Journal of Machine Learning Research 22 (2021) 1-53 Submitted 5/20; Published 7/21

Asymptotic Normality, Concentration, and Coverage of Generalized Posteriors

Jeffrey W. Miller JWMILLER@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

Editor: XuanLong Nguyen

Abstract

Generalized likelihoods are commonly used to obtain consistent estimators with attractive computational and robustness properties. Formally, any generalized likelihood can be used to define a generalized posterior distribution, but an arbitrarily defined "posterior" cannot be expected to appropriately quantify uncertainty in any meaningful sense. In this article, we provide sufficient conditions under which generalized posteriors exhibit concentration, asymptotic normality (Bernstein–von Mises), an asymptotically correct Laplace approximation, and asymptotically correct frequentist coverage. We apply our results in detail to generalized posteriors for a wide array of generalized likelihoods, including pseudolikelihoods in general, the Gaussian Markov random field pseudolikelihood, the fully observed Boltzmann machine pseudolikelihood, the Ising model pseudolikelihood, the Cox proportional hazards partial likelihood, and a median-based likelihood for robust inference of location. Further, we show how our results can be used to easily establish the asymptotics of standard posteriors for exponential families and generalized linear models. We make no assumption of model correctness so that our results apply with or without misspecification.

Keywords: Bayesian theory, consistency, misspecification, pseudolikelihood, robustness

1. Introduction

Many statistical estimation methods are based on maximizing a generalized likelihood function such as a pseudolikelihood, partial likelihood, or composite likelihood. Generalized likelihood functions are often advantageous in terms of computation or robustness while still having consistency guarantees, even though they do not necessarily correspond to the standard likelihood of a probabilistic model.

Formally, any generalized likelihood can be used to construct a generalized posterior proportional to the generalized likelihood times a prior. Generalized posteriors have been proposed based on a variety of generalized likelihoods, including composite likelihoods (Smith and Stephenson, 2009; Pauli et al., 2011; Ribatet et al., 2012; Friel, 2012), restricted likelihoods (Pettitt, 1983; Doksum and Lo, 1990; Hoff, 2007; Lewis et al., 2014), partial likelihoods (Raftery et al., 1996; Sinha et al., 2003; Kim and Kim, 2009; Ventura and Racugno, 2016), substitution likelihoods (Lavine, 1995; Dunson and Taylor, 2005), modular likelihoods (Liu et al., 2009; Jacob et al., 2017), quasi-likelihoods (Ventura et al., 2010), generalized method of moments likelihoods (Yin, 2009), loss-based likelihoods (Jiang and Tanner, 2008; Zhang, 2006; Bissiri et al., 2016), and more.

Generalized posteriors have been used in a range of applications, including spatial statistics (Ribatet et al., 2012), social networks (Friel, 2012), neural networks (Hyvärinen, 2006), protein modeling (Zhou and Schmidler, 2009), computer model emulators of physical processes (Liu et al., 2009), copula models (Hoff, 2007), survival analysis (Raftery et al., 1996), infrastructure networks (Bouranis et al., 2017), longitudinal studies (Yin, 2009), and survey sampling (Williams and Savitsky, 2018). Although various theoretical guarantees have been provided in various cases, bespoke theory has been needed to verify whether a given generalized posterior will be valid for statistical inference.

In this article, we provide new theoretical results on the asymptotic validity of generalized posteriors. We provide a range of sufficient conditions for concentration (Section 2), Bernstein–von Mises asymptotic normality and the Laplace approximation (Section 3), and asymptotic frequentist coverage of credible sets (Section 4) for generalized posteriors. For generalized posteriors derived from composite likelihoods—a large class covering essentially all the examples in this article—we informally discuss what can be expected in terms of consistency and coverage (Section 5). We show how our results can easily be applied to many standard posteriors, including i.i.d. exponential family models and (non-i.i.d.) generalized linear models for regression (Section 6). We then apply our results to generalized posteriors for an array of generalized likelihoods, including pseudolikelihoods in general, the Gaussian Markov random field pseudolikelihood, the fully observed Boltzmann machine pseudolikelihood, the Ising model pseudolikelihood, the Cox proportional hazards partial likelihood, and a median-based likelihood for robust inference of location (Section 7). Finally, we provide a discussion of previous work (Section 8).

1.1 Novelty and Overview of the Results

In some sense, new Bernstein–von Mises (BvM) theorems are never surprising since they only verify what we already expect to happen if things are sufficiently nice. Thus, the utility of a BvM result is directly related to the ease and generality with which it can be applied. The main novelty of this article is that we provide results that are not only general, but are also relatively easy to apply in practice.

More specifically, the results in this article are novel in the following respects: (a) we provide rigorous results on generalized posteriors for non-i.i.d. data without any assumption of model correctness (in fact, in our main results, we do not even require there to be a probability model—true or assumed), (b) we provide sufficient conditions that are relatively easy to verify when they hold, and (c) we apply our results to a number of non-trivial examples, providing precise and concrete sufficient conditions for each example.

Standard BvM theorems are only applicable to standard posteriors under correctly specified i.i.d. probabilistic models (Van der Vaart, 2000; Ghosh and Ramamoorthi, 2003). Kleijn and Van der Vaart (2012) generalize by establishing a Bernstein–von Mises theorem under misspecification, but their result still only applies to standard posteriors, and they focus mainly on the i.i.d. case. In contrast, our main results in Sections 2 and 3 do not involve a probability model at all and are applicable to arbitrary distributions of the form $\pi_n(\theta) \propto \exp(-n f_n(\theta))\pi(\theta)$, where π and the sequence of functions f_n are required to satisfy certain conditions. By treating the problem in this generality, we provide results for i.i.d. and non-i.i.d. cases with or without misspecification; see the examples in Sections 6 and 7.

Additionally, BvM theorems often only show that the total variation distance converges to zero in probability; in contrast, we prove it converges to zero almost surely.

For generalized posteriors, much of the previous work on asymptotic normality tends to rely on unspecified regularity conditions or only establishes weak convergence, that is, convergence in distribution (Doksum and Lo, 1990; Lazar, 2003; Greco et al., 2008; Pauli et al., 2011; Ribatet et al., 2012; Ventura and Racugno, 2016). In contrast, we show convergence in total variation distance and we provide rigorous results with all assumptions explicitly stated. Further, the usual regularity conditions in previous work include an assumption of concentration (Bernardo and Smith, 2000); in contrast, we prove concentration.

In general, we make no assumption of model correctness. However, to ensure that a generalized posterior is doing something reasonable, it is desirable to have a guarantee of consistency—that is, concentration at the true parameter—if the assumed model is correct or at least partially correct. To this end, in Section 5 we show that for any composite likelihood derived from a correct model, the resulting generalized posterior concentrates at the true parameter under fairly general conditions. Since many generalized likelihoods can be viewed as composite likelihoods, this establishes consistency in a wide range of cases. On the other hand, it is well-known that—except in special circumstances—the asymptotic frequentist coverage of composite likelihood-based posteriors is typically incorrect unless an adjustment is made (Pauli et al., 2011; Ribatet et al., 2012); see Section 5 for more details.

For each main result in Sections 2 and 3, we provide a range of alternative sufficient conditions, from more abstract to more concrete. The more abstract versions are more generally applicable, whereas the more concrete versions have conditions that are easier to verify when applicable. For instance, Theorem 4 is an abstract BvM theorem involving a quadratic representation condition; meanwhile, Theorem 5 is a more concrete BvM theorem involving conditions on derivatives that are roughly analogous to the conditions in classical BvM theorems. We also provide versions of the theorems based on convexity of f_n, see Theorems 3(3) and 5(2), which is usually easy to verify when it applies and simplifies the other required conditions.

See Section 8 for a detailed technical discussion of how our assumptions, results, and proof techniques compare with those in previous work.

2. Posterior Concentration

Theorem 2 is a general concentration result for generalized posteriors Π_n on a measurable space $(\Theta, \mathcal{A})$. The basic structure of the proof of Theorem 2 follows that of Schwartz's theorem (Schwartz, 1965; Ghosh and Ramamoorthi, 2003). Although Theorem 2 is useful for theoretical purposes, in practice, one typically needs to establish concentration on neighborhoods in a relevant topology on Θ. To this end, Theorem 3 provides a range of sufficient conditions for concentration on metric space neighborhoods of a point $\theta_0 \in \Theta$.

Condition 1 *Let $f_n : \Theta \to \mathbb{R}$ for $n \in \mathbb{N}$ be a sequence of functions on a probability space $(\Theta, \mathcal{A}, \Pi)$. For all n, assume $z_n < \infty$ where $z_n = \int_\Theta \exp(-n f_n(\theta))\Pi(d\theta)$, and define the probability measure*

$$\Pi_n(d\theta) = \exp(-n f_n(\theta))\Pi(d\theta)/z_n.$$

Throughout, all arbitrarily defined functions and sets are assumed to be measurable, and we denote $\mathbb{N} = \{1, 2, \ldots\}$. Here, $\exp(-nf_n(\theta))$ is interpreted as the "likelihood", possibly in some generalized sense, Π is the "prior", and Π_n is the "posterior".

Our main theorems in Sections 2 and 3 do not involve a probability model and do not even require that there be data. Instead, our results apply to arbitrary deterministic sequences of distributions Π_n satisfying certain conditions. Consequently, the mode of convergence in these theorems is not probabilistic in any sense. In the applications in Sections 6 and 7 that involve probability models, we show that the conditions hold with probability 1, and in this way we obtain almost sure convergence.

Theorem 2 *Assume Condition 1. If $\theta_0 \in \Theta$ and there exists $f : \Theta \to \mathbb{R}$ such that*

(1) *$f_n(\theta) \to f(\theta)$ as $n \to \infty$ for all $\theta \in \Theta$,*

(2) *$\Pi(A_\varepsilon) > 0$ for all $\varepsilon > 0$, where $A_\varepsilon = \{\theta \in \Theta : f(\theta) < f(\theta_0) + \varepsilon\}$, and*

(3) *$\liminf_n \inf_{\theta \in A_\varepsilon^c} f_n(\theta) > f(\theta_0)$ for all $\varepsilon > 0$,*

then $\Pi_n(A_\varepsilon) \to 1$ as $n \to \infty$, for any $\varepsilon > 0$.

See Section A for the proof. In Section 8, we discuss the interpretation of this result in relation to Schwartz's theorem. When Θ is a metric space, the collection of functions (f_n) is said to be *equicontinuous* if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $n \in \mathbb{N}$, $\theta, \theta' \in \Theta$, if $d(\theta, \theta') < \delta$ then $|f_n(\theta) - f_n(\theta')| < \varepsilon$. For a function $f : E \to \mathbb{R}$ where $E \subseteq \mathbb{R}^D$, we denote the gradient by $f'(\theta)$ (that is, $f'(\theta) = (\frac{\partial f}{\partial \theta_i}(\theta))_{i=1}^D \in \mathbb{R}^D$) and the Hessian by $f''(\theta)$ (that is, $f''(\theta) = (\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\theta))_{i,j=1}^D \in \mathbb{R}^{D \times D}$) when these derivatives exist. We use the following definition of convexity to allow the possibility that the domain $E \subseteq \mathbb{R}^D$ is non-convex: $f : E \to \mathbb{R}$ is *convex* if for all $\theta, \theta' \in E$ and all $t \in [0, 1]$ such that $t\theta + (1-t)\theta' \in E$, we have $f(t\theta + (1-t)\theta') \leq t f(\theta) + (1-t) f(\theta')$.

Theorem 3 *Assume Condition 1. Suppose (Θ, d) is a metric space and $\mathcal{A}$ is the resulting Borel sigma-algebra. Fix $\theta_0 \in \Theta$ and denote $N_\varepsilon = \{\theta \in \Theta : d(\theta, \theta_0) < \varepsilon\}$. If $\Pi(N_\varepsilon) > 0$ for all $\varepsilon > 0$, $f_n \to f$ pointwise on Θ for some $f : \Theta \to \mathbb{R}$, and any one of the following three sets of assumptions hold, then for any $\varepsilon > 0$, $\Pi_n(N_\varepsilon) \to 1$ as $n \to \infty$.*

(1) *f is continuous at θ_0 and $\liminf_n \inf_{\theta \in N_\varepsilon^c} f_n(\theta) > f(\theta_0)$ for all $\varepsilon > 0$.*

(2) *(f_n) is equicontinuous on some compact set $K \subseteq \Theta$, θ_0 is an interior point of K, $f(\theta) > f(\theta_0)$ for all $\theta \in K \setminus \{\theta_0\}$, and $\liminf_n \inf_{\theta \in K^c} f_n(\theta) > f(\theta_0)$.*

(3) *f_n is convex for each n, $\Theta \subseteq \mathbb{R}^D$ with the Euclidean metric, θ_0 is an interior point of Θ, and either*

   (a) *$f(\theta) > f(\theta_0)$ for all $\theta \in \Theta \setminus \{\theta_0\}$, or*

   (b) *f' exists in a neighborhood of θ_0, $f'(\theta_0) = 0$, and $f''(\theta_0)$ exists and is positive definite.*

Further, 2 ⇒1 and 3 ⇒1 under the assumptions of the theorem.

See Section A for the proof. Note that if Θ is compact, then case 2 with $K = \Theta$ simplifies to (f_n) being equicontinuous and $f(\theta) > f(\theta_0)$ for all $\theta \in \Theta \setminus \{\theta_0\}$. This can be used to prove consistency results based on classical Wald-type conditions such as in Ghosh and Ramamoorthi (2003, Section 1.3.4).

3. Asymptotic Normality and the Laplace Approximation

Theorem 4 establishes general sufficient conditions under which a generalized posterior exhibits asymptotic normality and an asymptotically correct Laplace approximation, along with concentration at θ_0. As in Section 2, $\pi(\theta)$ can be interpreted as the prior density and $\pi_n(\theta) \propto \exp(-n f_n(\theta))\pi(\theta)$ can be thought of as the "posterior" density. The points θ_n can be viewed as maximum generalized likelihood estimates. The proof of Theorem 4 is concise, but some of the conditions of the theorem are a bit abstract. Thus, we also provide Theorem 5 to give more concrete sufficient conditions which, when satisfied, are usually easier to verify. Theorem 5 is the main result used in the examples in the rest of the paper.

We remphasize that unlike previous work on BvM, the results in this section only involve conditions on f_n and π, and do not involve any assumptions at all regarding the data; indeed, we do not even require that there be any data. Thus, the limits in these theorems are not probabilistic in any sense—they are simply limits of deterministic sequences. When we apply the theorems to statistical models in Sections 6 and 7, we handle the randomness in the data by showing that the conditions of the theorems hold almost surely, which implies almost sure convergence.

We also highlight two supporting results that are employed in the proof of Theorem 5. Theorem 6 provides concrete sufficient conditions under which the quadratic representation (assumption 1) in Theorem 4 holds. Theorem 7 is a pure real analysis result on uniform convergence of f_n, f_n', and f_n'', which we believe is interesting in its own right.

Given $x_0 \in \mathbb{R}^D$ and $r > 0$, we write $B_r(x_0)$ to denote the open ball of radius r at x_0, that is, $B_r(x_0) = \{x \in \mathbb{R}^D : |x - x_0| < r\}$. We use $|\cdot|$ to denote the Euclidean norm. Given positive sequences (a_n) and (b_n), we write $a_n \sim b_n$ to denote that $a_n/b_n \to 1$ as $n \to \infty$. We write $\mathcal{N}(x \mid \mu, C)$ to denote the normal density with mean μ and covariance matrix C.

Theorem 4 *Fix $\theta_0 \in \mathbb{R}^D$ and let $\pi : \mathbb{R}^D \to \mathbb{R}$ be a probability density with respect to Lebesgue measure such that π is continuous at θ_0 and $\pi(\theta_0) > 0$. Let $f_n : \mathbb{R}^D \to \mathbb{R}$ for $n \in \mathbb{N}$ and assume:*

(1) *f_n can be represented as*

$$f_n(\theta) = f_n(\theta_n) + \tfrac{1}{2}(\theta - \theta_n)^{\mathsf{T}} H_n (\theta - \theta_n) + r_n(\theta - \theta_n) \tag{3.1}$$

where $\theta_n \in \mathbb{R}^D$ such that $\theta_n \to \theta_0$, $H_n \in \mathbb{R}^{D\times D}$ symmetric such that $H_n \to H_0$ for some positive definite H_0, and $r_n : \mathbb{R}^D \to \mathbb{R}$ has the following property: there exist $\varepsilon_0, c_0 > 0$ such that for all n sufficiently large, for all $x \in B_{\varepsilon_0}(0)$, we have $|r_n(x)| \le c_0 |x|^3$; and

(2) *for any $\varepsilon > 0$, $\liminf_n \inf_{\theta \in B_\varepsilon(\theta_n)^c} \big(f_n(\theta) - f_n(\theta_n)\big) > 0$.*

Then, defining $z_n = \int_{\mathbb{R}^D} \exp(-nf_n(\theta))\pi(\theta)d\theta$ *and* $\pi_n(\theta) = \exp(-nf_n(\theta))\pi(\theta)/z_n$*, we have*

$$\int_{B_\varepsilon(\theta_0)} \pi_n(\theta)d\theta \xrightarrow[n\to\infty]{} 1 \text{ for all } \varepsilon > 0, \tag{3.2}$$

that is, π_n *concentrates at* θ_0*,*

$$z_n \sim \frac{\exp(-nf_n(\theta_n))\pi(\theta_0)}{|\det H_0|^{1/2}} \Big(\frac{2\pi}{n}\Big)^{D/2} \tag{3.3}$$

as $n \to \infty$ *(Laplace approximation), and letting* q_n *be the density of* $\sqrt{n}(\theta - \theta_n)$ *when* $\theta \sim \pi_n$*,*

$$\int_{\mathbb{R}^D} \Big| q_n(x) - \mathcal{N}(x \mid 0, H_0^{-1}) \Big| dx \xrightarrow[n\to\infty]{} 0, \tag{3.4}$$

that is, q_n *converges to* $\mathcal{N}(0, H_0^{-1})$ *in total variation.*

See Section B for the proof. The virtue of Theorem 4 is not its technical depth—indeed, it is fairly straightforward to prove using the generalized dominated convergence theorem. Rather, the utility of this result is that it is formulated in such a way that it can be broadly applied to generalized posteriors. In Section 8, we compare Theorem 4 to previous Bernstein–von Mises results.

Roughly speaking, assumption 1 of Theorem 4 is that $f_n(\theta)$ can be approximated by a quadratic form in a neighborhood of θ_n. This is similar to a second-order Taylor expansion where the constants in the bound on the remainder (namely, ε_0 and c_0) need to work for all n sufficiently large; however, unlike in Taylor's theorem, differentiability of f_n is not assumed. Since $\theta_n \to \theta_0$, the idea of assumption 1 is that the log posterior density approaches a quadratic form near θ_0. The assumption that H_0 is positive definite is necessary to ensure that in the limit, the exponentiated quadratic form can be normalized to a probability density, namely, $\mathcal{N}(x \mid 0, H_0^{-1})$. Note that in the special case of a correctly specified i.i.d. probability model, H_0 typically coincides with the Fisher information matrix at θ_0.

Assumption 2 of Theorem 4 ensures that, asymptotically, the posterior puts negligible mass outside a neighborhood of θ_0, and thus, the locally normal part near θ_0 is all that remains in the limit. Assumption 2 is stronger than necessary, but it is not clear how to adapt the usual probabilistic separation conditions (such as uniformly consistent tests) to generalized posteriors, especially since we seek almost sure convergence.

Throughout, we use the Euclidean–Frobenius norms on vectors $v \in \mathbb{R}^D$, matrices $M \in \mathbb{R}^{D\times D}$, and tensors $T \in \mathbb{R}^{D^3}$, that is, $|v| = (\sum_i v_i^2)^{1/2}$, $\|M\| = (\sum_{i,j} M_{ij}^2)^{1/2}$, and $\|T\| = (\sum_{i,j,k} T_{ijk}^2)^{1/2}$. Convergence and boundedness for vectors, matrices, and tensors is defined with respect to these norms. A collection of functions $h_n : E \to F$, where F is a normed space, is *uniformly bounded* if the set $\{\|h_n(x)\| : x \in E, n \in \mathbb{N}\}$ is bounded, and is *pointwise bounded* if $\{\|h_n(x)\| : n \in \mathbb{N}\}$ is bounded for each $x \in E$. Let $f'''(\theta)$ denote the tensor of third derivatives, that is, $f'''(\theta) = (\frac{\partial^3 f}{\partial\theta_i \partial\theta_j \partial\theta_k}(\theta))_{i,j,k=1}^D \in \mathbb{R}^{D^3}$.

Theorem 5 *Let* $\Theta \subseteq \mathbb{R}^D$*. Let* $E \subseteq \Theta$ *be open (in* $\mathbb{R}^D$*) and bounded. Fix* $\theta_0 \in E$ *and let* $\pi : \Theta \to \mathbb{R}$ *be a probability density with respect to Lebesgue measure such that* π *is continuous at* θ_0 *and* $\pi(\theta_0) > 0$*. Let* $f_n : \Theta \to \mathbb{R}$ *have continuous third derivatives on* E*. Suppose* $f_n \to f$ *pointwise for some* $f : \Theta \to \mathbb{R}$*,* $f''(\theta_0)$ *is positive definite, and* (f_n''') *is uniformly bounded on* E*. If either of the following two assumptions is satisfied:*

(1) *$f(\theta) > f(\theta_0)$ for all $\theta \in K \setminus \{\theta_0\}$ and* $\liminf_n \inf_{\theta\in\Theta\setminus K} f_n(\theta) > f(\theta_0)$ *for some compact $K \subseteq E$ with θ_0 in the interior of K, or*

(2) *each f_n is convex and $f'(\theta_0) = 0$,*

then there is a sequence $\theta_n \to \theta_0$ such that $f_n'(\theta_n) = 0$ for all n sufficiently large, $f_n(\theta_n) \to f(\theta_0)$, Equation 3.2 (concentration at θ_0) holds, Equation 3.3 (Laplace approximation) holds, and Equation 3.4 (asymptotic normality) holds, where $H_0 = f''(\theta_0)$. Further, 2 ⇒1 under the assumptions of the theorem.

See Section B for the proof. While Theorem 4 is more general, Theorem 5 provides conditions that are easier to verify when applicable. The set E simply serves as a neighborhood of θ_0 on which f_n is well-behaved. The assumption that $f''(\theta_0)$ is positive definite ensures that f is locally convex at θ_0, but not necessarily globally convex. See Section 8 for comparison with previous work. The following result is used in the proof of Theorem 5.

Theorem 6 *Let $E \subseteq \mathbb{R}^D$ be open and convex, and let $\theta_0 \in E$. Let $f_n : E \to \mathbb{R}$ have continuous third derivatives, and assume:*

(1) *there exist $\theta_n \in E$ such that $\theta_n \to \theta_0$ and $f_n'(\theta_n) = 0$ for all n sufficiently large,*

(2) *$f_n''(\theta_0) \to H_0$ as $n \to \infty$ for some positive definite H_0, and*

(3) *(f_n''') is uniformly bounded.*

Letting $H_n = f_n''(\theta_n)$, assumption 1 of Theorem 4 is satisfied for all n sufficiently large.

See Section B for the proof. The main tool used in the proof of Theorem 5 is the following result, which provides somewhat more than we require. A collection of functions $h_n : E \to F$, where E and F are subsets of normed spaces, is *equi-Lipschitz* if there exists $c > 0$ such that for all $n \in \mathbb{N}$, $x, y \in E$, we have $\|h_n(x) - h_n(y)\| \leq c\|x - y\|$.

Theorem 7 (Regular convergence) *Let $E \subseteq \mathbb{R}^D$ be open, convex, and bounded. For $n \in \mathbb{N}$, let $f_n : E \to \mathbb{R}$ have continuous third derivatives, and suppose (f_n''') is uniformly bounded. If (f_n) is pointwise bounded, then (f_n), (f_n'), and (f_n'') are all equi-Lipschitz and uniformly bounded. If $f_n \to f$ pointwise for some $f : E \to \mathbb{R}$, then f' and f'' exist, $f_n \to f$ uniformly, $f_n' \to f'$ uniformly, and $f_n'' \to f''$ uniformly.*

Note that if $f_n \to f$ pointwise then (f_n) is pointwise bounded; thus, if $f_n \to f$ pointwise then we also get the equi-Lipschitz and uniform bounded result. See Section C for proof.

4. Coverage

For a generalized posterior to provide useful quantification of uncertainty, it is important that it be reasonably well-calibrated in terms of frequentist coverage. Ideally, we would like Π_n to have correct frequentist coverage in the sense that posterior credible sets of probability ρ have frequentist coverage ρ. Obviously, an arbitrarily chosen generalized posterior cannot be expected to have correct coverage. Thus, in Theorem 8, we provide simple conditions under which a generalized posterior has correct frequentist coverage, asymptotically.

Unfortunately, it seems that, like misspecified models, many common choices of generalized posterior do not exhibit correct coverage, even asymptotically. In Section 5.2, we discuss why this problem occurs in the context of composite likelihood-based posteriors, which is a very general class that includes nearly all of the examples in this paper. Due to this, we do not apply our main coverage result (Theorem 8) to the examples in Sections 6 and 7 for the simple reason that we do not expect it to hold, except under correct specification or in special circumstances. Nonetheless, we present the theorem here in order to help find those special circumstances when they do arise, and to provide a foundation for future work that may generalize upon this result. For instance, having correct coverage for each univariate component of the parameter, marginally, rather than jointly, is a less stringent property that would still be very useful in practice. Alternatively, one could aim for conservative coverage, which would be more achievable as well.

To interpret Theorem 8, we think of θ_n as a maximum generalized likelihood estimate, θ_0 as the "true" parameter we want to cover, Π_n as the generalized posterior distribution, S_n as a credible set of asymptotic probability ρ, Q_n as a centered and scaled version of Π_n, and R_n as a centered and scaled version of S_n. Roughly, the theorem says that if Q_n converges in total variation to the asymptotic distribution of $-\sqrt{n}(\theta_n - \theta_0)$, and R_n converges pointwise, then asymptotically, S_n contains the true parameter 100ρ percent of the time. In other words, if the conditions of the theorem hold, then asymptotically, Π_n has correct frequentist coverage in the sense that posterior credible sets of probability ρ have frequentist coverage ρ.

Typically, when things work out nicely, θ_n is $\sqrt{n}$-consistent and asymptotically normal and a BvM result holds for Π_n, in which case the result says that Π_n has correct coverage asymptotically if the covariance matrices of these two normal distributions are equal. In other words, if $\sqrt{n}(\theta_n - \theta_0) \xrightarrow{\mathrm{D}} \mathcal{N}(0, C_1)$ and $Q_n \to \mathcal{N}(0, C_2)$ in total variation distance, then Π_n has correct asymptotic frequentist coverage if $C_1 = C_2$. In this case, the only other condition is that R_n converges to a set R with finite nonzero Lebesgue measure, because it is guaranteed that $Q(\partial R) = 0$. (Note that if $X \sim \mathcal{N}(0, C_1)$ then $-X \sim \mathcal{N}(0, C_1)$ also.) This result is precisely what one would expect; thus, the purpose of the theorem is to make this rigorous under easy-to-verify conditions.

We give $\mathbb{R}^D$ the Euclidean topology and the resulting Borel sigma-algebra, $\mathcal{B}$, and we use $m(\cdot)$ to denote Lebesgue measure on $\mathbb{R}^D$. We write ∂R to denote the boundary of a set $R \in \mathbb{R}^D$, that is, $\partial R = \bar{R} \setminus R^\circ$, where $\bar{R}$ is the closure and R° is the interior of R. Given $R, R_1, R_2, \ldots \subseteq \mathbb{R}^D$, we write $R_n \to R$ to denote that for all $x \in \mathbb{R}^D$, $\mathbb{1}(x \in R_n) \to \mathbb{1}(x \in R)$ as $n \to \infty$. Define $d(x, A) = \inf_{y \in A} \|x - y\|$ for $x \in \mathbb{R}^D$ and $A \subseteq \mathbb{R}^D$.

Theorem 8 *Let $\theta_1, \theta_2, \ldots \in \mathbb{R}^D$ be a sequence of random vectors, and let $\theta_0 \in \mathbb{R}^D$ be fixed. Let $\Pi_1, \Pi_2, \ldots$ be a sequence of random probability measures on $\mathbb{R}^D$, possibly dependent on $\theta_1, \theta_2, \ldots$. Let $S_1, S_2, \ldots \subseteq \mathbb{R}^D$ be a sequence of random convex measurable sets such that $\Pi_n(S_n) \xrightarrow{\text{a.s.}} \rho$ for some fixed $\rho \in (0, 1)$. For $A \in \mathcal{B}$, define $Q_n(A) = \int \mathbb{1}\big(\sqrt{n}(\theta - \theta_n) \in A\big)\Pi_n(d\theta)$ and define $R_n = \{\sqrt{n}(\theta - \theta_n) : \theta \in S_n\}$. Suppose there is a fixed probability measure Q and a fixed set $R \subseteq \mathbb{R}^D$ such that*

(1) $-\sqrt{n}(\theta_n - \theta_0) \xrightarrow{\mathrm{D}} Q$ *as $n \to \infty$ (where $\xrightarrow{\mathrm{D}}$ denotes convergence in distribution),*

(2) $\sup_{A \in \mathcal{B}} |Q_n(A) - Q(A)| \xrightarrow{\text{a.s.}} 0$ *as $n \to \infty$ (that is, $Q_n \xrightarrow{\text{a.s.}} Q$ in total variation),*

(3) $R_n \xrightarrow{\text{a.s.}} R$ *as* $n \to \infty$ *(that is, a.s., for all* $x \in \mathbb{R}^D$, $\mathbb{1}(x \in R_n) \to \mathbb{1}(x \in R)$*), and*

(4) $Q(\partial R) = 0$ *and* $0 < m(R) < \infty$*, where* m *denotes Lebesgue measure on* $\mathbb{R}^D$.

Then $\mathbb{P}(\theta_0 \in S_n) \to \rho$ *as* $n \to \infty$.

See Section D for the proof. If Q has a density with respect to Lebesgue measure, then the condition that $Q(\partial R) = 0$ is automatically satisfied, since the assumptions imply that R is convex and thus $m(\partial R) = 0$. In Theorem 8, the assumption that the confidence sets S_n are convex is not essential. The only place that convexity is used is to ensure that the conclusion of Lemma 10 holds. Indeed, Theorem 8 still holds if $S_1, S_2, \ldots$ are not assumed to be convex and assumption 3 is replaced by the conclusion of Lemma 10 (that is, for any $\varepsilon > 0$, if $A = \{x \in \mathbb{R}^D : d(x, R^c) > \varepsilon\}$ and $B = \{x \in \mathbb{R}^D : d(x, R) \leq \varepsilon\}$ then for all n sufficiently large, $A \subseteq R_n \subseteq B$). We chose to state the theorem in this way because credible sets are often convex by construction, and pointwise convergence of R_n to R is considerably easier to verify than the conclusion of Lemma 10. The following lemmas are used in the proof of Theorem 8, and may be useful in their own right. See Section D for their proofs.

Lemma 9 *Let* $X_1, X_2, \ldots \in \mathbb{R}^D$ *be random vectors such that* $X_n \xrightarrow{\text{D}} X$ *for some random vector* X*. Let* $R_1, R_2, \ldots \subseteq \mathbb{R}^D$ *be random convex measurable sets, possibly dependent on* $X_1, X_2, \ldots$*. Assume there exists some fixed* $R \subseteq \mathbb{R}^D$ *with* $0 < m(R) < \infty$ *and* $\mathbb{P}(X \in \partial R) = 0$ *such that* $R_n \to R$ *almost surely as* $n \to \infty$*. Then* $\mathbb{P}(X_n \in R_n) \to \mathbb{P}(X \in R)$ *as* $n \to \infty$.

The probability $\mathbb{P}(X_n \in R_n)$ should be interpreted as $\int \mathbb{1}(X_n(\omega) \in R_n(\omega)) P(d\omega)$, that is, X_n and R_n are jointly integrated over and $\mathbb{P}(X_n \in R_n)$ is a non-random quantity.

Lemma 10 *Let* $R_1, R_2, \ldots \subseteq \mathbb{R}^D$ *be convex sets. Assume* $R_n \to R$ *for some* $R \subseteq \mathbb{R}^D$ *with* $0 < m(R) < \infty$*. For any* $\varepsilon > 0$*, if* $A = \{x \in \mathbb{R}^D : d(x, R^c) > \varepsilon\}$ *and* $B = \{x \in \mathbb{R}^D : d(x, R) \leq \varepsilon\}$ *then for all* n *sufficiently large,* $A \subseteq R_n \subseteq B$.

5. Composite Likelihood-based Posteriors

Composite likelihoods (CLs) (Lindsay, 1988) represent a large class of generalized likelihoods that encompasses essentially all of the examples in Sections 6 and 7. The theory of maximum composite likelihood estimation is well-established (Lindsay, 1988; Molenbergs and Verbeke, 2005; Varin et al., 2011). Theoretical results for CL-based generalized posteriors have been provided (Pauli et al., 2011; Ribatet et al., 2012; Ventura and Racugno, 2016; Greco et al., 2008; Lazar, 2003), subject to the caveats discussed in the introduction. The purpose of this section is to discuss how these previous results on CL-based generalized posteriors, or CL-posteriors for short, can be strengthened using our results in Sections 2 to 4. Roughly speaking, CL-posteriors derived from a correctly specified model can generally be expected to be consistent, but not necessarily correctly calibrated with respect to frequentist coverage.

Let y denote the full data set, which may take any form such as a sequence, a graph, a database, or any other data structure. Suppose $\{P_\theta : \theta \in \Theta\}$ is an assumed model for the distribution of y given θ, where $\Theta \subseteq \mathbb{R}^D$. For $j = 1, \ldots, k$, suppose $s_j(y)$ and $t_j(y)$ are functions of the data and, when $Y \sim P_\theta$, suppose the conditional distribution of $s_j(Y)$

given $t_j(Y)$ has density $p_\theta(s_j|t_j)$ with respect to a common dominating measure λ_j for all values of θ and t_j. Define the *composite likelihood* (Lindsay, 1988),

$$\mathcal{L}^{\mathrm{CL}}(\theta) = \prod_{j=1}^{k} p_\theta(s_j|t_j).$$

A few examples are given here and in Section 7; see Varin et al. (2011) for more examples.

Example 1 (i.i.d. likelihood) *If $y = (y_1, \ldots, y_n)$, $s_j(y) = y_j$, and $t_j(y) = 0$, then $\mathcal{L}^{\mathrm{CL}}(\theta) = \prod_{j=1}^{n} p_\theta(y_j)$ is simply the likelihood for an i.i.d. model.*

Example 2 (pseudolikelihood) *If $y = (y_1, \ldots, y_n)$, $s_j(y) = y_j$, and $t_j(y) = y_{-j} := (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n)$, then $\mathcal{L}^{\mathrm{CL}}(\theta) = \prod_{j=1}^{n} p_\theta(y_j|y_{-j})$ is a pseudolikelihood (Besag, 1975).*

Example 3 (restricted likelihood) *If $k = 1$, $t_1(y) = 0$, and $s_1(y)$ is an insufficient statistic, then $\mathcal{L}^{\mathrm{CL}}(\theta)$ is a restricted likelihood (Lewis et al., 2014). For instance, if $s_1(y)$ consists of ranks or selected quantiles, then $\mathcal{L}^{\mathrm{CL}}(\theta)$ is a rank likelihood (Pettitt, 1983; Hoff, 2007) or a quantile-based likelihood (Doksum and Lo, 1990), respectively.*

Due to the special structure of composite likelihoods, one can make some general observations about CL-posteriors of the form $\pi_n(\theta) \propto \mathcal{L}^{\mathrm{CL}}(\theta)\pi(\theta)$. First, a reassuring property is that if the model is correctly specified, then CL-posteriors are consistent under fairly general conditions; we discuss this next.

5.1 Consistency of CL-posteriors under Correct Specification

Throughout this article, we make no assumption of model correctness in the main results (Sections 2 to 4) or the applications (Sections 6 and 7). However, for interpretability, it is important to have a guarantee of consistency if the assumed model is correct or at least partially correct. Here, we show that in many cases of interest, if the model is correctly specified—or at least, if the conditional densities $p_\theta(s_j|t_j)$ are correctly specified—then the CL-posterior concentrates at the true parameter. The analogue of this result for maximum CL estimators is well-known (Lindsay, 1988; Varin et al., 2011); also see Pauli et al. (2011) and Ribatet et al. (2012).

First, observe that if $Y \sim P_{\theta_0}$, $S_j = s_j(Y)$, and $T_j = t_j(Y)$, then for all $\theta \in \Theta$,

$$\mathrm{E}\big(\log p_{\theta_0}(S_j|T_j)\big) \geq \mathrm{E}\big(\log p_\theta(S_j|T_j)\big) \tag{5.1}$$

because the conditional relative entropy $\mathrm{E}\big(\log(p_{\theta_0}(S_j|T_j)/p_\theta(S_j|T_j))\big)$ is nonnegative; this is referred to as the information inequality by Lindsay (1988). Now, suppose that for each $n \in \{1, 2, \ldots\}$, we have a data set Y^n, model $\{P_\theta^n : \theta \in \Theta\}$ (where Θ does not depend on n), and functions s_j^n, t_j^n for $j = 1, \ldots, k_n$. Further, suppose the assumed model is correct, such that $Y^n \sim P_{\theta_0}^n$ where the true parameter θ_0 is shared across all n. Define

$$f_n(\theta) = -\frac{1}{n}\log \mathcal{L}_n^{\mathrm{CL}}(\theta) = -\frac{1}{n}\sum_{j=1}^{k_n} \log p_\theta^n(S_j^n|T_j^n)$$

and $\pi_n(\theta) \propto \exp(-n f_n(\theta))\pi(\theta) = \mathcal{L}_n^{\mathrm{CL}}(\theta)\pi(\theta)$ where $S_j^n = s_j^n(Y^n)$ and $T_j^n = t_j^n(Y^n)$. In many cases of interest (see Sections 6 and 7), we have that with probability 1, for all $\theta \in \Theta$, $\lim_{n\to\infty} f_n(\theta) = f(\theta)$ where $f(\theta) = \lim_{n\to\infty} \mathrm{E}\big(f_n(\theta)\big)$. Then, by Equation 5.1, $f(\theta_0) \leq f(\theta)$ for all $\theta \in \Theta$, in other words, θ_0 is a minimizer of f. Further, in many cases, f has a unique minimizer, and π_n concentrates at the unique minimizer; in particular, this holds if the conditions of Theorem 3 or Theorem 5 are met. Therefore, in such cases, the CL-posterior π_n concentrates at the true parameter, θ_0.

5.2 Coverage of CL-posteriors under Correct Specification

Although CL-posteriors have appealing consistency properties, they do not generally have correct asymptotic frequentist coverage, except in special circumstances (Pauli et al., 2011; Ribatet et al., 2012). Continuing in the notation of Section 5.1, suppose $Y^n \sim P_{\theta_0}^n$, let $\pi_n(\theta) \propto \exp(-n f_n(\theta))\pi(\theta) = \mathcal{L}_n^{\mathrm{CL}}(\theta)\pi(\theta)$ be the CL-posterior, and let $\theta_n = \arg\max_\theta \mathcal{L}_n^{\mathrm{CL}}(\theta) = \arg\min_\theta f_n(\theta)$ be the maximum composite likelihood estimator. If Theorem 5 applies with probability 1, then $Q_n \xrightarrow{\text{a.s.}} \mathcal{N}(0, H_0^{-1})$ in total variation distance, where $H_0 = f''(\theta_0)$ and Q_n is the distribution of $\sqrt{n}(\theta - \theta_n)$ when $\theta \sim \pi_n$. This strengthens previous BvM results for CL-posteriors by showing almost sure convergence (rather than convergence in probability) with respect to total variation distance (rather than in the weak topology).

To use Theorem 8, we also need to know the asymptotic distribution of θ_n. The asymptotics of θ_n are well-known (Lindsay, 1988; Varin et al., 2011), but for completeness we provide an informal derivation (see below). Define $G_j^n = \nabla_\theta\big|_{\theta=\theta_0} \log p_\theta^n(S_j^n | T_j^n)$. It turns out that $-\sqrt{n}(\theta_n - \theta_0) \approx \mathcal{N}(0, A_n^{-1} J_n A_n^{-1})$ under regularity conditions, where $A_n = \frac{1}{n}\sum_{j=1}^{k_n} \mathrm{Cov}(G_j^n)$ and $J_n = \frac{1}{n}\mathrm{Cov}\big(\sum_{j=1}^{k_n} G_j^n\big)$. Typically, $A_n \to H_0$ and $J_n \to J_0$ for some J_0, so that

$$-\sqrt{n}(\theta_n - \theta_0) \xrightarrow{\mathrm{D}} \mathcal{N}(0, H_0^{-1} J_0 H_0^{-1}).$$

Hence, under typical conditions, the asymptotic distribution of $-\sqrt{n}(\theta_n - \theta_0)$ and the limit of Q_n are the same if and only if $H_0 = J_0$. Therefore, under these conditions, if $H_0 = J_0$ then the CL-posterior π_n has correct asymptotic frequentist coverage by Theorem 8. For instance, if for each n, $G_1^n, \ldots, G_{k_n}^n$ are pairwise uncorrelated, then $A_n = J_n$ and hence $H_0 = J_0$. However, in many cases of interest, $H_0 \neq J_0$ and the CL-posterior needs to be affinely transformed to have correct coverage (Ribatet et al., 2012; Pauli et al., 2011; Friel, 2012; Stoehr and Friel, 2015); also see Williams and Savitsky (2018) for a similar technique in survey sampling.

For completeness, here we provide a rough sketch of the derivation of the asymptotic distribution of θ_n; see Lindsay (1988) and Varin et al. (2011). By a first-order Taylor approximation applied to each entry of $f_n'(\theta) \in \mathbb{R}^D$, when θ_n is near θ_0 we have $0 = f_n'(\theta_n) \approx f_n'(\theta_0) + f_n''(\theta_0)(\theta_n - \theta_0)$, and thus, $-\sqrt{n}(\theta_n - \theta_0) \approx f_n''(\theta_0)^{-1}(\sqrt{n} f_n'(\theta_0))$, assuming $f_n''(\theta_0) \in \mathbb{R}^{D\times D}$ exists and is invertible and the error terms are negligible. When n is large, we typically have $f_n''(\theta_0) \approx \mathrm{E} f_n''(\theta_0)$ (for instance, due to a law of large numbers result), and thus, $f_n''(\theta_0) \approx \mathrm{E} f_n''(\theta_0) = \frac{1}{n}\sum_{j=1}^{k_n} \mathrm{E}(G_j^n G_j^{n\mathtt{T}}) = \frac{1}{n}\sum_{j=1}^{k_n} \mathrm{Cov}(G_j^n) = A_n$ since $\mathrm{E}(G_j^n) = 0$ and $\mathrm{E}\big(\nabla_\theta^2\big|_{\theta=\theta_0} \log p_\theta^n(S_j^n | T_j^n)\big) = -\mathrm{E}(G_j^n G_j^{n\mathtt{T}})$, as long as we can interchange the order of integrals and derivatives. Further, assuming a central limit theorem holds, $\sqrt{n} f_n'(\theta_0) = -\frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} G_j^n \approx \mathcal{N}(0, J_n)$ where $J_n = \frac{1}{n}\mathrm{Cov}\big(\sum_{j=1}^{k_n} G_j^n\big)$. Thus, under appropriate conditions, $-\sqrt{n}(\theta_n - \theta_0) \approx \mathcal{N}(0, A_n^{-1} J_n A_n^{-1})$.

6. Applications to Standard Posteriors

In this section, we illustrate how our results can be used to easily prove posterior concentration, the Laplace approximation, and asymptotic normality for standard models such as exponential families, linear regression, and generalized linear models including logistic regression and Poisson regression. We do not assume that the model is correctly specified; thus, this section can be compared to the misspecified setting of Kleijn and Van der Vaart (2012). Even in these standard models, our results go beyond the existing theory of Kleijn and Van der Vaart (2012) by showing almost sure convergence and employing conditions that we believe are easier to verify; see Section 8.5 for a detailed comparison. Further, these "toy" examples serve to illustrate our general results in familiar settings, enabling one to compare our assumptions with commonly used assumptions for these models.

6.1 Exponential Families

Consider an exponential family with density $q(y|\eta) = \exp(\eta^{\mathtt{T}} s(y) - \kappa(\eta))$ with respect to a sigma-finite Borel measure λ on $\mathcal{Y} \subseteq \mathbb{R}^d$ where $s : \mathcal{Y} \to \mathbb{R}^k$, $\eta \in \mathcal{E} \subseteq \mathbb{R}^k$, and $\kappa(\eta) = \log \int_{\mathcal{Y}} \exp(\eta^{\mathtt{T}} s(y))\lambda(dy)$. Any exponential family on $\mathbb{R}^d$ can be put in this form by choosing λ appropriately and possibly reparametrizing to η. Let $Q_\eta(E) = \int_E q(y|\eta)\lambda(dy)$ and denote $\mathrm{E}_\eta s(Y) = \int_{\mathcal{Y}} s(y) Q_\eta(dy)$. For any $m \in \mathbb{N}$, we give $\mathbb{R}^m$ the Euclidean metric and the resulting Borel sigma-algebra unless otherwise specified.

Condition 11 *Assume $q(y|\eta)$ is of the form above, $\mathcal{E} = \{\eta \in \mathbb{R}^k : |\kappa(\eta)| < \infty\}$, $\mathcal{E}$ is open, $\mathcal{E}$ is nonempty, and $\eta \mapsto Q_\eta$ is one-to-one (that is, η is identifiable).*

Theorem 12 (Exponential families) *Consider a family $q(y|\eta)$ satisfying Condition 11. Suppose $Y_1, Y_2, \ldots \in \mathcal{Y}$ are i.i.d. random vectors such that $\mathrm{E}s(Y_i) = \mathrm{E}_{\theta_0} s(Y)$ for some $\theta_0 \in \Theta := \mathcal{E}$. Then for any open ball E such that $\theta_0 \in E$ and $\bar{E} \subseteq \Theta$, $f_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log q(Y_i|\theta)$ satisfies the conditions of Theorem 5 with probability* 1.

Condition 11 is that the exponential family is full, regular, nonempty, identifiable, and in natural form; these are standard conditions that hold for most commonly used exponential families (Hoffman-Jørgensen, 1994; Miller and Harrison, 2014). Recall that the maximum likelihood estimate (MLE) is obtained by matching the expected sufficient statistics to the observed sufficient statistics. Thus, the assumption that $\mathrm{E}s(Y_i) = \mathrm{E}_{\theta_0} s(Y)$ for some θ_0 is simply assuming that this moment matching is possible, asymptotically. In many cases, this holds automatically since the moment space $\mathcal{M} := \{\mathrm{E}_\theta s(Y) : \theta \in \Theta\}$ is often equal to the full set of possible values of $\mathrm{E}s(Y_i)$, due to the fact that $\mathcal{M}$ is convex (e.g., Miller and Harrison, 2014, Prop. 19). Thus, while exceptions can occur, the result holds very generally.

Proof of Theorem 12 Note that $f_n(\theta) = \kappa(\theta) - \theta^{\mathtt{T}} S_n$ where $S_n = \frac{1}{n}\sum_{i=1}^n s(Y_i)$. By standard exponential family theory (e.g., Miller and Harrison, 2014, Prop. 19), κ is C^∞ (that is, κ has continuous derivatives of all order), κ is convex on Θ, $\kappa'(\theta) = \mathrm{E}_\theta s(Y)$, and $\kappa''(\theta)$ is symmetric positive definite for all $\theta \in \Theta$. Let $s_0 = \mathrm{E}s(Y_i)$. Since $s_0 = \mathrm{E}s(Y_i) = \mathrm{E}_{\theta_0} s(Y) = \kappa'(\theta_0)$

and $\kappa'(\theta_0)$ is finite (because κ is C^∞), $S_n \to s_0$ with probability 1 by the strong law of large numbers. Thus, letting $f(\theta) = \kappa(\theta) - \theta^{\mathtt{T}} s_0$, we have that with probability 1, for all $\theta \in \Theta$, $f_n(\theta) = \kappa(\theta) - \theta^{\mathtt{T}} S_n \to \kappa(\theta) - \theta^{\mathtt{T}} s_0 = f(\theta)$. (Note that due to the almost sure convergence of the sufficient statistics, we not only have that for all θ, with probability 1, $f_n(\theta) \to f(\theta)$, but we have the stronger consequence that with probability 1, for all θ, $f_n(\theta) \to f(\theta)$, which is needed for Theorem 5 to apply.) Let E be an open ball such that $\theta_0 \in E$ and $\bar{E} \subseteq \Theta$. Then $\kappa'''(\theta)$ is bounded on $\bar{E}$, since $\kappa'''(\theta)$ is continuous and $\bar{E}$ is compact. Hence, (f_n''') is uniformly bounded on E because $f_n'''(\theta) = \kappa'''(\theta)$. Therefore, with probability 1, $f_n \to f$ pointwise, f_n is convex and has continuous third derivatives on Θ, $f'(\theta_0) = \kappa'(\theta_0) - s_0 = 0$, $f''(\theta_0) = \kappa''(\theta_0)$ is positive definite, and (f_n''') is uniformly bounded on E. ■

6.2 Generalized Linear Models (GLMs)

First, we state a general theorem for GLMs, then we show how it applies to commonly used GLMs. Consider a regression model of the form $p(y_i \mid \theta, x_i) \propto_\theta q(y_i \mid \theta^{\mathtt{T}} x_i)$ for covariates $x_i \in \mathcal{X} \subseteq \mathbb{R}^D$ and coefficients $\theta \in \Theta \subseteq \mathbb{R}^D$, where $q(y|\eta) = \exp(\eta s(y) - \kappa(\eta))$ is a one-parameter exponential family satisfying Condition 11. Note that the proportionality is with respect to θ, not y_i. Assume Θ is open, Θ is convex, and $\theta^{\mathtt{T}} x \in \mathcal{E}$ for all $\theta \in \Theta$, $x \in \mathcal{X}$.

Theorem 13 (GLMs) *Suppose* $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathcal{X} \times \mathcal{Y}$ *i.i.d. such that:*

(1) $f'(\theta_0) = 0$ *for some* $\theta_0 \in \Theta$, *where* $f(\theta) = -\mathrm{E} \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$,

(2) $\mathrm{E}|X_i s(Y_i)| < \infty$ *and* $\mathrm{E}|\kappa(\theta^{\mathtt{T}} X_i)| < \infty$ *for all* $\theta \in \Theta$,

(3) *for all* $a \in \mathbb{R}^D$, *if* $a^{\mathtt{T}} X_i \stackrel{\text{a.s.}}{=} 0$ *then* $a = 0$, *and*

(4) *there is an open ball* $E \subseteq \mathbb{R}^D$ *such that* $\theta_0 \in E$, $\bar{E} \subseteq \Theta$, *and for all* $j, k, \ell \in \{1, \ldots, D\}$, $\mathrm{E}\big(\sup_{\theta \in \bar{E}} |\kappa'''(\theta^{\mathtt{T}} X_i) X_{ij} X_{ik} X_{i\ell}|\big) < \infty$.

Then for any open ball E satisfying assumption 4, $f_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$ *satisfies the conditions of Theorem 5 with probability* 1.

Assumption 1 of Theorem 13 is essentially that the MLE exists, asymptotically. Assumption 3 is that the support of the covariate vector X_i is not contained in any proper subspace of $\mathbb{R}^D$; this is necessary to ensure identifiability. When $\mathrm{E} X_i X_i^{\mathtt{T}}$ exists and is finite, assumption 3 is equivalent to the assumption that $\mathrm{E} X_i X_i^{\mathtt{T}}$ is non-singular, which is commonly used to ensure identifiability in GLMs (Van der Vaart, 2000, Example 16.8); alternatively, it is sometimes assumed that $\frac{1}{n}\sum_{i=1}^n X_i X_i^{\mathtt{T}}$ is non-singular for all n sufficiently large (Fahrmeir and Kaufmann, 1985). Assumptions 2 and 4 are moment conditions that are fairly easy to work with; for instance, if the covariates are bounded and $\mathrm{E}s(Y_i)$ exists, then assumptions 2 and 4 are always satisfied since κ is C^∞ smooth.

For comparison, traditional theorems on the asymptotic normality of the MLE in a GLM typically assume the model is correctly specified (whereas we do not), and they assume conditions on the observed Fisher information $n f_n''(\theta)$, such as divergence of the smallest eigenvalue of $n f_n''(\theta_0)$ and bounds on the variability of $f_n''(\theta)$ near θ_0 as $n \to \infty$ (Fahrmeir and Kaufmann, 1985). These Fisher information conditions are more closely analogous to assumption 1 of Theorem 4, which is implied by our result in Theorem 13. On the other hand, we show almost sure convergence of the posterior in TV distance, whereas Fahrmeir and Kaufmann (1985) only show convergence in distribution of the MLE.

Proof of Theorem 13 For all $\theta \in \Theta$, $f_n(\theta) = \frac{1}{n}\sum_{i=1}^n \kappa(\theta^\mathtt{T} X_i) - \theta^\mathtt{T} S_n$ where $S_n = \frac{1}{n}\sum_{i=1}^n X_i s(Y_i)$. Thus, $f_n(\theta)$ is C^∞ on Θ by the chain rule, since $\kappa(\eta)$ is C^∞ on $\mathcal{E}$. Further, $f_n(\theta)$ is convex since $\kappa(\eta)$ is convex. Noting that

$$f(\theta) = -\mathrm{E}\log q(Y_i \mid \theta^\mathtt{T} X_i) = \mathrm{E}(\kappa(\theta^\mathtt{T} X_i)) - \theta^\mathtt{T}\mathrm{E}(X_i s(Y_i)),$$

the assumed moment conditions (2) ensure that for all $\theta \in \Theta$, with probability 1, $f_n(\theta) \to f(\theta)$. This implies that with probability 1, for all $\theta \in \Theta$, $f_n(\theta) \to f(\theta)$, by the following argument. For any countable set $C \subseteq \Theta$, we have that with probability 1, for all $\theta \in C$, $f_n(\theta) \to f(\theta)$. Hence, letting C be a countable dense subset of Θ, and using the fact that each f_n is convex, we have that with probability 1, the limit $\tilde{f}(\theta) := \lim_n f_n(\theta)$ exists and is finite for all $\theta \in \Theta$ and $\tilde{f}$ is convex (Rockafellar, 1970, Theorem 10.8). Since f is also convex, then $\tilde{f}$ and f are continuous functions (Rockafellar, 1970, Theorem 10.1) that agree on a dense subset, so they are equal.

Choose E according to assumption 4. We show that with probability 1, (f_n''') is uniformly bounded on E. Fix $j, k, \ell \in \{1, \ldots, D\}$, and define $T(\theta, x) = \kappa'''(\theta^\mathtt{T} x) x_j x_k x_\ell$ for $\theta \in \Theta$, $x \in \mathcal{X}$. For all $x \in \mathcal{X}$, $\theta \mapsto T(\theta, x)$ is continuous, and for all $\theta \in \Theta$, $x \mapsto T(\theta, x)$ is measurable. Since $f_n'''(\theta)_{jk\ell} = \frac{1}{n}\sum_{i=1}^n T(\theta, X_i)$, assumption 4 implies that with probability 1, $(f_n'''(\theta)_{jk\ell})$ is uniformly bounded on $\bar{E}$, by the uniform law of large numbers (Ghosh and Ramamoorthi, 2003, Theorem 1.3.3). Letting $C_{jk\ell}(X_1, X_2, \ldots)$ be such a uniform bound for each j, k, ℓ, we have that with probability 1, for all $n \in \mathbb{N}$, $\theta \in \bar{E}$, $\|f_n'''(\theta)\|^2 = \sum_{j,k,\ell} f_n'''(\theta)_{jk\ell}^2 \leq \sum_{j,k,\ell} C_{jk\ell}(X_1, X_2, \ldots)^2 < \infty$. Thus, (f_n''') is a.s. uniformly bounded on $\bar{E}$, and hence on E.

By Theorem 7, $f''(\theta_0) \overset{\text{a.s.}}{=} \lim_{n\to\infty} f_n''(\theta_0) = \lim \frac{1}{n}\sum_{i=1}^n \kappa''(\theta_0^\mathtt{T} X_i) X_i X_i^\mathtt{T}$. Since this limit exists and is finite almost surely, then by the strong law of large numbers, the limit must be equal to the expectation (Kallenberg, 2002, Theorem 4.23), that is, $f''(\theta_0) = \mathrm{E}\big(\kappa''(\theta_0^\mathtt{T} X_i) X_i X_i^\mathtt{T}\big)$. Thus, $f''(\theta_0)$ is positive definite, since for all nonzero $a \in \mathbb{R}^D$, $a^\mathtt{T} f''(\theta_0) a = \mathrm{E}\big(\kappa''(\theta_0^\mathtt{T} X_i) a^\mathtt{T} X_i X_i^\mathtt{T} a\big) > 0$, by the fact that $\kappa''(\eta) > 0$ for all $\eta \in \mathcal{E}$ and by assumption 3, $a^\mathtt{T} X_i X_i^\mathtt{T} a = |a^\mathtt{T} X_i|^2$ is strictly positive with positive probability. ■

6.2.1 Linear Regression

The linear regression model is $p(y_i \mid \theta, x_i) = \mathcal{N}(y_i \mid \theta^\mathtt{T} x_i, \sigma^2)$ for $y_i \in \mathcal{Y} := \mathbb{R}$, $x_i \in \mathcal{X} := \mathbb{R}^D$, and $\theta \in \Theta := \mathbb{R}^D$. The model can equivalently be written as $p(y_i \mid \theta, x_i) \propto_\theta q(y_i \mid \theta^\mathtt{T} x_i)$ where $q(y|\eta) := \exp(\eta s(y) - \kappa(\eta))$ is a density with respect to $\lambda(dy) = \mathcal{N}(y \mid 0, \sigma^2)dy$ for $y \in \mathcal{Y}$ and $\eta \in \mathcal{E} := \mathbb{R}$, by defining $s(y) = y/\sigma^2$ and $\kappa(\eta) = \eta^2/(2\sigma^2)$.

Theorem 14 (Linear regression) *Suppose $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathcal{X}\times\mathcal{Y}$ i.i.d. such that:*

(1) $\mathrm{E}|X_i Y_i| < \infty$, $\mathrm{E}\|X_i X_i^\mathtt{T}\| < \infty$, *and*

(2) *for all $a \in \mathbb{R}^D$, if $a^\mathtt{T} X_i \overset{\text{a.s.}}{=} 0$ then $a = 0$.*

Then $\theta_0 := (\mathrm{E}X_iX_i^\mathtt{T})^{-1}\mathrm{E}X_iY_i$ *is well-defined and for any open ball* E *such that* $\theta_0 \in E$, $f_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log q(Y_i \mid \theta^\mathtt{T} X_i)$ *satisfies the conditions of Theorem 5 with probability* 1.

Assumption 1 is necessary to ensure that $\theta_0 := (\mathrm{E}X_iX_i^\mathtt{T})^{-1}\mathrm{E}X_iY_i$ is well-defined and assumption 2 is necessary to ensure identifiability, as in the case of Theorem 13. Since $\kappa(\eta) = \eta^2/(2\sigma^2)$, we have $f_n''(\theta) = \frac{1}{n}\sum_{i=1}^n X_iX_i^\mathtt{T}/\sigma^2$. Thus, for the traditional MLE conditions of Fahrmeir and Kaufmann (1985), it is not necessary to bound the variability of $f_n''(\theta)$, since $f_n''(\theta)$ does not depend on θ in the linear regression model. Hence, these traditional MLE conditions reduce to assuming divergence of the smallest eigenvalue of $\sum_{i=1}^n X_iX_i^\mathtt{T}$, which can be shown to be equivalent to assumption 2 of Theorem 14 when $X_1, X_2, \ldots$ are i.i.d.

Proof of Theorem 14 For any random vector $Z \in \mathbb{R}^k$, $\mathrm{E}|Z| < \infty$ if and only if $\mathrm{E}Z$ exists and is finite; likewise for matrices and tensors. Thus, $\mathrm{E}X_iY_i$ and $\mathrm{E}X_iX_i^\mathtt{T}$ exist and are finite. Further, $\mathrm{E}X_iX_i^\mathtt{T}$ is positive definite (and hence, invertible) since for all nonzero $a \in \mathbb{R}^D$, $a^\mathtt{T}(\mathrm{E}X_iX_i^\mathtt{T})a = \mathrm{E}|a^\mathtt{T}X_i|^2 > 0$. Condition 11 is easily checked: $\mathcal{E} = \{\eta \in \mathbb{R} : |\kappa(\eta)| < \infty\}$ since $\eta^2/(2\sigma^2) < \infty$ for all $\eta \in \mathbb{R}$, $\mathcal{E}$ is open and nonempty, and the mean of a normal distribution is identifiable. The GLM conditions are also straightforward to verify. Θ is open and convex, and $\theta^\mathtt{T}x \in \mathcal{E}$ for all $\theta \in \Theta$, $x \in \mathcal{X}$. Assumption 3 of Theorem 13 is satisfied by assumption, and assumption 4 of Theorem 13 is satisfied trivially since $\kappa'''(\eta) = 0$ for all $\eta \in \mathcal{E}$. Assumption 1 of Theorem 14 implies that assumption 2 of Theorem 13 holds, since $\mathrm{E}|X_i s(Y_i)| = \mathrm{E}|X_iY_i|/\sigma^2 < \infty$ and $\mathrm{E}|\kappa(\theta^\mathtt{T}X_i)| = \theta^\mathtt{T}(\mathrm{E}X_iX_i^\mathtt{T})\theta/(2\sigma^2) < \infty$. It is straightforward to verify that assumption 1 of Theorem 13 holds with $\theta_0 = (\mathrm{E}X_iX_i^\mathtt{T})^{-1}\mathrm{E}X_iY_i$. ■

6.2.2 Logistic Regression

The logistic regression model is $p(y_i \mid \theta, x_i) = \mathrm{Bernoulli}(y_i \mid \sigma(\theta^\mathtt{T}x_i))$ for $y_i \in \mathcal{Y} := \{0,1\}$, $x_i \in \mathcal{X} := \mathbb{R}^D$, and $\theta \in \Theta := \mathbb{R}^D$, where $\sigma(\eta) = 1/(1+e^{-\eta})$ for $\eta \in \mathcal{E} := \mathbb{R}$. Thus, $p(y_i \mid \theta, x_i) = q(y_i \mid \theta^\mathtt{T}x_i)$ where $q(y|\eta) := \exp(\eta y - \kappa(\eta))$ is a density with respect to $\lambda = \delta_0 + \delta_1$ for $y \in \mathcal{Y}$ and $\eta \in \mathcal{E}$, by defining $\kappa(\eta) = \log(1+e^\eta)$. Here, δ_y denotes the unit point mass at y.

Theorem 15 (Logistic regression) *Suppose* $(X_1,Y_1),(X_2,Y_2),\ldots \in \mathcal{X}\times\mathcal{Y}$ *i.i.d. such that:*

(1) $f'(\theta_0) = 0$ *for some* $\theta_0 \in \Theta$, *where* $f(\theta) = -\mathrm{E}\log q(Y_i \mid \theta^\mathtt{T}X_i)$,

(2) $\mathrm{E}|X_{ij}X_{ik}X_{i\ell}| < \infty$ *for all* $j,k,\ell \in \{1,\ldots,D\}$, *and*

(3) *for all* $a \in \mathbb{R}^D$, *if* $a^\mathtt{T}X_i \stackrel{\text{a.s.}}{=} 0$ *then* $a = 0$.

Then for any open ball $E \subseteq \Theta$ *such that* $\theta_0 \in E$, $f_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log q(Y_i \mid \theta^\mathtt{T}X_i)$ *satisfies the conditions of Theorem 5 with probability* 1.

Assumption 1 is essentially that the MLE exists, asymptotically, and assumption 3 is necessary for identifiability (see the remarks following Theorem 13); these are both very mild. Assumption 2 is a third-moment condition on the covariates, which we use to bound $f_n'''(\theta)$; this is more stringent, but is reasonable in many practical applications.

Proof Condition 11 is easily checked: $\mathcal{E} = \{\eta \in \mathbb{R} : |\kappa(\eta)| < \infty\}$, $\mathcal{E}$ is open and nonempty, and η is identifiable since $\sigma(\eta)$ is one-to-one. Trivially, Θ is open and convex, and $\theta^{\mathtt{T}} x \in \mathcal{E}$ for all $\theta \in \Theta$, $x \in \mathcal{X}$. Assumptions 1 and 3 of Theorem 13 are satisfied by assumptions 1 and 3 of Theorem 15, respectively. Assumption 4 of Theorem 13 is satisfied due to assumption 2 and the fact that $|\kappa'''(\eta)| \leq 3$ for all $\eta \in \mathcal{E}$, because $\kappa''' = \sigma(1-\sigma)(1-2\sigma)^2 - 2\sigma^2(1-\sigma)^2$ and $0 < \sigma(\eta) < 1$. Assumption 2 also implies that $\mathrm{E}|X_i| < \infty$, because $|X_i| \leq \sum_j |X_{ij}|$ and $\mathrm{E}|X_{ij}| < \infty$ for all j (Folland, 2013, 6.12). It follows that assumption 2 of Theorem 13 holds, since $\mathrm{E}|X_i Y_i| \leq \mathrm{E}|X_i| < \infty$ and $\mathrm{E}|\kappa(\theta^{\mathtt{T}} X_i)| \leq \log 2 + \mathrm{E}|\theta^{\mathtt{T}} X_i| \leq \log 2 + |\theta| \mathrm{E}|X_i| < \infty$, where we have used the inequality $|\kappa(\eta)| = \log(1+e^\eta) \leq \log 2 + |\eta|$ for $\eta \in \mathbb{R}$. ∎

6.2.3 Poisson Regression

The Poisson regression model is $p(y_i \mid \theta, x_i) = \mathrm{Poisson}(y_i \mid \exp(\theta^{\mathtt{T}} x_i))$ for $y_i \in \mathcal{Y} := \{0, 1, 2, \ldots\}$, $x_i \in \mathcal{X} := \mathbb{R}^D$, and $\theta \in \Theta := \mathbb{R}^D$. Thus, $p(y_i \mid \theta, x_i) \propto_\theta q(y_i \mid \theta^{\mathtt{T}} x_i)$ where $q(y|\eta) := \exp(\eta y - \kappa(\eta))$ is a density with respect to $\lambda := \sum_{y \in \mathcal{Y}} \delta_y / y!$ for $y \in \mathcal{Y}$ and $\eta \in \mathcal{E} := \mathbb{R}$, by defining $\kappa(\eta) = e^\eta$.

Theorem 16 (Poisson regression) *Suppose* $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathcal{X} \times \mathcal{Y}$ *i.i.d. such that:*

(1) $f'(\theta_0) = 0$ *for some* $\theta_0 \in \Theta$, *where* $f(\theta) = -\mathrm{E} \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$,

(2) $\mathrm{E}|X_i Y_i| < \infty$ *and* $\mathrm{E} \exp(c|X_i|) < \infty$ *for all* $c > 0$, *and*

(3) *for all* $a \in \mathbb{R}^D$, *if* $a^{\mathtt{T}} X_i \stackrel{\text{a.s.}}{=} 0$ *then* $a = 0$.

Then for any open ball $E \subseteq \Theta$ *such that* $\theta_0 \in E$, $f_n(\theta) := -\frac{1}{n} \sum_{i=1}^n \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$ *satisfies the conditions of Theorem 5 with probability* 1.

Proof As before, Condition 11 is easily checked: $\mathcal{E} = \{\eta \in \mathbb{R} : |\kappa(\eta)| < \infty\}$, $\mathcal{E}$ is open and nonempty, and η is identifiable. Trivially, Θ is open and convex, and $\theta^{\mathtt{T}} x \in \mathcal{E}$ for all $\theta \in \Theta$, $x \in \mathcal{X}$. Assumptions 1 and 3 of Theorem 13 are satisfied by assumptions 1 and 3 of Theorem 16. Assumption 2 of Theorem 13 is satisfied due to assumption 2 of Theorem 16, since for all $\theta \in \Theta$, $\mathrm{E}|\kappa(\theta^{\mathtt{T}} X_i)| = \mathrm{E} \exp(\theta^{\mathtt{T}} X_i) \leq \mathrm{E} \exp(|\theta||X_i|) < \infty$. For all $m \in \mathbb{N}$ and $j \in \{1, \ldots, D\}$, $\mathrm{E}|X_{ij}|^m \leq \mathrm{E}|X_i|^m = m! \mathrm{E}(|X_i|^m / m!) \leq m! \mathrm{E} \exp(|X_i|) < \infty$. Further, letting $r > 0$, $c = |\theta_0| + r$, and $E = B_r(\theta_0)$, we have that for all $\theta \in \bar{E}$, $\kappa'''(\theta^{\mathtt{T}} X_i) = \exp(\theta^{\mathtt{T}} X_i) \leq \exp(c|X_i|)$. Hence,

$$\mathrm{E}\big(\sup_{\theta \in \bar{E}} |\kappa'''(\theta^{\mathtt{T}} X_i) X_{ij} X_{ik} X_{i\ell}|\big) \leq \mathrm{E}\big(e^{c|X_i|} |X_{ij} X_{ik} X_{i\ell}|\big) \leq \Big(\mathrm{E} e^{4c|X_i|} \mathrm{E}|X_{ij}|^4 \mathrm{E}|X_{ik}|^4 \mathrm{E}|X_{i\ell}|^4\Big)^{1/4}$$

by Hölder's inequality (Folland, 2013, 6.2); thus, assumption 4 of Theorem 13 holds. ∎

7. Applications to Generalized Posteriors

In this section, we apply our results to a number of generalized posteriors of interest.

7.1 Pseudolikelihood-based Posteriors

Pseudolikelihood (Besag, 1975) is a powerful approach for many models in which the likelihood is difficult to compute due to intractability of the normalization constant. Instead of the standard likelihood $p(y_1, \ldots, y_n \mid \theta)$, the basic idea is to use a *pseudolikelihood* $\mathcal{L}(\theta) = \prod_{i=1}^n p(y_i \mid y_{-i}, \theta)$ where $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. Maximum pseudolikelihood estimates are used in many applications and have been shown to be consistent and asymptotically normal in a range of cases (Besag, 1975; Geman and Graffigne, 1986; Gidas, 1988; Comets, 1992; Jensen and Künsch, 1994; Mase, 1995; Liang and Yu, 2003; Hyvärinen, 2006). Usage of pseudolikelihoods for constructing generalized posteriors is much less common, perhaps due to concerns about the validity of the resulting posterior (but see Zhou and Schmidler, 2009; Bouranis et al., 2017; Pauli et al., 2011; Rydén and Titterington, 1998).

In this section, we provide sufficient conditions for concentration, asymptotic normality, and the Laplace approximation for a large class of pseudolikelihood-based posteriors. Specifically, we consider pseudolikelihoods in which each factor takes the form of a generalized linear model. We provide a general result for pseudolikelihoods in this class, and then consider three cases in particular: Gaussian Markov random fields (Section 7.2), fully visible Boltzmann machines (Section 7.3), and the Ising model on $\mathbb{Z}^m$ (Section 7.4). Any pseudolikelihood is a composite likelihood, so as discussed in Section 5, if the model is correct then we can expect consistency but not necessarily correct frequentist coverage.

Condition 17 *Suppose the data can be arranged in a sequence $y_1, y_2, \ldots \in \mathcal{Y} \subseteq \mathbb{R}^d$ and consider a pseudolikelihood of the form:*

$$\mathcal{L}_n^{\text{pseudo}}(\theta) \propto \prod_{i=1}^n q\big(y_i \mid \theta^{\mathtt{T}} \varphi_i(\vec{y})\big)$$

for $\theta \in \Theta \subseteq \mathbb{R}^D$, where $\varphi_i(\vec{y}) \in \mathcal{X} \subseteq \mathbb{R}^D$ is a function of $\vec{y} = (y_1, y_2, \ldots)$ and $q(y|\eta) = \exp(\eta s(y) - \kappa(\eta))$ is a one-parameter exponential family satisfying Condition 11 for $y \in \mathcal{Y}$, $\eta \in \mathcal{E}$. Assume Θ is open and convex, and $\theta^{\mathtt{T}} x \in \mathcal{E}$ for all $\theta \in \Theta$, $x \in \mathcal{X}$.

Theorem 18 *Assume the setup in Condition 17. Let $\vec{Y} = (Y_1, Y_2, \ldots)$ be a sequence of random vectors in $\mathcal{Y}$ and define $X_i = \varphi_i(\vec{Y})$. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ are identically distributed, but not necessarily independent. Define $f_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$ and $f(\theta) = -\mathrm{E} \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$ for $\theta \in \Theta$. Assume:*

(1) *for all $\theta \in \Theta$, $f(\theta)$ is finite and $f_n(\theta) \xrightarrow{\text{a.s.}} f(\theta)$ as $n \to \infty$,*

(2) *there exists $\theta_0 \in \Theta$ such that $f'(\theta_0) = 0$ and $f''(\theta_0) = \mathrm{E}\big(\kappa''(\theta_0^{\mathtt{T}} X_i) X_i X_i^{\mathtt{T}}\big)$,*

(3) *for all $a \in \mathbb{R}^D$, if $a^{\mathtt{T}} X_i \overset{\text{a.s.}}{=} 0$ then $a = 0$, and*

(4) *with probability 1, (f_n''') is uniformly bounded on some open ball $E \subseteq \Theta$ containing θ_0.*

For any E as in assumption 4, f_n satisfies the conditions of Theorem 5 with probability 1.

Proof As in the proof of Theorem 13, f_n is C^∞, f_n is convex, and by convexity, assumption 1 implies that with probability 1, for all $\theta \in \Theta$, $f_n(\theta) \to f(\theta)$. By Theorem 7,

$f''(\theta_0)$ exists and is finite. Thus, $f''(\theta_0)$ is positive definite since for all nonzero $a \in \mathbb{R}^D$, $a^\mathtt{T} f''(\theta_0) a = \mathrm{E}\big(\kappa''(\theta_0^\mathtt{T} X_i) a^\mathtt{T} X_i X_i^\mathtt{T} a\big) > 0$ by assumptions 2 and 3 and the fact that $\kappa''(\eta) > 0$. ∎

To explain the notation, the observed data consist of the first n elements of single random sequence $\vec{Y} = (Y_1, Y_2, \ldots)$, where each Y_i is a random vector. In the Gaussian Markov random field and Ising model examples in Sections 7.2 and 7.4 below, $\vec{Y}$ contains the values at the vertices of a single infinite graph, arranged as a sequence. Meanwhile, in the fully visible Boltzmann machine (Section 7.3), we have i.i.d. samples of graphs.

7.2 Gaussian Markov Random Fields

Gaussian Markov random fields (GMRFs) are widely used in spatial statistics and time-series (Banerjee et al., 2014). Let G be an infinite regular graph with vertices $v(1), v(2), \ldots$, and let $y_1, y_2, \ldots \in \mathbb{R}$ be variables associated with the vertices of G such that y_i is the value at $v(i)$. Consider a model in which the conditional distribution of y_i given y_{-i} is $p_\theta(y_i|y_{-i}) = \mathcal{N}(y_i \mid \theta^\mathtt{T}\varphi_i(\vec{y}), \gamma^{-1})$ where $\theta \in \Theta := \mathbb{R}^D$, $\varphi_i(\vec{y}) = (y_j : j \in N_i) \in \mathbb{R}^D$, and $N_i = \{j \in \mathbb{N} : v(j) \text{ is adjacent to } v(i)\}$. This leads to the pseudolikelihood (Besag, 1975)

$$\mathcal{L}_n^{\mathrm{GRF}}(\theta) = \prod_{i=1}^n p_\theta(y_i|y_{-i}) = \prod_{i=1}^n \mathcal{N}\big(y_i \mid \theta^\mathtt{T}\varphi_i(\vec{y}), \gamma^{-1}\big).$$

By defining $q(y|\eta) = \exp(\eta\gamma y - \kappa(\eta))$ for $y \in \mathbb{R}$ and $\eta \in \mathbb{R}$, where $\kappa(\eta) = \frac{1}{2}\gamma\eta^2$, this pseudolikelihood can be written as $\mathcal{L}_n^{\mathrm{GRF}}(\theta) \propto \prod_{i=1}^n q\big(y_i \mid \theta^\mathtt{T}\varphi_i(\vec{y})\big)$.

Theorem 19 *Let $\vec{Y} = (Y_1, Y_2, \ldots)$ be a sequence of random variables in $\mathbb{R}$ and define $X_i = (Y_j : j \in N_i) \in \mathbb{R}^D$ where N_i is defined as above. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ are identically distributed, but not necessarily independent. Assume:*

(1) $\frac{1}{n}\sum_{i=1}^n X_i Y_i \xrightarrow{\text{a.s.}} \mathrm{E}X_i Y_i \in \mathbb{R}^D$ *and* $\frac{1}{n}\sum_{i=1}^n X_i X_i^\mathtt{T} \xrightarrow{\text{a.s.}} \mathrm{E}X_i X_i^\mathtt{T} \in \mathbb{R}^{D\times D}$, *and*

(2) *for all $a \in \mathbb{R}^D$, if $a^\mathtt{T} X_i \stackrel{\text{a.s.}}{=} 0$ then $a = 0$.*

Then $\theta_0 := (\mathrm{E}X_i X_i^\mathtt{T})^{-1}\mathrm{E}X_i Y_i$ is well-defined and for any open ball E such that $\theta_0 \in E$, $f_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log q(Y_i \mid \theta^\mathtt{T} X_i)$ satisfies the conditions of Theorem 5 with probability 1.

Proof We apply Theorem 18. Let $f(\theta) = -\mathrm{E}\log q(Y_i \mid \theta^\mathtt{T} X_i) = \frac{1}{2}\gamma\theta^\mathtt{T}(\mathrm{E}X_i X_i^\mathtt{T})\theta - \gamma\theta^\mathtt{T}\mathrm{E}X_i Y_i$ for $\theta \in \mathbb{R}^D$. Thus, $f''(\theta) = \gamma(\mathrm{E}X_i X_i^\mathtt{T}) = \mathrm{E}\big(\kappa''(\theta^\mathtt{T} X_i) X_i X_i^\mathtt{T}\big)$ since $\kappa''(\eta) = \gamma$. By assumption 1, for all $\theta \in \mathbb{R}^D$, $f(\theta)$ is finite and $f_n(\theta) \xrightarrow{\text{a.s.}} f(\theta)$ as $n \to \infty$. As in the case of linear regression (Theorem 14), $EX_i X_i^\mathtt{T}$ is positive definite by assumption 2, $f'(\theta_0) = 0$, and (f_n''') is a.s. uniformly bounded on all of $\mathbb{R}^D$ since $\kappa'''(\eta) = 0$. ∎

The setup of Theorem 19 is quite general; note that the graph G may consist of a single connected component (such as the m-dimensional integer lattice $\mathbb{Z}^m$) or it may consist of many disconnected components, each of which could contain finitely many or infinitely many vertices. Further, the setup is that there is a single graph G, and more and more of the graph is observed as n grows; thus, it is necessary that G be infinite in order to obtain an asymptotic result, as we do in Theorem 19. The identically distributed assumption is quite general as well; for instance, it holds whenever the true distribution is stationary with respect to a set of transformations that can map $v(i)$ to $v(j)$ for any i, j. Thus, this assumption is reasonable since stationarity is commonly assumed (Banerjee et al., 2014; Lee et al., 2002; Kervrann and Heitz, 1995); also see Künsch (1981) for background. Assumption 1 of Theorem 19 is that a strong law of large numbers holds for $X_i Y_i$ and $X_i X_i^\mathsf{T}$; in Theorem 20 we show that this holds whenever the true distribution is a stationary, ergodic process on the integer lattice $\mathbb{Z}^m$, assuming a moment condition. Assumption 2 of Theorem 19 is simply that the support of the neighbor vector X_i is not contained in any proper subspace of $\mathbb{R}^D$; see Theorem 13 for further discussion of this non-degeneracy assumption.

Many applications in spatial statistics involve more complex models that do not satisfy the assumption of a regular graph with identically distributed neighborhoods (X_i, Y_i) (Ferreira and De Oliveira, 2007). While Theorem 19 could be extended to handle such generalizations, we chose to keep it relatively simple in order to capture the essential features of this class of models without being overburdened with details.

Theorem 20 *Suppose G is the m-dimensional lattice on $\mathbb{Z}^m$, and let $v : \mathbb{N} \to \mathbb{Z}^m$ be a bijection from $\mathbb{N}$ to $\mathbb{Z}^m$ such that $R(v(1)) \leq R(v(2)) \leq \cdots$ where $R(j) = \max\{|j_1|, \ldots, |j_m|\}$ for $j \in \mathbb{Z}^m$. Let $T_1, \ldots, T_m$ denote the shift transformations on $\mathbb{Z}^m$. Suppose $(Y_1, Y_2, \ldots)$ is a stochastic process such that the random field $(Y_{v^{-1}(j)} : j \in \mathbb{Z}^m)$ is stationary with respect to $T_1, \ldots, T_m$ and ergodic with respect to at least one of $T_1, \ldots, T_m$. If $\mathrm{E}|Y_1|^4 < \infty$, then assumption 1 of Theorem 19 holds.*

See Section E for the proof.

7.3 Fully Visible Boltzmann Machines

The Boltzmann machine is a stochastic recurrent neural network originally developed as a model of neural computation (Hinton and Sejnowski, 1983; Ackley et al., 1985). Maximum pseudolikelihood estimation has been shown to be consistent for fully visible Boltzmann machines (Hyvärinen, 2006). Here, we consider the corresponding pseudolikelihood-based generalized posteriors. To our knowledge, Theorem 21 is the first result establishing a Bernstein–von Mises theorem for this model.

Define $p_{A,b}(y) \propto \exp(y^\mathsf{T} A y + b^\mathsf{T} y)$ for $y \in \mathcal{Y} := \{-1, 1\}^d$, where $A \in \mathbb{R}^{d \times d}$ is a strictly upper triangular matrix and $b \in \mathbb{R}^d$. Given samples from $p_{A,b}$, inference for A and b is complicated by the intractability of the normalization constant $Z_{A,b} = \sum_{y \in \mathcal{Y}} \exp(y^\mathsf{T} A y + b^\mathsf{T} y)$ since $|\mathcal{Y}| = 2^d$ is very large when d is large. Observe that we can write

$$p_{A,b}(y_j | y_{-j}) \propto_{y_j} \exp\big(\textstyle\sum_{k=1}^{j-1} A_{kj} y_k y_j + \sum_{k=j+1}^{d} A_{jk} y_j y_k + b_j y_j\big) = \exp\big(y_j \theta^\mathsf{T} \varphi_j(y)\big) \tag{7.1}$$

where $\theta = \theta(A, b) \in \mathbb{R}^D$ is a $D = d + d(d-1)/2$ dimensional vector concatenating b and the strictly upper triangular entries of A, and $\varphi_j(y) \in \{-1, 0, 1\}^D$ is a function that does not depend on y_j. Thus, we have $p_{A,b}(y_j | y_{-j}) = q\big(y_j \mid \theta^\mathsf{T} \varphi_j(y)\big)$ by defining $q(y_j | \eta) = \exp(\eta y_j - \kappa(\eta))$ for $y_j \in \{-1, 1\}$ and $\eta \in \mathbb{R}$, where $\kappa(\eta) = \log(e^\eta + e^{-\eta})$. Now, suppose we have n

samples $y_1, \ldots, y_n \in \mathcal{Y} = \{-1, 1\}^d$ and for $\theta \in \Theta := \mathbb{R}^D$, consider the pseudolikelihood

$$\mathcal{L}_n^{\text{Boltz}}(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{d} p_{A,b}(y_{ij} | y_{i,-j}) = \prod_{i=1}^{n} \prod_{j=1}^{d} q\big(y_{ij} \mid \theta^{\mathtt{T}} \varphi_j(y_i)\big).$$

Theorem 21 *Let $Y_1, Y_2, \ldots \in \mathcal{Y}$ be i.i.d. random vectors and define $X_{ij} = \varphi_j(Y_i)$. Define $f(\theta) = -\sum_{j=1}^{d} \mathrm{E} \log q(Y_{ij} \mid \theta^{\mathtt{T}} X_{ij})$ for $\theta \in \Theta$. Assume:*

(1) *$f'(\theta_0) = 0$ for some $\theta_0 \in \Theta$, and*

(2) *for all nonzero $a \in \mathbb{R}^d$, $\mathrm{Var}(a^{\mathtt{T}} Y_i) > 0$.*

Then for any open ball E such that $\theta_0 \in E$, $f_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \log q\big(Y_{ij} \mid \theta^{\mathtt{T}} X_{ij}\big)$ satisfies the conditions of Theorem 5 with probability 1.

The assumptions of Theorem 21 are extremely mild and can be expected to typically hold in practice. Assumption 1 is simply that the maximum pseudolikelihood estimator exists, asymptotically – or more precisely, that the asymptotic pseudolikelihood function has a critical point. Assumption 2 is that there is no lower-dimensional affine subspace that contains Y_i almost surely; this is analogous to the non-degeneracy condition in Theorem 18.

Proof of Theorem 21 Observe that

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \kappa(\theta^{\mathtt{T}} X_{ij}) \;-\; \theta^{\mathtt{T}}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij} Y_{ij}\Big)$$

and $f(\theta) = \sum_{j=1}^{d} \mathrm{E}\kappa(\theta^{\mathtt{T}} X_{ij}) \;-\; \theta^{\mathtt{T}}\big(\sum_{j=1}^{d} \mathrm{E} X_{ij} Y_{ij}\big)$. As in the proof of Theorem 13, f_n is C^∞ and convex. Since $\{-1, 0, 1\}^D$ is a finite set, $\sup\big\{|\kappa(\theta^{\mathtt{T}} x)| : x \in \{-1, 0, 1\}^D\big\} < \infty$ for all $\theta \in \Theta$. Also, $|X_{ijk} Y_{ij}| \leq 1$, and thus, $f(\theta)$ is finite and $f_n(\theta) \xrightarrow{\text{a.s.}} f(\theta)$ by the strong law of large numbers. (Note that the standard strong law of large numbers applies here since the data consist of n i.i.d. samples from the Boltzmann machine, rather than the first n elements of a single sample as in the GMRF example.) Due to convexity, this implies that with probability 1, for all $\theta \in \Theta$, $f_n(\theta) \to f(\theta)$ as $n \to \infty$.

Let E be an open ball containing θ_0. Then for all $\theta \in E$, $|f_n'''(\theta)_{k\ell m}| \leq cd$ where $c = \sup\{|\kappa'''(\theta^{\mathtt{T}} x)| : x \in \{-1, 0, 1\}^D, \theta \in \bar{E}\}$, and $c < \infty$ because κ''' is continuous and $\bar{E}$ is compact. Thus, for all $\theta \in E$, $\|f_n'''(\theta)\|^2 = \sum_{k,\ell,m} |f_n'''(\theta)_{k\ell m}|^2 \leq c^2 d^2 D^3$. Hence, (f_n''') is uniformly bounded on E.

Now, we show that $f''(\theta_0)$ is positive definite. First, $f''(\theta_0) = \sum_{j=1}^{d} \mathrm{E}\big(\kappa''(\theta_0^{\mathtt{T}} X_{ij}) X_{ij} X_{ij}^{\mathtt{T}}\big)$ because differentiating under the integral sign is justified by the bounds $|\kappa(\eta)| \leq |\eta| + \log 2$, $|\kappa'(\eta)| \leq 1$, $|\kappa''(\eta)| \leq 2$, and $|X_{ijk}| \leq 1$ (Folland, 2013, 2.27). Let $\theta \in \mathbb{R}^D$ be nonzero and let A, b be the corresponding parameters such that $\theta = \theta(A, b)$. Then by Equation 7.1, $A^{\mathtt{T}} Y_i + A Y_i + b = (\theta^{\mathtt{T}} X_{i1}, \ldots, \theta^{\mathtt{T}} X_{id})^{\mathtt{T}} \in \mathbb{R}^d$. If $A \neq 0$, then $\mathrm{Var}(\theta^{\mathtt{T}} X_{ij'}) > 0$ for some j' by assumption 2, and hence, $\theta^{\mathtt{T}} f''(\theta_0)\theta = \sum_{j=1}^{d} \mathrm{E}\big(\kappa''(\theta_0^{\mathtt{T}} X_{ij}) |\theta^{\mathtt{T}} X_{ij}|^2\big) > 0$ because $\kappa''(\eta) > 0$ and $\mathbb{P}(|\theta^{\mathtt{T}} X_{ij'}| > 0) > 0$. Meanwhile, if $A = 0$, then $b_{j'} \neq 0$ for some j' (because $\theta \neq 0$), and again $\theta^{\mathtt{T}} f''(\theta_0)\theta > 0$ because $|\theta^{\mathtt{T}} X_{ij'}| = |b_{j'}| > 0$. Therefore, $f''(\theta_0)$ is positive definite. ∎

7.4 Ising Model

The Ising model is a classical model of ferromagnetism in statistical mechanics and has gained widespread use in many other applications such as spatial statistics (Banerjee et al., 2014) and image processing (Geman and Geman, 1984). Pseudolikelihood-based posteriors for the Ising model and Potts model, more generally, have been used by Zhou and Schmidler (2009) for protein modeling.

Consider the m-dimensional integer lattice $\mathbb{Z}^m$ and let $v : \mathbb{N} \to \mathbb{Z}^m$ be a bijection from $\mathbb{N}$ to $\mathbb{Z}^m$. Let $y_1, y_2, \ldots \in \mathcal{Y} := \{-1, 1\}$ be variables associated with the points of $\mathbb{Z}^m$ such that y_i is the value at $v(i)$. The Ising model is a Markov random field with singleton potentials $\exp(\theta_1 y_i)$ for each $i \in \mathbb{N}$ and pairwise potentials $\exp(\theta_2 y_i y_j)$ for each pair $i, j \in \mathbb{N}$ such that $v(i)$ and $v(j)$ are adjacent in $\mathbb{Z}^m$, that is, such that $|v(i) - v(j)| = 1$. This motivates the use of the pseudolikelihood (Besag, 1975),

$$\mathcal{L}_n^{\text{Ising}}(\theta) = \prod_{i=1}^{n} \frac{\exp(\theta_1 y_i + \theta_2 \sum_{j \in N_i} y_i y_j)}{\sum_{y \in \mathcal{Y}} \exp(\theta_1 y + \theta_2 \sum_{j \in N_i} y y_j)}$$

for $\theta \in \Theta := \mathbb{R}^2$, where $N_i = \{j \in \mathbb{N} : v(j) \text{ is adjacent to } v(i)\}$. By defining $q(y|\eta) = \exp(\eta y - \kappa(\eta))$ for $y \in \{-1, 1\}$ and $\eta \in \mathbb{R}$, where $\kappa(\eta) = \log(e^\eta + e^{-\eta})$, the Ising model pseudolikelihood can be written as $\mathcal{L}_n^{\text{Ising}}(\theta) = \prod_{i=1}^n q(y_i \mid \theta_1 + \theta_2 \sum_{j \in N_i} y_j)$.

Theorem 22 *Let $\vec{Y} = (Y_1, Y_2, \ldots)$ be a sequence of random variables in $\{-1, 1\}$ and define $X_i = \big(1, \sum_{j \in N_i} Y_j\big)^{\mathtt{T}} \in \mathbb{R}^2$. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ are identically distributed, but not necessarily independent. Define $f_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$ and $f(\theta) = -\mathrm{E} \log q(Y_i \mid \theta^{\mathtt{T}} X_i)$ for $\theta \in \Theta$. Assume:*

1. *for all $\theta \in \Theta$, $f_n(\theta) \xrightarrow{\text{a.s.}} f(\theta)$ as $n \to \infty$,*
2. *$f'(\theta_0) = 0$ for some $\theta_0 \in \Theta$, and*
3. $\operatorname{Var}\big(\sum_{j \in N_i} Y_j\big) > 0$.

Then for any open ball E such that $\theta_0 \in E$, f_n satisfies the conditions of Theorem 5 with probability 1.

Assumption 1 is that a strong law of large numbers holds for the log-likelihood terms. In Theorem 23, we show that assumption 1 holds whenever the true distribution is a stationary, ergodic process on $\mathbb{Z}^m$ satisfying a certain moment condition. Assumption 2 is that a maximum pseudolikelihood estimate exists, asymptotically. Assumption 3 is simply that the distribution of the neighbors is not degenerate, in the sense that their support is not restricted to an affine subspace orthogonal to the vector $(1, \ldots, 1)^{\mathtt{T}}$.

Proof of Theorem 22 We apply Theorem 18. Define $\mathcal{X} = \big\{(1, z)^{\mathtt{T}} : z \in \{-2m, \ldots, 2m\}\big\}$, noting that $X_i \in \mathcal{X}$. It is easy to check that Condition 17 holds. For all $\theta \in \Theta$, $f(\theta)$ is finite since $|\mathcal{X} \times \mathcal{Y}| < \infty$. If $a^{\mathtt{T}} X_i \overset{\text{a.s.}}{=} 0$ then $a = 0$, since $a^{\mathtt{T}} X_i = a_1 + a_2 \sum_{j \in N_i} Y_j$ and $\operatorname{Var}\big(\sum_{j \in N_i} Y_j\big) > 0$. Let E be an open ball containing θ_0, and let $c = \sup\{|\kappa'''(\theta^{\mathtt{T}} x)| : x \in \mathcal{X}, \theta \in \bar{E}\}$. Then $c < \infty$ since κ''' is continuous, $|\mathcal{X}|$ is finite, and $\bar{E}$ is compact. Therefore, for all $\theta \in E$, $|f_n'''(\theta)_{jk\ell}| \leq \frac{1}{n}\sum_{i=1}^n |\kappa'''(\theta^\mathtt{T} X_i) X_{ij} X_{ik} X_{i\ell}| \leq c(2m)^3$, and thus, (f_n''') is a.s. uniformly bounded on E. Finally, $f''(\theta_0) = \mathrm{E}\big(\kappa''(\theta_0^\mathtt{T} X_i) X_i X_i^\mathtt{T}\big)$ because differentiating under the integral sign is justified by the bounds $|\kappa(\eta)| \leq |\eta| + \log 2$, $|\kappa'(\eta)| \leq 1$, $|\kappa''(\eta)| \leq 2$, and $|X_{ij}| \leq 2m$ (Folland, 2013, 2.27). ∎

Theorem 23 *Let $v : \mathbb{N} \to \mathbb{Z}^m$ be a bijection such that $R(v(1)) \leq R(v(2)) \leq \cdots$ where $R(j) = \max\{|j_1|, \ldots, |j_m|\}$ for $j \in \mathbb{Z}^m$. Let $T_1, \ldots, T_m$ denote the shift transformations on $\mathbb{Z}^m$. Suppose $(Y_1, Y_2, \ldots)$ is a stochastic process such that the random field $(Y_{v^{-1}(j)} : j \in \mathbb{Z}^m)$ is stationary with respect to $T_1, \ldots, T_m$ and ergodic with respect to at least one of $T_1, \ldots, T_m$. Assume that $\mathrm{Var}\big(\log q(Y_1 \mid \theta_1 + \theta_2 \sum_{j \in N_1} Y_j)\big) < \infty$ for all $\theta \in \Theta$. Then assumption 1 of Theorem 22 holds.*

The proof is the same as Theorem 20, except with $Z_i = \log q(Y_i \mid \theta_1 + \theta_2 \sum_{j \in N_i} Y_j) - \mathrm{E}\big(\log q(Y_i \mid \theta_1 + \theta_2 \sum_{j \in N_i} Y_j)\big)$.

7.5 Cox Proportional Hazards Model

The Cox proportional hazards model (Cox, 1972) is widely used for survival analysis. The proportional hazards model assumes the hazard function for subject i is $\lambda_0(y)\exp(\theta^\mathtt{T} x_i)$ for $y \geq 0$, where $\lambda_0(y) \geq 0$ is a baseline hazard function shared by all subjects, $x_i \in \mathbb{R}^D$ is a vector of covariates for subject i, and $\theta \in \mathbb{R}^D$ is a vector of coefficients. To perform inference for θ in a way that does not require any modeling of λ_0 and elegantly handles censoring, Cox (1972, 1975) proposed using the *partial likelihood*,

$$\mathcal{L}_n^{\mathrm{Cox}}(\theta) = \prod_{i=1}^n \left(\frac{\exp(\theta^\mathtt{T} x_i)}{\sum_{j=1}^n \exp(\theta^\mathtt{T} x_j) \mathbb{1}(y_j \geq y_i)} \right)^{z_i}$$

where $y_i \geq 0$ is the outcome time for subject i and $z_i \in \{0, 1\}$ indicates whether y_i is an observed event time ($z_i = 1$) or a right-censoring time ($z_i = 0$). When $z_i = 1$, the ith factor in the partial likelihood can be interpreted as the conditional probability that subject i has an event at time y_i, given the risk set $\{j : y_j \geq y_i\}$ (the set of subjects that have not yet had an event or been censored up until time y_i) and given that some subject has an event at time y_i. See Efron (1977) for an intuitive explanation of the Cox partial likelihood based on a discrete approximation. Formally, the Cox partial likelihood coincides with the likelihood of a certain generalized linear model with categorical outcomes, however, asymptotic analysis is complicated by the dependencies between the factors of the partial likelihood. A number of authors have studied the asymptotics of the Cox partial likelihood; we mention, in particular, the result of Lin and Wei (1989) on asymptotic normality of the maximum partial likelihood estimator for the Cox model under misspecification.

The generalized posterior $\pi_n(\theta) \propto \mathcal{L}_n^{\mathrm{Cox}}(\theta)\pi(\theta)$ based on the Cox partial likelihood has been considered by several authors (Raftery et al., 1996; Sinha et al., 2003; Kim and Kim, 2009; Ventura and Racugno, 2016). Sinha et al. (2003) show that π_n approximates the standard posterior under a semiparametric Bayesian model, extending the results of Kalbfleisch (1978). Here, we provide sufficient conditions for π_n to exhibit concentration, asymptotic normality, and an asymptotically correct Laplace approximation.

Theorem 24 *Suppose* $(X, Y, Z), (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$ *are i.i.d., where* $X \in \mathcal{X} \subseteq \mathbb{R}^D$*,* $Y \geq 0$*, and* $Z \in \{0, 1\}$*. Define* $f(\theta) = \mathrm{E}\big(h_Y(\theta) Z\big) - \theta^\mathtt{T} \mathrm{E}(XZ)$ *for* $\theta \in \Theta := \mathbb{R}^D$ *where* $h_y(\theta) = \log \mathrm{E}\big(\exp(\theta^\mathtt{T} X) \mathbb{1}(Y \geq y)\big)$*. Assume:*

(1) $\mathcal{X}$ *is bounded,*

(2) *the c.d.f. of* Y *is continuous on* $\mathbb{R}$*,*

(3) $\mathbb{P}(Z = 1) > 0$ *and* $\mathrm{Var}(a^\mathtt{T} X) > 0$ *for all nonzero* $a \in \mathbb{R}^D$*,*

(4) $\mathbb{P}(Y \geq y \mid X = x) > 0$ *for all* $x \in \mathcal{X}$*,* $y \geq 0$*, and*

(5) $f'(\theta_0) = 0$ *for some* $\theta_0 \in \mathbb{R}^D$*.*

Then for any open ball E *such that* $\theta_0 \in E$*,* $f_n(\theta) := -\frac{1}{n} \log \mathcal{L}_n^{\mathrm{Cox}}(\theta) - \frac{1}{n} \sum_{i=1}^n Z_i \log n$ *satisfies the conditions of Theorem 5 with probability* 1*.*

See Section E for the proof. Note that $\exp(-n f_n(\theta)) \propto \mathcal{L}_n^{\mathrm{Cox}}(\theta)$ since $\frac{1}{n} \sum_{i=1}^n Z_i \log n$ does not depend on θ; the purpose of introducing this term is so that f_n converges. For interpretation, in words, the assumptions of Theorem 24 are that: (1) the covariates are bounded, (2) the time outcome is a continuous random variable, (3) the probability of observing an uncensored outcome is nonzero, and there is no lower-dimensional affine subspace that always contains the covariates (which is necessary for identifiability), (4) the survival function is nonzero, and (5) the maximum partial likelihood estimate exists, asymptotically (or more precisely, the asymptotic partial likelihood function has a critical point). These conditions are fairly mild and can be expected to hold in many practical applications.

7.6 Median-based Posterior for a Location Parameter

Suppose we wish to perform robust Bayesian inference for the parameter θ of a location family model $G_\theta(x) = G(x - \theta)$ where G is a cumulative distribution function (c.d.f.) on $\mathbb{R}$. If G is misspecified, then the posterior on θ can be poorly behaved, and may even fail to converge at all. For instance, if G_θ is the c.d.f. of $\mathcal{N}(\theta, \sigma^2)$ and the data are $X_1, X_2, \ldots$ i.i.d. $\sim$ Cauchy$(0, 1)$, then the posterior on θ is concentrated near $\frac{1}{n} \sum_{i=1}^n X_i$ when n is large, but $\frac{1}{n} \sum_{i=1}^n X_i \sim$ Cauchy$(0, 1)$; thus, the posterior does not converge to any fixed value.

Doksum and Lo (1990) propose to use the conditional distribution of θ given the sample median (or some other robust estimate of location) to perform robust Bayesian inference for θ. More precisely, let $M(x_{1:n})$ be a sample median of $x_{1:n} = (x_1, \ldots, x_n)$ and assume G_θ has a density g_θ. Then when n is odd,

$$\begin{aligned} p\big(\theta \,\big|\, M(X_{1:n}) = m\big) &\propto \pi(\theta)\, p(M(X_{1:n}) = m \mid \theta) \\ &\propto \pi(\theta) g_\theta(m) G_\theta(m)^{(n-1)/2} \big(1 - G_\theta(m)\big)^{(n-1)/2} \\ &= \pi(\theta) \exp\Big(\tfrac{1}{2}(n-1) \log G(m - \theta)(1 - G(m - \theta)) + \log g_\theta(m)\Big) \end{aligned}$$

where π is the prior on θ. Here, the conditional densities are under the model in which $\theta \sim \pi$ and $X_1, \ldots, X_n | \theta$ i.i.d. $\sim G_\theta$. Doksum and Lo (1990) show that $p(\theta \mid M(X_{1:n}) = M(x_{1:n}))$ and generalizations thereof have desirable properties as robust posteriors for θ; in particular, they provide consistency and asymptotic normality results.

With this as motivation, consider the generalized posterior $\pi_n(\theta) \propto \pi(\theta)\exp(-nf_n(\theta))$ where $f_n(\theta) = -\frac{1}{2}\log G(m_n-\theta)(1-G(m_n-\theta))$ and $m_n = M(x_{1:n})$; this approximates $p(\theta \mid M(X_{1:n}) = m_n)$ and is somewhat simpler to analyze. The following theorem strengthens the Doksum and Lo (1990) asymptotic normality result by showing convergence in total variation distance, rather than convergence in the weak topology. Further, our conditions are simpler, but we do assume greater regularity of G and we only consider the median.

Theorem 25 *Suppose $G : \mathbb{R} \to (0,1)$ is a c.d.f. such that G''' exists and is continuous, $G(-x) = 1 - G(x)$ for all $x \in \mathbb{R}$, $(\log G)''(x) \le 0$ for all $x \in \mathbb{R}$, and $(\log G)''(0) < 0$. If $\theta_0 \in \mathbb{R}$ and $m_1, m_2, \ldots \in \mathbb{R}$ such that $\theta_0 = \lim_{n\to\infty} m_n$, then for any open ball E containing θ_0, $f_n(\theta) := -\frac{1}{2}\log G(m_n-\theta)(1-G(m_n-\theta))$ satisfies the conditions of Theorem 5 on $\mathbb{R}$.*

Proof By the chain rule, $f_n(\theta)$ has a continuous third derivative since $\log(x)$ and $G(x)$ have continuous third derivatives and $G(x) \in (0,1)$. Define $f(\theta) = -\frac{1}{2}\log G(\theta_0-\theta)(1-G(\theta_0-\theta))$ for $\theta \in \mathbb{R}$. Then for all $\theta \in \mathbb{R}$, $f_n(\theta) \to f(\theta)$ as $n \to \infty$ since $m_n \to \theta_0$, $\log(x)$ and $G(x)$ are continuous, and $G(x) \in (0,1)$. Further,

$$\begin{aligned}
f(\theta) &= -\tfrac{1}{2}\log G(\theta_0-\theta) - \tfrac{1}{2}\log G(\theta-\theta_0), \\
f'(\theta) &= \tfrac{1}{2}(\log G)'(\theta_0-\theta) - \tfrac{1}{2}(\log G)'(\theta-\theta_0), \\
f''(\theta) &= -\tfrac{1}{2}(\log G)''(\theta_0-\theta) - \tfrac{1}{2}(\log G)''(\theta-\theta_0).
\end{aligned}$$

Thus, $f'(\theta_0) = 0$ and $f''(\theta_0) = -(\log G)''(0) > 0$. Similarly, $f_n''(\theta) = -\frac{1}{2}(\log G)''(m_n-\theta) - \frac{1}{2}(\log G)''(\theta-m_n) \ge 0$ since $(\log G)''(x) \le 0$. Thus, f_n is convex. Finally, for any bounded open interval E containing θ_0, (f_n''') is uniformly bounded on E by Proposition 26 with $h(\theta,s) = -\frac{1}{2}\log G(s-\theta)G(\theta-s)$, $K = \bar{E}$, and $S = [\inf m_n, \sup m_n] \subseteq \mathbb{R}$. ■

In cases where $f_n(\theta) = h(\theta, s_n)$ for some finite-dimensional statistic s_n, the following simple proposition can make it easy to verify the uniform boundedness condition.

Proposition 26 *Let $K \subseteq \mathbb{R}^D$ and $S \subseteq \mathbb{R}^d$ be compact sets. Suppose $f_n(\theta) = h(\theta, s_n)$ for $\theta \in K$, $n \in \mathbb{N}$, where $h : K \times S \to \mathbb{R}$ and $s_1, s_2, \ldots \in S$. If $(\partial^3 h/\partial\theta_i\partial\theta_j\partial\theta_k)(\theta,s)$ exists and is continuous on $K \times S$ for all $i,j,k \in \{1,\ldots,D\}$, then (f_n''') is uniformly bounded on K.*

Proof Let $h'''(\theta,s)$ denote the tensor of third derivatives with respect to θ, and let $c = \sup\{\|h'''(\theta,s)\| : \theta \in K, s \in S\}$. For all $\theta \in K$, $n \in \mathbb{N}$, we have $\|f_n'''(\theta)\| = \|h'''(\theta,s_n)\| \le c$, and $c < \infty$ since $(\theta,s) \mapsto \|h'''(\theta,s)\|$ is continuous and $K \times S$ is compact. ■

8. Previous Work

In this section, we provide a discussion comparing our assumptions, results, and proof techniques with those in previous work. Our discussion focuses primarily on asymptotic normality (Bernstein–von Mises), and we discuss work on posterior consistency in Section 8.7.

8.1 Overview of Previous BvM Results

The origins of the Bernstein–von Mises (BvM) theorem go back to Laplace (1810), Bernstein (1917), and Von Mises (1931). Rigorous formulations of the theorem were developed by Le Cam (1953), Bickel and Yahav (1969), Walker (1969), and Dawid (1970). These works employed "classical conditions" involving second, third, or even fourth-order derivatives of the log-likelihood; see the texts by Lehmann and Casella (2006) and Ghosh and Ramamoorthi (2003).

Le Cam (1970) discovered that the classical differentiability assumptions could be replaced by a less stringest condition referred to as differentiability in quadratic mean (DQM), which yields the benefits of a quadratic expansion while only requiring a certain first-order derivative; also see Le Cam (1986), Pollard (1997), Van der Vaart (2000), and Le Cam and Yang (2000) for background. Le Cam and Schwartz (1960) and Schwartz (1965) developed the assumption of the existence of uniformly consistent tests (UCTs) as a way of guaranteeing that θ_0 is distinguishable. In combination, the DQM and UCT assumptions form the basis for an elegant BvM theorem in the i.i.d. setting (Van der Vaart, 2000).

The works listed above focus on the canonical setting of a correctly specified, i.i.d. probabilistic model in which the dimension of parameter is fixed and finite. Going beyond this canonical setting, a number of authors have provided extensions of the theory. For instance, BvM theorems have been established for non-i.i.d. models such as Markov processes (Borwanker et al., 1971) and the Cox proportional hazards model with a prior on the baseline hazard function (Kim, 2006). More recently, Kleijn and Van der Vaart (2012) provide a BvM for cases in which the assumed model is misspecified, focusing primarily on the i.i.d. setting. Bochkina and Green (2014) provide an interesting BvM result when the true parameter is on the boundary of the parameter space, and their result is also applicable under misspecification.

For semiparametric and nonparametric models, BvM results have been established by a number of authors (Shen, 2002; Kim and Lee, 2004; Leahu, 2011; Castillo and Nickl, 2013; Bickel and Kleijn, 2012; Castillo and Rousseau, 2013). A very general result is provided by Panov and Spokoiny (2015), who establish a finite-sample BvM theorem for non-i.i.d. semiparametric models under misspecification, allowing the dimension of the parameter to grow with the sample size. While the theorem of Panov and Spokoiny (2015) is very general, their conditions are quite abstract and may be challenging for non-experts to employ.

For more references of early BvM contributions, see Bernardo and Smith (2000).

8.2 Categories of BvM Conclusions and Conditions

While all BvM results show that "the posterior converges to a normal distribution", each result can be placed along various axes in terms of the strength of the conclusions obtained and the generality of the conditions assumed.

Strength of the conclusions. First, the topology with respect to which convergence is shown to occur is usually either the weak topology (that is, convergence in distribution) or the "strong topology", that is, the metric topology induced by total variation (TV) distance. Second, in either topology, one can prove convergence in probability or almost sure convergence. In our results, we prove almost sure convergence in TV distance, which is the stronger conclusion in both respects.

Generality of the conditions. We group the conditions commonly assumed in BvM theorems as follows. The key conditions fall into two categories:

(A) regularity of the log-likelihood or generalized log-likelihood, and

(B) separation conditions enabling θ_0 to be distinguished.

Other conditions often assumed are that:

(C) the limiting Hessian at θ_0 is positive definite,

(D) the prior density is continuous and positive at θ_0,

(E) the dimension of parameter is fixed and finite,

(F) a consistent root of the likelihood exists (such as the MLE),

(G) the posterior arises from a probabilistic model,

(H) the data are independent and identically distributed, and

(I) the model is correctly specified.

8.3 Abstract BvM Conditions

To discuss how our Theorem 4 relates to the existing literature, we compare with Van der Vaart (2000, Theorem 10.1) (VdV, for short) as an example of a modern BvM theorem.

VdV assumes conditions G, H, and I (that is, the posterior arises from a correctly specified, i.i.d. probabilistic model), whereas Theorem 4 does not require these conditions—our results hold in misspecified, non-i.i.d. settings and do not even require that the posterior arise from a probability model. Theorem 4 and VdV both assume condition C (the limiting Hessian at θ_0, which coincides with the Fisher information matrix in VdV, is positive definite), condition D (the prior is continous and positive at θ_0), and condition E (the parameter dimension is fixed and finite). Meanwhile, neither theorem explicitly assumes condition F (consistent root), but both work with a sequence θ_n that converges to θ_0.

In category A (regularity conditions), VdV assumes differentiability in quadratic mean (DQM) whereas Theorem 4 assumes the quadratic approximation in assumption 1. DQM implies a particular quadratic expansion referred to as *local asymptotic normality* (LAN) (Van der Vaart, 2000, Theorem 7.2). The LAN property is roughly similar to assumption 1 of Theorem 4, except that it is centered at θ_0 and the remainder is formulated probabilistically (Pollard, 1997). In Theorem 4, having a deterministic (rather than probabilistic) bound on the remainder facilitates applications to generalized posteriors, since it decouples the deterministic convergence result from the (possibly complex) distribution of the data.

In category B (separation conditions), VdV assumes there exists a sequence of uniformly consistent tests (UCTs) for $H_0 : \theta = \theta_0$ versus $H_1 : \|\theta - \theta_0\| \geq \varepsilon$, for every $\varepsilon > 0$. Correspondingly, Theorem 4 employs assumption 2. The UCT assumption is less stringent in the i.i.d. setting, but not as broadly applicable in general; see our discussion of Schwartz's theorem in Section 8.7.

Finally, VdV establishes convergence in probability with respect to TV distance, whereas Theorem 4 enables us to obtain almost sure convergence in TV. Thus, while our assumptions in categories A and B may be stronger than VdV's in the correctly specified i.i.d. model setting, our conditions apply to much more general settings and we also obtain the stronger conclusion of almost sure convergence. Exploring whether DQM and UCT-like conditions can be extended to generalized posteriors is a potential area for future work.

8.4 Concrete BvM Conditions

To relate our Theorem 5 to the existing literature, we compare with Ghosh and Ramamoorthi (2003, Theorem 1.4.2) (G&R, for short) as an example of a BvM theorem employing more concrete, classical conditions.

G&R assume conditions F, G, H, and I (the posterior arises from a correctly specified, i.i.d. probabilistic model with a consistent root), whereas Theorem 5 does not assume any of these conditions. Theorem 5 and G&R both assume conditions C, D, and E.

In category A (regularity), G&R assume the third derivatives of the log-likelihood terms are dominated, uniformly over a neighborhood of θ_0, by an integrable function of the data. Similarly, Theorem 5 assumes the third derivatives of f_n are uniformly bounded in a neighborhood of θ_0. In the i.i.d. setting, the G&R domination condition is slightly weaker, but it is not clear how to extend it to arbitrary generalized posteriors. In category A, G&R also assume that the densities have common support, that the first and second derivatives of the log-likelihood at θ_0 are integrable, and that differentiation under the integral sign is justified; meanwhile, we only additionally require that f_n converges pointwise and has continuous third derivatives. In both G&R and Theorem 5, the role of these category A conditions is to bound the error term in a second-order Taylor expansion as in assumption 1 of Theorem 4; this is formalized on page 37 of G&R and in our Theorem 6. This establishes that near θ_0, the log posterior density approaches a quadratic form. Note that condition C is necessary to ensure that, when exponentiated, the limiting quadratic form can be normalized to a probability density.

In category B (separation), G&R assume that for any $\delta > 0$, there exists $\varepsilon > 0$ such that $\mathbb{P}\big(\inf_{|\theta-\theta_0|>\delta}(f_n(\theta) - f_n(\theta_0)) \geq \varepsilon\big) \longrightarrow 1$ as $n \to \infty$, where the probability $\mathbb{P}(\cdot)$ is with respect to the randomness in f_n due to the data. Meanwhile, in Theorem 5, we employ assumption 1, which is more stringent but is helpful in obtaining almost sure convergence rather than just convergence in probability. The role of the category B conditions is to ensure that negligible mass is placed outside a neighborhood of θ_0, asymptotically.

Like VdV, G&R prove convergence in probability with respect to TV distance, whereas Theorem 5 enables us to obtain a.s. convergence in TV. Thus, overall, we obtain a stronger and more general conclusion while assuming fewer conditions. In the special case of correctly specified i.i.d. models, our third derivative bounds and our separation condition are more stringent than the corresponding G&R conditions, however, our conditions extend readily to generalized posteriors.

8.5 BvM under Misspecification

Relatively recently, Kleijn and Van der Vaart (2012) extended the BvM theory to handle misspecification, that is, to apply to cases in which the assumed model is incorrect; also see Bochkina and Green (2014). Kleijn and Van der Vaart (2012, Theorem 2.1) (K&V, for short) establish a BvM result assuming (i) the LAN property holds at rate δ_n and (ii) the posterior concentrates (in probability) at θ_0 at the same rate, δ_n. Additionally, like our

Theorem 4, K&V assume conditions C, D, and E above. (Note that our H_0 corresponds to their V_{θ^*}.) Further, K&V assume condition G (whereas Theorem 4 does not), however, this assumption might not be essential for their proof.

In category A, condition (i) of K&V is roughly similar to assumption 1 of Theorem 4, as discussed in Section 8.3. In category B, condition (ii) of K&V roughly corresponds to assumption 2 of Theorem 4, except that K&V assume concentration at a particular rate (to match the LAN rate), whereas our condition does not require a rate. On the other hand, (ii) is less stringent in the sense that it only involves posterior probabilities and convergence in probability, whereas we require strict separation of θ_0 in terms of the values of $f_n(\theta)$.

To facilitate application of their theorem, Kleijn and Van der Vaart (2012) focus in particular on the i.i.d. setting, providing concrete conditions under which (i) and (ii) hold, assuming differentiability, a Lipschitz condition, a second-order Taylor expansion, non-singular Fisher information, and existence of UCTs. These conditions generalize VdV to the misspecified setting.

Regarding the strength of the conclusions, like VdV, K&V show convergence in probability (in TV), whereas our results show a.s. convergence (in TV). Thus, while our conditions are stronger in some respects, they are weaker in other respects, and we obtain a stronger conclusion in a more general setting.

8.6 BvM for Generalized Posteriors

Much previous work on generalized posteriors relies on Bernardo and Smith (2000, Proposition 5.14) to establish asymptotic normality; for instance, see Lazar (2003), Greco et al. (2008), Pauli et al. (2011), and Ventura and Racugno (2016). Thus, to relate our results to this literature, we compare our Theorem 4 to Bernardo and Smith (2000, Proposition 5.14) (B&S, for short), which is originally due to Chen (1985).

First, in terms of the strength of the conclusions, B&S only show convergence in distribution (that is, convergence in the weak topology), whereas Theorem 4 shows convergence in TV distance, which is much stronger. On the other hand, both theorems consider an arbitrary deterministic sequence of distributions (playing the role of generalized posteriors indexed by n), and thus, both are conducive for establishing almost sure BvM results.

In category A (regularity), B&S assume that (i) f_n is twice differentiable, (ii) the smallest eigenvalue of $n f_n''(\theta_n)$ tends to ∞ as $n \to \infty$, and (iii) for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all n sufficiently large, for all $\theta \in B_\delta(\theta_n)$, the Hessian $f_n''(\theta)$ satisfies $I - A(\varepsilon) \preceq f_n''(\theta) f_n''(\theta_n)^{-1} \preceq I + A(\varepsilon)$ where I is the identity matrix and $A(\varepsilon)$ is a symmetric positive semidefinite matrix whose largest eigenvalue tends to 0 as $\varepsilon \to 0$. Here, $A \preceq B$ denotes that $B - A$ is positive definite. Meanwhile, Theorem 4 employs assumption 1, which does not require any differentiability, but may require stronger control on the remainder term in Equation 3.1 compared to (iii) above (which B&S use to derive a quadratic approximation). B&S's eigenvalue condition in (ii) above is related to our assumption that $H_n \to H_0$ positive definite; indeed, the latter implies the former if $r_n''(0) \to 0$ as $n \to \infty$, where r_n is the remainder term in Equation 3.1.

In category B (separation), B&S assume that (iv) f_n has a strict local minimum θ_n, and (v) for any $\delta > 0$, the posterior probability of $B_\delta(\theta_n)$ converges to 1 as $n \to \infty$. Meanwhile, Theorem 4 obtains posterior concentration as a conclusion rather than assuming it, but

Theorem 4 does employ assumption 2. Both theorems assume condition E (the dimension of parameter is fixed and finite), however, neither theorem requires any of the additional conditions F, G, H, or I. Theorem 4 assumes condition D, while B&S does not since their assumptions are placed directly on the posterior density.

Overall, while Theorem 4 assumes more stringent conditions than B&S in terms of remainder control and separation, Theorem 4 does not require differentiability and yields a considerably stronger result in terms of TV distance rather than convergence in distribution.

8.7 Posterior Consistency

Doob (1949) used martingales to prove a very general result on posterior consistency for correctly specified i.i.d. models (also see Miller, 2018), however, it seems difficult to extend this proof technique, especially to generalized posteriors. Further, Doob's theorem is only guaranteed to hold on a set of prior probability 1. Schwartz (1965) established a powerful theorem on posterior consistency based on the UCT assumption along with an assumption that the prior puts positive mass in Kullback–Leibler neighborhoods of θ_0; see Ghosh and Ramamoorthi (2003, Theorem 4.4.1) for a clear exposition. Schwartz's theorem forms the basis for many modern results on posterior consistency in nonparametric Bayesian models (Ghosal, 2010). Schwartz's approach improves upon Doob's theorem by guaranteeing consistency at all points, and it is also conducive to generalization. Other notable early results on posterior consistency are due to Le Cam (1953) and Freedman (1963).

To interpret our posterior consistency theorems in the context of these well-known results, we first relate our Theorem 2 to Schwartz's theorem as presented by Ghosh and Ramamoorthi (2003, Theorem 4.4.1). Schwartz considers the setting of an i.i.d. probability model with densities $p(x|\theta)$, and assumes the data X_i are i.i.d. from some P_0. In this setting, $f_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p(X_i|\theta)$ and $f_n(\theta) \to f(\theta) = \mathrm{E}_{P_0}(-\log p(X_i|\theta))$ almost surely by the strong law of large numbers (SLLN), assuming $\mathrm{E}_{P_0}|\log p(X_i|\theta)| < \infty$. Further, if the model is correctly specified, so that $P_0 = P_{\theta_0}$ for some θ_0, then $f(\theta) - f(\theta_0)$ equals the Kullback–Leibler divergence, $D(P_{\theta_0}\|P_\theta)$.

Thus, in the correctly specified i.i.d. setting, the interpretation of Theorem 2 is as follows: assumption 1 essentially amounts to assuming a SLLN holds, assumption 2 is that the prior puts positive mass on Kullback–Leibler neighborhoods of θ_0 (just like Schwartz assumes), and roughly speaking, assumption 3 is that outside neighborhoods of θ_0, the log-likelihood does not get too close to the log-likelihood at θ_0 when n is sufficiently large. The main difference between Theorem 2 and Schwartz's theorem is that Theorem 2 employs assumption 3 instead of the UCT condition. In the i.i.d. setting, it seems inevitable that the UCT assumption is less stringent than assumption 3, however, it is not obvious how to extend the UCT approach to our setting of arbitrary generalized posteriors where we do not even assume there exists a data distribution. Thus, although in the i.i.d. setting, our condition is more stringent, we obtain the benefit of much broader applicability in general.

8.8 Proof Techniques

The proofs in this paper involve several new or non-standard techniques. First, in our main results in Sections 2 and 3, a major shift in technique is to study deterministic sequences of posterior distributions, rather than the usual approach of studying sequences of random posterior distributions obtained from random data. By showing that the conditions of the theorems hold with probability 1, we obtain almost sure convergence. This device enables one to separate the problem into a real analysis part (involving asymptotics of functions) and a probability part (involving the randomness in the data), which is particularly useful when considering generalized posteriors and misspecified models. This decoupling technique may facilitate future work on each part, separately. Bernardo and Smith (2000, Proposition 5.14) and Chen (1985) also considered deterministic sequences of distributions, but only for showing weak convergence rather than in total variation.

The proof of Theorem 2 has the same core structure as the proof of Schwartz's theorem (Ghosh and Ramamoorthi, 2003, Theorem 4.4.1), however, to handle generalized posteriors, we use assumption 3 rather than the UCT assumption in order to employ the deterministic sequence technique and enable application to generalized posteriors.

The proof technique for Theorem 4 differs from previous BvM proofs in some key respects — specifically, this formulation of the conditions facilitates a succinct proof using the generalized dominated convergence theorem. Further, we use the deterministic sequence technique described above. On the other hand, certain aspects of the proof are adapted from Ghosh and Ramamoorthi (2003, Theorem 1.4.2), such as how we break up the integral into regions.

The proof technique for Theorem 5 involves innovations as well, encapsulated primarily in Theorem 7. Specifically, regularity properties of f, f', f'', f_n, f_n', and f_n'' are obtained via Theorem 7, rather than having to be assumed. This simplifies several aspects of the proof, such as interchanging the order of derivatives and limits or expectations. Another advantage of our proof technique is that there is no need for a common support condition, which is sometimes assumed (Ghosh and Ramamoorthi, 2003, Theorem 1.4.2), because we deal with f_n directly, rather than with a probability model.

Acknowledgments

I would like to thank David Dunson, Matthew Harrison, Natalia Bochkina, and Basilis Gidas for helpful conversations.

Appendix A. Proofs of Concentration Results

Proof of Theorem 2 Let $\varepsilon > 0$. Define $\mu_n(E) = \int_E e^{-nf_n(\theta)}\Pi(d\theta)$ for $E \subseteq \Theta$. Recall that $\mu_n(\Theta) = z_n < \infty$ by assumption. For any $\beta \in \mathbb{R}$,

$$1 - \Pi_n(A_\varepsilon) = \Pi_n(A_\varepsilon^c) = \frac{\mu_n(A_\varepsilon^c)}{\mu_n(\Theta)} = \frac{e^{n(f(\theta_0)+\beta)}\mu_n(A_\varepsilon^c)}{e^{n(f(\theta_0)+\beta)}\mu_n(\Theta)},$$

so prove the result, it suffices to show that for some β, the numerator is bounded and the denominator goes to ∞.

First, consider the numerator. Assumption 3 implies that there exists $\beta > 0$ such that for all n sufficiently large, $\inf_{\theta \in A_\varepsilon^c} f_n(\theta) \geq f(\theta_0) + \beta$. Then for all n sufficiently large, for all $\theta \in A_\varepsilon^c$, we have $\exp\big(-n(f_n(\theta) - f(\theta_0) - \beta)\big) \leq 1$. Hence, for all n sufficiently large,

$$e^{n(f(\theta_0)+\beta)}\mu_n(A_\varepsilon^c) = \int_{A_\varepsilon^c} \exp\big(-n(f_n(\theta) - f(\theta_0) - \beta)\big)\Pi(d\theta) \leq \int_{A_\varepsilon^c} \Pi(d\theta) \leq 1.$$

Now, consider the denominator. For any $\theta \in A_{\beta/2}$, $f_n(\theta) - f(\theta_0) - \beta \longrightarrow f(\theta) - f(\theta_0) - \beta < -\beta/2 < 0$, and thus, $\exp\big(-n(f_n(\theta) - f(\theta_0) - \beta)\big) \longrightarrow \infty$ as $n \to \infty$. Therefore, by Fatou's lemma,

$$\liminf_{n\to\infty} e^{n(f(\theta_0)+\beta)}\mu_n(A_{\beta/2}) = \liminf_{n\to\infty} \int_{A_{\beta/2}} \exp\big(-n(f_n(\theta) - f(\theta_0) - \beta)\big)\Pi(d\theta) = \infty$$

since $\Pi(A_{\beta/2}) > 0$. Hence, $e^{n(f(\theta_0)+\beta)}\mu_n(\Theta) \to \infty$ since $\mu_n(\Theta) \geq \mu_n(A_{\beta/2})$. ■

Lemma 27 *Suppose $\Theta \subseteq \mathbb{R}^D$, $E \subseteq \Theta$ is convex and open in $\mathbb{R}^D$, and $\theta_0 \in E$. Let $f_n : \Theta \to \mathbb{R}$ be convex, and assume $f_n \to f$ pointwise on E for some $f : E \to \mathbb{R}$.*

(1) *If f' exists in a neighborhood of θ_0, $f'(\theta_0) = 0$, and $f''(\theta_0)$ exists and is positive definite, then $f(\theta) > f(\theta_0)$ for all $\theta \in E \setminus \{\theta_0\}$.*

(2) *If $f(\theta) > f(\theta_0)$ for all $\theta \in E \setminus \{\theta_0\}$, then $\liminf_n \inf_{\theta \in \Theta \setminus B_\varepsilon(\theta_0)} f_n(\theta) > f(\theta_0)$ for any $\varepsilon > 0$.*

Proof (1) As the pointwise limit of convex functions on a convex open set, f is convex on E (Rockafellar, 1970, 10.8). Let $R > 0$ such that $f'(\theta)$ exists for all $\theta \in B_R(\theta_0)$. Let $u \in \mathbb{R}^D$ with $|u| = 1$, and define $g(r) = f(\theta_0 + ru)$ for $r \in [0, R)$. Then $g'(r) = f'(\theta_0 + ru)^\mathtt{T} u$ and $g''(0) = u^\mathtt{T} f''(\theta_0)u$. Since

$$\frac{g'(r)}{r} = \frac{g'(r) - g'(0)}{r} \xrightarrow[r\to 0]{} g''(0) = u^\mathtt{T} f''(\theta_0)u > 0,$$

then $g'(r) > 0$ for all $r > 0$ sufficiently small, say, all $r \in (0, \varepsilon]$. Then for any $s \in (0, \varepsilon]$, we have

$$f(\theta_0 + su) - f(\theta_0) = g(s) - g(0) = \int_0^s g'(r)dr > 0. \tag{A.1}$$

Meanwhile, for any $s > \varepsilon$ such that $\theta_0 + su \in E$, we have

$$\frac{1}{s}(f(\theta_0 + su) - f(\theta_0)) \geq \frac{1}{\varepsilon}(f(\theta_0 + \varepsilon u) - f(\theta_0)) > 0$$

by the convexity of f and by Equation A.1 with $s = \varepsilon$. Hence, for any $s > 0$ such that $\theta_0 + su \in E$, $f(\theta_0 + su) > f(\theta_0)$. Since u is arbitrary, the result follows.

(2) By Rockafellar (1970, 10.8), $f_n \to f$ uniformly on any compact subset of E, and f is convex on E. Further, f is continuous on E, as a convex function on a convex open set (Rockafellar, 1970, Theorem 10.1). Let $\varepsilon > 0$ small enough that the ε-sphere $S_\varepsilon = \{\theta \in \mathbb{R}^D : |\theta - \theta_0| = \varepsilon\}$ is contained in E. Let $\alpha_n = \inf_{\theta \in S_\varepsilon} f_n(\theta) - f_n(\theta_0)$ and $\alpha = \inf_{\theta \in S_\varepsilon} f(\theta) - f(\theta_0)$. By uniform convergence, $\alpha_n \to \alpha$. Note that $\alpha > 0$, as the minimum of the continuous positive function $f(\theta) - f(\theta_0)$ on the compact set S_ε. For any $\theta \in \Theta \setminus B_\varepsilon(\theta_0)$, letting ξ_θ be the point of S_ε on the line from θ to θ_0, we have, by the convexity of f_n,

$$f_n(\theta) - f_n(\theta_0) \geq |\theta - \theta_0| \frac{f_n(\xi_\theta) - f_n(\theta_0)}{|\xi_\theta - \theta_0|} \geq \alpha_n$$

whenever $\alpha_n \geq 0$, since $|\theta - \theta_0| \geq |\xi_\theta - \theta_0|$. Since $\alpha_n \to \alpha > 0$, then for all n sufficiently large, for all $\theta \in \Theta \setminus B_\varepsilon(\theta_0)$, $f_n(\theta) \geq f_n(\theta_0) + \alpha_n \longrightarrow f(\theta_0) + \alpha$. Therefore, $\liminf_n \inf_{\theta \in \Theta \setminus B_\varepsilon(\theta_0)} f_n(\theta) \geq f(\theta_0) + \alpha > f(\theta_0)$. Note that this also implies the same inequality for any $\varepsilon' > \varepsilon$. ■

Proof of Theorem 3

(Part 1) Defining A_ε as in Theorem 2, it suffices to show that

(a) for any $\varepsilon > 0$ there exists $\delta > 0$ such that $A_\delta \subseteq N_\varepsilon$, and

(b) for any $\delta > 0$ there exists $\varepsilon' > 0$ such that $N_{\varepsilon'} \subseteq A_\delta$,

since for any $\varepsilon > 0$, choosing δ by (a), we have $\Pi_n(N_\varepsilon) \geq \Pi_n(A_\delta)$; meanwhile, for any $\delta > 0$, choosing ε' by (b), we have $\Pi(A_\delta) \geq \Pi(N_{\varepsilon'}) > 0$ and $\liminf_n \inf_{\theta \in A_\delta^c} f_n(\theta) \geq \liminf_n \inf_{\theta \in N_{\varepsilon'}^c} f_n(\theta) > f(\theta_0)$, and hence, by Theorem 2, $\Pi_n(A_\delta) \to 1$.

(a) Let $\varepsilon > 0$. Pointwise convergence and the liminf condition imply $\inf_{\theta \in N_\varepsilon^c} f(\theta) > f(\theta_0)$, hence, letting $\delta = \inf_{\theta \in N_\varepsilon^c} f(\theta) - f(\theta_0)$, we have $\delta > 0$ and $A_\delta \subseteq N_\varepsilon$.

(b) Let $\delta > 0$. By the continuity of f at θ_0, choose $\varepsilon' > 0$ such that $|f(\theta) - f(\theta_0)| < \delta$ for all $\theta \in N_{\varepsilon'}$. Then for any $\theta \in N_{\varepsilon'}$, $f(\theta) < f(\theta_0) + \delta$, hence, $\theta \in A_\delta$.

(Part 2) We show that 2 implies 1. By Lemma 36, $f_n \to f$ uniformly on K. Consequently, $f|_K$ is continuous, as the uniform limit of continuous functions (Rudin, 1976, 7.12). In particular, f is continuous at θ_0, since θ_0 is an interior point of K. For any $\varepsilon > 0$,

$$\liminf_n \inf_{\theta \in K \setminus N_\varepsilon} f_n(\theta) = \inf_{\theta \in K \setminus N_\varepsilon} f(\theta) > f(\theta_0),$$

the first step holding since $f_n \to f$ uniformly on K, and the second step since $f|_K$ is continuous, $K \setminus N_\varepsilon$ is compact, and $f(\theta) > f(\theta_0)$ for all $\theta \in K \setminus \{\theta_0\}$. Therefore, since $N_\varepsilon^c \subseteq (K \setminus N_\varepsilon) \cup K^c$,

$$\liminf_n \inf_{\theta \in N_\varepsilon^c} f_n(\theta) \geq \liminf_n \min\Big\{ \inf_{\theta \in K \setminus N_\varepsilon} f_n(\theta),\ \inf_{\theta \in K^c} f_n(\theta) \Big\} > f(\theta_0).$$

(Part 3) We show that 3 implies 1. Denote $B_\varepsilon = \{\theta \in \mathbb{R}^D : |\theta - \theta_0| < \varepsilon\}$. Let $r > 0$ small enough that $B_r \subseteq \Theta$. As the pointwise limit of convex functions, f is convex, and thus, it is continuous on B_r (Rockafellar, 1970, 10.1). By Lemma 27 with $E = B_r$, in either case (a) or (b), we have

$$\liminf_n \inf_{\theta\in\Theta\setminus B_\varepsilon} f_n(\theta) > f(\theta_0)$$

for any $\varepsilon > 0$. Since $\Theta \setminus B_\varepsilon = \Theta \setminus N_\varepsilon = N_\varepsilon^c$, this proves the result. ■

Appendix B. Proofs of Asymptotic Normality Results

Lemma 28 *Let $\theta_n \in \mathbb{R}^D$ such that $\theta_n \to \theta_0$ for some $\theta_0 \in \mathbb{R}^D$, let π_n be a density with respect to Lebesgue measure on $\mathbb{R}^D$, and let q_n be the density of $\sqrt{n}(\theta - \theta_n)$ when $\theta \sim \pi_n$. If $\int |q_n(x) - q(x)|dx \longrightarrow 0$ for some probability density q, then π_n concentrates at θ_0.*

Proof Let Π_n, Q_n, and Q denote the probability measures corresponding to π_n, q_n, and q, respectively. For any $\varepsilon > 0$ and $\delta > 0$,

$$Q_n(B_\delta(0)) = \Pi_n(B_{\delta/\sqrt{n}}(\theta_n)) \leq \Pi_n(B_\varepsilon(\theta_0))$$

for all n sufficiently large. Hence, since $Q_n \to Q$ in total variation,

$$Q(B_\delta(0)) = \lim_n Q_n(B_\delta(0)) \leq \liminf_n \Pi_n(B_\varepsilon(\theta_0)).$$

Taking the limit as $\delta \to \infty$ shows that $\lim_n \Pi_n(B_\varepsilon(\theta_0)) = 1$. ■

Proof of Theorem 4 Note that $q_n(x) = \pi_n(\theta_n + x/\sqrt{n})n^{-D/2}$. Define

$$\begin{aligned} g_n(x) &= \exp\big(-n[f_n(\theta_n + x/\sqrt{n}) - f_n(\theta_n)]\big)\pi(\theta_n + x/\sqrt{n}) \\ &= q_n(x)e^{nf_n(\theta_n)}n^{D/2}z_n, \end{aligned} \tag{B.1}$$

recalling that $z_n < \infty$ by assumption, and define

$$g_0(x) = \exp(-\tfrac{1}{2}x^{\mathtt{T}} H_0 x)\pi(\theta_0).$$

Let $\alpha \in (0, \lambda)$, where λ is the smallest eigenvalue of H_0. Let $\varepsilon > 0$ small enough that $\varepsilon < \alpha/(2c_0)$, $\varepsilon < \varepsilon_0$, and $\pi(\theta) \leq 2\pi(\theta_0)$ for all $\theta \in B_{2\varepsilon}(\theta_0)$ (which we can do since π is continuous at θ_0). Let $\delta = \liminf_n \inf_{\theta\in B_\varepsilon(\theta_n)^c} \big(f_n(\theta) - f_n(\theta_n)\big)$, noting that $\delta > 0$ by assumption. Letting $A_n = H_n - \alpha I$ and $A_0 = H_0 - \alpha I$, define

$$\begin{aligned} h_n(x) &= \begin{cases} \exp(-\tfrac{1}{2}x^{\mathtt{T}} A_n x)2\pi(\theta_0) & \text{if } |x| < \varepsilon\sqrt{n}, \\ e^{-n\delta/2}\pi(\theta_n + x/\sqrt{n}) & \text{if } |x| \geq \varepsilon\sqrt{n}, \end{cases} \\ h_0(x) &= \exp(-\tfrac{1}{2}x^{\mathtt{T}} A_0 x)2\pi(\theta_0). \end{aligned}$$

We will show that

(a) $g_n \to g_0$ and $h_n \to h_0$ pointwise,

(b) $\int h_n \to \int h_0$,

(c) $g_n = |g_n| \leq h_n$ for all n sufficiently large, and

(d) $g_n, g_0, h_n, h_0 \in L^1$ for all n sufficiently large.

By the generalized dominated convergence theorem, this will imply that $\int g_n \to \int g_0$ and $\int |g_n - g_0| \to 0$ (e.g., Folland, 2013, exercises 2.20, 2.21). Supposing this for the moment, we show how the result follows. Since $\int q_n = 1$, by Equation B.1 we have

$$e^{nf_n(\theta_n)} n^{D/2} z_n = \int g_n \longrightarrow \int g_0 = \pi(\theta_0) \frac{(2\pi)^{D/2}}{|H_0|^{1/2}}, \tag{B.2}$$

where $|H_0| = |\det H_0|$, and hence,

$$z_n \sim \frac{e^{-nf_n(\theta_n)} \pi(\theta_0)}{|H_0|^{1/2}} \Big(\frac{2\pi}{n}\Big)^{D/2}$$

as $n \to \infty$; this proves Equation 3.3. For any $a_n \to a \in \mathbb{R}$, we have $\int |a_n g_n - a g_0| \to 0$ since

$$\int |a_n g_n - a g_0| \leq \int |a_n g_n - a_n g_0| + \int |a_n g_0 - a g_0| = |a_n| \int |g_n - g_0| + |a_n - a| \int |g_0| \longrightarrow 0.$$

Thus, letting $1/a_n = e^{nf_n(\theta_n)} n^{D/2} z_n$ and $1/a = \pi(\theta_0) \frac{(2\pi)^{D/2}}{|H_0|^{1/2}}$, we have $a_n \to a$ by Equation B.2, and thus,

$$\int \Big| q_n(x) - \frac{|H_0|^{1/2}}{(2\pi)^{D/2}} \exp(-\tfrac{1}{2} x^{\mathtt{T}} H_0 x) \Big| dx \longrightarrow 0,$$

proving Equation 3.4. Equation 3.2 (concentration at θ_0) follows by Lemma 28, since $\theta_n \to \theta_0$. It remains to show (a)–(d) above.

(a) Fix $x \in \mathbb{R}^D$. First, consider h_n. For all n sufficiently large, $|x| < \varepsilon\sqrt{n}$, and thus,

$$h_n(x) = \exp(-\tfrac{1}{2} x^{\mathtt{T}} A_n x) 2\pi(\theta_0) \longrightarrow \exp(-\tfrac{1}{2} x^{\mathtt{T}} A_0 x) 2\pi(\theta_0) = h_0(x)$$

since $A_n \to A_0$. Now, for g_n, first note that $\pi(\theta_n + x/\sqrt{n}) \to \pi(\theta_0)$ since π is continuous at θ_0 and $\theta_n \to \theta_0$, $x/\sqrt{n} \to 0$. By the assumed representation of f_n (Equation 3.1),

$$n(f_n(\theta_n + x/\sqrt{n}) - f_n(\theta_n)) = \tfrac{1}{2} x^{\mathtt{T}} H_n x + n r_n(x/\sqrt{n}) \longrightarrow \tfrac{1}{2} x^{\mathtt{T}} H_0 x$$

since $H_n \to H_0$ and for all n sufficiently large (to ensure that $|x/\sqrt{n}| < \varepsilon_0$ and the assumed bound on r_n holds),

$$|n r_n(x/\sqrt{n})| \leq n c_0 |x/\sqrt{n}|^3 = c_0 |x|^3/\sqrt{n} \to 0 \tag{B.3}$$

as $n \to \infty$. Hence, $g_n(x) \to g_0(x)$.

(b) By the definition of h_n, letting $B_n = B_{\varepsilon\sqrt{n}}(0)$,

$$\textstyle\int h_n = \int_{B_n} \exp(-\frac{1}{2}x^{\mathtt{T}} A_n x) 2\pi(\theta_0) dx + \int_{B_n^c} e^{-n\delta/2}\pi(\theta_n + x/\sqrt{n}) dx.$$

Since $A_n \to A_0$ and A_0 is positive definite, then for all n sufficiently large, A_n is also positive definite and the first term equals

$$2\pi(\theta_0)\frac{(2\pi)^{D/2}}{|A_n|^{1/2}}\mathbb{P}(|A_n^{-1/2}Z| < \varepsilon\sqrt{n}) \longrightarrow 2\pi(\theta_0)\frac{(2\pi)^{D/2}}{|A_0|^{1/2}} = \textstyle\int h_0$$

where $Z \sim \mathcal{N}(0, I)$. The second term goes to zero, since it is nonnegative and upper bounded by

$$\int_{\mathbb{R}^D} e^{-n\delta/2}\pi(\theta_n + x/\sqrt{n}) dx = e^{-n\delta/2} n^{D/2} \longrightarrow 0,$$

using the fact that $\pi(\theta_n + x/\sqrt{n})n^{-D/2}$ is the density of $X = \sqrt{n}(\theta - \theta_n)$ when $\theta \sim \pi$.

(c) For all n sufficiently large, $|\theta_n - \theta_0| < \varepsilon$, the bound on r_n applies, and $\inf_{\theta \in B_\varepsilon(\theta_n)^c} f_n(\theta) - f_n(\theta_n) > \delta/2$. Let n large enough that these hold, and let $x \in \mathbb{R}^D$. If $|x| \geq \varepsilon\sqrt{n}$, then $f_n(\theta_n + x/\sqrt{n}) - f_n(\theta_n) > \delta/2$, and thus,

$$g_n(x) \leq e^{-n\delta/2}\pi(\theta_n + x/\sqrt{n}) = h_n(x).$$

Meanwhile, if $|x| < \varepsilon\sqrt{n}$, then $\pi(\theta_n + x/\sqrt{n}) \leq 2\pi(\theta_0)$ (by our choice of ε, since $|(\theta_n + x/\sqrt{n}) - \theta_0| \leq |\theta_n - \theta_0| + |x/\sqrt{n}| < 2\varepsilon$), and

$$n(f_n(\theta_n + x/\sqrt{n}) - f_n(\theta_n)) = \tfrac{1}{2}x^{\mathtt{T}} H_n x + n r_n(x/\sqrt{n}) \geq \tfrac{1}{2}x^{\mathtt{T}} H_n x - \tfrac{1}{2}\alpha x^{\mathtt{T}} x = \tfrac{1}{2}x^{\mathtt{T}} A_n x$$

since $|n r_n(x/\sqrt{n})| \leq c_0|x|^3/\sqrt{n} \leq c_0\varepsilon|x|^2 \leq \frac{1}{2}\alpha|x|^2$, by the fact that $|x/\sqrt{n}| < \varepsilon < \varepsilon_0$ and $\varepsilon < \alpha/(2c_0)$. Therefore,

$$g_n(x) \leq \exp(-\tfrac{1}{2}x^{\mathtt{T}} A_n x) 2\pi(\theta_0) = h_n(x).$$

(d) Since H_0 and A_0 are positive definite, $\int g_0$ and $\int h_0$ are finite. By (b) and (c), since $\int h_n \to \int h_0 < \infty$, we have $\int g_n \leq \int h_n < \infty$ for all n sufficiently large. Measurability of g_n and h_n follows from measurability of f_n and π. ■

Proof of Theorem 5 Without loss of generality, we may assume E is convex, since otherwise we can choose $E' \subseteq E$ to be an open ball around θ_0, and proceed with E' in place of E throughout the proof. First, we show that under case 2, the conditions for case 1 hold. By Lemma 27(1), $f(\theta) > f(\theta_0)$ for all $\theta \in E \setminus \{\theta_0\}$ since f' exists on E by Theorem 7. Letting $K = \overline{B_\varepsilon(\theta_0)}$ where $\varepsilon > 0$ is small enough that $K \subseteq E$, we have $\liminf_n \inf_{\theta \in \Theta \setminus K} f_n(\theta) > f(\theta_0)$ by Lemma 27(2). Thus, it suffices to prove the result under case 1.

Consider case 1. Extend π, f_n, and f to all of $\mathbb{R}^D$ by defining $\pi(\theta) = 0$ and $f(\theta) = f_n(\theta) = f(\theta_0) + 1$ for all $\theta \in \mathbb{R}^D \setminus \Theta$. Then all the conditions of Theorem 5 (under case 1) still hold with $\mathbb{R}^D$ in place of Θ. We will show that:

(a) (f_n) is equicontinuous on E, and $f_n''(\theta_0) \to f''(\theta_0)$ as $n \to \infty$,

(b) there exist $\theta_n \in E$ such that $\theta_n \to \theta_0$ and $f_n'(\theta_n) = 0$ for all n sufficiently large, and

(c) $f_n(\theta_n) \to f(\theta_0)$.

Assuming (a)–(c) for the moment, we show how the result follows. Letting $H_0 = f''(\theta_0)$, the conditions of Theorem 6 are satisfied, and thus, assumption 1 of Theorem 4 is satisfied for all n sufficiently large. Assumption 2 of Theorem 4 holds, since for all $\varepsilon > 0$,

$$\begin{aligned}\liminf_n \inf_{\theta \in B_\varepsilon(\theta_n)^c} (f_n(\theta) - f_n(\theta_n)) &= \Big(\liminf_n \inf_{\theta \in B_\varepsilon(\theta_n)^c} f_n(\theta)\Big) - f(\theta_0) \\ &\geq \Big(\liminf_n \inf_{\theta \in B_{\varepsilon/2}(\theta_0)^c} f_n(\theta)\Big) - f(\theta_0) > 0\end{aligned}$$

the first step holding by (c), the second step since $\theta_n \to \theta_0$ and thus $B_{\varepsilon/2}(\theta_0) \subseteq B_\varepsilon(\theta_n)$ for all n sufficiently large, and the third step by the implication 2 $\Rightarrow$ 1 in Theorem 3. Thus, the conditions of Theorem 4 are satisfied (except possibly for some initial sequence of n's, which can be ignored since the conclusions are asymptotic in nature), establishing Equation 3.2 (concentration at θ_0), Equation 3.3 (the Laplace approximation), and Equation 3.4 (asymptotic normality). To complete the proof, we establish (a), (b), and (c).

(a) By Theorem 7, (f_n) is equi-Lipschitz (hence, equicontinuous) on E and $f_n'' \to f''$ uniformly on E.

(b) Let $\varepsilon > 0$ small enough that $S_\varepsilon \subseteq K$ where $S_\varepsilon = \{\theta \in \mathbb{R}^D : |\theta - \theta_0| = \varepsilon\}$. By Theorem 7, f is continuous on E (since f' exists on E). Thus, f attains its minimum on the compact set S_ε, and since $f(\theta) > f(\theta_0)$ on S_ε, we have $\inf_{\theta \in S_\varepsilon} f(\theta) > f(\theta_0)$. For each n, since f_n is continuous on E, its minimum over the set $\overline{B_\varepsilon(\theta_0)}$ is attained at one or more points; define θ_n^ε to be such a minimizer. Since $f_n \to f$ uniformly on E (by Theorem 7), then for all n sufficiently large, any such minimizer cannot be in S_ε (since $\inf_{\theta \in S_\varepsilon} f(\theta) > f(\theta_0)$). Hence, for all sufficiently small $\varepsilon > 0$, for all n sufficiently large, we have $\theta_n^\varepsilon \in B_\varepsilon(\theta_0)$ and (by Lemma 38) $f_n'(\theta_n^\varepsilon) = 0$.

Thus, we can choose a sequence $\varepsilon_n > 0$ such that (a) $\varepsilon_n \to 0$ and (b) for all n sufficiently large, $\theta_n^{\varepsilon_n} \in B_{\varepsilon_n}(\theta_0)$ and $f_n'(\theta_n^{\varepsilon_n}) = 0$. Therefore, letting $\theta_n = \theta_n^{\varepsilon_n}$, we have $\theta_n \to \theta_0$ and $f_n'(\theta_n) = 0$ for all n sufficiently large.

(c) We have $|f_n(\theta_n) - f(\theta_0)| \leq |f_n(\theta_n) - f_n(\theta_0)| + |f_n(\theta_0) - f(\theta_0)| \to 0$, the first term going to zero since $\theta_n \to \theta_0$ and (f_n) is equi-Lipschitz on E, and the second term since $f_n \to f$ pointwise. ■

For tensors $S, T \in \mathbb{R}^{D^3}$, define the inner product $\langle S, T\rangle = \sum_{i,j,k} S_{ijk} T_{ijk}$ (noting that this is just the dot product of the vectorized versions of S and T). For $x \in \mathbb{R}^D$, define $x^{\otimes 3} = x \otimes x \otimes x = \big(x_i x_j x_k\big)_{i,j,k=1}^D \in \mathbb{R}^{D^3}$, and note that $\|x^{\otimes 3}\| = |x|^3$.

Proof of Theorem 6 By Lemma 37, (f_n'') is equi-Lipschitz. Thus,

$$\|f_n''(\theta_n) - H_0\| \leq \|f_n''(\theta_n) - f_n''(\theta_0)\| + \|f_n''(\theta_0) - H_0\| \leq C|\theta_n - \theta_0| + \|f_n''(\theta_0) - H_0\| \longrightarrow 0,$$

and hence, $H_n \to H_0$. Let $C_0 = \sup_n \sup_{\theta \in E} \|f_n'''(\theta)\|$. Let n large enough that $f_n'(\theta_n) = 0$. For $\theta \in E$, by Taylor's theorem,

$$f_n(\theta) = f_n(\theta_n) + \tfrac{1}{2}(\theta - \theta_n)^\intercal f_n''(\theta_n)(\theta - \theta_n) + r_n(\theta - \theta_n)$$

where $r_n(\theta - \theta_n) = \frac{1}{6}\langle f_n'''(t_n(\theta)), (\theta - \theta_n)^{\otimes 3}\rangle$, and $t_n(\theta)$ is a point on the line between θ and θ_n. Then by Cauchy–Schwarz,

$$|r_n(\theta - \theta_n)| \leq \tfrac{1}{6}\|f_n'''(t_n(\theta))\|\|(\theta - \theta_n)^{\otimes 3}\| \leq \tfrac{1}{6}C_0\|(\theta - \theta_n)^{\otimes 3}\| = \tfrac{1}{6}C_0|\theta - \theta_n|^3. \tag{B.4}$$

Choose $\varepsilon_0 > 0$ small enough that $B_{2\varepsilon_0}(\theta_0) \subseteq E$, and choose $c_0 = C_0/6$. For all n sufficiently large, $|\theta_n - \theta_0| \leq \varepsilon_0$ and hence for all $x \in B_{\varepsilon_0}(0)$, we have $\theta_n + x \in B_{2\varepsilon_0}(\theta_0) \subseteq E$; thus, setting $\theta = \theta_n + x$ in Equation B.4 yields $|r_n(x)| \leq c_0|x|^3$. ∎

Appendix C. Proof of Regular Convergence Theorem

Lemma 29 *Let $E \subseteq \mathbb{R}^D$ be open. If $f_n : E \to \mathbb{R}$ has continuous second derivatives, (f_n) is pointwise bounded, and (f_n'') is uniformly bounded, then (f_n') is pointwise bounded.*

Proof Let $C = \sup\{\|f_n''(x)\| : n \in \mathbb{N}, x \in E\} < \infty$. Fix $x \in E$, and let $\varepsilon > 0$ small enough that $B_{2\varepsilon}(x) \subseteq E$. By Taylor's theorem, for any $u \in \mathbb{R}^D$ with $|u| = 1$,

$$f_n(x + \varepsilon u) = f_n(x) + \varepsilon f_n'(x)^{\mathtt{T}} u + \tfrac{1}{2}\varepsilon^2 u^{\mathtt{T}} f_n''(z) u$$

for some z on the line between x and $x + \varepsilon u$, and therefore,

$$|f_n'(x)^{\mathtt{T}} u| \leq (1/\varepsilon)|f_n(x + \varepsilon u) - f_n(x)| + \tfrac{1}{2}\varepsilon C$$

since $|u^{\mathtt{T}} f_n''(z) u| \leq \|f_n''(z)\||u|^2 \leq C$. Thus, $\{f_n'(x)^{\mathtt{T}} u : n \in \mathbb{N}\}$ is bounded, for any u with $|u| = 1$. Applying this to each element of the standard basis, we see $f_n'(x)$ is bounded. ∎

Lemma 30 *Let $E \subseteq \mathbb{R}^D$ be open. If $f_n : E \to \mathbb{R}$ has continuous third derivatives, (f_n) is pointwise bounded, and (f_n''') is uniformly bounded, then (f_n'') is pointwise bounded.*

Proof Let $C = \sup_n \sup_{x \in E} \|f_n'''(x)\| < \infty$. Fix $x \in E$, and let $\varepsilon > 0$ small enough that $\overline{B_\varepsilon(x)} \subseteq E$. By Taylor's theorem, for any $u \in \mathbb{R}^D$ with $|u| = 1$,

$$f_n(x + \varepsilon u) = f_n(x) + \varepsilon f_n'(x)^{\mathtt{T}} u + \tfrac{1}{2}\varepsilon^2 u^{\mathtt{T}} f_n''(x) u + \tfrac{1}{6}\varepsilon^3 \langle f_n'''(z^+), u^{\otimes 3}\rangle$$

for some z^+ on the line between x and $x + \varepsilon u$. Likewise,

$$f_n(x - \varepsilon u) = f_n(x) - \varepsilon f_n'(x)^{\mathtt{T}} u + \tfrac{1}{2}\varepsilon^2 u^{\mathtt{T}} f_n''(x) u - \tfrac{1}{6}\varepsilon^3 \langle f_n'''(z^-), u^{\otimes 3}\rangle$$

for some z^- on the line between x and $x - \varepsilon u$. Adding these two equations gives

$$f_n(x + \varepsilon u) + f_n(x - \varepsilon u) = 2f_n(x) + \varepsilon^2 u^{\mathtt{T}} f_n''(x) u + \tfrac{1}{6}\varepsilon^3 \langle f_n'''(z^+) - f_n'''(z^-),\, u^{\otimes 3}\rangle.$$

For any tensor $T \in \mathbb{R}^{D^3}$, $|\langle T, u^{\otimes 3}\rangle| \leq \|T\|\|u^{\otimes 3}\| = \|T\|$, by the Cauchy–Schwarz inequality. Therefore,

$$|u^{\mathtt{T}} f_n''(x) u| \leq (1/\varepsilon^2)|f_n(x + \varepsilon u) + f_n(x - \varepsilon u) - 2f_n(x)| + \tfrac{1}{3}\varepsilon C.$$

Thus, since (f_n) is pointwise bounded, this implies that $\{u^{\mathsf{T}} f_n''(x) u : n \in \mathbb{N}\}$ is bounded, for any u with $|u| = 1$. Let $u_1, \ldots, u_k \in \mathbb{R}^D$, with $|u_i| = 1$, such that $u_1 u_1^{\mathsf{T}}, \ldots, u_k u_k^{\mathsf{T}}$ is a basis for the vector space $V \subseteq \mathbb{R}^{D \times D}$ of symmetric matrices. (This is possible since $\operatorname{span}\{uu^{\mathsf{T}} : |u| = 1\} = V$ by the spectral decomposition theorem.) With $\langle A, B\rangle := \sum_{i,j} A_{ij} B_{ij}$, V is an inner product space. Since $\{u_i^{\mathsf{T}} f_n''(x) u_i : n \in \mathbb{N}\}$ is bounded for each i, and $u_i^{\mathsf{T}} f_n''(x) u_i = \langle u_i u_i^{\mathsf{T}}, f_n''(x)\rangle$, then by Lemma 31, $\{f_n''(x) : n \in \mathbb{N}\}$ is bounded. Since x is arbitrary, (f_n'') is pointwise bounded. ∎

Lemma 31 *Suppose V is a finite-dimensional inner product space over $\mathbb{R}$, and let $e_1, \ldots, e_k \in V$ be a basis. If $S \subseteq V$ such that $\{\langle e_i, x\rangle : x \in S\}$ is bounded for each $i = 1, \ldots, k$, then S is bounded.*

Proof Let G be the Gram matrix of (e_i), i.e., $G_{ij} = \langle e_i, e_j\rangle$. Note that G is positive definite, since for any $a \in \mathbb{R}^k$,

$$a^{\mathsf{T}} G a = \sum_{i,j} a_i a_j G_{ij} = \sum_{i,j} \langle a_i e_i, a_j e_j\rangle = \big\langle \textstyle\sum_i a_i e_i, \sum_j a_j e_j \big\rangle = \|\textstyle\sum_i a_i e_i\|^2 \geq 0, \tag{C.1}$$

with equality if and only if $\sum_i a_i e_i = 0$, that is, if and only if $a = 0$ (since (e_i) is a linearly independent set). For $x \in V$, define $a(x) \in \mathbb{R}^k$ by the property that $\sum_i a_i(x) e_i = x$ (noting that $a(x)$ always exists and is unique, since (e_i) is a basis). Define $b(x) \in \mathbb{R}^k$ such that $b_i(x) = \langle e_i, x\rangle$. Then for any $x \in V$,

$$b_i(x) = \big\langle e_i, \textstyle\sum_j a_j(x) e_j \big\rangle = \sum_j a_j(x) \langle e_i, e_j\rangle = \sum_j a_j(x) G_{ij},$$

and thus, $b(x) = G a(x)$. Hence, $a(x) = G^{-1} b(x)$, so by Equation C.1,

$$\|x\|^2 = a(x)^{\mathsf{T}} G a(x) = b(x)^{\mathsf{T}} G^{-1} b(x) \leq \|G^{-1}\| |b(x)|^2.$$

By assumption, $\{|b(x)| : x \in S\}$ is bounded, hence, $\{\|x\| : x \in S\}$ is bounded. ∎

Lemma 32 *Let $E \subseteq \mathbb{R}^D$ be open, convex, and bounded. Let $f_n : E \to \mathbb{R}$ have continuous second derivatives. If $f_n \to f$ pointwise for some $f : E \to \mathbb{R}$, and (f_n'') is uniformly bounded, then f' exists and is continuous, and $f_n' \to f'$ uniformly.*

Proof First, we show that (f_n') converges pointwise. Let $C = \sup_n \sup_{x \in E} \|f_n''(x)\| < \infty$. Let $x \in E$, and let $\varepsilon > 0$ small enough that $\overline{B_\varepsilon(x)} \subseteq E$. Then for any $u \in \mathbb{R}^D$ with $|u| = 1$, for any m, n, by applying Taylor's theorem to $f_m - f_n$,

$$f_m(x + \varepsilon u) - f_n(x + \varepsilon u) = f_m(x) - f_n(x) + (f_m'(x) - f_n'(x))^{\mathsf{T}} (\varepsilon u) + \tfrac{1}{2} (\varepsilon u)^{\mathsf{T}} (f_m''(z) - f_n''(z))(\varepsilon u)$$

for some z on the line between x and $x + \varepsilon u$. Thus,

$$|(f_m'(x) - f_n'(x))^{\mathsf{T}} u| \leq \frac{1}{\varepsilon} |f_m(x + \varepsilon u) - f_n(x + \varepsilon u)| + \frac{1}{\varepsilon} |f_m(x) - f_n(x)| + \tfrac{1}{2} \varepsilon \|f_m''(z) - f_n''(z)\|.$$

The first two terms on the right go to zero as $m, n \to \infty$ (by pointwise convergence of f_n), and $\|f''_m(z) - f''_n(z)\| \leq \|f''_m(z)\| + \|f''_n(z)\| \leq 2C$, therefore, $\limsup_{m,n\to\infty} |(f'_m(x) - f'_n(x))^{\mathtt{T}} u| \leq \varepsilon C$. Since ε can be arbitrarily small, $|(f'_m(x) - f'_n(x))^{\mathtt{T}} u| \to 0$ as $m, n \to \infty$. Choosing $u = (1, 0, 0, \ldots, 0)^{\mathtt{T}}$, then $u = (0, 1, 0, \ldots, 0)^{\mathtt{T}}$, and so on, this implies $|f'_m(x) - f'_n(x)| \to 0$ as $m, n \to \infty$, and hence, $f'_n(x)$ converges.

Next, by Lemma 37, (f'_n) is equi-Lipschitz, and hence, equicontinuous. Thus, in fact, (f'_n) converges uniformly, by Lemma 36. Finally, we show that f' exists and $f'_n \to f'$ uniformly; it will follow that f' is continuous, as the limit of a uniformly convergent sequence of continuous functions.

Let $C_{mn} = \sup_{x\in E} |f'_m(x) - f'_n(x)|$. Then $C_{mn} \to 0$ as $m, n \to \infty$, by uniform convergence. To establish the result, it suffices to show that for any $x_0 \in E$, $f'(x_0)$ exists and $f'_n(x_0) \to f'(x_0)$. Fix $x_0 \in E$, and let $B = B_\varepsilon(x_0) \setminus \{x_0\}$ where $\varepsilon > 0$ is small enough that $B \subseteq E$. For $x \in B$, define $\varphi_n(x) = (f_n(x) - f_n(x_0))/|x - x_0|$ and $\varphi(x) = (f(x) - f(x_0))/|x - x_0|$, noting that $\varphi_n \to \varphi$ pointwise. For any $x \in B$, by Taylor's theorem applied to $f_m - f_n$,

$$f_m(x) - f_n(x) = f_m(x_0) - f_n(x_0) + (f'_m(z) - f'_n(z))^{\mathtt{T}}(x - x_0)$$

for some z on the line between x and x_0, and hence,

$$|\varphi_m(x) - \varphi_n(x)| \leq |f'_m(z) - f'_n(z)| \leq C_{mn} \longrightarrow 0$$

as $m, n \to \infty$. Therefore, $\varphi_n \to \varphi$ uniformly (on B) (by e.g., Rudin, 1976, 7.8).

Now, define $\psi_n(x) = f'_n(x_0)^{\mathtt{T}}(x - x_0)/|x - x_0|$ and $\psi(x) = v^{\mathtt{T}}(x - x_0)/|x - x_0|$ for $x \in B$, where $v = \lim_n f'_n(x_0)$. Since $|\psi_n(x) - \psi(x)| \leq |f'_n(x_0) - v| \to 0$ as $n \to \infty$, then $\psi_n \to \psi$ uniformly as well. Hence, $|\varphi_n - \psi_n| \to |\varphi - \psi|$ uniformly (on B).

By the definition of the derivative $f'_n(x_0)$,

$$|\varphi_n(x) - \psi_n(x)| = \frac{|f_n(x) - f_n(x_0) - f'_n(x_0)^{\mathtt{T}}(x - x_0)|}{|x - x_0|} \xrightarrow[x\to x_0]{} 0.$$

Therefore (by e.g., Rudin, 1976, 7.11),

$$\begin{aligned} 0 = \lim_{n\to\infty} \lim_{x\to x_0} |\varphi_n(x) - \psi_n(x)| = \lim_{x\to x_0} \lim_{n\to\infty} |\varphi_n(x) - \psi_n(x)| &= \lim_{x\to x_0} |\varphi(x) - \psi(x)| \\ &= \lim_{x\to x_0} \frac{|f(x) - f(x_0) - v^{\mathtt{T}}(x - x_0)|}{|x - x_0|}. \end{aligned}$$

Hence, $f'(x_0)$ exists and equals $v = \lim_n f'_n(x_0)$. ∎

Proof of Theorem 7 First, suppose (f_n) is pointwise bounded. By Lemma 37 with $k = 3$, (f''_n) is equi-Lipschitz, and by Lemma 30, (f''_n) is pointwise bounded. Thus, since E is bounded, it follows that (f''_n) is uniformly bounded. Therefore, by Lemma 37 with $k = 2$, (f'_n) is equi-Lipschitz, and by Lemma 29, (f'_n) is pointwise bounded. Thus, likewise, (f'_n) is uniformly bounded. And lastly, applying Lemma 37 with $k = 1$, we have that (f_n) is equi-Lipschitz, and hence, uniformly bounded, since it is pointwise bounded by assumption.

Now, assume $f_n \to f$ pointwise. Then in fact, $f_n \to f$ uniformly, by Lemma 36, since (f_n) is equi-Lipschitz (as just established), and hence, equicontinuous. By Lemma 32, f'

exists and $f_n' \to f'$ uniformly. To complete the proof, we show that f'' exists and $f_n'' \to f''$ uniformly. For any $i \in \{1, \ldots, D\}$, if we define $h_n(x) = f_n'(x)_i$ and $h(x) = f'(x)_i$, then $h_n \to h$ pointwise and (h_n'') is uniformly bounded (since (f_n''') is uniformly bounded and $\|h_n''(x)\| \leq \|f_n'''(x)\|$); hence, by Lemma 32, h' exists and is continuous, and $h_n' \to h'$ uniformly. Since this holds for each coordinate i, then f'' exists, and $f_n'' \to f''$ uniformly. ∎

Appendix D. Proofs of Coverage Results

Proof of Theorem 8 Letting $X_n = -\sqrt{n}(\theta_n - \theta_0)$ and $X \sim Q$,

$$\mathbb{P}(\theta_0 \in S_n) \overset{\text{(a)}}{=} \mathbb{P}(\sqrt{n}(\theta_0 - \theta_n) \in R_n) = \mathbb{P}(X_n \in R_n) \xrightarrow{\text{(b)}} \mathbb{P}(X \in R) = Q(R)$$

where step (a) is by the definition of R_n, and (b) is by Lemma 9, using assumptions 1 ($X_n \xrightarrow{\text{D}} X$), 3, and 4. To see that $Q(R) = \rho$, note that $\Pi_n(S_n) \xrightarrow{\text{a.s.}} \rho$ by assumption and also $\Pi_n(S_n) = Q_n(R_n) \xrightarrow{\text{a.s.}} Q(R)$ since

$$\begin{aligned}
|Q_n(R_n) - Q(R)| &\leq |Q_n(R_n) - Q(R_n)| + |Q(R_n) - Q(R)| \\
&\leq \sup_{A \in \mathcal{B}} |Q_n(A) - Q(A)| + |Q(R_n) - Q(R)| \xrightarrow{\text{a.s.}} 0
\end{aligned}$$

by assumption 2 and assumption 3 plus the dominated convergence theorem (Folland, 2013, Theorem 2.24). ∎

Proof of Lemma 9 For each $k = 1, 2, \ldots$, define $A_k = \{x \in \mathbb{R}^D : d(x, R^c) > 1/k\}$ and $B_k = \{x \in \mathbb{R}^D : d(x, R) \leq 1/k\}$. Note that A_k is open and B_k is closed since $x \mapsto d(x, R)$ and $x \mapsto d(x, R^c)$ are continuous. For any k, by Lemma 10 we have that with probability 1, for all n sufficiently large, $A_k \subseteq R_n \subseteq B_k$. Thus, with probability 1, $\liminf_n \big(\mathbb{1}(X_n \in R_n) - \mathbb{1}(X_n \in A_k)\big) \geq \liminf_n \inf_x \big(\mathbb{1}(x \in R_n) - \mathbb{1}(x \in A_k)\big) \geq 0$. It follows that

$$\liminf_n \mathrm{E}\big(\mathbb{1}(X_n \in R_n) - \mathbb{1}(X_n \in A_k)\big) \geq \mathrm{E} \liminf_n \big(\mathbb{1}(X_n \in R_n) - \mathbb{1}(X_n \in A_k)\big) \geq 0$$

by Fatou's lemma applied to $\mathbb{1}(X_n \in R_n) - \mathbb{1}(X_n \in A_k) + 1$. (The $+1$ is added to make the function nonnegative, so that Fatou's lemma applies directly.) If $\liminf_n (a_n - b_n) \geq 0$ then $\liminf a_n = \liminf(a_n - b_n + b_n) \geq \liminf(a_n - b_n) + \liminf b_n \geq \liminf b_n$. Therefore, $\liminf_{n\to\infty} \mathbb{P}(X_n \in A_k) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \in R_n)$. Similarly, by reverse Fatou's lemma,

$$\limsup_n \mathrm{E}\big(\mathbb{1}(X_n \in R_n) - \mathbb{1}(X_n \in B_k)\big) \leq \mathrm{E} \limsup_n \big(\mathbb{1}(X_n \in R_n) - \mathbb{1}(X_n \in B_k)\big) \leq 0,$$

and therefore, $\limsup_n \mathbb{P}(X_n \in R_n) \leq \limsup_n \mathbb{P}(X_n \in B_k)$. Hence, by the portmanteau theorem (Dudley, 2002, Theorem 11.1.1), for all k,

$$\begin{aligned}
\mathbb{P}(X \in A_k) &\leq \liminf_n \mathbb{P}(X_n \in A_k) \leq \liminf_n \mathbb{P}(X_n \in R_n) \\
&\leq \limsup_n \mathbb{P}(X_n \in R_n) \leq \limsup_n \mathbb{P}(X_n \in B_k) \leq \mathbb{P}(X \in B_k).
\end{aligned}$$

Taking limits as $k \to \infty$ and using the fact that $\bigcup_{k=1}^{\infty} A_k = R^\circ$ and $\bigcap_{k=1}^{\infty} B_k = \bar{R}$, we have $\mathbb{P}(X \in R^\circ) = \lim_k \mathbb{P}(X \in A_k) \leq \liminf_n \mathbb{P}(X_n \in R_n) \leq \limsup_n \mathbb{P}(X_n \in R_n) \leq \lim_k \mathbb{P}(X \in B_k) = \mathbb{P}(X \in \bar{R})$ by Folland (2013, Theorem 1.8). Further, $\mathbb{P}(X \in R^\circ) = \mathbb{P}(X \in R) = \mathbb{P}(X \in \bar{R})$ since $\mathbb{P}(X \in \partial R) = 0$. Therefore, $\lim_n \mathbb{P}(X_n \in R_n) = \mathbb{P}(X \in R)$. ■

Proof of Lemma 10 First, we establish some initial facts. It is straightforward to check that R is convex. R° is nonempty since $m(\bar{R}) \geq m(R) > 0$ and $m(\partial R) = 0$ (Lang, 1986). It follows that R, A, and B are bounded. For any open cube E such that $\bar{E} \subseteq R$, we have $E \subseteq R_n$ for all n sufficiently large, since $\mathbb{1}(x \in R_n) \to \mathbb{1}(x \in R)$ for each corner x of the cube E.

Next, we show that $A \subseteq R_n$ for all n sufficiently large. For each $x \in \bar{A}$, let E_x be a nonempty open cube centered at x such that $\bar{E}_x \subseteq R$. Then $\{E_x : x \in \bar{A}\}$ is an open cover of $\bar{A}$. Since $\bar{A}$ is compact, there is a finite subcover $E_{x_1}, \ldots, E_{x_k}$. Thus, for all n sufficiently large, $A \subseteq \bar{A} \subseteq \bigcup_{i=1}^{k} E_{x_i} \subseteq R_n$.

Now, we show that $R_n \subseteq B$ for all n sufficiently large. Let $S_\delta = \{x \in \mathbb{R}^D : d(x, R) = \delta\}$ for $\delta > 0$. Let $E \subseteq R$ be a nonempty open cube such that $E \subseteq R_n$ for all n sufficiently large. For each $x \in S_{\varepsilon/2}$, define $C_x = \bigcup_{t>1}\{tx + (1-t)z : z \in E\}$. Then C_x is open, as a union of open sets. Note that $y \in C_x$ if and only if $x = sy + (1-s)z$ for some $s \in (0,1)$, $z \in E$, i.e., if and only if x is a (strict) convex combination of y and some point of E. Thus, $\{C_x : x \in S_{\varepsilon/2}\}$ is an open cover of S_ε (since for any $y \in S_\varepsilon$, the line between y and any $z \in E$ must pass through $S_{\varepsilon/2}$ by the intermediate value theorem applied to $s \mapsto d(sx + (1-s)z, R)$). Since S_ε is compact, there is a finite subcover $C_{x_1}, \ldots, C_{x_k}$ for some $x_1, \ldots, x_k \in S_{\varepsilon/2}$. Since $x_i \in R^c$ for each $i = 1, \ldots, k$, there exists N such that for all $n \geq N$, $x_1, \ldots, x_k \in R_n^c$ and $E \subseteq R_n$. Then for all $n \geq N$, by the convexity of R_n, we have $S_\varepsilon \subseteq \bigcup_{i=1}^{k} C_{x_i} \subseteq R_n^c$ and hence $R_n \subseteq B$. ■

Appendix E. Proofs for Generalized Applications

Proof of Theorem 20 Define $B_r = \{j \in \mathbb{Z}^m : R(j) \leq r\}$, that is, $B_r = \{-r, \ldots, -1, 0, 1, \ldots, r\}^m$ for $r \in \mathbb{N}$. Let $r_n = R(v(n))$, $L_n = |B_{r_n - 1}|$, and $M_n = |B_{r_n}|$, for $n \in \mathbb{N}$. Observe that $L_n < n \leq M_n$ since $R(v(1)) \leq R(v(2)) \leq \cdots$. Further, $M_n/n \to 1$ as $n \to \infty$, since

$$1 \leq \frac{M_n}{n} \leq \frac{M_n}{L_n} = \frac{|B_{r_n}|}{|B_{r_n-1}|} = \Big(\frac{2r_n + 1}{2r_n - 1}\Big)^m \longrightarrow 1$$

as $n \to \infty$ since $r_n \to \infty$.

Fix $k, \ell \in \{1, \ldots, m\}$, and define $Z_i = Y_i X_{ik} - \mathrm{E}(Y_i X_{ik})$ or $Z_i = X_{ik} X_{i\ell} - \mathrm{E}(X_{ik} X_{i\ell})$ where X_i is defined as in Theorem 19; the proof is the same in either case. Then $Z_1, Z_2, \ldots$ are identically distributed, and in fact, by Lemma 33, $(Z_{v^{-1}(j)} : j \in \mathbb{Z}^m)$ is stationary with respect to $T_1, \ldots, T_m$ and ergodic with respect to at least one of $T_1, \ldots, T_m$. Note that $\mathrm{E}Z_i = 0$ and $\mathrm{E}|Z_i|^2 = \mathrm{Var}(Z_i) < \infty$, since for all i, j, $\mathrm{Var}(Y_i Y_j) \leq \mathrm{E}|Y_i Y_j|^2 \leq (\mathrm{E}|Y_i|^4 \mathrm{E}|Y_j|^4)^{1/2} < \infty$ by the Cauchy–Schwarz inequality.

To prove the result, we need to show that $\frac{1}{n}\sum_{i=1}^{n} Z_i \to 0$ almost surely as $n \to \infty$. The key part is showing that (A) $\frac{1}{M_n}\sum_{i=1}^{M_n} Z_i \to 0$ by the multivariate ergodic theorem; the remainder of the proof is showing that (B) the difference between $\frac{1}{n}\sum_{i=1}^n Z_i$ and $\frac{1}{M_n}\sum_{i=1}^{M_n} Z_i$ is negligible.

(A) For the first part, letting $\mathcal{F}_k$ denote the invariant sigma-field with respect to T_k,

$$\begin{aligned}\frac{1}{M_n}\sum_{i=1}^{M_n} Z_i &= \frac{1}{|B_{r_n}|}\sum_{j\in B_{r_n}} Z_{v^{-1}(j)}\\ &= \frac{1}{(2r_n+1)^m}\sum_{j_1=-r_n}^{r_n}\cdots\sum_{j_m=-r_n}^{r_n} Z_{v^{-1}(j_1,\ldots,j_m)} \qquad \text{(E.1)}\\ &\xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}^{\mathcal{F}_m}\cdots\mathrm{E}^{\mathcal{F}_1}Z_1 = \mathrm{E}(Z_1 \mid \cap_k\mathcal{F}_k) = \mathrm{E}Z_1 = 0\end{aligned}$$

by applying the multivariate ergodic theorem (Kallenberg, 2002, Theorems 10.12 and 10.13) to each of the 2^m orthants of $\mathbb{Z}^m$. A few clarifying remarks. The subset $\{j \in \mathbb{Z}^m : \min\{|j_1|,\ldots,|j_m|\} = 0\}$ can be handled by shifting each orthant to ensure that, collectively, they form a partition of $\mathbb{Z}^m$. Kallenberg (2002, Theorem 10.12) shows that the limit is $\mathrm{E}^{\mathcal{F}_m}\cdots\mathrm{E}^{\mathcal{F}_1}Z_1$, and since $T_1,\ldots,T_m$ are commutative, we have $\mathrm{E}^{\mathcal{F}_m}\cdots\mathrm{E}^{\mathcal{F}_1}Z_1 = \mathrm{E}(Z_1 \mid \cap_k\mathcal{F}_k)$ by Kallenberg (2002, Theorem 10.13). Ergodicity with respect to T_k means that the corresponding invariant sigma-field $\mathcal{F}_k$ is trivial, that is, the probability of any set $A \in \mathcal{F}_k$ is either 0 or 1. So by assumption, at least one of $\mathcal{F}_1,\ldots,\mathcal{F}_m$ is trivial, and hence, $\cap_k\mathcal{F}_k$ is trivial as well. Therefore, $\mathrm{E}(Z_1 \mid \cap_k\mathcal{F}_k) = \mathrm{E}Z_1 = 0$.

(B) Now, for the remainder of the proof,

$$\Big|\frac{1}{M_n}\sum_{i=1}^{n} Z_i\Big| \le \Big|\frac{1}{M_n}\sum_{i=1}^{n} Z_i - \frac{1}{M_n}\sum_{i=1}^{M_n} Z_i\Big| + \Big|\frac{1}{M_n}\sum_{i=1}^{M_n} Z_i\Big|.$$

As for the second term, we have $\frac{1}{M_n}\sum_{i=1}^{M_n} Z_i \to 0$ a.s. by Equation E.1. As for the first term, we have

$$\Big|\frac{1}{M_n}\sum_{i=1}^{n} Z_i - \frac{1}{M_n}\sum_{i=1}^{M_n} Z_i\Big| \le \frac{1}{M_n}\sum_{i=n+1}^{M_n}|Z_i| \le \frac{1}{M_n}\sum_{i=L_n+1}^{M_n}|Z_i| = c_{r_n}W_{r_n}$$

where $c_r = |B_r \setminus B_{r-1}|/|B_r|$, $W_r = \frac{1}{|S_r|}\sum_{i\in S_r}|Z_i|$, and $S_r = \{L_n+1,\ldots,M_n\}$. If we can show that $c_{r_n}W_{r_n} \to 0$ almost surely as $n\to\infty$, then this will prove the result, since then $\frac{1}{n}\sum_{i=1}^n Z_i = \big(\frac{M_n}{n}\big)\frac{1}{M_n}\sum_{i=1}^n Z_i \to 0$ a.s., using the fact that $M_n/n \to 1$ as shown above.

We show that $c_{r_n}W_{r_n} \to 0$ using the Borel–Cantelli lemma. For all $\varepsilon > 0$, $r\in\mathbb{N}$, by Markov's inequality we have

$$\mathbb{P}(c_rW_r \ge \varepsilon) = \mathbb{P}(W_r^2 \ge (\varepsilon/c_r)^2) \le (c_r/\varepsilon)^2\mathrm{E}(W_r^2) \le (c_r/\varepsilon)^2\mathrm{Var}(Z_1), \qquad \text{(E.2)}$$

where the last step holds since $\mathrm{E}(W_r^2) \le \mathrm{E}\big(\frac{1}{|S_r|}\sum_{i\in S_r}|Z_i|^2\big) = \mathrm{E}|Z_1|^2 = \mathrm{Var}(Z_1)$ by Jensen's inequality. Now, for all $r\in\mathbb{N}$, we have the bound

$$c_r = \frac{(2r+1)^m - (2r-1)^m}{(2r+1)^m} \le \frac{(2r-1)^{m-1}3^m}{(2r+1)^m} \le \frac{3^m}{r} \qquad \text{(E.3)}$$

where the first inequality holds by the following application of the binomial theorem, taking $x = 2r - 1$ and $y = 2$: for all $m \in \mathbb{N}$, $x \geq 1$, $y \geq 0$,

$$(x+y)^m - x^m = \sum_{k=1}^{m} \binom{m}{k} x^{m-k} y^k \leq x^{m-1} \sum_{k=0}^{m} \binom{m}{k} y^k = x^{m-1}(1+y)^m.$$

Therefore, combining Equations E.2 and E.3, we have that for all $\varepsilon > 0$,

$$\sum_{r=1}^{\infty} \mathbb{P}(c_r W_r \geq \varepsilon) \leq \sum_{r=1}^{\infty} (c_r/\varepsilon)^2 \mathrm{Var}(Z_1) \leq (3^m/\varepsilon)^2 \mathrm{Var}(Z_1) \sum_{r=1}^{\infty} \frac{1}{r^2} < \infty.$$

Hence, by the Borel–Cantelli lemma, for all $\varepsilon > 0$, $\mathbb{P}(\mathrm{limsup}_r\, c_r W_r \leq \varepsilon) = 1$. Therefore, $\mathbb{P}(\lim c_r W_r = 0) = \mathbb{P}(\cap_{k=1}^{\infty} \{\mathrm{limsup}_r\, c_r W_r \leq 1/k\}) = 1$, that is, $c_r W_r \to 0$ a.s. as $r \to \infty$. ■

In the following lemma, we adopt the notational conventions and definitions of Kallenberg (2002), page 181.

Lemma 33 *Let ξ be a random element in S with distribution μ, and let $T : S \to S$ be μ-preserving (that is, $T(\xi) \overset{d}{=} \xi$). Suppose $f : S \to S$ is a measurable function such that $f \circ T = T \circ f$. Then $T(f(\xi)) \overset{d}{=} f(\xi)$, that is, T preserves the distribution of $f(\xi)$. If, further, ξ is T-ergodic, then $f(\xi)$ is T-ergodic.*

Proof The first part is immediate, since $T(f(\xi)) = f(T(\xi)) \overset{d}{=} f(\xi)$. Suppose ξ is T-ergodic. In other words, suppose that for any measurable set A such that $T^{-1}(A) = A$, we have $\mathbb{P}(\xi \in A) \in \{0, 1\}$. (This is also equivalent to saying that the T-invariant sigma-algebra is trivial under μ.) To show that $f(\xi)$ is T-ergodic, let A such that $T^{-1}(A) = A$. Then $T^{-1}(f^{-1}(A)) = f^{-1}(T^{-1}(A)) = f^{-1}(A)$ since $f \circ T = T \circ f$, so we have $\mathbb{P}(f(\xi) \in A) = \mathbb{P}(\xi \in f^{-1}(A)) \in \{0, 1\}$. Hence, $f(\xi)$ is T-ergodic. ■

In the proof of Theorem 20, we apply Lemma 33 in the following way. Suppose ξ is a real-valued stochastic process on $\mathbb{Z}^m$, that is, $\xi = (\xi(i_1, \ldots, i_m) : i \in \mathbb{Z}^m)$ where $\xi(i_1, \ldots, i_m)$ is a real-valued random variable. Suppose T_k is the shift transformation in coordinate k, that is, $T_k \xi = T_k(\xi) = (\xi(i_1, \ldots, i_k + 1, \ldots, i_m) : i \in \mathbb{Z}^m)$. Let $\varphi(\xi) \in \mathbb{R}$ be a measurable function of ξ, and define $f(\xi) = (\varphi(T_1^{j_1} \cdots T_m^{j_m} \xi) : j \in \mathbb{Z}^m)$. Then $f(T_k \xi) = (\varphi(T_1^{j_1} \cdots T_k^{j_k+1} \cdots T_m^{j_m} \xi) : j \in \mathbb{Z}^m) = T_k f(\xi)$ and thus, $f \circ T_k = T_k \circ f$. Hence, if ξ is stationary with respect to T_k (that is, $T_k \xi \overset{d}{=} \xi$) and ξ is T_k-ergodic, then by Lemma 33, $f(\xi)$ is stationary with respect to T_k and is T_k-ergodic.

Proof of Theorem 24 For $\theta \in \mathbb{R}^D$,

$$f_n(\theta) = -\frac{1}{n} \log \mathcal{L}_n^{\mathrm{Cox}}(\theta) - \frac{1}{n} \sum_{i=1}^{n} Z_i \log n = \frac{1}{n} \sum_{i=1}^{n} H_{Y_i}^n(\theta) Z_i - \theta^{\mathsf{T}} \big(\tfrac{1}{n} \textstyle\sum_{i=1}^n X_i Z_i\big)$$

where $H_y^n(\theta) = \log\big(\frac{1}{n} \sum_{j=1}^n \exp(\theta^{\mathsf{T}} X_j) \mathbb{1}(Y_j \geq y)\big)$. Note that f_n is C^∞, as a composition of C^∞ functions. Further, f_n is convex on $\mathbb{R}^D$, since $H_{Y_i}^n(\theta)$ is convex by Lemma 41 with $\mu = \frac{1}{n} \sum_{j: Y_j \geq Y_i} \delta_{X_j}$. By Lemma 35, $f''(\theta_0)$ is positive definite.

By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^n X_i Z_i \xrightarrow{\text{a.s.}} \mathrm{E}(XZ)$ as $n \to \infty$, and by Lemma 34, for all $\theta \in \mathbb{R}^D$, $\mathrm{E}|h_Y(\theta)Z| < \infty$ and $\frac{1}{n}\sum_{i=1}^n H^n_{Y_i}(\theta) Z_i \xrightarrow{\text{a.s.}} \mathrm{E}\big(h_Y(\theta)Z\big)$ as $n \to \infty$. Therefore, for all $\theta \in \mathbb{R}^D$, with probability 1, $f_n(\theta) \to f(\theta)$. Due to convexity, this implies that with probability 1, for all $\theta \in \mathbb{R}^D$, $f_n(\theta) \to f(\theta)$.

Let $m = \sup\{|x| : x \in \mathcal{X}\} < \infty$. Then by Lemma 41, $\big|(\partial^3/\partial\theta_j\partial\theta_k\partial\theta_\ell) H^n_{Y_i}(\theta)\big| \leq (2m)^3 = 8m^3$ for all $\theta \in \mathbb{R}^D$. Thus, $\|f_n'''(\theta)\|^2 = \sum_{j,k,\ell} \big|(\partial^3/\partial\theta_j\partial\theta_k\partial\theta_\ell) f_n(\theta)\big|^2 \leq D^3(8m^3)^2$ for all $\theta \in \mathbb{R}^D$, $n \in \mathbb{N}$. Hence, (f_n''') is a.s. uniformly bounded on all of $\mathbb{R}^D$. Thus, for any open ball E containing θ_0, the conditions of Theorem 5 are satisfied with probability 1. ■

Note that $H^n_{Y_1}(\theta), H^n_{Y_2}(\theta), \ldots$ are not i.i.d., which is why the next lemma is not trivial.

Lemma 34 *Suppose $(X,Y,Z), (X_1,Y_1,Z_1), (X_2,Y_2,Z_2), \ldots$ are i.i.d., where $X \in \mathcal{X} \subseteq \mathbb{R}^D$, $Y \geq 0$, and $Z \in \{0,1\}$. Define $h_y(\theta) = \log \mathrm{E}\big(\exp(\theta^\mathtt{T} X)\mathbb{1}(Y \geq y)\big)$ and $H^n_y(\theta) = \log\big(\frac{1}{n}\sum_{j=1}^n \exp(\theta^\mathtt{T} X_j)\mathbb{1}(Y_j \geq y)\big)$ for $\theta \in \mathbb{R}^D$, $y \geq 0$. If $\mathcal{X}$ is bounded and the c.d.f. of Y is continuous on $\mathbb{R}$, then for all $\theta \in \mathbb{R}^D$, $\mathrm{E}|h_Y(\theta)Z| < \infty$ and*

$$\frac{1}{n}\sum_{i=1}^n H^n_{Y_i}(\theta) Z_i \xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}\big(h_Y(\theta)Z\big).$$

Proof Let $F(y) = \mathbb{P}(Y \leq y)$, $c^* = \sup\{y \in \mathbb{R} : F(y) < 1\}$, and $m = \sup\{|x| : x \in \mathcal{X}\} < \infty$. Since $|X| \leq m$ and F is continuous, $\mathrm{E}|h_Y(\theta)Z| \leq m|\theta| - \mathrm{E}\log(1 - F(Y)) = m|\theta| + 1$ because $F(Y) \sim \mathrm{Uniform}(0,1)$. Fix $\theta \in \mathbb{R}^D$ and define $g(y) = h_y(\theta)$ and $G_n(y) = H^n_y(\theta)$.

First, we show that for all $c \in (0, c^*)$,

$$\sup_{y \in [0,c]} |G_n(y) - g(y)| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{E.4}$$

Let S be a countable dense subset of $[0,c]$ such that $0, c \in S$. For all $y \in S$, $G_n(y) \xrightarrow{\text{a.s.}} g(y) \in \mathbb{R}$ by the strong law of large numbers since $0 < \mathrm{E}(e^{\theta^\mathtt{T} X}\mathbb{1}(Y \geq y)) \leq e^{m|\theta|} < \infty$. Next, G_n is a non-increasing function on $[0,c]$ (that is, if $0 \leq y < y' \leq c$ then $G_n(y) \geq G_n(y')$) since $y \mapsto \mathbb{1}(Y_j \geq y)$ is non-increasing. Further, $g(y)$ is continuous on $[0,c]$ by the dominated convergence theorem, since $|e^{\theta^\mathtt{T} X}\mathbb{1}(Y \geq y)| \leq e^{m|\theta|}$ and $\mathbb{P}(Y = y) = 0$ by the continuity of F. Thus, with probability 1, for all n sufficiently large, G_n is finite on $[0,c]$ since $G_n(0) \xrightarrow{\text{a.s.}} g(0)$ and $G_n(c) \xrightarrow{\text{a.s.}} g(c)$. It follows that $\sup_{y\in[0,c]} |G_n(y) - g(y)| \xrightarrow{\text{a.s.}} 0$ by Lemma 39.

Second, we show that for all $c \in (0, c^*)$,

$$\frac{1}{n}\sum_{i=1}^n G_n(Y_i) Z_i \mathbb{1}(Y_i \leq c) \xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}\big(g(Y) Z \mathbb{1}(Y \leq c)\big). \tag{E.5}$$

To see this, observe that by Equation E.4,

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n G_n(Y_i) Z_i \mathbb{1}(Y_i \leq c) - \frac{1}{n}\sum_{i=1}^n g(Y_i) Z_i \mathbb{1}(Y_i \leq c)\Big| \\
&\qquad \leq \frac{1}{n}\sum_{i=1}^n |G_n(Y_i) - g(Y_i)| \mathbb{1}(Y_i \leq c) \leq \sup_{y\in[0,c]} |G_n(y) - g(y)| \xrightarrow[n\to\infty]{\text{a.s.}} 0
\end{aligned}$$

and $\frac{1}{n}\sum_{i=1}^n g(Y_i)Z_i\mathbb{1}(Y_i \le c) \xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}\big(g(Y)Z\mathbb{1}(Y\le c)\big)$ by the strong law of large numbers.

Third, we show that for all $c \in (0, c^*)$,

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i\mathbb{1}(Y_i > c)\right| \overset{\text{a.s.}}{\le} m|\theta|p_c - p_c\log p_c + p_c \tag{E.6}$$

where $p_c = \mathbb{P}(Y > c)$. This follows from the fact that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i\mathbb{1}(Y_i > c)\right| &\le \frac{1}{n}\sum_{i=1}^n |G_n(Y_i)|\mathbb{1}(Y_i > c)\\ &\le \frac{1}{n}\sum_{i=1}^n \Big(m|\theta| - \log\big(\tfrac{1}{n}\textstyle\sum_{j=1}^n \mathbb{1}(Y_j \ge Y_i)\big)\Big)\mathbb{1}(Y_i > c)\\ &\overset{\text{a.s.}}{=} m|\theta|K_n/n - \frac{1}{n}\sum_{k=1}^{K_n}\log(k/n)\\ &\xrightarrow[n\to\infty]{\text{a.s.}} m|\theta|p_c - \int_0^{p_c}(\log x)dx = m|\theta|p_c - p_c\log p_c + p_c\end{aligned}$$

where $K_n = \sum_{i=1}^n \mathbb{1}(Y_i > c)$, using that $\mathbb{P}(Y_i = Y_j) = 0$ for $i \ne j$ by continuity of F.

Now, we put these pieces together to obtain the result. Writing $\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i = \frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i\mathbb{1}(Y_i \le c) + \frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i\mathbb{1}(Y_i > c)$, for all $c \in (0, c^*)$ we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i - \mathrm{E}(g(Y)Z)\right| &\le \left|\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i\mathbb{1}(Y_i\le c) - \mathrm{E}\big(g(Y)Z\mathbb{1}(Y\le c)\big)\right|\\ &\quad + \Big|\mathrm{E}\big(g(Y)Z\mathbb{1}(Y\le c)\big) - \mathrm{E}(g(Y)Z)\Big| + \left|\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i\mathbb{1}(Y_i > c)\right|,\end{aligned}$$

and therefore, by Equations E.5 and E.6,

$$\begin{aligned}&\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i - \mathrm{E}(g(Y)Z)\right|\\ &\quad\overset{\text{a.s.}}{\le} \big|\mathrm{E}\big(g(Y)Z\mathbb{1}(Y\le c)\big) - \mathrm{E}(g(Y)Z)\big| + m|\theta|p_c - p_c\log p_c + p_c.\end{aligned} \tag{E.7}$$

Let $c_1, c_2, \ldots \in (0, c^*)$ such that $c_k \to c^*$. Then $p_{c_k} \to p_{c^*} = 0$ by continuity of F, and thus, $m|\theta|p_{c_k} - p_{c_k}\log p_{c_k} + p_{c_k} \to 0$ as $k\to\infty$. Further, $\mathrm{E}\big(g(Y)Z\mathbb{1}(Y\le c_k)\big) \to \mathrm{E}(g(Y)Z)$ by the dominated convergence theorem, since $|g(Y)Z\mathbb{1}(Y\le c_k)| \le |g(Y)Z|$, $\mathrm{E}|g(Y)Z| < \infty$, and $\mathbb{1}(Y\le c_k) \overset{\text{a.s.}}{\to} 1$ as $k\to\infty$. Applying Equation E.7 to each c_k and taking limits as $k\to\infty$, we have that $\limsup_{n\to\infty}\big|\frac{1}{n}\sum_{i=1}^n G_n(Y_i)Z_i - \mathrm{E}(g(Y)Z)\big| = 0$ almost surely. ■

Lemma 35 *Under the conditions of Theorem 24, $f''(\theta)$ is positive definite for all $\theta\in\mathbb{R}^D$.*

Proof Recall that $f(\theta) = \mathrm{E}\big(h_Y(\theta)Z\big) - \theta^{\mathsf{T}}\mathrm{E}(XZ)$ where $h_y(\theta) = \log\mathrm{E}(e^{\theta^{\mathsf{T}}X}\mathbb{1}(Y\ge y))$ for $\theta\in\mathbb{R}^D$. First, we put $h_y(\theta)$ in the form of $\kappa(\theta)$ in Lemma 41 by noting that $h_y(\theta) =$

$\log \mathrm{E}(e^{\theta^\mathsf{T} X}\mathbb{P}(Y \geq y \mid X)) = \log \int \exp(\theta^\mathsf{T} x)\mu_y(dx)$ where $\mu_y(dx) = \mathbb{P}(Y \geq y \mid X = x)P(dx)$ and P is the distribution of X (Dudley, 2002, 10.2.1-10.2.2). Let $m = \sup\{|x| : x \in \mathcal{X}\} < \infty$. We have $|h_y(\theta)| < \infty$ for all $\theta \in \mathbb{R}^D$ and all $y \geq 0$ because $\exp(-m|\theta|) \leq \exp(\theta^\mathsf{T} X) \leq \exp(m|\theta|)$, and thus $-\infty < -m|\theta| + \log \mathbb{P}(Y \geq y) \leq h_y(\theta) \leq m|\theta| + \log \mathbb{P}(Y \geq y) < \infty$ due to assumptions 1 and 4 of Theorem 24.

For any given $\theta \in \mathbb{R}^D$ and $y \geq 0$, following Lemma 41, we define a probability measure $\tilde{P} = \tilde{P}_{\theta,y}$ on $\mathcal{X}$ by $\tilde{P}(dx) = \exp(\theta^\mathsf{T} x - h_y(\theta))\mathbb{P}(Y \geq y \mid X = x)P(dx)$. Note that P and $\tilde{P}$ are mutually absolutely continuous since $\exp(\theta^\mathsf{T} x - h_y(\theta))\mathbb{P}(Y \geq y \mid X = x)$ is strictly positive for all $x \in \mathcal{X}$. By Lemma 41, $h_y'(\theta) = \mathrm{E}(\tilde{X})$ and $h_y''(\theta) = \mathrm{Cov}(\tilde{X})$ where $\tilde{X} \sim \tilde{P}$. We claim that for any nonzero $a \in \mathbb{R}^D$, $a^\mathsf{T} h_y''(\theta) a > 0$. To see this, suppose $a \in \mathbb{R}^D$ such that $a^\mathsf{T} h_y''(\theta)a = 0$. Since $a^\mathsf{T} h_y''(\theta)a = \mathrm{Var}(a^\mathsf{T}\tilde{X})$, it follows that $\mathbb{P}(a^\mathsf{T}\tilde{X} = \mathrm{E}(a^\mathsf{T}\tilde{X})) = 1$. But then $\mathbb{P}(a^\mathsf{T} X = \mathrm{E}(a^\mathsf{T}\tilde{X})) = 1$ since $P \ll \tilde{P}$. Hence, $a^\mathsf{T} X$ is a.s. equal to a constant, so $\mathrm{Var}(a^\mathsf{T} X) = 0$, which implies $a = 0$ by assumption 3 of Theorem 24.

To justify differentiating under the expectation in $\mathrm{E}(h_Y(\theta)Z)$, we apply Folland (2013, Theorem 2.27b) using the following bounds. First, $\mathrm{E}|h_Y(\theta)Z| < \infty$ by Lemma 34. Next, $|\tilde{X}| \leq m$ because $\tilde{P}$ is supported on $\mathcal{X}$. Thus, $|\frac{\partial}{\partial\theta_j} h_y(\theta)z| = |\mathrm{E}(\tilde{X}_j)z| \leq \mathrm{E}|\tilde{X}_j| \leq \mathrm{E}|\tilde{X}| \leq m$ and $|\frac{\partial^2}{\partial\theta_j\partial\theta_k} h_y(\theta)z| = |\mathrm{Cov}(\tilde{X}_j, \tilde{X}_k)z| \leq \mathrm{E}|\tilde{X}_j||\tilde{X}_k| + \mathrm{E}|\tilde{X}_j|\mathrm{E}|\tilde{X}_k| \leq 2m^2$ for $z \in \{0,1\}$.

Hence, $f''(\theta) = \mathrm{E}\big(h_Y''(\theta)Z\big)$, and we have that for any nonzero $a \in \mathbb{R}^D$, $a^\mathsf{T} f''(\theta)a = \mathrm{E}\big(a^\mathsf{T} h_Y''(\theta)aZ\big) > 0$ because $a^\mathsf{T} h_Y''(\theta)a > 0$ and $\mathbb{P}(Z = 1) > 0$ due to assumption 3 of Theorem 24. Therefore, $f''(\theta)$ is positive definite. ■

Appendix F. Supporting Results

This section contains miscellaneous supporting results used in the proofs. A metric space E is *totally bounded* if for any $\delta > 0$, there exist $x_1, \ldots, x_k \in E$, for some $k \in \mathbb{N}$, such that $E = \bigcup_{i=1}^k \{x \in E : d(x, x_i) < \delta\}$. In particular, any bounded subset of a Euclidean space is totally bounded.

Lemma 36 *Suppose $h_n : E \to F$ for $n \in \mathbb{N}$, where E is a totally bounded metric space and F is a normed space. If (h_n) converges pointwise and is equicontinuous, then it converges uniformly.*

Proof Let $\varepsilon > 0$. Choose $\delta > 0$ by equicontinuity, so that for any $n \in \mathbb{N}$, $x, y \in E$, if $d(x, y) < \delta$ then $\|h_n(x) - h_n(y)\| < \varepsilon$. Choose $x_1, \ldots, x_k \in E$ by totally boundedness, and by pointwise convergence, let N such that for all $m, n > N$, for all $i \in \{1, \ldots, k\}$, $\|h_m(x_i) - h_n(x_i)\| < \varepsilon$. Then, for any $x \in E$, there is some $i \in \{1, \ldots, k\}$ such that $d(x, x_i) < \delta$, and thus

$$\|h_m(x) - h_n(x)\| \leq \|h_m(x) - h_m(x_i)\| + \|h_m(x_i) - h_n(x_i)\| + \|h_n(x_i) - h_n(x)\| < 3\varepsilon$$

for any $m, n > N$. Therefore, (h_n) converges uniformly (by e.g., Rudin, 1976, 7.8). ■

When all the kth order partial derivatives of f exist, let $f^{(k)}(x)$ denote the k-way tensor of kth derivatives; in particular, $f^{(1)} = f'$, $f^{(2)} = f''$, and so on. When these derivatives are continuous, the order of differentiation does not matter (Rudin, 1976, exercise 9.29).

Lemma 37 *Let $E \subseteq \mathbb{R}^D$ be open and convex, and let $f_n : E \to \mathbb{R}$ for $n \in \mathbb{N}$. For any $k \in \mathbb{N}$, if each f_n has continuous kth-order derivatives and $(f_n^{(k)})$ is uniformly bounded, then $(f_n^{(k-1)})$ is equi-Lipschitz.*

Proof First, we prove the case of $k = 1$. Let $C = \sup_n \sup_{x \in E} |f_n'(x)| < \infty$. By Taylor's theorem, for any $n \in \mathbb{N}$, $x, y \in E$, $f_n(x) = f_n(y) + f_n'(z)^{\mathtt{T}}(x - y)$ for some z on the line between x and y, and therefore,

$$|f_n(x) - f_n(y)| \leq |f_n'(z)|\,|x - y| \leq C|x - y|.$$

Thus, (f_n) is equi-Lipschitz.

For notational clarity, we prove the case of $k = 3$, and observe that the extension from this to the general case is immediate. For any $i, j \in \{1, \ldots, D\}$, if we define $h_n(x) = f_n''(x)_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f_n(x)$, then (h_n') is uniformly bounded (since $|h_n'(x)| \leq \|f_n'''(x)\|$ and (f_n''') is uniformly bounded), and hence, (h_n) is equi-Lipschitz by the case of $k = 1$ just proven. Thus, (f_n'') is equi-Lipschitz, since if C_{ij} is the equi-Lipschitz constant for entry (i, j), then

$$\|f_n''(x) - f_n''(y)\|^2 = \sum_{i,j} |f_n''(x)_{ij} - f_n''(y)_{ij}|^2 \leq C^2 |x - y|^2$$

where $C^2 = \sum_{i,j} C_{ij}^2$. ■

Lemma 38 *Let $B \subseteq \mathbb{R}^D$ be open and let $f : B \to \mathbb{R}$ be differentiable. If $x_0 \in B$ such that $f(x) \geq f(x_0)$ for all $x \in B$, then $f'(x_0) = 0$.*

Proof For any $u \in \mathbb{R}^D$ with $|u| = 1$, $f'(x_0)^{\mathtt{T}} u = \lim_{\varepsilon \to 0}(f(x_0 + \varepsilon u) - f(x_0)) \geq 0$. If $f'(x_0) \neq 0$, then choosing $u = -f'(x_0)/|f'(x_0)|$, we have $0 \leq f'(x_0)^{\mathtt{T}} u = -|f'(x_0)| < 0$, a contradiction. ■

Lemma 39 *Let $a, b \in \mathbb{R}$ such that $a < b$, let $g : [a, b] \to \mathbb{R}$ be continuous, and for $n \in \mathbb{N}$, let $g_n : [a, b] \to \mathbb{R}$ be a non-increasing function. If there is a dense subset $S \subseteq [a, b]$ such that $a, b \in S$ and $g_n(y) \to g(y)$ for all $y \in S$, then $\sup_{y \in [a,b]} |g_n(y) - g(y)| \longrightarrow 0$ as $n \to \infty$.*

Lemma 39 is straightforward to verify, so we omit the proof. Lemmas 40 and 41 are standard well-known results, but we provide precise statements and proofs for completeness. We write S° to denote the interior of S.

Lemma 40 *Let μ be a Borel measure on $\mathbb{R}^D$ and define $G(\theta) = \int_{\mathbb{R}^D} \exp(\theta^{\mathtt{T}} x)\mu(dx)$ for $\theta \in \mathbb{R}^D$. Let $S = \{\theta \in \mathbb{R}^D : G(\theta) < \infty\}$. Then G is C^∞ on S° and for all $\theta \in S^\circ$, $k \in \{0, 1, 2, \ldots\}$, $i_1, \ldots, i_k \in \{1, \ldots, D\}$, we have*

$$\frac{\partial}{\partial \theta_{i_1}} \cdots \frac{\partial}{\partial \theta_{i_k}} G(\theta) = \int x_{i_1} \cdots x_{i_k} \exp(\theta^{\mathtt{T}} x)\mu(dx). \tag{F.1}$$

Proof We proceed by induction. By construction, for all $\theta \in S^\circ$, $\int |e^{\theta^\mathsf{T} x}|\mu(dx) < \infty$ and Equation F.1 holds when $k = 0$. Fix $i_1, \ldots, i_k \in \{1, \ldots, D\}$ and suppose that for all $\theta \in S^\circ$, $\int |x_{i_1} \cdots x_{i_k} e^{\theta^\mathsf{T} x}|\mu(dx) < \infty$ and Equation F.1 holds. Let $j \in \{1, \ldots, D\}$ and $\theta_0 \in S^\circ$. Define $u = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^D$ where the 1 is in the jth position. Choose $\varepsilon > 0$ such that $\theta_0 + tu \in S^\circ$ for all $t \in [-2\varepsilon, 2\varepsilon]$. Define $f(x, t) = x_{i_1} \cdots x_{i_k} e^{(\theta_0 + tu)^\mathsf{T} x}$ and $F(t) = \int f(x, t)\mu(dx)$ for $x \in \mathbb{R}^D$, $t \in [-2\varepsilon, 2\varepsilon]$. Note that $\int |f(x, t)|\mu(dx) < \infty$ for all $t \in [-2\varepsilon, 2\varepsilon]$ by the induction hypothesis. Define $g(x) = |f(x, 2\varepsilon)|/\varepsilon + |f(x, -2\varepsilon)|/\varepsilon$. It is straightforward to verify that $|\frac{\partial f}{\partial t}(x, t)| = |x_j f(x, t)| \leq g(x)$ for all $x \in \mathbb{R}^D$, $t \in [-\varepsilon, \varepsilon]$, by using the inequality $|x_j| \leq e^{\varepsilon|x_j|}/\varepsilon$. Further, $\int |g(x)|\mu(dx) < \infty$ by the induction hypothesis. Therefore, F is differentiable and $F'(t) = \int \frac{\partial f}{\partial t}(x, t)\mu(dx)$ for all $t \in (-\varepsilon, \varepsilon)$ by Folland (2013, Theorem 2.27b).

Putting these pieces together, we have

$$\begin{aligned}
\frac{\partial}{\partial \theta_j}\bigg|_{\theta=\theta_0} \frac{\partial}{\partial \theta_{i_1}} \cdots \frac{\partial}{\partial \theta_{i_k}} G(\theta) &= \frac{\partial}{\partial \theta_j}\bigg|_{\theta=\theta_0} \int x_{i_1} \cdots x_{i_k} \exp(\theta^\mathsf{T} x)\mu(dx) \\
&= \frac{\partial}{\partial t}\bigg|_{t=0} \int f(x, t)\mu(dx) = F'(0) = \int \frac{\partial f}{\partial t}(x, 0)\mu(dx) \\
&= \int x_j f(x, 0)\mu(dx) = \int x_j x_{i_1} \cdots x_{i_k} \exp(\theta_0^\mathsf{T} x)\mu(dx)
\end{aligned}$$

and $\int |x_j x_{i_1} \cdots x_{i_k} e^{\theta_0^\mathsf{T} x}|\mu(dx) = \int |\frac{\partial f}{\partial t}(x, 0)|\mu(dx) \leq \int |g(x)|\mu(dx) < \infty$. Since $j \in \{1, \ldots, D\}$ and $\theta_0 \in S^\circ$ are arbitrary, this completes the induction step. ■

Lemma 41 *Let μ be a Borel measure on $\mathbb{R}^D$ and define $\kappa(\theta) = \log \int_{\mathbb{R}^D} \exp(\theta^\mathsf{T} x)\mu(dx)$ for $\theta \in \mathbb{R}^D$. Let $\Theta = \{\theta \in \mathbb{R}^D : |\kappa(\theta)| < \infty\}$, and define $P_\theta(A) = \int_A \exp(\theta^\mathsf{T} x - \kappa(\theta))\mu(dx)$ for $\theta \in \Theta$ and $A \subseteq \mathbb{R}^D$ Borel measurable. Then Θ is a convex set and κ is convex on Θ. Further, for all θ in the interior of Θ, for all $i, j, k \in \{1, \ldots, D\}$, if $X \sim P_\theta$ then*

(1) $\dfrac{\partial \kappa}{\partial \theta_i}(\theta) = \mathrm{E}(X_i)$,

(2) $\dfrac{\partial^2 \kappa}{\partial \theta_i \partial \theta_j}(\theta) = \mathrm{E}\big((X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j)\big) = \mathrm{Cov}(X_i, X_j)$, *and*

(3) $\dfrac{\partial^3 \kappa}{\partial \theta_i \partial \theta_j \partial \theta_k}(\theta) = \mathrm{E}\big((X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j)(X_k - \mathrm{E}X_k)\big)$.

More succinctly, items 1 and 2 state that $\kappa'(\theta) = \mathrm{E}(X)$ and $\kappa''(\theta) = \mathrm{Cov}(X)$ where $X \sim P_\theta$.

Proof Convexity of Θ and κ is a straightforward application of Hölder's inequality. Define $G(\theta) = \int \exp(\theta^\mathsf{T} x)\mu(dx)$ for $\theta \in \mathbb{R}^D$. By Lemma 40, G is C^∞ on the interior of Θ and its partial derivatives are given by Equation F.1. The identities in items 1 - 3 are straightforward to derive using Equation F.1 and the chain rule. ■

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2014.

J. M. Bernardo and A. F. Smith. *Bayesian Theory*. John Wiley & Sons, 2000.

S. Bernstein. *Theory of Probability (Russian)*. Moscow, 1917.

J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

P. J. Bickel and B. J. Kleijn. The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237, 2012.

P. J. Bickel and J. A. Yahav. Some contributions to the asymptotic theory of Bayes solutions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11(4):257–276, 1969.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

N. A. Bochkina and P. J. Green. The Bernstein–von Mises theorem and nonregular models. *The Annals of Statistics*, 42(5):1850–1878, 2014.

J. Borwanker, G. Kallianpur, and B. P. Rao. The Bernstein-von Mises theorem for Markov processes. *The Annals of Mathematical Statistics*, pages 1241–1253, 1971.

L. Bouranis, N. Friel, and F. Maire. Efficient Bayesian inference for exponential random graph models by correcting the pseudo-posterior distribution. *Social Networks*, 50:98–108, 2017.

I. Castillo and R. Nickl. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41(4):1999–2028, 2013.

I. Castillo and J. Rousseau. A General Bernstein–von Mises Theorem in semiparametric models. *arXiv preprint arXiv:1305.4482*, 2013.

C.-F. Chen. On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(3):540–546, 1985.

F. Comets. On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *The Annals of Statistics*, pages 455–468, 1992.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

A. Dawid. On the limiting normality of posterior distributions. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 67, pages 625–633. Cambridge University Press, 1970.

K. A. Doksum and A. Y. Lo. Consistent and robust Bayes procedures for location based on partial information. *The Annals of Statistics*, 18(1):443–453, 1990.

J. L. Doob. Application of the theory of martingales. In *Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin – 3 Juillet, 1948)*, pages 23–27.

Paris CNRS, 1949.

R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2002.

D. B. Dunson and J. A. Taylor. Approximate Bayesian inference for quantiles. *Nonparametric Statistics*, 17(3):385–400, 2005.

B. Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368, 1985.

M. A. Ferreira and V. De Oliveira. Bayesian reference analysis for Gaussian Markov random fields. *Journal of Multivariate Analysis*, 98(4):789–812, 2007.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications.* John Wiley & Sons, 2013.

D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963.

N. Friel. Bayesian inference for Gibbs random fields using composite likelihoods. In *Proceedings of the Winter Simulation Conference.* Winter Simulation Conference, 2012.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721, 1984.

S. Geman and C. Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pages 1496–1517. Berkeley, CA, 1986.

S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*, chapter 2, pages 35–79. Cambridge University Press, 2010.

J. K. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics.* Springer, 2003.

B. Gidas. Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 129–145. Springer, 1988.

L. Greco, W. Racugno, and L. Ventura. Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference*, 138(5):1258–1270, 2008.

G. E. Hinton and T. J. Sejnowski. Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 448–453, 1983.

P. D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283, 2007.

J. Hoffman-Jørgensen. *Probability With a View Towards Statistics*, volume 2. Chapman & Hall, 1994.

A. Hyvärinen. Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

P. E. Jacob, L. M. Murray, C. C. Holmes, and C. P. Robert. Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*, 2017.

J. L. Jensen and H. R. Künsch. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46 (3):475–486, 1994.

W. Jiang and M. A. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207–2231, 2008.

J. D. Kalbfleisch. Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):214–221, 1978.

O. Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

C. Kervrann and F. Heitz. A Markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics. *IEEE Transactions on Image Processing*, 4(6):856–862, 1995.

Y. Kim. The Bernstein–von Mises theorem for the proportional hazard model. *The Annals of Statistics*, 34(4):1678–1700, 2006.

Y. Kim and D. Kim. Bayesian partial likelihood approach for tied observations. *Journal of Statistical Planning and Inference*, 139(2):469–477, 2009.

Y. Kim and J. Lee. A Bernstein–von Mises theorem in the nonparametric right-censoring model. *The Annals of Statistics*, 32(4):1492–1512, 2004.

B. Kleijn and A. Van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

H. Künsch. Thermodynamics and statistical analysis of Gaussian random fields. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 58(3):407–421, 1981.

R. Lang. A note on the measurability of convex sets. *Archiv der Mathematik*, 47(1):90–92, 1986.

P.-S. Laplace. Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur applications aux probabilités. *Memoires de l'Academie des Sciences de Paris*, 1810.

M. Lavine. On an approximate likelihood for quantiles. *Biometrika*, 82(1):220–222, 1995.

N. A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90(2):319–326, 2003.

L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. California Pub. Statist.*, 1:277–330, 1953.

L. Le Cam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828, 1970.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 1986.

L. Le Cam and L. Schwartz. A necessary and sufficient condition for the existence of consistent estimates. *The Annals of Mathematical Statistics*, 31(1):140–150, 1960.

L. Le Cam and G. L. Yang. *Asymptotics in statistics: Some basic concepts*. Springer Science & Business Media, 2000.

H. Leahu. On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electronic Journal of Statistics*, 5:373–404, 2011.

H. K. Lee, D. M. Higdon, Z. Bi, M. A. Ferreira, and M. West. Markov random field models for high-dimensional parameters in simulations of fluid flow in porous media.

Technometrics, 44(3):230–241, 2002.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

J. R. Lewis, S. N. MacEachern, and Y. Lee. Bayesian restricted likelihood methods. *Technical report 878, The Ohio State University*, 2014.

G. Liang and B. Yu. Maximum pseudo likelihood estimation in network tomography. *IEEE Transactions on Signal Processing*, 51(8):2043–2053, 2003.

D. Y. Lin and L.-J. Wei. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239, 1988.

F. Liu, M. Bayarri, and J. Berger. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, 2009.

S. Mase. Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes. *The Annals of Applied Probability*, 5(3):603–612, 1995.

J. W. Miller. A detailed treatment of Doob's theorem. *arXiv preprint arXiv:1801.03122*, 2018.

J. W. Miller and M. T. Harrison. Inconsistency of Pitman–Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

G. Molenbergs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer Science & Business Media, 2005.

M. Panov and V. Spokoiny. Finite sample Bernstein–von Mises theorem for semiparametric problems. *Bayesian Analysis*, 10(3):665–710, 2015.

F. Pauli, W. Racugno, and L. Ventura. Bayesian composite marginal likelihoods. *Statistica Sinica*, pages 149–164, 2011.

A. Pettitt. Likelihood based inference using signed ranks for matched pairs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 45(2):287–296, 1983.

D. Pollard. Another look at differentiability in quadratic mean. In *Festschrift for Lucien Le Cam*, pages 305–314. Springer, 1997.

A. E. Raftery, D. Madigan, and C. T. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics*, 5:323–349, 1996.

M. Ribatet, D. Cooley, and A. C. Davison. Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, pages 813–845, 2012.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill New York, 1976.

T. Rydén and D. Titterington. Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, 7(2):194–211, 1998.

L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, 97(457):222–235, 2002.

D. Sinha, J. G. Ibrahim, and M.-H. Chen. A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90(3):629–641, 2003.

E. L. Smith and A. G. Stephenson. An extended Gaussian max-stable process model for spatial extremes. *Journal of Statistical Planning and Inference*, 139(4):1266–1275, 2009.

J. Stoehr and N. Friel. Calibration of conditional composite likelihood for Bayesian inference on Gibbs random fields. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38. JMLR, 2015.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

L. Ventura and W. Racugno. Pseudo-likelihoods for Bayesian inference. In *Topics on Methodological and Applied Statistical Inference*, pages 205–220. Springer, 2016.

L. Ventura, S. Cabras, and W. Racugno. Default prior distributions from quasi-and quasi-profile likelihoods. *Journal of Statistical Planning and Inference*, 140(11):2937–2942, 2010.

R. Von Mises. Wahrscheinlichkeitsrechnung. 1931.

A. M. Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(1):80–88, 1969.

M. R. Williams and T. D. Savitsky. Bayesian uncertainty estimation under complex sampling. *arXiv preprint arXiv:1807.11796*, 2018.

G. Yin. Bayesian generalized method of moments. *Bayesian Analysis*, 4(2):191–207, 2009.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

X. Zhou and S. C. Schmidler. Bayesian parameter estimation in Ising and Potts models: A comparative study with applications to protein modeling. *Technical report, Duke University*, 2009.