\sup_{w^*\in\mathscr{W}} \int_{\Gamma_n} \|P(t)\| \exp\{c^* \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)\} \big|\exp\{R_{nw}(t)\} - 1\big| \phi_V(t)dt = o_p(1).$$

Proof. Let $\mathscr{I}_{\boldsymbol{n}}$ be the left hand side of the lemma statement and let $0 < c < 1$. We will show that $\mathscr{I}_{\boldsymbol{n}} = O_p(c)$. Taking $b = R_{nw}(t)/c$ and $j = 0$ in B.3 $C_V \mathscr{I}_{\boldsymbol{n}}/c$ is bounded by

$$\int_{\Gamma_n} \|P(t)\| \exp\{c^* \bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) + |R_{nw}(t)/c| - t^\top Vt/2\} dt. \tag{28}$$

By B.7, for sufficiently large n, (28) is bounded above by $\int_{\Gamma_n} \|P(t)\| \exp\{c^* \bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) - t^\top Vt/4\} dt$ with probability one, which does not depend on c and is $O_p(1)$ by B.6. Therefore, it follows that $\mathscr{I}_{\boldsymbol{n}} = O_p(c)$; let $c \downarrow 0$. □

Lemma B.9. *For any polynomial function P,*

$$\sup_{w^*\in\mathscr{W}} \int_{\Gamma_n} \|P(t)\| \, |\pi_{nw}(t) - \pi_0| \exp\{c_\alpha^3 \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)\} \phi_V(t)dt = o_p(1). \tag{29}$$

Proof. Note that by assumption D, for some $0 < C < \infty$, $|\pi_{nw}(t) - \pi_0| \leq \big|\pi_0 - \pi\{\theta_{0n}(w)\}\big| + C\|t\|/\tilde{\alpha}_n$, where the first right hand side term is $o(1)$, and does not depend on t. Therefore, letting $\mathscr{I}_{\boldsymbol{n}}(t) = \|P(t)\| \exp\{c_{\alpha}^3 \bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t)\}\phi_V(t)dt$, the left hand side in (29) is bounded above by $o(1)\int_{\Gamma_n} \mathscr{I}_{\boldsymbol{n}}(t)dt + \int_{\Gamma_n} \|t\| \mathscr{I}_{\boldsymbol{n}}(t)dt/\tilde{\alpha}_n$, which is $o(1)O_p(1) + O_p(1/\tilde{\alpha}_n) = o_p(1)$ by B.6. □

Lemma B.10. *For any polynomial function P and constants $c_1, c_2 > 0$,*

$$\exp(\gamma_n) \int_{\Gamma_n^c} \|P(t)\| \exp\big\{c_1 \bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) - c_2 t^{\mathsf{T}} V t\big\} dt = o_p(1).$$

Proof. Follows immediately from B.6. □

Lemma B.11. *For any polynomial function P and constants $c_1, c_2, c_3 \geq 0$ with $c_2 + c_3 > 0$, and any sequences $\{\rho_n\}$ and $\{\tilde{\beta}_n\}$ with $1/\rho_n = O(1)$ and $\tilde{\beta}_n = O(1)$,*

$$\boldsymbol{I^c} = \exp(\gamma_n) \int_{\Gamma_n^{c*}} \|P(t)\| \exp\big[\rho_n^2 \{c_1 \tilde{\beta}_n \bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t) + c_2 Q_{nw}(t) - c_3 t^{\mathsf{T}} V t\}\big] dt = o_p(1).$$

Proof. Let $\bar{\boldsymbol{S}}^{\boldsymbol{c}}_{\boldsymbol{nw}} = \sup_{t \in \Gamma_n^c} \bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t)/\|t\|$. Then for any $\epsilon^*, \epsilon > 0$,

$$\mathbb{P}(\boldsymbol{I^c} > \epsilon^*) \leq \mathbb{P}(\bar{\boldsymbol{S}}^{\boldsymbol{c}}_{\boldsymbol{nw}} > \epsilon) + \mathbb{P}(\bar{\boldsymbol{S}}^{\boldsymbol{c}}_{\boldsymbol{nw}} \leq \epsilon, \boldsymbol{I^c} > \epsilon^*). \tag{30}$$

RHS1 in (30) is $o(1)$ by B.5. For RHS2, note that if $\bar{\boldsymbol{S}}^{\boldsymbol{c}}_{\boldsymbol{nw}} \leq \epsilon$ then for some $\bar{\beta} < \infty$,

$$\boldsymbol{I^c} \leq \exp(\gamma_n) \int_{\Gamma_n^{c*}} \|P(t)\| \exp\big[\rho_n^2 \{c_1 \bar{\beta} \epsilon \|t\| + c_2 Q_{nw}(t) - c_3 t^{\mathsf{T}} V t\}\big] dt.$$

By B.4,

$$\begin{aligned}
\sup_{t \in \Gamma_n^{c*}} \big\{c_1 \bar{\beta}\epsilon\|t\| + c_2 Q_{nw}(t) - c_3 t^{\mathsf{T}} V t\big\} &\leq \sup_{t \in \Gamma_n^{c*}} \big\{c_1 \bar{\beta}\epsilon\|t\| - c_3 t^{\mathsf{T}} V t - c_2 \min(c_q \tilde{\alpha}_n^2, t^{\mathsf{T}} V t/4)\big\} \\
&\leq \sup_{t \in \Gamma_n^{c*}} \big\{c_1 \bar{\beta}\epsilon\|t\| - (c_3 + c_2/4)\lambda^- \|t\|^2\big\} + \sup_{t \in \Gamma_n^{c*}} \big(c_1 \bar{\beta}\epsilon\|t\| - c_3 \lambda_- \|t\|^2 - c_2 c_q \tilde{\alpha}_n^2\big),
\end{aligned}$$

which, for some $c > 0$ and all sufficiently large n, is bounded by $-c\gamma_n^2$. So,

$$\lim_{n \to \infty} \mathbb{P}(\bar{\boldsymbol{S}}^{\boldsymbol{c}}_{\boldsymbol{nw}} \leq \epsilon, \boldsymbol{I^c} > \epsilon^*) \leq \lim_{n \to \infty} \mathbb{1}\Big\{\exp(\gamma_n - c\gamma_n^2) \int_{\Gamma_n^{c*}} \|P(t)\| dt > \epsilon^*\Big\} = 0.$$ □

Lemma B.12. *For any integer $j \geq 0$,*

$$\sup_{w^* \in \mathscr{W}} \left\| \int t^j \exp\{c_{\alpha}^3 \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)\}\big[\pi_{nw}(t) \exp\{Q_{nw}(t)\} - \pi_0 \exp\big(-t^{\mathsf{T}} V t/2\big)\}\big] dt \right\| = o_p(1).$$

Proof. Let $\boldsymbol{k_1}(t,w^*) = t^j\pi_{nw}(t)\exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + Q_{nw}(t)\}$ and $\boldsymbol{k_2}(t,w^*) = t^j\pi_0\times \exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) - t^\intercal Vt/2\}$. Noting that by assumption D $\boldsymbol{k_1}(t,w^*) = 0$ for $t \in \Gamma_n^c - \Gamma_n^{c*}$, we have (omitting arguments)

$$\sup\left\|\int(\boldsymbol{k_1}-\boldsymbol{k_2})\right\| \le \sup\left\|\int_{\Gamma_n}(\boldsymbol{k_1}-\boldsymbol{k_2})\right\| + \sup\left\|\int_{\Gamma_n^{c*}}\boldsymbol{k_1}\right\| + \sup\left\|\int_{\Gamma_n^c}\boldsymbol{k_2}\right\|. \tag{31}$$

The first right hand side term in (31) is $o_p(1)$ by B.8 and B.9, the second term is $o_p(1)$ by B.11, and the last term is $o_p(1)$ by B.10. □

Proof of part (i) of theorem 1. Note that for $\tilde{\alpha}_n = \alpha_n$ by (17) and B.12,

$$\alpha_n\{\hat{\boldsymbol{\theta}}^*(w^*,w) - \theta_{0n}(w)\} = \frac{\int t\pi_{nw}(t)\exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + Q_{nw}(t)\}dt}{\int \pi_{nw}(t)\exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + Q_{nw}(t)\}dt} = \frac{\int t\exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\}\phi_V(t)dt}{\int \exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\}\phi_V(t)dt} + o_p(1),$$

uniformly in $w^* \in \mathscr{W}$: note here that in view of B.5 and B.6 the integral in the middle is finite with probability approaching one. Apply B.2 together with the continuous mapping theorem. Divide both sides by c_α^2. □

Appendix C. Part (ii) of theorem 1

Lemma C.1. *Let* $\tilde{\gamma} > 0$, $\tilde{\Gamma} = \{t \in \mathbb{R}^d : \|t\| \le \tilde{\gamma}\}$, $\rho_n = \alpha_n^2/n^{2/3}$, *and* $\tilde{\boldsymbol{L}}_{\boldsymbol{nw}}(t,w^*) = \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + Q_{nw}(t)$. *Let further,*

$$\boldsymbol{X_n}(w^*) = \frac{\int_{\tilde{\Gamma}} t\pi_{nw}(t)\exp\{\rho_n\tilde{\boldsymbol{L}}_{\boldsymbol{nw}}(t,w^*)\}dt}{\int_{\tilde{\Gamma}} \pi_{nw}(t)\exp\{\rho_n\tilde{\boldsymbol{L}}_{\boldsymbol{nw}}(t,w^*)\}dt},$$

and for all $w^* \in \mathscr{W}$, *let* $\mathbb{X}(w^*) = \mathbb{X} = \operatorname{argmax}_{t\in\tilde{\Gamma}}\mathbb{C}(t)$. *Then* $\boldsymbol{X_n} \rightsquigarrow \mathbb{X}$ *in* $\mathbb{L}^\infty(\mathscr{W})$.

Proof. We will use the Skorokhod representation theorem and the fact that the collection of continuous functions on a compact set equipped with the sup norm is separable. Let $\Omega = \tilde{\Gamma}\times\mathscr{W}$ and μ be the Lebesgue measure. By B.2 and the fact that $\sup_{(t,w^*)\in\Omega}\left|Q_{nw}(t) + t^\intercal Vt/2\right| = o(1)$, we have the weak convergence of $\tilde{\boldsymbol{L}}_{\boldsymbol{nw}}$ to $\mathbb{C}$ in $\mathbb{L}^\infty(\Omega)$. Since the support of $\mathbb{C}$, which is the collection of continuous functions on the compact set Ω, is separable (with the sup norm), the Skorokhod representation theorem implies that there exist $\tilde{\boldsymbol{L}}^{\boldsymbol{*}}_{\boldsymbol{nw}}$ and $\mathbb{C}^{\boldsymbol{*}}$ with the same distributions as $\tilde{\boldsymbol{L}}_{\boldsymbol{nw}}$ and $\mathbb{C}$ for which

$$\sup_{(t,w^*)\in\Omega}\left|\tilde{\boldsymbol{L}}^{\boldsymbol{*}}_{\boldsymbol{nw}}(t,w^*) - \mathbb{C}^{\boldsymbol{*}}(t)\right| = o_{\text{a.s.}}(1). \tag{32}$$

We will now establish that

$$\sup_{w^*\in\mathscr{W}} \left\| \boldsymbol{X_n^*}(w^*) - \mathbb{X}^* \right\| = o_{\text{a.s}}(1), \tag{33}$$

where $\mathbb{X}^* = \operatorname{argmax}_{t\in\tilde{\Gamma}} \mathbb{C}^*(t)$ and

$$\boldsymbol{X_n^*}(w^*) = \frac{\int_{\tilde{\Gamma}} t\pi_{nw}(t)\exp\{\rho_n \boldsymbol{\tilde{L}^*_{nw}}(t,w^*)\}dt}{\int_{\tilde{\Gamma}} \pi_{nw}(t)\exp\{\rho_n \boldsymbol{\tilde{L}^*_{nw}}(t,w^*)\}dt}.$$

First, let $\tilde{L}^*_{nw}, \mathbb{C}^*$ be sample paths of $\boldsymbol{\tilde{L}^*_{nw}}, \mathbb{C}^*$ for which (32) holds. Let for arbitrary $c > 0$ and $\varkappa : \Omega \mapsto \mathbb{R}$, $T(\varkappa, c) = \{(t,w^*) \in \Omega : \left|\varkappa(t,w^*) - \bar{\mathbb{C}}^*\right| \le c\}$, where $\bar{\mathbb{C}}^* = \max_{t\in\tilde{\Gamma}} \mathbb{C}^*(t)$. Let further for arbitrary sets $T_1, T_2 \subset \Omega$, $d^*(T_1,T_2) = \mu(T_1 - T_2) + \mu(T_2 - T_1)$. We first establish that

$$d^*\{T(\tilde{L}^*_{nw}, c), T(\mathbb{C}^*, c)\} = o(1). \tag{34}$$

Let $T_{1n}(c) = T(\tilde{L}^*_{nw}, c) - T(\mathbb{C}^*, c)$ and $T_{2n}(c) = T(\mathbb{C}^*, c) - T(\tilde{L}^*_{nw}, c)$. We show that $\mu\{T_{2n}(c)\} = o(1)$; establishing that $\mu\{T_{1n}(c)\} = o(1)$ is analogous.

Let for arbitrary $c^* > 0$, $T_n^*(c^*) = \{(t,w^*) \in \Omega : \left|\tilde{L}^*_{nw}(t,w^*) - \mathbb{C}^*(t)\right| \le c^*\}$. Defining complements relative to Ω, we note that

$$\mu\{T_{2n}(c)\} = \mu\{T_{2n}(c) \cap T_n^*(c^*)\} + \mu\{T_{2n}(c) \cap T_n^{*c}(c^*)\} \le \\ \mu\{T_{2n}(c) \cap T_n^*(c^*)\} + \mu\{T_n^{*c}(c^*)\} = \mu\{T_{2n}(c) \cap T_n^*(c^*)\} + o(1),$$

by (32). Further, by the triangle inequality,

$$T_{2n}(c) \cap T_n^*(c^*) \subset T_n^{**}(c,c^*) = \{(t,w^*) \in \Omega : c \le |\tilde{L}^*_{nw}(t,w^*) - \bar{\mathbb{C}}^*| \le c + c^*\}.$$

Now, by (32),

$$\lim_{c^*\downarrow 0} \lim_{n\to\infty} \mu\{T_n^{**}(c,c^*)\} = \lim_{c^*\downarrow 0} \mu\left[\{(t,w^*) \in \Omega : c \le \left|\mathbb{C}^*(t) - \bar{\mathbb{C}}^*\right| \le c + c^*\}\right] = \\ \mu\left[\{(t,w^*) \in \Omega : \left|\mathbb{C}^*(t) - \bar{\mathbb{C}}^*\right| = c\}\right] = 0,$$

because $\mathbb{C}^*$ is everywhere continuous and nowhere differentiable. So (34) holds.

Finally, note that for $j = 0, 1$, $\bar{\pi}$ defined in assumption D, and some $C < \infty$,

$$\sup_{w^*\in\mathscr{W}} \int_{T^c(\tilde{L}^*_{nw}, c)} \|t\|^j |\pi_{nw}(t)| \exp\left[\rho_n\{\tilde{L}^*_{nw}(t,w^*) - \bar{\mathbb{C}}^*\}\right]dt \le C\exp(-\rho_n c)\bar{\pi} \int_{\tilde{\Gamma}} \|t\|^j dt = o(1),$$

by the compactness of $\tilde{\Gamma}$ and divergence of ρ_n. Thus,

$$\sup_{w^*\in\mathscr{W}} \frac{\int_{\tilde{\Gamma}} t\pi_{nw}(t)\exp\{\rho_n \tilde{L}^*_{nw}(t,w^*)\}dt}{\int_{\tilde{\Gamma}} \pi_{nw}(t)\exp\{\rho_n \tilde{L}^*_{nw}(t,w^*)\}dt} \leq \text{ess sup}\, T(\tilde{L}^*_{nw}, c) + o(1) = \text{ess sup}\, T(\mathbb{C}^*, c) + o(1),$$

where the second equality is due to the uniform convergence of $\tilde{L}^*_{nw}$ to $\mathbb{C}^*$. Repeat the arguments to obtain

$$\inf_{w^*\in\mathscr{W}} \frac{\int_{\tilde{\Gamma}} t\pi_{nw}(t)\exp\{\rho_n \tilde{L}^*_{nw}(t,w^*)\}dt}{\int_{\tilde{\Gamma}} \pi_{nw}(t)\exp\{\rho_n \tilde{L}^*_{nw}(t,w^*)\}dt} \geq \text{ess inf}\, T(\tilde{L}^*_{nw}, c) + o(1) = \text{ess inf}\, T(\mathbb{C}^*, c) + o(1).$$

Then note that $\lim_{c\downarrow 0}$ ess inf $T(\mathbb{C}^*, c) = \lim_{c\downarrow 0}$ ess sup $T(\mathbb{C}^*, c) = \text{argmax}_{t\in\tilde{\Gamma}}\, \mathbb{C}^*(t)$, where the last equality follows from Kim and Pollard (1990, lemma 2.6).

So we have established (33). Hence for any bounded and continuous function $\mathfrak{f} : \mathbb{L}^\infty(\mathscr{W}) \mapsto \mathbb{R}$ we have by the dominated convergence theorem that $\mathbb{E}\mathfrak{f}(\boldsymbol{X_n}) = \mathbb{E}\mathfrak{f}(\boldsymbol{X_n^*}) \to \mathbb{E}\mathfrak{f}(\mathbb{X}^*) = \mathbb{E}\mathfrak{f}(\mathbb{X})$. □

Lemma C.2. *Let $\rho_n, \tilde{\Gamma}$ be defined as in C.1 and $\tilde{\Gamma}^c_n = \mathfrak{T}_n - \tilde{\Gamma}$. Then for any $\epsilon > 0$ and $j = 0, 1$,*

$$\lim_{\tilde{\gamma}\to\infty}\lim_{n\to\infty} \mathbb{P}\left(\frac{\sup_{w^*\in\mathscr{W}} \int_{\tilde{\Gamma}^c_n} \|t\|^j \pi_{nw}(t)\exp\big[\rho_n\big\{\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + Q_{nw}(t)\big\}\big]dt}{\inf_{w^*\in\mathscr{W}} \int \pi_{nw}(t)\exp\big[\rho_n\big\{\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + Q_{nw}(t)\big\}\big]dt} > \epsilon\right) = 0. \quad (35)$$

Proof. Denote the numerator and denominator in (35) $\boldsymbol{I_N}$ and $\boldsymbol{I_D}$, respectively. Note that for any $\bar{c} > 0$,

$$\mathbb{P}\big(\boldsymbol{I_N}/\boldsymbol{I_D} > \epsilon\big) \leq \mathbb{P}\big\{\boldsymbol{I_N} > \epsilon\exp(-\bar{c}\rho_n)\big\} + \mathbb{P}\big\{\boldsymbol{I_D} \leq \exp(-\bar{c}\rho_n)\big\}. \quad (36)$$

We first work on the first right hand side term in (36). For $c > 0$ let $\boldsymbol{Z_n}(\tilde{\gamma}, c) = \big\{t \in \tilde{\Gamma}^c_n : \sup_{w^*\in\mathscr{W}}|\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)| \leq c\|t\|^2\big\}$. For $c^* = 2\sup_{\theta\in\Theta}\|\theta\| < \infty$, we have $\sup_{t\in\boldsymbol{Z_n}(\tilde{\gamma},c)}\|t\| \leq c^*\sqrt[3]{n}$, such that by B.4 for sufficiently small c,

$$\begin{aligned}
&\sup_{t\in\boldsymbol{Z_n}(\tilde{\gamma},c)}\sup_{w^*\in\mathscr{W}} \exp\big[\rho_n\big\{\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*) + Q_{nw}(t)\big\}\big] \leq \\
&\qquad \sup_{t\in\boldsymbol{Z_n}(\tilde{\gamma},c)} \exp\big[\rho_n\big\{c\|t\|^2 - \min\big(c_q n^{2/3}, \lambda_-\|t\|^2/4\big)\big\}\big] \leq \\
&\qquad \exp\big\{\alpha_n^2(cc^{*2} - c_q)\big\} + \sup_{t\in\boldsymbol{Z_n}(\tilde{\gamma},c)} \exp\big\{\rho_n\|t\|^2(c - \lambda_-/4)\big\} \leq \\
&\qquad \exp\big\{\rho_n(c - \lambda_-/4)/\tilde{\gamma}^2\big\} + o\big\{\bar{c}\alpha_n^2\big\} = o\big\{\exp(-\bar{c}\rho_n)\big\}, \quad (37)
\end{aligned}$$

if $\bar{c} > 0$ is chosen sufficiently small. Further,

$$\lim_{\tilde{\gamma}\to\infty}\lim_{n\to\infty} \mathbb{P}\big\{t \in \tilde{\Gamma}^c_n - \boldsymbol{Z_n}(\tilde{\gamma}, c)\big\} \leq \lim_{\tilde{\gamma}\to\infty}\lim_{n\to\infty} \mathbb{P}\left\{\sup_{\|t\|>\tilde{\gamma}} \frac{\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(t)}{\|t\|^2} > c\right\} = 0, \quad (38)$$

by B.5. Combining (37) and (38) implies that the first right hand side term in (36) is $o(1)$.

Now the second right hand side term in (36). Let $0 < \tilde{\epsilon} \leq \sqrt{\bar{c}/4\lambda^+}$ be some constant to be manipulated later. Note that for $\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(\tilde{\epsilon}) = \sup_{\|t\| \leq \tilde{\epsilon}} \sup_{w^* \in \mathscr{W}} |\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t, w^*)|$,

$$\mathbb{P}\big\{\boldsymbol{I_D} \leq \exp(-\bar{c}\rho_n)\big\} \leq \mathbb{P}\big\{\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(\tilde{\epsilon}) > \bar{c}/2\big\} + \mathbb{P}\big\{\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(\tilde{\epsilon}) \leq \bar{c}/2,\ \boldsymbol{I_D} \leq \exp(-\bar{c}\rho_n)\big\}. \quad (39)$$

Using assumption C, the second right hand side term in (39) is for sufficiently large n and $\underline{\pi} > 0$ bounded by

$$\mathbb{1}\Big[\underline{\pi} \int_{\|t\| \leq \tilde{\epsilon}} \exp\big\{\rho_n(-\bar{c}/2 - \lambda^+ \tilde{\epsilon}^2)\big\} dt \leq \exp(-\bar{c}\rho_n)\Big] \leq \mathbb{1}\big\{\underline{\pi}\tilde{\epsilon} \exp(\rho_n \bar{c}/4) \leq 1\big\} = o(1).$$

Finally, for the first right hand side term in (39) note that by B.2,

$$\lim_{\tilde{\epsilon} \to 0} \lim_{n \to \infty} \mathbb{P}\big\{\bar{\boldsymbol{S}}_{\boldsymbol{nw}}(\tilde{\epsilon}) > \bar{c}/2\big\} = \lim_{\tilde{\epsilon} \to 0} \mathbb{P}\Big\{\sup_{\|t\| \leq \tilde{\epsilon}} |\mathbb{G}(t)| > \bar{c}/2\Big\} = 0,$$

since $\mathbb{G}(0) = 0$ by definition. □

Proof of part (ii) of theorem 1. For $\rho_n = \alpha_n^2/n^{2/3}$, $j = 0, 1$, and any set $\Gamma \in \mathbb{R}^d$, let $\boldsymbol{I_j}(\Gamma) = \int_\Gamma t^j \pi_{nw}(t) \exp\{\rho_n \tilde{\boldsymbol{L}}_{\boldsymbol{nw}}(t, w^*)\} dt$, and $\boldsymbol{I_j} = \boldsymbol{I_j}(\mathbb{R}^d)$. Then for $\tilde{\Gamma}, \tilde{\Gamma}_n^c$ defined in C.1 and C.2 by assumption D and (17) using $t = \sqrt[3]{n}\{\theta - \theta_{0n}(w)\}$,

$$\sqrt[3]{n}\{\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*, w) - \theta_{0n}(w)\} = \frac{\boldsymbol{I_1}}{\boldsymbol{I_0}} = \frac{\boldsymbol{I_1}(\tilde{\Gamma}) + \boldsymbol{I_1}(\tilde{\Gamma}_n^c)}{\boldsymbol{I_0}(\tilde{\Gamma}) + \boldsymbol{I_0}(\tilde{\Gamma}_n^c)} = \frac{\boldsymbol{I_1}(\tilde{\Gamma})}{\boldsymbol{I_0}(\tilde{\Gamma})}\left\{1 - \frac{\boldsymbol{I_0}(\tilde{\Gamma}_n^c)}{\boldsymbol{I_0}}\right\} + \frac{\boldsymbol{I_1}(\tilde{\Gamma}_n^c)}{\boldsymbol{I_0}}.$$

Apply C.1 and C.2. □

Appendix D. Part (iii) of theorem 1

Lemma D.1. $\int \pi_{nw}(t) \exp\{Q_{nw}(t)\} dt = \pi_0 C_V + o(1)$.

Proof. The difference between left hand side and right hand side can be expanded as

$$C_V \int_{\Gamma_n} \pi_{nw}(t)\big[\exp\{R_{nw}(t)\} - 1\big]\phi_V(t) dt + C_V \int_{\Gamma_n} \{\pi_{nw}(t) - \pi_0\}\phi_V(t) dt + \int_{\Gamma_n^c} \pi_{nw}(t) \exp\{Q_{nw}(t)\} dt - C_V \pi_0 \int_{\Gamma_n^c} \phi_V(t) dt, \quad (40)$$

where all four terms are $o(1)$ by B.7, assumption D, B.11, and the fact that $1/\gamma_n = o(1)$, respectively. □

Lemma D.2. *For* $\mathscr{V}_N = \pi_0^2 \iint ts^\top H(t,s)\phi_V(t)\phi_V(s)dtds$, $\int \pi_{nw}(t)t\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\phi_V(t)dt$ *converges weakly in* $\mathbb{L}^\infty(\mathscr{W})$ *to a flat limit process whose marginals have a* $N(0,\mathscr{V}_N)$*–distribution.*

Proof. By B.2 and assumption D, as a process of w^*,

$$\int \pi_{nw}(t)t\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\phi_V(t)dt \rightsquigarrow \pi_0 \int t\mathbb{G}(t)\phi_V(t)dt \sim N(0,\mathscr{V}_N). \qquad \square$$

Lemma D.3. *For* $j = 0, 1$, *if* $\beta_n = o(1)$,

$$\sup_{w^*\in\mathscr{W}} \left\| \int \pi_{nw}(t)t^j \left[\exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\} - \sum_{s=0}^{j} \{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\}^s \right] \exp\{Q_{nw}(t)\}dt \right\| = O_p(\beta_n^{j+1}).$$

Proof. By B.3 with $c = \beta_n$ and $b = \tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)$, the left hand side is bounded by $\beta_n^{j+1}\boldsymbol{I}$, where $\boldsymbol{I} = \int |\pi_{nw}(t)|\,\|t\|^j \exp\{\bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t)c_t + Q_{nw}(t)\}dt$. So it suffices to show that $\boldsymbol{I} = O_p(1)$. Now, for any $0 < C^* < \infty$,

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(\boldsymbol{I} > C) \leq \lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}\left\{\boldsymbol{I} > C,\ \sup_{t\in\mathbb{R}^d} \bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t) \leq C^*\right\}$$
$$+ \limsup_{n\to\infty} \mathbb{P}\left\{\sup_{t\in\mathbb{R}^d} \bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t) > C^*\right\}. \qquad (41)$$

The first right hand side term in (41) is by B.4 for some polynomial function P bounded by

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{1}\left[\exp(C^*) \int_{\mathfrak{T}_n} \|P(t)\| \exp\{C^* c_t - \min(t^\top Vt/4, c_q\tilde{\alpha}_n^2)\}dt > C\right] = 0,$$

since c_t cannot grow faster than $\tilde{\alpha}_n$ by the definition of $\mathfrak{T}_n$, and the normal distribution has infinitely many moments. Finally, by B.2, the second right hand side term in (41) equals $\mathbb{P}\{\sup_{t\in\mathbb{R}^d} \mathbb{G}^*(t) > C^*\}$. Let $C^* \to \infty$. $\square$

Lemma D.4. *For any polynomial function* P, $\int_{\mathfrak{T}_n} \|P(t)\| \bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t) \left|\exp\{R_{nw}(t)\} - 1\right| \phi_V(t)dt = o_p(1)$.

Proof. Note that $\sup_{t\in\mathbb{R}^d} \bar{\boldsymbol{S}}^*_{\boldsymbol{nw}}(t) = O_p(1)$ by B.6, so it suffices to show that

$$\begin{cases} \int_{\Gamma_n} \|P(t)\| \left|\exp\{R_{nw}(t)\} - 1\right| \phi_V(t)dt &= o_p(1), \\ \int_{\Gamma_n^{c*}} \|P(t)\| \exp\{Q_{nw}(t)\}dt &= o_p(1), \\ \int_{\Gamma_n^{c*}} \|P(t)\|\phi_V(t)dt &= o_p(1), \end{cases}$$

which are established in B.8 and B.11. $\square$

Proof of part (iii) of theorem 1. Define

$$\mathbb{B}_{nw} = \beta_n^{-1}\alpha_n^{d+1}\int\{\theta-\theta_{0n}(w)\}\pi(\theta)\exp\{\alpha_n^2 Q(\theta,\eta_0+w/\mathfrak{r}_n)\}d\theta, \tag{42}$$

$$\mathbb{D}_{\boldsymbol{nw}}(w^*) = \alpha_n^d\int\pi(\theta)\exp\Big[\alpha_n^2\{\boldsymbol{S_n}(\theta,\eta_0+w^*/\mathfrak{r}_n)+Q(\theta,\eta_0+w/\mathfrak{r}_n)\}\Big]d\theta. \tag{43}$$

By (17) we get

$$\sqrt{n/\alpha_n}\{\hat{\boldsymbol{\theta}}^{\boldsymbol{*}}(w^*,w)-\theta_{0n}(w)\} = \frac{\beta_n^{-1}\int t\pi_{nw}(t)\exp\{\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)+Q_{nw}(t)\}dt}{\int\pi_{nw}(t)\exp\{\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)+Q_{nw}(t)\}dt}. \tag{44}$$

Denote the right hand side numerator in (44) $\mathbb{N}_{\boldsymbol{nw}}(w^*)$ and note that the denominator equals $\mathbb{D}_{\boldsymbol{nw}}(w^*)$.Then $\mathbb{D}_{\boldsymbol{nw}}(w^*)-C_V\pi_0$ is equal to

$$\int\pi_{nw}(t)\Big[\exp\{\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\}-1\Big]\exp\{Q_{nw}(t)\}dt+\Big[\int\pi_{nw}(t)\exp\{Q_{nw}(t)\}dt-C_V\pi_0\Big],$$

where the first term is $o_p(1)$, uniformly in w^*, by D.3 and the second term is $o_p(1)$ by D.1. So $\mathbb{D}_{\boldsymbol{nw}}(w^*) = C_V\pi_0+o_p(1)$, uniformly in w^*.

For $\mathbb{N}_{\boldsymbol{nw}}(w^*)$, noting that $\mathbb{B}_{nw} = \beta_n^{-1}\int t\pi_{nw}(t)\exp\{Q_{nw}(t)\}dt$, we have

$$\mathbb{N}_{\boldsymbol{nw}}(w^*)-\mathbb{B}_{nw} = \frac{1}{\beta_n}\int\pi_{nw}(t)t\big[\exp\{\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\}-1-\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\big]\exp\{Q_{nw}(t)\}dt+$$
$$C_V\int\pi_{nw}(t)t\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\big[\exp\{R_{nw}(t)\}-1\big]\phi_V(t)dt+C_V\int\pi_{nw}(t)t\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\phi_V(t)dt,$$

where the first two right hand side terms are $o_p(1)$, uniformly in w^*, by D.3 and D.4. Hence, $\mathbb{N}_{\boldsymbol{nw}}(w^*)-\mathbb{B}_{nw}\rightsquigarrow C_V\pi_0 N(0,\mathscr{V})$, as a process of w^*. □

Appendix E. Part (iv) of theorem 1

Lemma E.1.

$$\int_{\Gamma_n} t\pi_{nw}(t)\exp\Big[\exp\{R_{nw}(t)\}-\sum_{p=0}^{q}\frac{R_{nw}^p(t)}{p!}\Big]\phi_V(t)dt = O\{(\gamma_n^3/\alpha_n)^{q+1}\} = o(1). \tag{45}$$

Proof. First note that for $s_n^* = \sup_{t\in\Gamma_n}\big|R_{nw}(t)\big|$ and by the definition of γ_n in appendix A,

$$s_n^*\le\gamma_n^2\sup_{\|t\|\le\gamma_n}\big\|\partial_{\theta\theta^\intercal}Q\{\theta_{0n}(w)+t/\alpha_n,\eta_0+w/\mathfrak{r}_n\}+V\big\| = O\big(\gamma_n^3/\alpha_n\big) = o(1),$$

by assumption C and the definition of γ_n in appendix A. The length of the left hand side in (45) is equal to

$$\left\| \sum_{p=q+1}^{\infty} \int_{\Gamma_n} \pi_{nw}(t) t \frac{R_{nw}^p(t)}{p!} \phi_V(t) dt \right\| \le (s_n^*)^{q+1} \exp(s_n^*) \int_{\Gamma_n} \|\pi_{nw}(t) t\| \phi_V(t) dt = O\{(s_n^*)^{q+1}\}.$$

□

Proof of part (iv) of theorem 1. Since $\mathbb{B}_{nw} = \beta_n^{-1} \int t\pi_{nw}(t) \exp\{Q_{nw}(t)\} dt$, it suffices to show that

$$\int t\pi_{nw}(t) \exp\{R_{nw}(t)\} \phi_V(t) dt = \sum_{\tau=0}^{q} \frac{b_{q\tau}^*}{\alpha_n^{\tau}} + o(\alpha_n^{-q}). \tag{46}$$

By B.11 and E.1, the left hand side in (46) equals

$$\sum_{p=0}^{q} \frac{1}{p!} \int_{\Gamma_n} \pi_{nw}(t) t R_{nw}^p(t) \phi_V(t) dt + o\big(\alpha_n^{-q}\big). \tag{47}$$

Let $D_{Q,\delta;nw}(t)/\alpha_n^{\delta}$ and $D_{\pi\delta;nw}(t)/\alpha_n^{\delta}$ be the order δ terms in the Taylor expansions of $Q\{\theta_{0n}(w) + t/\alpha_n, \eta_0 + w^*/\mathfrak{r}_n\}$ and $\pi\{\theta_{0n}(w) + t/\alpha_n\}$ around $\big(\theta_{0n}(w), \eta_0 + w^*/\mathfrak{r}_n\big)$ and $\theta_{0n}(w)$, respectively. Then, by assumptions C and D and B.1 and the mean value theorem, $R_{nw}(t)$ and $\pi_{nw}(t)$ have expansions such that

$$R_{nw}(t) = \sum_{\delta=1}^{q} \frac{D_{Q,\delta+2;nw}(t)}{\alpha_n^{\delta}} + \frac{D_{Q\Delta nw}^*(t)}{\alpha_n^{q}}, \qquad \pi_{nw}(t) = \sum_{\delta=0}^{q} \frac{D_{\pi\delta;nw}(t)}{\alpha_n^{\delta}} + \frac{D_{\pi qnw}^*(t)}{\alpha_n^{q}}, \tag{48}$$

where the remainders $D_{Q\Delta nw}^*(t)$ and $D_{\pi qnw}^*(t)$ satisfy

$$\sup_{t\in\Gamma_n\setminus\{0\}} \big\{ \big|D_{Q\Delta;nw}^*(t)\big| / \|t\|^{\Delta} \big\} = o(1), \qquad \sup_{t\in\Gamma_n\setminus\{0\}} \big\{ \big|D_{\pi q;nw}^*(t)\big| / \|t\|^{q} \big\} = o(1). \tag{49}$$

Further, by the definition of γ_n, assumptions C and D and B.1,

$$\begin{cases} \sup_{t\in\Gamma_n} \big|D_{Q;\delta+2;nw}(t) - D_{Q;\delta+2}(t)\big| / \alpha_n^{\delta} = O\big(\gamma_n^{\delta+2}/\mathfrak{r}_n \alpha_n^{\delta}\big) = o\big(\alpha_n^{-q}\big), & \delta = 1, \ldots, q, \\ \sup_{t\in\Gamma_n} \big|D_{\pi\delta;nw}(t) - D_{\pi\delta}(t)\big| / \alpha_n^{\delta} = O\big(\gamma_n^{\delta}/\mathfrak{r}_n \alpha_n^{\delta}\big) = o\big(\alpha_n^{-q}\big), & \delta = 0, \ldots, q, \end{cases} \tag{50}$$

Now, (48) to (50) and the fact that $\int \|P(t)\| \phi_V(t) dt < \infty$ for any polynomial function P imply that (47) equals

$$\sum_{p=0}^{q} \frac{1}{p!} \int_{\Gamma_n} \left\{ \sum_{\delta=0}^{q} \frac{D_{\pi\delta}(t)}{\alpha_n^{\delta}} \right\} \left\{ \sum_{\delta=1}^{q} \frac{D_{Q,\delta+2}(t)}{\alpha_n^{\delta}} \right\}^p t\phi_V(t) dt + o\big(\alpha_n^{-q}\big). \tag{51}$$

Let $\mathscr{M}_{pq}$ be the collection of vectors $m = (m_1, \cdots, m_q)$ consisting of nonnegative integers for which $\sum_{j=1}^{q} m_j = p$,such that $\mathscr{M}^*_{pqs} \subset \mathscr{M}_{pq}$ for all s. Then the first term in (51) is by the multinomial theorem equal to

$$\sum_{p=0}^{q} \int_{\Gamma_n} \Big\{ \sum_{\delta=0}^{q} \frac{D_{\pi\delta}(t)}{\alpha_n^{\delta}} \Big\} \sum_{m \in \mathscr{M}_{pq}} \prod_{\delta=1}^{q} \Big[\Big\{ \frac{D_{Q;\delta+2}}{\alpha_n^{\delta}} \Big\}^{m_\delta} \frac{1}{m_\delta !} \Big] t \phi_V(t) dt =$$
$$\sum_{p=0}^{q} \sum_{m \in \mathscr{M}_{pq}} \sum_{j=0}^{q} \frac{1}{\alpha_n^{j+\sum_{\delta=1}^{q} m_\delta \delta}} \int_{\Gamma_n} D_{\pi j}(t) \Big\{ \prod_{\delta=1}^{q} \frac{D_{Q,\delta+2}^{m_\delta}(t)}{m_\delta !} \Big\} t \phi_V(t) dt =$$
$$\sum_{j=0}^{q} \sum_{s=0}^{q-j} \frac{b_{qsj}}{\alpha_n^{s+j}} + O\big(\alpha_n^{-q-1}\big),$$

where $\int_{\Gamma_n^c} \cdot$ is negligible by B.10. □

Appendix F. Theorem 2

Lemma F.1. $\int \{D_{\pi 1}(t) + \pi_0 D_{Q3}(t)\} t \phi_V(t) dt = 0.$

Proof. Note that $D_{\pi 1}(t) = -(\pi_0/2) \sum_{s=1}^{d} \big\{ \partial_{\theta_s} \operatorname{vec}\big(\partial_{\theta\theta^\intercal} Q\big) \big\}^\intercal (\theta_0) \operatorname{vec}\big(V^{-1}\big) t_s$. Therefore, letting V^{pm} be the (p, m) element of V^{-1} and $\mathscr{K}_{pm\delta} = \partial_{\theta_p \theta_m \theta_\delta} Q(\theta_0)$, the j^{th} element of the left hand side of the lemma statement is

$$\int \pi_0 D_{Q3}(t) t_j \phi_V(t) dt + \int D_{\pi 1}(t) t_j \phi_V(t) dt =$$
$$\frac{\pi_0}{6} \sum_{p,m,\delta=1}^{d} \mathscr{K}_{pm\delta} \int t_p t_m t_\delta t_j \phi_V(t) dt - \frac{\pi_0}{2} \sum_{p,m,\delta=1}^{d} \mathscr{K}_{pm\delta} V^{m\delta} \int t_p t_j \phi_V(t) dt =$$
$$\frac{\pi_0}{6} \sum_{p,m,\delta=1}^{d} \mathscr{K}_{pm\delta} \big(V^{pm} V^{\delta j} + V^{p\delta} V^{mj} + V^{pj} V^{m\delta} \big) - \frac{\pi_0}{2} \sum_{p,m,\delta=1}^{d} \mathscr{K}_{pm\delta} V^{m\delta} V^{pj} = 0,$$

where the second equality is due to Isserlis's theorem for higher order moments of the multivariate normal distribution. □

Proof of theorem 2. Let $\hat{\boldsymbol{\pi}}_{\boldsymbol{nw}}(t) = \hat{\boldsymbol{\pi}}\{\theta_{0n}(w) + t/\tilde{\alpha}_n\}$, where $\tilde{\alpha}_n$ is chosen as in theorem 1. We first establish that if $q = 1$ then theorem 1 still holds if π is replaced with $\hat{\boldsymbol{\pi}}$, albeit that (since $q = 1$) $\mathbb{B}_{nw}$ is replaced with

$$\hat{\mathbb{B}}_{\boldsymbol{nw}} = \int \{D_{\hat{\boldsymbol{\pi}} 1}(t) + \hat{\boldsymbol{\pi}}_{\boldsymbol{0}} D_{Q3}(t)\} \phi_V(t) dt / c_\alpha^4 \hat{\boldsymbol{\pi}}_{\boldsymbol{0}} + o_p(1). \tag{52}$$

For this, we need to allow for randomness of $\hat{\boldsymbol{\pi}}_{\boldsymbol{nw}}$ in the proofs involving π_{nw}, notably those of B.9, B.12, C.1, C.2 and D.1 to D.3 and appendix E. The argument is essentially the same in all cases, so we will use B.9 as a representative example. We need to show that

$$\sup_{w^*\in\mathscr{W}}\int_{\Gamma_n}\|P(t)\|\;|\hat{\boldsymbol{\pi}}_{\boldsymbol{nw}}(t)-\pi_0|\exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\}\phi_V(t)dt=o_p(1). \tag{53}$$

Denote the left hand side in (53) by $\hat{\boldsymbol{I}}$ and let $\hat{\boldsymbol{\sigma}}=\sup_{\theta\in\Theta}\big\|\partial_\theta\hat{\boldsymbol{\pi}}(\theta)\big\|$. Then for any $\epsilon>0$,

$$\mathbb{P}(\hat{\boldsymbol{I}}>\epsilon)\le\mathbb{P}(\hat{\boldsymbol{I}}>\epsilon,\ \hat{\boldsymbol{\sigma}}\le\bar{\pi}_1)+\mathbb{P}(\hat{\boldsymbol{\sigma}}>\bar{\boldsymbol{\pi}}_{\boldsymbol{1}})=\mathbb{P}(\hat{\boldsymbol{I}}>\epsilon,\ \hat{\boldsymbol{\sigma}}\le\bar{\pi}_1)+o(1). \tag{54}$$

Since for all $t\in\Gamma_n$, $|\hat{\boldsymbol{\pi}}_{\boldsymbol{nw}}(t)-\pi_0|\le\hat{\boldsymbol{\sigma}}\big\{\gamma_n/\tilde{\alpha}_n+\|\theta_{0n}(w)-\theta_0\|\big\}$, RHS1 in (54) is bounded by

$$\mathbb{P}\left[\bar{\pi}_1\{\gamma_n/\tilde{\alpha}_n+\|\theta_{0n}(w)-\theta_0\|\}\sup_{w^*\in\mathscr{W}}\int_{\Gamma_n}\|P(t)\|\;\exp\{c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}(t,w^*)\}\phi_V(t)dt>\epsilon\right]=o(1),$$

by B.9. Since both the numerator and the denominator on the left hand side in (52) are linear in $\hat{\boldsymbol{\pi}}_{\boldsymbol{0}}$, $\boldsymbol{\partial_\theta}\hat{\pi}(\boldsymbol{\theta_0})$, which are consistent for π_0, $\partial_\theta\pi(\theta_0)$, respectively, the stated result holds. □

Appendix G. Theorem 3

Lemma G.1. *Let $\mathfrak{d}_H^*(\theta,\eta)=\mathfrak{d}_H\big\{\theta,\Theta_0(\eta)\big\}$, where $\mathfrak{d}_H$ is the Hausdorff distance. The correspondence (mapping) $\bar{\Theta}^*:N\times\mathbb{R}\to\Theta$ given by $\bar{\Theta}^*(\eta,c)=\big\{\theta\in\Theta:\mathfrak{d}_H^*(\theta,\eta)\ge c\big\}$ is continuous in η in the sense of Berge (1963, page 109).*

Proof. We first show that the correspondence $\Theta_0:N\to\Theta$ given by $\Theta_0(\eta)=\operatorname{argmax}_{\theta\in\Theta}\tilde{Q}(\theta,\eta)$ is continuous in η. Upper hemicontinuity of Θ_0 follows by the maximum theorem (Berge, 1963, page 116) in view of the continuity of $\tilde{Q}$, because the correspondence $\Theta_0^*:N\to\Theta$ given by $\Theta_0^*(\eta)=\Theta$ is continuous due to the compactness of Θ. For lower hemicontinuity, let $C=\max_{\theta\in\Theta,\eta\in N}\big\|\partial_\eta\tilde{Q}(\theta,\eta)\big\|<\infty$ by assumption K such that $\sup_{\theta\in\Theta}\big|\tilde{Q}(\theta,\tilde{\eta})-\tilde{Q}(\theta,\eta)\big|\le C\|\eta-\tilde{\eta}\|$. It then suffices to show that for all $\tilde{\eta}\in N$, all $\tilde{\theta}\in\Theta_0(\tilde{\eta})$, and all $\epsilon>0$, $\|\eta-\tilde{\eta}\|<\epsilon/2C$ implies $Q(\tilde{\theta},\eta)>-\epsilon$ (recall that $Q(\theta,\eta)=\tilde{Q}(\theta,\eta)-\max_{\theta\in\Theta}\tilde{Q}(\theta,\eta)$). Take $\theta\in\Theta_0(\eta)$ and note that

$$Q(\tilde{\theta},\eta)=\tilde{Q}(\tilde{\theta},\eta)-\tilde{Q}(\theta,\eta)=\big\{\tilde{Q}(\tilde{\theta},\eta)-\tilde{Q}(\tilde{\theta},\tilde{\eta})\big\}-Q(\theta,\tilde{\eta})+\big\{\tilde{Q}(\theta,\tilde{\eta})-\tilde{Q}(\theta,\eta)\big\}>-\epsilon.$$

So $\Theta_0(\eta)$ is a continuous correspondence and hence by the maximum theorem, $\mathfrak{d}_H^*(\theta,\eta)$ is continuous in η and $\bar{\Theta}^*(\eta,c)$ is an upper hemicontinuous correspondence in η. It remains to be shown that $\bar{\Theta}^*(\eta,c)$ is also lower hemicontinuous in η. Choose an arbitrary $\bar{\eta}\in N$ and $\bar{\theta}\in\bar{\Theta}^*(\bar{\eta},c)$.

Now, for any sequence $\{\eta_j\}$ converging to $\bar{\eta}$,

$$\min_{\breve{\theta}\in\Theta_0(\eta_j)} \|\bar{\theta}-\breve{\theta}\| \to \min_{\breve{\theta}\in\Theta_0(\bar{\eta})} \|\bar{\theta}-\breve{\theta}\| = c^* \geq c. \tag{55}$$

Take θ_j^* from the set of minimizers on the left hand side in (55) and set $\theta_j = \theta_j^* + c^*(\bar{\theta}-\theta_j^*)/\|\bar{\theta}-\theta_j^*\|$. By assumption I, $\Theta_0(\eta_j)$ is convex and hence $\theta_j \in \bar{\Theta}^*(\eta_j, c)$. Finally, $\|\theta_j - \bar{\theta}\| = \big|\|\theta_j^* - \bar{\theta}\| - c^*\big| \to 0$ by (55). □

Lemma G.2.

$$\sup_{\eta^*,\eta\in N} \frac{\int \mathfrak{d}_H^*(\theta,\eta)\pi(\theta)\exp\{\alpha_n^2 \boldsymbol{S_n}(\theta,\eta^*) + \alpha_n^2 Q(\theta,\eta)\}d\theta}{\int \pi(\theta)\exp\{\alpha_n^2 \boldsymbol{S_n}(\theta,\eta^*) + \alpha_n^2 Q(\theta,\eta)\}d\theta} = o_p(1). \tag{56}$$

Proof. Denote the left hand side in (56) by $\boldsymbol{I}$. Then for any $\epsilon^*, \epsilon > 0$,

$$\mathbb{P}(\boldsymbol{I} > 2\epsilon^*) \leq \mathbb{P}\Big\{\sup_{\theta\in\Theta}\sup_{\eta^*\in N} |\boldsymbol{S_n}(\theta,\eta^*)| > \epsilon\Big\} + \mathbb{P}\Big\{\boldsymbol{I} > \epsilon^*, \sup_{\theta\in\Theta}\sup_{\eta^*\in N} |\boldsymbol{S_n}(\theta,\eta^*)| \leq \epsilon\Big\},$$

where the first right hand side term is $o(1)$ by assumptions F, G, and M, and the uniform law of large numbers (e.g. van der Vaart and Wellner (1996, theorem 2.4.3)). We will deal with the second right hand side term below.

For any $c > 0$ and $\eta \in N$, let $\bar{\Theta}(\eta, c) = \{\theta \in \Theta : \mathfrak{d}_H^*(\theta,\eta) \leq c\}$, which is compact, and let $\bar{\Theta}^c(\eta, c)$ be its complement relative to Θ. Let further $c_q = c_q(c) = -\sup_{\eta\in N}\sup_{\theta\in\bar{\Theta}^c(\eta,c)} Q(\theta,\eta) > 0$. Note that c_q is finite because $\bar{\Theta}^c(\eta,c) \subset \bar{\Theta}^*(\eta,c)$ and $\sup_{\theta\in\bar{\Theta}^*(\eta,c)} Q(\theta,\eta)$ is continuous in η by the maximum theorem and G.1. Also, c_q cannot be equal to zero since $\bar{\Theta}^c(\eta,c) \subset \bar{\Theta}^*(\eta, c/2)$, which is compact and continuous (as a correspondence) in η by G.1, such that for some $\eta^* \in N$ and some $\theta^* \in \bar{\Theta}^c(\eta^*, c)$, $Q(\theta^*, \eta^*) = 0$, which is at odds with the definition of $\bar{\Theta}^c(\eta, c)$. Now, by assumption H, for $\lambda^\dagger = \max_{\eta\in N}\max_{\theta\in\Theta}\lambda^+(\theta,\eta)$, some $\boldsymbol{C} < \infty$ and sufficiently large n,

$$\inf_{\eta\in N}\int_{\bar{\Theta}(\eta,c)} \pi(\theta)\exp\{-\alpha_n^2\epsilon + \alpha_n^2 Q(\theta,\eta)\}d\theta \geq$$
$$\exp(-\alpha_n^2\epsilon)\inf_{\eta\in N}\int_{\bar{\Theta}(\eta,c)} \pi(\theta)\exp\{-\alpha_n^2\lambda^\dagger\|\theta-\theta_0(\eta)\|^2\}d\theta > \exp(-2\alpha_n^2\epsilon).$$

Finally, if $\sup_{\theta\in\Theta}\sup_{\eta^*\in N}|\boldsymbol{S_n}(\theta,\eta^*)| \leq \epsilon$ then

$$\boldsymbol{I} \leq c + \sup_{\eta\in N}\frac{\int_{\bar{\Theta}^c(\eta,c)} \mathfrak{d}_H^*(\theta,\eta)\pi(\theta)\exp\{\alpha_n^2\epsilon + \alpha_n^2 Q(\theta,\eta)\}d\theta}{\int \pi(\theta)\exp\{-\alpha_n^2\epsilon + \alpha_n^2 Q(\theta,\eta)\}d\theta} \leq$$

$$c + \exp(3\alpha_n^2\epsilon) \sup_{\eta\in N} \int_{\bar{\Theta}^c(\eta,c)} \exp\{\alpha_n^2 Q(\theta,\eta)\} d\theta \leq c + C\exp\{\alpha_n^2(3\epsilon - c_q)\} = c + o(1),$$

provided that ϵ is chosen less than $c_q/3$. Finally, let $c \downarrow 0$. □

Lemma G.3. *Suppose that for any fixed $w \in \mathscr{W}$ and some sequence $\{\mathfrak{r}_n\}$,*

$$\mathfrak{r}_n\{\hat{\boldsymbol{\theta}}^*(w^*, w) - \theta_0\} \rightsquigarrow \mathbb{Z} + \mathscr{C}w, \tag{57}$$

in $\mathbb{L}^\infty(\mathscr{W})$, where $\mathbb{Z}$ does not depend on w^*, w. Then $\mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0) \xrightarrow{d} -\mathscr{C}^{-1}\mathbb{Z}$.

Proof. Note that $\hat{\boldsymbol{w}} = \mathfrak{r}_n(\hat{\boldsymbol{\eta}} - \eta_0)$ is for sufficiently large n with probability approaching one a solution to $\hat{\boldsymbol{\theta}}^*(\hat{\boldsymbol{w}}, \hat{\boldsymbol{w}}) = \theta_0$. Let $\hat{\boldsymbol{\psi}}(w^*, w) = \mathfrak{r}_n\{\hat{\boldsymbol{\theta}}^*(w^*, w) - \theta_0\}$. By Chebyshev's order inequality (Steele, 2004, problem 5.2) and assumption K, $\hat{\boldsymbol{\theta}}^*$ is increasing in every element of w. Hence, for any closed hypercube $\mathscr{W}_c \subset \mathscr{W}$ using element–wise inequalities and assumption M,

$$\mathbb{P}(\hat{\boldsymbol{w}} \in \mathscr{W}_c) = \mathbb{P}\{\exists w^* \in \mathscr{W}_c : \hat{\boldsymbol{\psi}}(w^*, w^*) = 0\} = \\ \mathbb{P}\Big\{\inf_{w^*\in\mathscr{W}_c} \hat{\boldsymbol{\psi}}(w^*, w^*) \leq 0 \leq \sup_{w^*\in\mathscr{W}_c} \hat{\boldsymbol{\psi}}(w^*, w^*)\Big\}. \tag{58}$$

From (58) it follows that if $\underline{w}, \bar{w}$ represent respectively the vectors of minima and maxima of $\mathscr{W}_c$ then

$$\mathbb{P}\Big\{\sup_{w^*\in\mathscr{W}} \hat{\boldsymbol{\psi}}(w^*, \underline{w}) \leq 0 \leq \inf_{w^*\in\mathscr{W}} \hat{\boldsymbol{\psi}}(w^*, \bar{w})\Big\} \leq \mathbb{P}(\hat{\boldsymbol{w}} \in \mathscr{W}_c) \leq \\ \mathbb{P}\Big\{\inf_{w^*\in\mathscr{W}} \hat{\boldsymbol{\psi}}(w^*, \underline{w}) \leq 0 \leq \sup_{w^*\in\mathscr{W}} \hat{\boldsymbol{\psi}}(w^*, \bar{w})\Big\}. \tag{59}$$

By (57) the limits of the majorant and minorant sides in (59) are equal and again by (59) hence

$$\lim_{n\to\infty} \mathbb{P}(\hat{\boldsymbol{w}} \in \mathscr{W}_c) = \lim_{n\to\infty} \mathbb{P}\{\hat{\boldsymbol{\psi}}(0, \underline{w}) \leq 0 \leq \hat{\boldsymbol{\psi}}(0, \bar{w})\} = \mathbb{P}(\mathbb{Z} \in -\mathscr{C}\mathscr{W}_c) = \mathbb{P}\big(-\mathscr{C}^{-1}\mathbb{Z} \in \mathscr{W}_c\big).$$

□

Proof of theorem 3. For part (i) note that $\mathfrak{d}_H\{\theta_0(\eta_0), \Theta_0(\eta)\} = 0$ has a unique solution at η_0 by assumption L. Since $\mathfrak{d}_H\{\theta_0(\eta_0), \Theta_0(\eta)\}$ is continuous in $\eta \in N$ by the maximum theorem and G.1, where N is compact, we know that for any $\epsilon > 0$ there is an $\epsilon^* > 0$ such that

$$\mathbb{P}\big(\|\hat{\boldsymbol{\eta}} - \eta_0\| \geq \epsilon\big) \leq \mathbb{P}\big[\mathfrak{d}_H\{\theta_0(\eta_0), \Theta_0(\hat{\boldsymbol{\eta}})\} \geq 2\epsilon^*\big] \leq \\ \mathbb{P}\big[\mathfrak{d}_H\{\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}), \Theta_0(\hat{\boldsymbol{\eta}})\} \geq \epsilon^*\big] + \mathbb{P}\{\|\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}) - \theta_0(\eta_0)\| \geq \epsilon^*\}. \tag{60}$$

The first right hand side term in (60) is $o(1)$ by G.2 and the second right hand side term can by definition be no greater than $\mathbb{P}\{\|\hat{\boldsymbol{\theta}}(\eta_0)-\theta_0(\eta_0)\|\geq\epsilon^*\}=o(1)$, again by G.2.

Now, observe that for any divergent sequence $\{\mathfrak{r}_n\}$ and $\hat{\boldsymbol{w}}=\mathfrak{r}_n(\hat{\boldsymbol{\eta}}-\eta_0)$ we have (with probability approaching one),

$$\mathfrak{r}_n\{\hat{\boldsymbol{\theta}}^*(w^*,w)-\theta_0\}=\mathfrak{r}_n\{\hat{\boldsymbol{\theta}}^*(w^*,w)-\theta_{0n}(w)\}+\mathfrak{r}_n\{\theta_0(\eta_0+w/\mathfrak{r}_n)-\theta_0(\eta_0)\}. \tag{61}$$

For (ii), suppose that the convergence rate $\mathfrak{r}_n$ of $\hat{\boldsymbol{\eta}}$ is slower than $\sqrt[3]{n}$. By G.3 it suffices to show that for any $w\in\mathscr{W}$, the second right hand side term in (61) converges to $\mathscr{C}w$ for any $w\in\mathscr{W}$, which was established in B.1, and that the first right hand side term converges to zero uniformly in $w^*\in\mathscr{W}$. For the second requirement, replace $\sqrt[3]{n},\tilde{\alpha}_n$ in appendix C by $\mathfrak{r}_n$ and replace $\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}$ by $\sqrt{\mathfrak{r}_n^3/n}\,\tilde{\boldsymbol{S}}_{\boldsymbol{nw}}$, such that $\mathbb{X}=0$ a.s., where $\mathbb{X}$ is as defined in C.1. Consequently, following the same steps as in (ii) of theorem 1 at the end of appendix C delivers the desired result.

Finally, the remaining results follow immediately from the corresponding parts of theorem 1 using $\mathfrak{r}_n=\sqrt[3]{n}$ for parts (iii) and (iv) and $\mathfrak{r}_n=\sqrt{n/\alpha_n}$ for (v). □

Appendix H. Theorem 4

In this section μ_V is the probability measure $N(0,V^{-1})$ on $(\mathbb{R}^d,\mathscr{B})$, where $\mathscr{B}$ is the Borel σ–algebra.

Lemma H.1. *For all sufficiently large n, $\limsup_{\|t\|\to\infty}|\hat{\mathbb{G}}(t)|/c_t=0$ with probability one.*

Proof. We will show that for any ϵ, $\mathbb{P}\{\limsup_{\|t\|\to\infty}|\hat{\mathbb{G}}(t)|/c_t>\epsilon\}=0$. By the Borel–Cantelli lemma, it suffices to show that

$$\sum_{s=1}^{\infty}\mathbb{P}\Big\{\sup_{s-1\leq\|t\|<s}|\hat{\mathbb{G}}(t)|/c_t>\epsilon\Big\}<\infty. \tag{62}$$

Repeating the latter part of the proof of B.5, the left hand side of (62) is by assumption P bounded above by

$$\begin{aligned}\sum_{s=1}^{\infty}\mathbb{P}\Big\{\sup_{\|t\|\leq1}|\hat{\mathbb{G}}(t)|>\epsilon\sqrt{s}\Big\}&\leq\frac{1}{\epsilon^4}\mathbb{E}\mathbb{E}\Big\{\sup_{\|t\|\leq1}\hat{\mathbb{G}}^{\boldsymbol{4}}(t)\big|\hat{\boldsymbol{H}}\Big\}\sum_{s=1}^{\infty}\frac{1}{s^2}\\&\leq\frac{C}{\epsilon^4}\mathbb{E}\Big[\mathbb{E}\Big\{\sup_{\|t\|\leq1}|\hat{\mathbb{G}}(t)|\big|\hat{\boldsymbol{H}}\Big\}\Big]^4,\end{aligned} \tag{63}$$

where C is a generic constant; the last inequality follows from van der Vaart and Wellner (1996, proposition A.2.4) and the fact that $\sum_{s=1}^{\infty}(1/s^2) < \infty$. Now, by van der Vaart and Wellner (1996, theorem 2.2.8), for some $0 < \tilde{C} < \infty$, $\mathbb{E}\{\sup_{\|t\|\leq 1} |\hat{\mathbb{G}}(t)| | \hat{\boldsymbol{H}}\} \leq \tilde{C} \sup_{\|t\|=1} \sqrt{\hat{\boldsymbol{H}}(t,t)}$, because $\hat{\mathbb{G}}$ given $\hat{\boldsymbol{H}}$ is sub–Gaussian with respect to the semimetric $d_{VW}^2(t,s) = \hat{\boldsymbol{H}}(t,t) + \hat{\boldsymbol{H}}(s,s) - 2\hat{\boldsymbol{H}}(t,s) = \hat{\boldsymbol{H}}(t-s,t-s) \leq \|t-s\| \sup_{\|t\|=1} \hat{\boldsymbol{H}}(t,t)$ by assumption P. Therefore, the right hand side of (63) is bounded by $\tilde{C}^4\mathbb{E}\{\sup_{\|t\|=1} \hat{\boldsymbol{H}}^{\mathbf{2}}(t,t)\}$, which is finite by assumption Q. □

Lemma H.2. $\hat{\mathbb{G}} \rightsquigarrow \mathbb{G}$ *in* $\mathbb{L}^2(\mathbb{R}^d, \mathscr{B}, \mu_V)$.

Proof. We use Cremers and Kadelka (1986, theorem 2). Note first that $\hat{\mathbb{G}}$ and $\mathbb{G}$ are contained in $\mathbb{L}^2(\mathbb{R}^d, \mathscr{B}, \mu_V)$ by H.1 and similar arguments using $\mathbb{E}\{\sup_{\|t\|\leq 1} |\mathbb{G}(t)|\} < \infty$. Further, convergence in distribution of $\hat{\mathbb{G}}$ to $\mathbb{G}$ for finite marginals holds by construction. So, by theorem 2 of Cremers and Kadelka, it suffices to show that for $\boldsymbol{T} \sim N(0, V^{-1})$ independent of $\hat{\mathbb{G}}$,

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{E}\big[|\hat{\mathbb{G}}(\boldsymbol{T})| \mathbb{1}\{|\hat{\mathbb{G}}(\boldsymbol{T})| > C\}\big] = 0.$$

For this purpose, note that

$$\limsup_{n\to\infty} \mathbb{E}\big[|\hat{\mathbb{G}}(\boldsymbol{T})| \mathbb{1}\{|\hat{\mathbb{G}}(\boldsymbol{T})| > C\}\big] \leq \limsup_{n\to\infty} \mathbb{E}\hat{\mathbb{G}}^{\mathbf{2}}(\boldsymbol{T})/C = \limsup_{n\to\infty} \mathbb{E}\hat{\boldsymbol{H}}(\boldsymbol{T},\boldsymbol{T})/C,$$

where $\limsup\{\mathbb{E}\hat{\boldsymbol{H}}(\boldsymbol{T},\boldsymbol{T})\}^2 \leq \mathbb{E}\|\boldsymbol{T}\|^2 \limsup \mathbb{E}\{\sup_{\|t\|=1} \hat{\boldsymbol{H}}^{\mathbf{2}}(t,t)\} < \infty$ by assumption Q. □

Lemma H.3. $\hat{\mathbb{G}} \rightsquigarrow \mathbb{G}$ *in* $\mathbb{L}^\infty(\mathscr{C})$*, where* $\mathscr{C}$ *is an arbitrary compact subset of* $\mathbb{R}^d$.

Proof. By construction, any finite marginals of $\hat{\mathbb{G}}$ converge in distribution to the corresponding finite marginals of $\mathbb{G}$. Therefore, it suffices to show the stochastic equicontinuity of $\hat{\mathbb{G}}$. For this purpose, we will show that for any sequence $\delta_n \downarrow 0$ and for $t, s \in \mathscr{C}$,

$$\sup_{\|t-s\|<\delta_n} \big|\hat{\mathbb{G}}(t) - \hat{\mathbb{G}}(s)\big| = o_p(1). \tag{64}$$

Let $\hat{\mathbb{H}}(t,s)$ be the covariance matrix of $\big(\hat{\mathbb{G}}(t)\ \hat{\mathbb{G}}(s)\big)^{\mathsf{T}}$. Then, letting $\boldsymbol{Z}$ be a two dimensional standard normal random vector, we have for $\chi = [1, -1]^{\mathsf{T}}$,

$$\Big\{\mathbb{E} \sup_{\|t-s\|<\delta_n} \big|\hat{\mathbb{G}}(t) - \hat{\mathbb{G}}(s)\big|\Big\}^2 = \Big\{\mathbb{E} \sup_{\|t-s\|<\delta_n} \big|\chi^{\mathsf{T}}\hat{\mathbb{H}}^{\mathbf{1/2}}(t,s)\boldsymbol{Z}\big|\Big\}^2$$
$$\leq \mathbb{E} \sup_{\|t-s\|<\delta_n} \chi^{\mathsf{T}}\hat{\mathbb{H}}(t,s)\chi = \mathbb{E} \sup_{\|t-s\|<\delta_n} \hat{\boldsymbol{H}}(t-s,t-s) \leq \delta_n \mathbb{E} \sup_{\|t\|=1} \hat{\boldsymbol{H}}(t,t). \tag{65}$$

where the last inequality and equality follow from assumption P. Then, it follows from assumption Q that the last expression in (65) is $O(\delta_n) = o(1)$. Now, use the Markov inequality to obtain (64). □

Lemma H.4. *For any polynomial P and for $j, \tilde{j} \in \{0, 1\}$,*

$$\int \|P(t)\| \hat{\mathbb{G}}^j(t) \exp\{\tilde{j}\hat{\mathbb{G}}(t)\}\phi_V(t)dt \xrightarrow{d} \int \|P(t)\| \mathbb{G}^j(t) \exp\{\tilde{j}\mathbb{G}(t)\}\phi_V(t)dt.$$

Proof. We only consider the case $j = \tilde{j} = 1$; the other cases are similar. Let $\mathscr{E} = \big\{\kappa \in \mathbb{L}^2(\mathbb{R}^d, \mathscr{B}, \mu_V) : \limsup_{\|t\|\to\infty} |\kappa(t)|/c_t = 0\big\}$. Since the functional $\mathfrak{M} : \mathscr{E} \to \mathbb{R}$ given by $\mathscr{M}(\kappa) = \int \|P(t)\|\kappa(t) \exp\big\{\kappa(t)\big\} d\mu_V(t)$ is continuous, the stated result then follows from H.1 and H.2 and the continuous mapping theorem (van der Vaart and Wellner, 1996, theorem 1.11.1). □

Proof of theorem 4. The first half of the theorem statement is a reformulation of parts of theorem 1 with implicit definition of $\mathbb{J}_{\boldsymbol{\alpha}}$ in each case, so we only need to establish that $\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}}$ has the same limiting distribution as the corresponding $\mathbb{J}_{\boldsymbol{\alpha}}(0)$ in all three cases.

First, suppose that $\sqrt[3]{n}/\alpha_n = o(1)$. Then by H.3 for $\hat{\mathbb{C}}(t) = \hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2$, $\hat{\mathbb{C}} \rightsquigarrow \mathbb{C}$ in $\mathbb{L}^\infty(\mathscr{C})$ for any compact subset $\mathscr{C}$ of $\mathbb{R}^d$. Then follow the same steps as in appendix C using $\hat{\mathbb{C}}$ in lieu of $\tilde{\boldsymbol{L}}_{\boldsymbol{nw}}$

Next, suppose that $\alpha_n = c_\alpha^2 \sqrt[3]{n}$. The maximum and minimum in the definition of $\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}}$ and the theorem statement have the effect of multiplying both sides by a constant, so suppose without loss of generality that $c_\alpha > 1$. Then carry out the substitution $t \leftarrow \beta_n^{2/3} t$ to obtain

$$\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}} = \int t \exp\big\{c_\alpha^3 \hat{\mathbb{G}}(t)\big\}\phi_{\hat{\boldsymbol{V}}}(t)dt \Big/ c_\alpha^2 \int \exp\big\{c_\alpha^3 \hat{\mathbb{G}}(t)\big\}\phi_{\hat{\boldsymbol{V}}}(t)dt.$$

Apply H.4 plus a minor correction for the fact that $\hat{\boldsymbol{V}}$ depends on data to obtain the limit distribution indicated in (19).

Finally suppose that $\alpha_n = o(\sqrt[3]{n})$. Use the substitution $t \leftarrow \beta_n^{2/3} t$ to obtain

$$\hat{\boldsymbol{\Psi}}_{\boldsymbol{n}} = \int t \exp\big\{\beta_n \hat{\mathbb{G}}(t)\big\}\phi_{\hat{\boldsymbol{V}}}(t)dt \Big/ \beta_n \int \exp\big\{\beta_n \hat{\mathbb{G}}(t)\big\}\phi_{\hat{\boldsymbol{V}}}(t)dt.$$

Now, for $j = 0, 1$ and some $c \in [0, 1]$,

$$\int \Big(\frac{t}{\beta_n}\Big)^j \exp\big\{\beta_n \hat{\mathbb{G}}(t)\big\}\phi_{\hat{\boldsymbol{V}}}(t)dt = \int \big\{t\hat{\mathbb{G}}(t)\big\}^j \phi_{\hat{\boldsymbol{V}}}(t)dt + \frac{\beta_n}{(j+1)!} \int t^j \hat{\mathbb{G}}^{\boldsymbol{j+1}}(t) \exp\big\{c\beta_n \hat{\mathbb{G}}(t)\big\}\phi_{\hat{\boldsymbol{V}}}(t)dt. \quad (66)$$

The second right hand side term in (66) is for some constant C with probability approaching one bounded in norm by $C\beta_n \int \|t\|^j |\hat{\mathbb{G}}(t)|^j \exp\{\hat{\mathbb{G}}(t) - \lambda_- \|t\|^2/4\} dt = o(1)O_p(1) = o_p(1)$ by H.1 and H.2. For $j = 0$ the first right hand side in (66) equals one and for $j = 1$ we have after a minor correction for the fact that $\hat{V}$ depends on data, using H.4 that $\int t\hat{\mathbb{G}}(t)\phi_{\hat{V}}(t)dt \xrightarrow{d} \int t\mathbb{G}(t)\phi_V(t)dt \sim N(0, \mathscr{V})$. □