The Annals of Statistics
2024, Vol. 52, No. 6, 2714–2737
https://doi.org/10.1214/24-AOS2429

SKEWED BERNSTEIN–VON MISES THEOREM AND SKEW-MODAL APPROXIMATIONS

BY DANIELE DURANTE[a], FRANCESCO POZZA[b] AND BOTOND SZABO[c]

Department of Decision Sciences and Institute for Data Science and Analytics, Bocconi University,
[a]*daniele.durante@unibocconi.it,* [b]*francesco.pozza2@unibocconi.it,* [c]*botond.szabo@unibocconi.it*

Gaussian deterministic approximations are routinely employed in Bayesian statistics to ease inference when the target posterior is intractable. While these approximations are justified, in asymptotic regimes, by Bernstein–von Mises type results, in practice the expected Gaussian behavior might poorly represent the actual shape of the target posterior, thus affecting approximation accuracy. Motivated by these considerations, we derive an improved class of closed-form and valid deterministic approximations of posterior distributions that arise from a novel treatment of a third-order version of the Laplace method yielding approximations within a tractable family of skew-symmetric distributions. Under general assumptions accounting for misspecified models and non-i.i.d. settings, such a family of approximations is shown to have a total variation distance from the target posterior whose convergence rate improves by at least one order of magnitude the one achieved by the Gaussian from the classical Bernstein–von Mises theorem. Specializing this result to the case of regular parametric models shows that the same accuracy improvement can be also established for the posterior expectation of polynomially bounded functions. Unlike available higher-order approximations based on, for example, Edgeworth expansions, our results prove that it is possible to derive closed-form and valid densities which provide a more accurate, yet similarly tractable, alternative to Gaussian approximations of the target posterior, while inheriting its limiting frequentist properties. We strengthen these arguments by developing a practical skew-modal approximation for both joint and marginal posteriors which preserves the guarantees of its theoretical counterpart by replacing the unknown model parameters with the corresponding maximum a posteriori estimate. Simulation studies and real-data applications confirm that our theoretical results closely match the empirical gains observed in practice.

1. Introduction. Modern Bayesian statistics often relies on deterministic approximations in order to facilitate inference in those challenging, yet routine, situations where the target posterior is intractable (e.g., Tierney and Kadane (1986), Minka (2001), Rue, Martino and Chopin (2009), Blei, Kucukelbir and McAuliffe (2017)). A natural option to enforce the desired tractability is to constrain the approximating distribution within a suitable family which facilitates the evaluation of functionals of interest for inference. To this end, both classical solutions, such as the approximation of posterior distributions induced by the Laplace method (e.g., Bishop ((2006), Chapter 4.4)), and state-of-the-art strategies, including, for example, Gaussian variational Bayes (Opper and Archambeau (2009)) and standard implementations of expectation-propagation (Minka (2001)), employ Gaussian approximations. These further appear, either as the final solution or as a key building-block, also in several routinely implemented alternatives, such as mean-field variational Bayes (Blei, Kucukelbir and McAuliffe (2017)) and integrated nested Laplace approximation (INLA) (Rue, Martino

Received October 2023; revised April 2024.
MSC2020 subject classifications. Primary 62F15; secondary 62E20, 62E17.
Key words and phrases. Bernstein–von Mises theorem, deterministic approximation, skew-symmetric distribution.

and Chopin (2009)). See also Wang and Blei (2013), Chopin and Ridgway (2017), Durante and Rigon (2019), Ray and Szabó (2022) and Vehtari et al. (2020), among others, for further examples illustrating the relevance of Gaussian approximations.

From a theoretical perspective, the choice of the Gaussian family to approximate the posterior distribution is justified, in asymptotic regimes, by Bernstein–von Mises type results. In its classical formulation (e.g., Laplace (1810), Bernstein (1917), Von Mises (1931), LeCam (1953), Le Cam and Yang (1990), van der Vaart (1998)), the Bernstein–von Mises theorem states that, in sufficiently regular parametric models, the posterior distribution converges in total variation (TV) distance, with probability tending to one under the law of the data, to a Gaussian distribution. The expectation of this limiting Gaussian is a known function of the true data-generative parameter, or any efficient estimator of this quantity, such as the maximum likelihood estimator, while the variance is the inverse of the Fisher information. Extensions of the Bernstein–von Mises theorem to more complex settings have also been made in recent years. Relevant contributions along these directions include, among others, generalizations to high-dimensional regimes (Boucheron and Gassiat (2009), Spokoiny and Panov (2021)), along with in-depth treatments of misspecified (Kleijn and van der Vaart (2012)) and irregular (Bochkina and Green (2014)) models. Semiparametric settings have also been addressed (Bickel and Kleijn (2012), Castillo and Rousseau (2015)). In the nonparametric context, Bernstein–von Mises type results do not hold in general, but the asymptotic Gaussianity can be proved for weak Sobolev spaces via a multiscale analysis (Castillo and Nickl (2014)).

Besides providing crucial advances in the understanding of the limiting frequentist properties of posterior distributions, the above Bernstein–von Mises type results have also substantial implications in the design and in the theoretical justification of practical Gaussian deterministic approximations for intractable posterior distributions from, for example, the Laplace method (Kasprzak, Giordano and Broderick (2022)), variational Bayes (VB) (Wang and Blei (2019), Katsevich and Rigollet (2024)) and expectation-propagation (EP) (Dehaene and Barthelmé (2018)). Such a direction has led to important results. Nonetheless, in practical situations the Gaussian approximation may lack the required flexibility to closely match the actual shape of the target posterior of interest, thereby undermining accuracy when inference is based on such an approximation. In fact, as illustrated via two representative real-data clinical applications (see Section 5.2, and Appendices E5–E6 in the Supplementary Material, Durante, Pozza and Szabo (2024)), the error in posterior mean estimation of the Gaussian approximation supported by the classical Bernstein–von Mises theorem is nonnegligible not only in a study with a low sample size $n = 27$ and $d = 3$ parameters, but also in a higher-dimensional application with $n = 333$ and $d \approx n/2.5$. Both regimes often occur in routine implementations. These results further clarify that the issues encountered by the Gaussian approximation are mainly due to the inability of capturing the nonnegligible skewness often displayed by the actual posterior in these settings. Such an asymmetric shape is inherent to routinely studied posterior distributions. For example, Durante (2019), Fasano and Durante (2022) and Anceschi et al. (2023) have recently proved that, under a broad class of priors which includes multivariate normals, the posterior distribution induced by probit, multinomial probit and tobit models belongs to a skewed generalization of the Gaussian distribution known as unified skew-normal (SUN) (Arellano-Valle and Azzalini (2006)). More generally, available extensions of Gaussian deterministic approximations which account, either explicitly or implicitly, for skewness (see, e.g., Rue, Martino and Chopin (2009), Challis and Barber (2012), Fasano, Durante and Zanella (2022)) have shown evidence of improved empirical accuracy relative to their Gaussian counterparts. Nonetheless, these approximations are often model-specific and general justifications relying on Bernstein–von Mises type results are not available yet. In fact, in-depth theory and methods for skewed approximations are either lacking or are tailored to specific models and priors (Fasano, Durante and Zanella (2022)).

In this article, we address the above gaps by deriving an improved class of closed-form, valid and theoretically supported skewed approximations of generic posterior distributions. Such a class arises from a novel treatment of a higher-order version of the Laplace method which replaces the third-order term with a suitable univariate cumulative distribution function (cdf) satisfying mild regularity conditions. As clarified in Section 2.1, this perspective yields tractable approximations that crucially belong to the broad and known skew-symmetric family (e.g., Ma and Genton (2004)). More specifically, these approximations can be readily obtained by direct perturbation of the density of a multivariate Gaussian via a suitably defined univariate cdf evaluated at a cubic function of the parameter. This implies that the proposed class of approximations admits straightforward i.i.d. sampling schemes which facilitate direct Monte Carlo evaluation of any functional of interest for posterior inference. These are crucial advancements relative to other higher-order studies relying on Edgeworth-type, or other, representations (see, e.g., Johnson (1970), Weng (2010), Kolassa and Kuffner (2020), and references therein), which consider arbitrarily truncated versions of infinite expansions that do not necessarily correspond to closed-form and valid densities, even after normalization—for example, the density approximation is not guaranteed to be nonnegative (e.g., Kolassa and Kuffner ((2020), Remark 11)). This undermines the methodological and practical impact of current higher-order results which still fail to provide a natural, valid and general alternative to Gaussian deterministic approximations that can be readily employed in practice. In contrast, our novel results prove that a previously unexplored treatment of specific higher-order expansions can actually lead to valid, practical and theoretically supported approximations, thereby opening the avenues to extend such a perspective to orders even higher than the third one; see also our final discussion in Section 6.

Section 2.2 clarifies that the proposed class of skew-symmetric approximations has also a strong theoretical support in terms of accuracy improvements relative to its Gaussian counterpart. More specifically, in Theorem 2.1 we prove that such a newly proposed class has a TV distance from the target posterior distribution whose rate of convergence improves by at least one order of magnitude the one attained by the Gaussian from the classical Bernstein–von Mises theorem. Crucially, this result is derived under general assumptions which account for both misspecified models and non-i.i.d. settings. This yields an important refinement of standard Bernstein–von Mises type results clarifying that it is possible to derive closed-form and valid densities which are expected to provide, in practice, a more accurate, yet similarly tractable, alternative to Gaussian approximations of the target posterior of interest, while inheriting its limiting frequentist properties. In Section 2.3 these general results are further specialized to, possibly non-i.i.d. and misspecified, regular parametric models, where $n \to \infty$ and the dimension d of the parameter space is fixed. Under this practically relevant setting, we show that the proposed skew-symmetric approximation can be explicitly derived as a function of the log-prior and log-likelihood derivatives. Moreover, we prove that by replacing the Gaussian approximation from the Bernstein–von Mises theorem with such a newly derived alternative yields a remarkable improvement in the rates of order $\sqrt{n}$, up to a poly-log term. This gain is shown to hold not only for the TV distance from the target posterior, but also for the error in approximating the posterior expectation of polynomially bounded functions (e.g., posterior moments).

The methodological impact of the theory in Section 2 is strengthened in Section 4 through the development of a readily applicable plug-in version for the newly proposed class of skew-symmetric approximations derived in Section 2.3. This is obtained by replacing the unknown true data-generating parameter in the theoretical construction with the corresponding maximum a posteriori estimate, or any other efficient estimator. The resulting solution is named skew-modal approximation and, under mild conditions, is shown to achieve the same improved rates of its theoretical counterpart, both in terms of the TV distance from the target

posterior, and with respect to the error in approximating the posterior expectation of polynomially bounded functions. In such a practically relevant setting, we further refine the theoretical analysis through the derivation of nonasymptotic bounds for the TV distance among the skew-modal approximation and the target posterior. These bounds are guaranteed to vanish also when the dimension d grows with n, as long as $d \ll n^{1/3}$, up to a poly-log term. Interestingly, such a condition is related to those required either for d (e.g., Panov and Spokoiny (2015)) or for the notion of effective dimension $\tilde{d}$ (Spokoiny and Panov (2021), Spokoiny (2025)) in recent high-dimensional analyses of the Gaussian Laplace approximation. However, unlike these studies, the bounds we derive vanish with n, up to a poly-log term, rather than $\sqrt{n}$, for any given dimension. These advancements enable also the derivation of a novel lower bound for the TV distance among the Gaussian Laplace approximation and the posterior. This bound still vanishes with $\sqrt{n}$, under suitable conditions. Such a result strengthens the proposed skew-modal solution whose upper bound vanishes with n, up to a poly-log term. When the focus is not on the joint posterior but rather on its marginals, we further derive in Section 4.2 skew-modal approximations for such marginals that inherit the same theoretical guarantees, while scaling up computation.

The superior empirical performance of the proposed class of skew-symmetric approximations and the practical consequences of our theoretical results on the improved rates are illustrated through simulation studies and two real-data applications in Sections 3 and 5, and in Appendix E of the Supplementary Material. These analyses demonstrate that the remarkable theoretical improvements encoded in the asymptotic rates we derive closely match the empirical behavior observed in practice even in finite, possibly small, sample size regimes. This gain translates into noticeable empirical accuracy improvements relative to the Gaussian-modal approximation from the Laplace method. Even more, in the real-data applications the proposed skew-modal approximation displays a competitive performance also with respect to more sophisticated state-of-the-art Gaussian and non-Gaussian approximations from both VB and EP (e.g., Minka (2001), Blei, Kucukelbir and McAuliffe (2017), Chopin and Ridgway (2017), Durante and Rigon (2019), Fasano, Durante and Zanella (2022)).

As discussed in the concluding remarks within Section 6, the above results stimulate future advancements for refining the accuracy of other popular Gaussian approximations from, for example, VB and EP, via the inclusion of skewness. To this end, our contribution provides the foundations to achieve such a goal, and suggests that a natural and tractable class where to search for these improved approximations would be still the skew-symmetric family. Extensions to higher-order expansions beyond the third term are also discussed as directions of future research. Finally, notice that although the nonasymptotic bounds we derive for the skew-modal approximation in Section 4 yield refined theoretical results that can be readily proved for the general skew-symmetric class within Section 2, the practical consequences of nonasymptotic bounds and the associated constants is an ongoing area of research even for basic Gaussian approximations (see, e.g., Kasprzak, Giordano and Broderick (2022), and the references therein).

Proofs, technical lemmas and additional results can be found in the Supplementary Material (Durante, Pozza and Szabo (2024)).

1.1. *Notation.* Let $X^n = \{X_i\}_{i=1}^n$, $n \in \mathbb{N}$ denote a sequence of random variables with true unknown distribution P_0^n. Moreover, let $\mathcal{P}_\Theta = \{P_\theta^n, \theta \in \Theta\}$, with $\Theta \subseteq \mathbb{R}^d$, be a parametric family of distributions. In the following, we will assume that there exists a common σ-finite measure μ^n which dominates P_0^n as well as all the measures P_θ^n, and we denote by p_0^n and p_θ^n the two corresponding density functions. The Kullback–Leibler (KL) projection $P_{\theta_*}^n$ of P_0^n on $\mathcal{P}_\Theta$ is defined as $P_{\theta_*}^n = \operatorname{argmin}_{P_\theta^n \in \mathcal{P}_\Theta} \mathrm{KL}(P_0^n \| P_\theta^n)$, where $\mathrm{KL}(P_0^n \| P_\theta^n)$ denotes the KL divergence between P_0^n and P_θ^n. The log-likelihood of the, possibly misspecified, model is

$\ell(\theta) = \ell(\theta, X^n) = \log p_\theta^n(X^n)$. The prior and posterior distributions are denoted by $\Pi(\cdot)$ and $\Pi_n(\cdot)$, whereas the corresponding densities are indicated with $\pi(\cdot)$ and $\pi_n(\cdot)$, respectively.

As mentioned in Section 1, our results rely on higher-order expansions and derivatives. To this end, we characterize operations among vectors, matrices and arrays in a compact manner by adopting the index notation along with the Einstein's summation convention (e.g., Pace and Salvan ((1997), p. 335)). More specifically, the inner product $Z^\intercal a$ between the generic random vector $Z \in \mathbb{R}^d$, with components Z_s for $s = 1, \ldots, d$, and the vector of coefficients $a \in \mathbb{R}^d$ having elements a_s for $s = 1, \ldots, d$, is expressed as $a_s Z_s$, with the sum being implicit in the repetition of the indexes. Similarly, if B is a $d \times d$ matrix with entries b_{st} for $s, t = 1, \ldots, d$, the quadratic form $Z^\intercal B Z$ is expressed as $b_{st} Z_s Z_t$. The generalization to operations involving arrays with higher dimensions is obtained under the same reasoning.

Leveraging the above notation, the score vector evaluated at θ_* is defined as

$$\ell_{\theta_*}^{(1)} = \big[\ell_s^{(1)}(\theta)\big]_{|\theta=\theta_*} = \big[(\partial/\partial\theta_s)\ell(\theta)\big]_{|\theta=\theta_*} \in \mathbb{R}^d,$$

whereas, the second, third and fourth order derivatives of $\ell(\theta)$, still evaluated at θ_*, are

$$\ell_{\theta_*}^{(2)} = \big[\ell_{st}^{(2)}(\theta)\big]_{|\theta=\theta_*} = \big[\partial/(\partial\theta_s\partial\theta_t)\ell(\theta)\big]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d},$$

$$\ell_{\theta_*}^{(3)} = \big[\ell_{stl}^{(3)}(\theta)\big]_{|\theta=\theta_*} = \big[\partial/(\partial\theta_s\partial\theta_t\partial\theta_l)\ell(\theta)\big]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d\times d},$$

$$\ell_{\theta_*}^{(4)} = \big[\ell_{stlk}^{(4)}(\theta)\big]_{|\theta=\theta_*} = \big[\partial/(\partial\theta_s\partial\theta_t\partial\theta_l\partial\theta_k)\ell(\theta)\big]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d\times d\times d},$$

where all the above indexes and those in the subsequent definitions go from 1 to d. Moreover, denote by $J_{\theta_*} = [j_{st}] = -[\ell_{\theta_*,st}^{(2)}] \in \mathbb{R}^{d\times d}$ and $I_{\theta_*} = [i_{st}] = [\mathbb{E}_0^n j_{st}] \in \mathbb{R}^{d\times d}$, the observed and expected Fisher information, respectively, where $\mathbb{E}_0^n$ is the expectation with respect to P_0^n. In addition, let

$$\log \pi_{\theta_*}^{(1)} = \big[\log \pi(\theta)_s^{(1)}\big]_{|\theta=\theta_*} = \big[\partial/(\partial\theta_s)\log\pi(\theta)\big]_{|\theta=\theta_*} \in \mathbb{R}^d,$$

$$\log \pi_{\theta_*}^{(2)} = \big[\log \pi(\theta)_{st}^{(2)}\big]_{|\theta=\theta_*} = \big[\partial/(\partial\theta_s\partial\theta_t)\log\pi(\theta)\big]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d},$$

be the first two derivatives of the log-prior density, evaluated at θ_*.

The Euclidean norm of the vector $a \in \mathbb{R}^d$ is denoted by $\|a\|$, whereas, for a generic $d \times d$ matrix B, the notation $|B|$ indicates its determinant, while $\lambda_{\text{MIN}}(B)$ and $\lambda_{\text{MAX}}(B)$ its minimum and maximum eigenvalue, respectively. Furthermore, $u \wedge v$ and $u \vee v$ correspond to $\min\{u, v\}$ and $\max\{u, v\}$, respectively. For two positive sequences u_n, v_n we employ $u_n \lesssim v_n$ if there exists a universal positive constant C such that $u_n \le C v_n$. When $u_n \lesssim v_n$ and $v_n \lesssim u_n$ are satisfied simultaneously, we write $u_n \asymp v_n$. Finally, $u_n \ll v_n$ means that $u_n/v_n = o(1)$.

2. The skewed Bernstein–von Mises theorem. This section presents our first important contribution. In particular, Section 2.1 shows that, for Bayesian models satisfying a refined version of the local asymptotic normality (LAN) condition (e.g., van der Vaart (1998), Kleijn and van der Vaart (2012)), a previously unexplored treatment of a third-order version of the Laplace method yields a novel, closed-form and valid approximation of the target posterior distribution. Crucially, this approximation is shown to belong to the tractable skew-symmetric (SKS) family (e.g., Ma and Genton (2004)). Focusing on such a novel class of SKS approximations, we prove in Section 2.2 that the n-indexed sequence of TV distances between this class and the target posterior has a rate which improves by at least one order of magnitude the one achieved under the classical Bernstein–von Mises theorem based on Gaussian approximations. This skewed Bernstein–von Mises type result is proved under general assumptions that account for misspecified models and non-i.i.d. settings. Section 2.3 then specializes this result to the relevant context of regular parametric models with $n \to \infty$ and fixed d. In this

setting we prove that the improvement in rates over the Bernstein–von Mises theorem is by a factor of order $\sqrt{n}$, up to a poly-log term. Such a result is shown to hold not only for the TV distance from the posterior, but also for the error in approximating the posterior expectation of polynomially bounded functions.

Let $\delta_n \to 0$ be a generic norming rate governing the posterior contraction toward θ_*. Consistent with standard Bernstein–von Mises type theory (see, e.g., van der Vaart (1998), Kleijn and van der Vaart (2012)), consider the re-parametrization $h = \delta_n^{-1}(\theta - \theta_*) \in \mathbb{R}^d$. Moreover, let $F(\cdot) : \mathbb{R} \to [0,1]$ denote any univariate cumulative distribution function (cdf) which satisfies $F(-x) = 1 - F(x)$ and $F(x) = 1/2 + \eta x + O(x^2)$, $x \to 0$, for some $\eta \in \mathbb{R}$. Then, the class of SKS approximating densities $p^n_{\mathrm{SKS}}(h)$ we derive and study has the general form

$$p^n_{\mathrm{SKS}}(h) = 2\phi_d(h; \xi, \Omega) w(h - \xi) = 2\phi_d(h; \xi, \Omega) F\big(\alpha_\eta(h - \xi)\big), \tag{1}$$

with $P^n_{\mathrm{SKS}}(S) = \int_S p^n_{\mathrm{SKS}}(h)\, dh$ denoting the associated cdf. In (1), $\phi_d(\cdot; \xi, \Omega)$ corresponds to the density of a d-variate Gaussian having mean vector ξ and covariance matrix Ω, while the function $w(h - \xi) \in (0, 1)$ is responsible for inducing skewness, and takes form $w(h - \xi) = F(\alpha_\eta(h - \xi))$, where $\alpha_\eta(\cdot) : \mathbb{R}^d \to \mathbb{R}$ denotes a third order odd polynomial depending on the parameter that regulates the expansion of $F(\cdot)$, that is, η.

Crucially, equation (1) not only ensures that $p^n_{\mathrm{SKS}}(h)$ is a valid and closed-form density, but also that such a density belongs to the tractable and known skew-symmetric class (Azzalini and Capitanio (2003), Ma and Genton (2004)). This follows directly by the definition of SKS densities, after noticing that $\alpha_\eta(\cdot)$ is an odd function, and $\phi_d(\cdot; \xi, \Omega)$ is symmetric about ξ (e.g., Azzalini and Capitanio ((2003), Proposition 1)). Therefore, in contrast to higher-order studies of posterior distributions based on Edgeworth or other type of expansions (e.g., Johnson (1970), Weng (2010), Kolassa and Kuffner (2020)), our theoretical and methodological results focus on a family of closed-form and valid approximating densities that are essentially as tractable as multivariate Gaussians, both in terms of evaluation of the corresponding density, and i.i.d. sampling. More specifically, let $z_0 \in \mathbb{R}^d$ and $z_1 \in [0, 1]$ denote samples from a d-variate Gaussian having density $\phi_d(z_0; 0, \Omega)$ and from a uniform with support $[0, 1]$, respectively. Then, adapting results in, for example, Wang, Boyer and Genton (2004), a sample from the SKS distribution with density as in (1) can be readily obtained via

$$\xi + \operatorname{sgn}\big(F(\alpha_\eta(z_0)) - z_1\big) z_0.$$

Therefore, sampling from the proposed SKS approximation is essentially as tractable as simulating realizations from a d-variate Gaussian.

As clarified within Sections 2.3 and 4, the SKS approximation in equation (1) is not only interesting from a theoretical perspective, but has also relevant methodological consequences and direct applicability. This is because, when specializing the general theory in Section 2.2 to, possibly misspecified and non-i.i.d., regular parametric models where $n \to \infty$, d is fixed and $\delta_n^{-1} = \sqrt{n}$, we can show that the quantities defining $p^n_{\mathrm{SKS}}(h)$ in (1) can be expressed as closed-form functions of the log-prior and log-likelihood derivatives at θ_*.

In particular, let $u_t = (\ell^{(1)}_{\theta_*} + \log \pi^{(1)}_{\theta_*})_t / \sqrt{n}$ for $t = 1, \ldots, d$, then, as it will be clarified in Section 2.3, we have

$$\begin{aligned} \xi &= [n(J^{-1}_{\theta_*})_{st} u_t], \qquad \Omega^{-1} = [j_{st}/n - (\xi_l \ell^{(3)}_{\theta_*, stl}/n)/\sqrt{n}], \\ \alpha_\eta(h - \xi) &= \{1/(12\eta\sqrt{n})\}(\ell^{(3)}_{\theta_*, stl}/n)\{(h - \xi)_s (h - \xi)_t (h - \xi)_l + 3(h - \xi)_s \xi_t \xi_l\}. \end{aligned} \tag{2}$$

Interestingly, in this case, the first factor on the right hand side of equation (1) closely resembles the limiting Gaussian density with mean vector $\ell^{(1)}_{\theta_*}/\sqrt{n}$ and covariance matrix $(I_{\theta_*}/n)^{-1}$ from the classical Bernstein–von Mises theorem which, however, fails to incorporate skewness. To this end, the symmetric component in (1) is perturbed through a skewness-inducing

mechanism regulated by $F(\alpha_\eta(h-\xi))$ to obtain a valid asymmetric density having tractable normalizing constant. As shown in Section 4, this solution admits a directly applicable practical counterpart, which can be obtained by replacing $F(\cdot)$ and θ_* in (1)–(2), with routine-use univariate cdfs such as, for instance, $\Phi(\cdot)$, and with the maximum a posteriori estimate of θ, respectively. This results in a practical and novel skew-modal approximation that can be shown to preserve the same guarantees of improved accuracy of its theoretical counterpart.

2.1. *Constructive derivation of the skew-symmetric approximating density.* Prior to stating in Section 2.2 the general skewed Bernstein–von Mises theorem that supports the proposed class of SKS approximations, let us first focus on providing a constructive derivation of such a class through a novel treatment of a third-order extension of the Laplace method. To simplify notation, we consider the univariate case with $d=1$ and $\delta_n^{-1}=\sqrt{n}$. The extension of these derivations to $d>1$ and to the general setting we consider in Theorem 2.1 follow as a direct adaptation of the reasoning for the univariate case; see Sections 2.2–2.3.

As a first step towards deriving the approximating density $p^n_{\mathrm{SKS}}(h)$, notice that the posterior for $h=\sqrt{n}(\theta-\theta_*)$ can be expressed as

$$\pi_n(h) \propto \frac{p^n_{\theta_*+h/\sqrt{n}}(X^n)}{p^n_{\theta_*}}\frac{\pi(\theta_*+h/\sqrt{n})}{\pi(\theta_*)}, \tag{3}$$

since $p^n_{\theta_*}(X^n)$ and $\pi(\theta_*)$ do not depend on h, and $\theta=\theta_*+h/\sqrt{n}$.

Let j_{θ_*} be the scalar counterpart of J_{θ_*} defined in Section 1.1, then, under suitable regularity conditions discussed in Sections 2.2–2.3 below, the third-order Taylor expansion for the logarithm of the likelihood ratio in equation (3) is

$$\log\frac{p^n_{\theta_*+h/\sqrt{n}}(X^n)}{p^n_{\theta_*}} = \frac{\ell^{(1)}_{\theta_*}}{\sqrt{n}}h - \frac{1}{2}\frac{j_{\theta_*}}{n}h^2 + \frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\theta_*}}{n}h^3 + O_{P^n_0}(n^{-1}), \tag{4}$$

whereas the first-order Taylor expansion of the log-prior ratio is

$$\log\frac{\pi(\theta_*+h/\sqrt{n})}{\pi(\theta_*)} = \frac{\log\pi^{(1)}_{\theta_*}}{\sqrt{n}}h + O(n^{-1}). \tag{5}$$

Combining (4) and (5) it is possible to reformulate the right hand side of equation (3) as

$$\frac{p^n_{\theta_*+h/\sqrt{n}}(X^n)}{p^n_{\theta_*}}\frac{\pi(\theta_*+h/\sqrt{n})}{\pi(\theta_*)} = \exp\left(uh - \frac{1}{2}\frac{j_{\theta_*}}{n}h^2 + \frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\theta_*}}{n}h^3\right) + O_{P^n_0}(n^{-1}), \tag{6}$$

where $u=(\ell^{(1)}_{\theta_*}+\log\pi^{(1)}_{\theta_*})/\sqrt{n}$.

Notice that, up to a multiplicative constant, the Gaussian density arising from the classical Bernstein–von Mises theorem can be obtained by neglecting all terms in (4)–(5) which converge to zero in probability. These correspond to the contribution of the prior, the difference between the observed and expected Fisher information, and the term associated to the third-order log-likelihood derivative. Maintaining these quantities would surely yield an improved accuracy, but it is unclear whether a valid and similarly tractable density can be still identified. In fact, current solutions (e.g., Johnson (1970)) consider approximations based on the sum among a Gaussian density and additional terms in the Taylor expansion. However, as for related alternatives arising from Edgeworth-type expansions (e.g., Weng (2010), Kolassa and Kuffner (2020)), there is no guarantee that these constructions provide valid densities.

As a first result we prove below that a valid and tractable approximating density can be, in fact, derived from the above expansions and belongs to the SKS class. To this end, let $\omega=1/v$

with $v = j_{\theta_*}/n - (\xi \ell^{(3)}_{\theta_*}/n)/\sqrt{n}$ and $\xi = n(j_{\theta_*})^{-1}u$, and note that, by replacing h^3 in the right hand side of (6) with $(h - \xi + \xi)^3$, the exponential term can be rewritten as proportional to

$$\phi(h;\xi,\omega)\exp\big(\{1/(6\sqrt{n})\}(\ell^{(3)}_{\theta_*}/n)\{(h-\xi)^3 + 3(h-\xi)\xi^2\}\big). \tag{7}$$

At this point, recall that, for $x \to 0$, we can write $\exp(x) = 1 + x + O(x^2)$ and $2F(x) = 1 + 2\eta x + O(x^2)$, for some $\eta \in \mathbb{R}$, where $F(\cdot)$ is the univariate cdf introduced in (1). Therefore, leveraging the similarity among these two expansions and the fact that the exponent in (7) is an odd function of $(h - \xi)$ about 0, of order $O_{P_0^n}(n^{-1/2})$, it follows that (7) is equal to

$$2\phi(h;\xi,\omega)F\big(\alpha_\eta(h-\xi)\big) + O_{P_0^n}\big(n^{-1}\big),$$

with $\alpha_\eta(h-\xi)$ defined as in (2) for a univariate setting. The above expression coincides with the univariate case of the SKS density in (1), up to an $O_{P_0^n}(n^{-1})$ term. The extension of the above derivations to the multivariate case yields the general $p^n_{\mathrm{SKS}}(h)$ in (1) with parameters as in (2). Section 2.2 extends, and supports theoretically, this construction in general settings.

2.2. *The general theorem.* Section 2.1 shows that a suitable treatment of the cubic terms in the Taylor expansion of the log-posterior can yield a higher-order, yet valid, SKS approximating density. This solution is expected to improve the accuracy of the classical second-order Gaussian approximation, while avoiding known issues of polynomial approximations, such as regions with negative mass (see, e.g., McCullagh ((2018), p. 154)). In this section, we clarify that the derivations in Section 2.1 can be applied generally to obtain provably accurate SKS approximations in a variety of settings, provided that the posterior contraction is governed by a generic norming rate $\delta_n \to 0$, and that some reasonable regularity conditions are met. In particular, Theorem 2.1 requires Assumptions 1–4 below. For convenience, let us also introduce the notation $M_n = \sqrt{c_0 \log(1/\delta_n)}$, with $c_0 > 0$ a constant to be specified later.

ASSUMPTION 1. The Kullback–Leibler projection $\theta_* \in \Theta$ is unique.

ASSUMPTION 2. There exists a sequence of $d \times 1$ random vectors $\Delta^n_{\theta_*} = O_{P_0^n}(1)$, a sequence of $d \times d$ random matrices $V^n_{\theta_*} = [v^n_{st}]$ with $v^n_{st} = O_{P_0^n}(1)$, and also a sequence of $d \times d \times d$ random arrays $a^{(3),n}_{\theta_*} = [a^{(3),n}_{\theta_*,stl}]$ with $a^{(3),n}_{\theta_*,stl} = O_{P_0^n}(1)$, so that

$$\log \frac{p^n_{\theta_*+\delta_n h}}{p^n_{\theta_*}}(X^n) - h_s v^n_{st}\Delta^n_{\theta_*,t} + \frac{1}{2}v^n_{st}h_s h_t - \frac{\delta_n}{6}a^{(3),n}_{\theta_*,stl}h_s h_t h_l = r_{n,1}(h),$$

with $r_{n,1} := \sup_{h\in K_n}|r_{n,1}(h)| = O_{P_0^n}(\delta_n^2 M_n^{c_1})$, for some positive constant c_1, where $K_n = \{h : \|h\| < M_n\}$. In addition, there are two positive constants η_1^* and η_2^* such that the event $A_{n,0} = \{\lambda_{\mathrm{MIN}}(V^n_{\theta_*}) > \eta_1^*\} \cap \{\lambda_{\mathrm{MAX}}(V^n_{\theta_*}) < \eta_2^*\}$, holds with $P_0^n A_{n,0} = 1 - o(1)$.

ASSUMPTION 3. There exists a d-dimensional vector $\log \pi^{(1)} = [\log \pi^{(1)}_s]$ such that

$$\log\big[\pi(\theta_* + \delta_n h)/\pi(\theta_*)\big] - \delta_n h_s \log \pi^{(1)}_s = r_{n,2}(h),$$

with $\log \pi^{(1)}_s = O(1)$ and $r_{n,2} := \sup_{h\in K_n}|r_{n,2}(h)| = O(\delta_n^2 M_n^{c_2})$ for some constant $c_2 > 0$.

ASSUMPTION 4. It holds $\lim_{\delta_n\to 0} P_0^n\{\Pi_n(\|\theta - \theta_*\| > M_n\delta_n) < \delta_n^2\} = 1$.

Albeit general, Assumptions 1–4 provide reasonable conditions that extend those commonly considered to derive classical Bernstein–von Mises type results. Moreover, as clarified in Section 2.3, these assumptions translate, under regular parametric models, into natural and

explicit conditions on the behavior of the log-likelihood and log-prior. In particular, Assumption 1 is mild and can also be found, for example, in Kleijn and van der Vaart (2012). Together with Assumption 4, it guarantees that, asymptotically, the posterior distribution concentrates in the region where the two expansions in Assumptions 2 and 3 hold with negligible remainders. Notice that Assumptions 2 and 4 naturally extend those found in theoretical studies of Bernstein–von Mises type results (e.g., Kleijn and van der Vaart (2012)) to a third-order construction, which further requires quantification of rates. Assumption 3 provides, instead, an additional condition relative to those found in the classical theory. This is because, unlike for second-order Gaussian approximations, the log-prior enters the SKS construction through its first derivative; see Section 2.1. To this end, Assumption 3 imposes natural regularity conditions on the prior. Interestingly, such a need to include a careful study for the behavior of the prior density is also useful in forming the bases to naturally extend our proofs and theory to the general high-dimensional settings considered in Spokoiny and Panov (2021) and Spokoiny (2025) for the classical Gaussian Laplace approximation, where the prior effect has a critical role in controlling the behavior of the third and fourth-order components of the log-posterior. Motivated by these results, Section 4 further derives nonasymptotic bounds for the practical skew-modal approximation, which are guaranteed to vanish also when d grows with n, as long as this growth in the dimension is such that $d \ll n^{1/3}$ up to a poly-log term. See Remark 4.2 for a detailed discussion.

Under the above assumptions, Theorem 2.1 supports theoretically the proposed SKS approximation by stating a novel skewed Bernstein–von Mises type result. This result establishes that in general contexts, covering also misspecified models and non-i.i.d. settings, it is possible to derive a SKS approximation, with density as in (1), whose TV distance from the target posterior has a rate that improves by at least one order of magnitude the one achieved by the classical Gaussian counterpart from the Bernstein–von Mises theorem. By approaching the posterior at a provably faster rate, the proposed solution is therefore expected to provide, in practice, a more accurate alternative to Gaussian approximations of such a posterior, while inheriting its limiting frequentist properties. To this end, Theorem 2.1 shall not be interpreted as a theoretical result aimed at providing novel or alternative frequentist support to Bayesian inference. Rather, it represents an important refinement of the classical Bernstein–von Mises theorem which guides and supports the derivation of improved deterministic approximations to be used in practice as tractable, yet accurate, alternatives to the intractable posterior. The practical impact of these results is illustrated in the empirical studies within Sections 3 and 5. Such studies clarify that the theoretical improvements encoded in the rates we derive directly translate into remarkable accuracy gains of the proposed class of SKS approximations in finite-sample studies.

THEOREM 2.1. *Let* $h = \delta_n^{-1}(\theta - \theta_*)$, *and define* $M_n = \sqrt{c_0 \log \delta_n^{-1}}$, *with* $c_0 > 0$. *Then, under Assumptions 1–4, it holds*

$$\|\Pi_n(\cdot) - P^n_{\mathrm{SKS}}(\cdot)\|_{\mathrm{TV}} = O_{P_0^n}(M_n^{c_3}\delta_n^2), \tag{8}$$

where $c_3 > 0$, *and* $P^n_{\mathrm{SKS}}(\cdot)$ *is the cdf of the* SKS *density* $p^n_{\mathrm{SKS}}(h)$ *in* (1) *with parameters*

$$\xi = \big[\Delta^n_{\theta_*,s} + \delta_n\big((V^n_{\theta_*})^{-1}\big)_{st}\log\pi_t^{(1)}\big], \qquad \Omega^{-1} = \big[v^n_{st} - \delta_n a^{(3),n}_{\theta_*,stl}\xi_l\big],$$

$$\alpha_\eta(h-\xi) = (\delta_n/12\eta)a^{(3),n}_{\theta_*,stl}\big\{(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3(h-\xi)_s\xi_t\xi_l\big\}.$$

The function $F(\cdot)$ *entering the definition of* $p^n_{\mathrm{SKS}}(h)$ *in* (1) *is any univariate cdf which satisfies* $F(-x) = 1 - F(x)$ *and* $F(x) = 1/2 + \eta x + O(x^2)$, *for some* $\eta \in \mathbb{R}$, *when* $x \to 0$.

REMARK 2.2. Under related assumptions and a simpler proof, it is possible to show that the order of convergence for the Bernstein–von Mises theorem based on limiting Gaussians is $O_{P_0^n}(M_n^{c_4}\delta_n)$, for some $c_4 > 0$. Thus, Theorem 2.1 guarantees that by relying on suitably derived SKS approximations with density as in (1), it is possible to improve the rates of the classical Bernstein–von Mises result by a multiplicative factor of δ_n. This follows directly from the fact that the SKS approximation is able to include terms of order δ_n that are present in the Taylor expansion of the log-posterior but are neglected in the Gaussian limit. This allows an improved redistribution of the mass in the high posterior probability region, thereby yielding increased accuracy in characterizing the shape of the target posterior. As illustrated in Sections 3 and 5, this correction yields remarkable accuracy improvements in practice.

REMARK 2.3. Theorem 2.1 holds for a broad class of SKS approximating distributions as long as the univariate cdf $F(\cdot)$ entering the skewing factor in (1) satisfies $F(-x) = 1 - F(x)$ and $F(x) = 1/2 + \eta x + O(x^2)$ for some $\eta \in \mathbb{R}$, when $x \to 0$. These conditions are mild and add flexibility in the selection of $F(\cdot)$. Relevant and practical examples of possibile choices for $F(\cdot)$ are the cdf of the standard Gaussian distribution, $\Phi(\cdot)$, and the inverse logit function, $g(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$. Both satisfy $F(-x) = 1 - F(x)$, and the associated Taylor expansions are $\Phi(x) = 1/2 + x/\sqrt{2\pi} + O(x^3)$ and $g(x) = 1/2 + x/4 + O(x^3)$, respectively, for $x \to 0$. Interestingly, when $F(\cdot) = \Phi(\cdot)$, the resulting skew-symmetric approximation belongs to the well-studied sub-class of generalized skew-normal (GSN) distributions (Ma and Genton (2004)), which provide the most natural extension of multivariate skew-normals (Azzalini and Capitanio (2003)) to more flexible skew-symmetric representations. Due to this, Sections 3 and 5 focus on assessing the empirical performance of such a noticeable example.

Before discussing the proof of Theorem 2.1, let us highlight an interesting aspect regarding the interplay between skew-symmetric and Gaussian approximations that can be deduced from our theoretical studies. In particular, notice that Theorem 2.1 states results in terms of approximation of the whole posterior distribution under the TV distance. This implies, as a direct consequence of the definition of such a distance, that the same rates hold also for the absolute error in approximating the posterior expectation of any bounded function. As shown later in Corollary 2.5, such an improvement can also be proved, under mild additional conditions, for the approximation of the posterior expectation of any function bounded by a polynomial (e.g., posterior moments). According to Remark 2.2 these rates cannot be achieved in general under a Gaussian approximation. Nonetheless, as stated in Lemma 2.4, for some specific functionals the classical Gaussian approximation can actually attain the same rates of its skewed version. This result follows from the skew-symmetric distributional invariance with respect to even functions (Wang, Boyer and Genton (2004)). Such a property implies that the SKS approximation $2\phi_d(h; \xi, \Omega)F(\alpha_\eta(h - \xi))$ and its Gaussian component $\phi_d(h; \xi, \Omega)$ yield the same level of accuracy in estimating the posterior expected value of functions that are symmetric with respect to the location parameter ξ. Thus, our results provide also a novel explanation of the phenomenon observed in Spokoiny and Panov (2021) and Spokoiny (2025), where the quality of the Gaussian approximation, in high-dimensional models, increases by one order of magnitude when evaluated on Borel sets which are centrally symmetric with respect to the location ξ (see, e.g., Spokoiny ((2025), Theorem 3.4)). Nonetheless, as clarified in Theorem 2.1 and Remark 2.2, Gaussian approximations remain still unable to attain the same rates of the SKS counterparts in the estimation of generic functionals. Relevant examples are highest posterior density intervals which are often of interest in practice and will be nonsymmetric by definition whenever the posterior is skewed.

LEMMA 2.4. *Let $2\phi_d(h;\xi,\Omega)F(\alpha_\eta(h-\xi))$, with $\xi\in\mathbb{R}^d$ and $\Omega\in\mathbb{R}^{d\times d}$, denote a skew-symmetric approximation of $\pi_n(h)$ and let $G:\mathbb{R}^d\to\mathbb{R}$ be an even function. If both $\int G(h-\xi)\pi_n(h)\,dh$ and $\int G(h-\xi)2\phi_d(h;\xi,\Omega)F(\alpha_\eta(h-\xi))\,dh$ are finite, it holds*

$$\int G(h-\xi)\{\pi_n(h)-2\phi_d(h;\xi,\Omega)F(\alpha_\eta(h-\xi))\}\,dh=\int G(h-\xi)\{\pi_n(h)-\phi_d(h;\xi,\Omega)\}\,dh.$$

As clarified in the Supplementary Material, Lemma 2.4 follows as a direct consequence of Proposition 6 in Wang, Boyer and Genton (2004).

The proof of Theorem 2.1 is reported in the Supplementary Material and extends to SKS approximating distributions the reasoning behind the derivation of general Bernstein–von Mises type results (e.g., Kleijn and van der Vaart (2012)). Nonetheless, as mentioned before, the need to derive sharper rates which establish a higher approximation accuracy, relative to Gaussian limiting distributions, requires a number of additional technical lemmas and refined arguments ensuring a tighter control of the error terms in the expansions behind Theorem 2.1. Notice also that, in addressing these aspects, it is not sufficient to rely on standard theory for higher-order approximations. In fact, unlike for current results, Theorem 2.1 establishes improved rates for a valid class of approximating densities. This means that, beside replacing the second-order expansion of the log-posterior with a third-order one, it is also necessary to carefully control the difference between such an expansion and the class of SKS distributions.

2.3. *Skew-symmetric approximations in regular parametric models.* Theorem 2.1 states a general result under broad assumptions. In this section we strengthen the methodological and practical impact of such a result by specializing the analysis to the context of, possibly misspecified and non-i.i.d., regular parametric models with d fixed and $\delta_n=n^{-1/2}$. The focus on this practically relevant setting clarifies that Assumptions 1–4 can be readily verified under standard explicit conditions on the log-likelihood and log-prior derivatives, which in turn enter the definition of the SKS parameters ξ, Ω and $\alpha_\eta(h-\xi)$. This allows direct and closed-form derivation of $p^n_{\mathrm{SKS}}(h)$ in routine implementations of deterministic approximations for intractable posteriors induced by broad classes of parametric models. As stated in Corollary 2.5, in this setting the resulting SKS approximating density achieves a remarkable improvement in the rates by a $\sqrt{n}$ factor, up to a poly-log term, relative to the classical Gaussian approximation. This accuracy gain can be proved both for the TV distance from the target posterior and also for the absolute error in the approximation for the posterior expectation of general polynomially bounded functions (e.g., moments), with finite prior expectation.

Prior to stating Corollary 2.5, let us introduce a number of explicit assumptions that allow to specialize the general theory in Section 2.2 to the setting with d fixed and $\delta_n=n^{-1/2}$. As discussed in the following, Assumptions 5–8 provide natural and verifiable conditions which ensure that the general Assumptions 2–4 are met, thereby allowing Theorem 2.1 to be applied, and specialized, to the regular parametric models setting.

ASSUMPTION 5. Define $\ell^{(4)}_{\theta_*,stlk}(h):=\ell^{(4)}_{stlk}(\theta_*+h/\sqrt{n})$. Then, the log-likelihood of the, possibly misspecified, model is four times differentiable at θ_* with

$$\ell^{(1)}_{\theta_*,s}=O_{P_0^n}(n^{1/2}),\qquad \ell^{(2)}_{\theta_*,st}=O_{P_0^n}(n),\qquad \ell^{(3)}_{\theta_*,stl}=O_{P_0^n}(n)\quad \text{for } s,t,l=1,\dots,d,$$

and $\sup_{h\in K_n}|\ell^{(4)}_{\theta_*,stlk}(h)|=O_{P_0^n}(n)$, for $s,t,l,k=1,\dots,d$.

ASSUMPTION 6. All the entries of the expected Fisher information matrix satisfy $i_{st}=O(n)$ while $j_{st}/n-i_{st}/n=O_{P_0^n}(n^{-1/2})$, for $s,t=1,\dots,d$. Moreover, there exist two positive constants η_1 and η_2 such that $\lambda_{\mathrm{MIN}}(I_{\theta_*}/n)>\eta_1$ and $\lambda_{\mathrm{MAX}}(I_{\theta_*}/n)<\eta_2$.

ASSUMPTION 7. The log-prior density $\log \pi(\theta)$ is two times continuously differentiable in a neighborhood of θ_*, and $0 < \pi(\theta_*) < \infty$.

ASSUMPTION 8. For every sequence $M_n \to \infty$ there exists a positive constant $c_5 > 0$ such that $\lim_{n\to\infty} P_0^n\{\sup_{\|\theta-\theta_*\|>M_n/\sqrt{n}}\{(\ell(\theta)-\ell(\theta_*))/n\} < -c_5 M_n^2/n\} = 1$.

Assumptions 5–6 are mild and considered standard in classical frequentist theory (e.g., Pace and Salvan ((1997), p. 347)). In the Supplementary Material (Lemma A.1) we show that these conditions allow to control with precision the error in the Taylor approximation of the log-likelihood. Assumption 7 is also mild and is satisfied by several routinely used priors. This condition allows to consider a first-order Taylor expansion for the log-prior of the form

$$\log \pi(\theta) = \log \pi(\theta_*) + \log \pi^{(1)}_{\theta_*,s} h_s/\sqrt{n} + r_{n,2}(h), \tag{9}$$

with $r_{n,2} := \sup_{h\in K_n} |r_{n,2}(h)| = O(M_n^2/n)$. Finally, Assumption 8 is required to control the rate of contraction of the, possibly misspecified, posterior inside K_n. In other versions of Bernstein–Von Mises type results, such an assumption is usually replaced by conditions on the existence of a suitable sequence of tests. Sufficient conditions for the correctly specified case can be found in van der Vaart (1998). In the misspecified setting, assumptions ensuring the existence of these tests have been derived by Kleijn and van der Vaart (2012). Another possible option is to assume, for every $\delta > 0$, the presence of a constant $c_\delta > 0$ such that

$$\lim_{n\to\infty} P_0^n\Big\{\sup_{\|\theta-\theta_*\|>\delta}\{(\ell(\theta)-\ell(\theta_*))/n\} < -c_\delta\Big\} = 1. \tag{10}$$

In the misspecified setting, (10) is considered by Koers, Szabo and van der Vaart (2023). Assumption 8 is a slightly more restrictive version of (10). In fact, Lemma A.3 in the Supplementary Material shows that it is implied by mild and readily verifiable sufficient conditions.

Under Assumption 1 and 5–8, Corollary 2.5 clarifies that Theorem 2.1 holds for a general class of SKS distributions yielding TV rates in approximating the exact posterior of order $O_{P_0^n}(M_n^{c_6}/n)$, with $c_6 > 0$ and $M_n = \sqrt{c_0 \log n}$. Furthermore, the SKS parameters are defined as explicit functions of the log-prior and log-likelihood derivatives. As stated in equation (12) of Corollary 2.5, the same rates can be derived also for the absolute error in approximating the posterior expectation of general polynomially bounded functions (e.g., moments).

COROLLARY 2.5. *Let $h = \sqrt{n}(\theta - \theta_*)$, and define $M_n = \sqrt{c_0 \log n}$, with $c_0 > 0$. Then, under Assumptions 1 and 5–8, it holds*

$$\big\|\Pi_n(\cdot) - P^n_{\mathrm{SKS}}(\cdot)\big\|_{\mathrm{TV}} = O_{P_0^n}\big(M_n^{c_6}/n\big), \tag{11}$$

where $c_6 > 0$, and $P^n_{\mathrm{SKS}}(\cdot)$ is the cdf of the SKS *density $p^n_{\mathrm{SKS}}(h)$ in (1) with parameters*

$$\xi = \big[n\big(J^{-1}_{\theta_*}\big)_{st} u_t\big], \qquad \Omega^{-1} = \big[j_{st}/n - \big(\xi_l \ell^{(3)}_{\theta_*,stl}/n\big)/\sqrt{n}\big],$$

$$\alpha_\eta(h-\xi) = \big\{1/(12\eta\sqrt{n})\big\}\big(\ell^{(3)}_{\theta_*,stl}/n\big)\big\{(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3(h-\xi)_s\xi_t\xi_l\big\},$$

where $u_t = (\ell^{(1)}_{\theta_*} + \log \pi^{(1)}_{\theta_*})_t/\sqrt{n}$ for $t = 1, \ldots, d$. The function $F(\cdot)$ entering the definition of $p^n_{\mathrm{SKS}}(h)$ in (1) denotes any univariate cdf which satisfies $F(-x) = 1 - F(x)$ and $F(x) = 1/2 + \eta x + O(x^2)$, for some $\eta \in \mathbb{R}$, when $x \to 0$. In addition, let $G : \mathbb{R}^d \to R$ be a function satisfying $|G(h)| \lesssim \|h\|^r$. If the prior is such that $\int \|h\|^r \pi(\theta_* + h/\sqrt{n})\, dh < \infty$ then

$$\int G(h)\big|\pi_n(h) - p^n_{\mathrm{SKS}}(h)\big|\, dh = O_{P_0^n}\big(M_n^{c_6+r}/n\big), \tag{12}$$

with $p^n_{\mathrm{SKS}}(h)$ denoting the skew-symmetric approximating density defined above.

REMARK 2.6. As for Theorem 2.1, under conditions similar to those required by Corollary 2.5, it is possible to show that the TV distance between the posterior and the Gaussian approximation studied in the classical Bernstein–von Mises theorem is $O_{P_0^n}(M_n^{c_7}/\sqrt{n})$ for some fixed $c_7 > 0$. Therefore, the improvement in rates achieved by the proposed SKS approximation is by a $\sqrt{n}$ factor, up to a poly-log term. As illustrated in Figure E.1 of the Supplementary Material, this implies that the SKS solution is expected to substantially improve, in practice, the accuracy of the classical Gaussian in approximating the target posterior, while inheriting its limiting frequentist properties. Intuitively, the rates we derive suggest that the proposed SKS approximation can possibly attain with a $n \approx \sqrt{\bar{n}}$ sample size the same accuracy obtained by its Gaussian counterpart with a sample size of $\bar{n}$. The empirical studies in Sections 3 and 5 confirm such an intuition, which is further strengthened in Section 4 through the derivation of nonasymptotic upper bounds for the practical skew-modal approximation, along with novel lower bounds for the classical Gaussian from the Laplace method.

REMARK 2.7. Equation (12) confirms that the improved rates hold also when the focus is on the error in approximating the posterior expectation of generic polynomially bounded functions. More specifically, notice that, by direct application of standard properties of integrals, the proof of equation (12) in the Supplementary Material, implies

$$\left| \int G(h)\pi_n(h)\,dh - \int G(h)\,p^n_{\mathrm{SKS}}(h)\,dh \right| = O_{P_0^n}\big(M_n^{c_6+r}/n\big). \tag{13}$$

This clarifies that the skewed Bernstein–von Mises type result in (11) has important methodological and practical consequences that point toward remarkable improvements in the approximation of posterior functionals of direct interest for inference (e.g., moments).

The proof of equation (12) can be found in the Supplementary Material. As for the main result in (11), it is sufficient to apply Theorem 2.1, after ensuring that its Assumptions 1–4 are implied by 1 and 5–8. Appendix A.1 in the Supplementary Material presents two key results (see Lemmas A.1 and A.2) which address this point. Appendix A.2 introduces instead simple and verifiable conditions which ensure the validity of Assumption 8.

3. Empirical results. Here we provide empirical evidence for the improved accuracy of the proposed SKS approximation in Corollary 2.5 (S-BVM), relative to its Gaussian counterpart (BVM) from the classical Bernstein–von Mises theorem in regular parametric models. In Section 3.1 we consider, in particular, a correctly specified setting and defer to Appendix E.2.1 in the Supplementary Material the analysis of a misspecified case. In both studies the focus is not only on assessing the superior performance of the new SKS approximation, with $F(\cdot) = \Phi(\cdot)$, but also on quantifying whether the improvements encoded in the rates we derived under asymptotic arguments find empirical evidence also in finite samples. To this end, S-BVM and BVM are compared both in terms of the TV distance from the posterior and also with respect to the absolute error in approximating the posterior mean. These two measures illustrate the practical implications of the rates derived in (11) and (12). Since for the illustrative studies we consider the target posterior can be derived in closed form, the TV distances $\mathrm{TV}^n_{\mathrm{BVM}} = (1/2)\int_{\mathbb{R}} |\pi_n(h) - p^n_{\mathrm{GAUSS}}(h)|\,dh$ and $\mathrm{TV}^n_{\text{S-BVM}} = (1/2)\int_{\mathbb{R}} |\pi_n(h) - p^n_{\mathrm{SKS}}(h)|\,dh$ can be evaluated numerically, for every size n, via standard routines in R. The same holds also for the errors in posterior mean approximation $\mathrm{FMAE}^n_{\mathrm{BVM}} = |\int_{\mathbb{R}} h\{\pi_n(h) - p^n_{\mathrm{GAUSS}}(h)\}\,dh|$ and $\mathrm{FMAE}^n_{\text{S-BVM}} = |\int_{\mathbb{R}} h\{\pi_n(h) - p^n_{\mathrm{SKS}}(h)\}\,dh|$.

Notice that, as for other versions of the Bernstein–von Mises theorem, also our theoretical results in Sections 2.2 and 2.3 require knowledge of the KL minimizer between the true data-generating process and the parametric family $\mathcal{P}_\Theta$. Since θ_* is unknown in practice, in Section 4 we address this point via a plug-in version of the SKS approximation in Corollary 2.5, which replaces θ_* with its maximum a posteriori estimate. This yields a readily applicable skew-modal approximation with similar theoretical and empirical support.

TABLE 1

Empirical comparison, averaged over 50 *replicated studies, between the classical* (BVM) *and skewed* (S-BVM) *Bernstein–von Mises theorem in the correctly specified exponential example. The first table shows, for different sample sizes ranging from* $n = 10$ *to* $n = 1500$, *the log-*TV *distances* (TV) *and log-approximation errors for the posterior mean* (FMAE) *under* BVM *and* S-BVM. *The bold values indicate the best performance for each* n. *The second table shows, for different sample sizes ranging from* $n = 10$ *to* $n = 100$, *the sample size* $\bar{n}$ *required by the classical Gaussian* BVM *to achieve the same* TV *and* FMAE *attained by the* SKS *approximation with that* n

	$n = 10$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 1500$
$\log \text{TV}^n_{\text{BVM}}$	−1.67	−2.50	−2.82	−3.59	−3.98	−4.18
$\log \text{TV}^n_{\text{S-BVM}}$	**−2.53**	**−3.86**	**−4.41**	**−5.76**	**−5.58**	**−6.58**
$\log \text{FMAE}^n_{\text{BVM}}$	−0.90	−1.77	−1.97	−2.85	−3.21	−3.33
$\log \text{FMAE}^n_{\text{S-BVM}}$	**−1.07**	**−2.81**	**−3.74**	**−6.14**	**−7.09**	**−7.42**

	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 50$	$n = 75$	$n = 100$
$\bar{n} : \text{TV}^{\bar{n}}_{\text{BVM}} = \text{TV}^n_{\text{S-BVM}}$	55	120	250	350	820	1690	2470
$\bar{n} : \text{FMAE}^{\bar{n}}_{\text{BVM}} = \text{FMAE}^n_{\text{S-BVM}}$	15	25	70	110	380	1050	2280

3.1. *Exponential model.* Let $X_i \overset{\text{iid}}{\sim} \text{EXP}(\theta_0)$, for $i = 1, \ldots, n$, where $\text{EXP}(\theta_0)$ denotes the exponential distribution with rate parameter $\theta_0 = 2$. In the following, we consider a correctly specified model having exponential likelihood and an $\text{EXP}(1)$ prior for θ. As clarified in Appendix E.2 of the Supplementary Material, such a Bayesian model verifies all the conditions of Corollary 2.5, and hence, the induced posterior admits a SKS approximation with parameters $\xi = \theta_0^2(n/\theta_0 - \sum_{i=1}^n x_i - 1)/\sqrt{n}$, $\Omega = 1/(\theta_0^{-2} - 2\theta_0^{-1}\{1/\theta_0 - (\sum_{i=1}^n x_i)/n - 1/n\})$ and $\alpha_\eta(h - \xi) = \{\sqrt{2\pi}/(6\sqrt{n}\theta_0^3)\}\{(h - \xi)^3 + 3(h - \xi)\xi^2\}$.

Table 1 compares the accuracy of the novel S-BVM and classial BVM, under growing sample size and in replicated experiments. More specifically, we consider 50 different simulated datasets with $\theta_0 = 2$ and sample size $n_{\text{TOT}} = 1500$. Then, within each of these 50 experiments, we derive the target posterior under different subsets of data $x_1, \ldots, x_n$, with growing sample size n, and then compare the accuracy of the two approximations under the TV and FMAE measures introduced previously. The first part of Table 1 displays, for each n, these two measures averaged across the 50 replicated experiments under both S-BVM and BVM. The empirical results confirm that the SKS approximation yields remarkable improvements over the Gaussian counterpart for any n. Such an empirical finding clarifies that the $\sqrt{n}$ improvement encoded in the rates we derived, is visible also in finite, even small, sample size settings. This suggests that the theory in Sections 2.2–2.3 is informative also in practice, and motivates the adoption of the SKS approximation in place of the Gaussian one. Such a result is strengthened in the second part of Table 1, which shows that to attain the same accuracy of the SKS approximation with a given n, the classical Gaussian counterpart requires a sample size $\bar{n}$ higher by approximately one order of magnitude. The analysis of a misspecified exponential model in the Supplementary Material (Appendix E.2.1) confirms these conclusions.

4. Skew-modal approximation. As for standard theoretical derivations of Bernstein–von Mises type results, also our theory in Section 2 studies approximating densities whose parameters depend on the minimizer θ_* of $\text{KL}(P_0^n || P_\theta^n)$ for $\theta \in \Theta$, that coincides with θ_0 when the model is correctly specified. This quantity is unknown in practice. Hence, to provide an effective approximation which can be implemented in practical contexts, it is necessary to replace θ_* with a suitable estimate. To this end, in Section 4.1 we consider a simple,

yet effective, plug-in version of the SKS density in Corollary 2.5 which replaces θ_* with its maximum a posteriori (MAP) estimator $\hat{\theta} = \operatorname{argmax}_{\theta\in\Theta}\{\ell(\theta) + \log \pi(\theta)\}$, without losing the theoretical accuracy guarantees. Note that, in general, θ_* can be replaced by any efficient estimator. However, by relying on the MAP several quantities simplify, giving raise to a tractable and accurate solution, which is named skew-modal approximation. See Section 4.2 for a theoretically supported, yet more scalable, skew-modal approximation of posterior marginals.

4.1. *Skew-modal approximation and theoretical guarantees.* Consistent with the above discussion, let us consider the plug-in version of $p^n_{\text{SKS}}(h)$ in (1), where θ_* is replaced by the MAP. This yields the SKS density, for the rescaled parameter $\hat{h} = \sqrt{n}(\theta - \hat{\theta}) \in \mathbb{R}^d$, defined as

$$\hat{p}^n_{\text{SKS}}(\hat{h}) = 2\phi_d(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h}) = 2\phi_d(\hat{h}; 0, \hat{\Omega})F\big(\hat{\alpha}_\eta(\hat{h})\big), \tag{14}$$

with $\hat{\Omega} = (\hat{V}^n)^{-1}$, $\hat{V}^n = [j_{\hat{\theta},st}/n] \in \mathbb{R}^{d\times d}$, and $\hat{\alpha}_\eta(\hat{h}) = \{1/(12\eta\sqrt{n})\}(\ell^{(3)}_{\hat{\theta},stl}/n)\hat{h}_s\hat{h}_t\hat{h}_l \in \mathbb{R}$.

Relative to the expression for $p^n_{\text{SKS}}(h)$ in (1), the location $\hat{\xi}$ is zero in (14), since replacing θ_* with the MAP inside u_t in Corollary 2.5 gives zero by definition of MAP. This implies also that in the expression for Ω^{-1} in Corollary 2.5 the term including the third order derivative disappears. Therefore, (14) does not introduce further complications in terms of positive-definiteness and nonnegativity of the precision matrix relative to those of the classical Laplace approximation.

Equation (14) provides a practical skewed approximation of the target posterior with symmetric component centered at its mode. As such, this solution is referred to as skew-modal approximation. In order to provide theoretical guarantees for this practical version, similar to those in Corollary 2.5, while further refining these guarantees through novel nonasymptotic bounds, let us introduce two mild assumptions in addition to those in Section 2.3.

ASSUMPTION 9. For every $M_n \to \infty$, the event $\hat{A}_{n,0} = \{\|\hat{\theta} - \theta_*\| \le M_n\sqrt{d}/\sqrt{n}\}$ satisfies $P^n_0(\hat{A}_{n,0}) > 1 - \hat{\epsilon}_{n,0}$ for some sequence $\{\hat{\epsilon}_{n,0}\}_{n=1}^\infty$ converging to zero as $n \to \infty$.

ASSUMPTION 10. There exist two positive constants $\bar{\eta}_1$ and $\bar{\eta}_2$ such that the event

$$\hat{A}_{n,1} = \big\{\lambda_{\text{MIN}}\big(\hat{\Omega}^{-1}\big) > \bar{\eta}_1\big\} \cap \big\{\lambda_{\text{MAX}}\big(\hat{\Omega}^{-1}\big) < \bar{\eta}_2\big\},$$

holds with a probability $P^n_0(\hat{A}_{n,1}) > 1 - \hat{\epsilon}_{n,1}$ for a suitable sequence $\{\hat{\epsilon}_{n,1}\}_{n=1}^\infty$ converging to zero as $n \to \infty$. Moreover, there exist positive constants $\delta > 0$ and $L_3 > 0$, $L_4 > 0$, $L_{\pi,2} > 0$ such that, for $B_\delta(\hat{\theta}) := \{\theta \in \Theta : \|\hat{\theta} - \theta\| < \delta\}$, the event

$$\begin{aligned}\hat{A}_{n,2} = &\Big\{\sup_{\theta\in B_\delta(\hat{\theta})} \big\|\log \pi^{(2)}(\theta)\big\| < L_{\pi,2}\Big\} \cap \Big\{\sup_{\theta\in B_\delta(\hat{\theta})} \big\|\ell^{(3)}(\theta)/n\big\| < L_3\Big\} \\ &\cap \Big\{\sup_{\theta\in B_\delta(\hat{\theta})} \big\|\ell^{(4)}(\theta)/n\big\| < L_4\Big\},\end{aligned}$$

holds with a probability $P^n_0(\hat{A}_{n,2}) > 1 - \hat{\epsilon}_{n,2}$, for some suitable sequence $\{\hat{\epsilon}_{n,2}\}_{n=1}^\infty$ converging to zero as $n \to \infty$, where $\|\cdot\|$ represents the spectral norm.

Assumption 9 is mild and holds generally in regular parametric problems. This assumption ensures that the MAP is in a suitably small neighborhood of θ_*, where the centering took place in Corollary 2.5. Condition 10 is a similar and arguably not stronger version of the assumptions for the Laplace method described in Kass, Tierney and Kadane (1990). Notice also that under Assumption 9, condition 10 replaces Assumptions 5–6, and requires the upper bound to hold in a neighborhood of θ_*. These conditions ensure uniform control on the

difference between the log-likelihood ratio and its third-order Taylor expansion. Based on these additional conditions we provide an asymptotic result for the skew-modal approximation in (14), similar to Corollary 2.5. The proof can be found in the Supplementary Material and follows as a direct consequence of a more refined nonasymptotic bound we derive for $\|\Pi_n(\cdot) - \hat{P}^n_{\mathrm{SKS}}(\cdot)\|_{\mathrm{TV}}$; see Remark 4.2.

THEOREM 4.1. *Let* $\hat{h} = \sqrt{n}(\theta - \hat{\theta})$, *and define* $M_n = \sqrt{c_0 \log n}$, *with* $c_0 > 0$. *If Assumptions* 1, 7–8, *and* 9–10 *are met, then the posterior for* $\hat{h}$ *satisfies*

$$\big\|\Pi_n(\cdot) - \hat{P}^n_{\mathrm{SKS}}(\cdot)\big\|_{\mathrm{TV}} = O_{P_0^n}\big(M_n^{c_8}/n\big), \tag{15}$$

for some $c_8 > 0$, *where* $\hat{P}^n_{\mathrm{SKS}}(S) = \int_S \hat{p}^n_{\mathrm{SKS}}(\hat{h})\, d\hat{h}$ *for* $S \subset \mathbb{R}^d$ *with* $\hat{p}^n_{\mathrm{SKS}}(\hat{h})$ *defined as in* (14). *In addition, let* $G : \mathbb{R}^d \to \mathbb{R}$ *be a function satisfying* $|G(\hat{h})| \lesssim \|\hat{h}\|^r$. *If the prior is such that* $\int \|\hat{h}\|^r \pi(\hat{\theta} + \hat{h}/\sqrt{n})\, d\hat{h} < \infty$ *then*

$$\int G(\hat{h})\big|\pi_n(\hat{h}) - \hat{p}^n_{\mathrm{SKS}}(\hat{h})\big|\, d\hat{h} = O_{P_0^n}\big(M_n^{c_8+r}/n\big). \tag{16}$$

REMARK 4.2. As discussed above, Theorem 4.1 is a direct consequence of a more refined nonasymptotic upper bound for $\|\Pi_n(\cdot) - \hat{P}^n_{\mathrm{SKS}}(\cdot)\|_{\mathrm{TV}}$ that we derive in Appendix D of the Supplementary Material. In particular, as in recently derived nonasymptotic results for the Gaussian Laplace approximation (e.g., Spokoiny and Panov (2021), Spokoiny (2025)), it is possible to keep track of the constants and the dimension dependence also within our derivations, to show that on an event with high probability (approaching 1), it holds

$$\big\|\Pi_n(\cdot) - \hat{P}^n_{\mathrm{SKS}}(\cdot)\big\|_{\mathrm{TV}} \le C M_n^{c_8} d^3/n, \tag{17}$$

for some positive constant C not depending on d and n; see Theorem D.1 and Remark D.2. Hence, the rates in (15) follow directly from (17), when d is kept fixed and $n \to \infty$. More importantly, the above bound vanishes also when d grows with n, as long as $d \ll n^{1/3}$, up to a poly-log term. Although our original focus is not specific to high-dimensional regimes, it shall be emphasized that this growth for d is interestingly in line with those required either for d (Panov and Spokoiny (2015)) or for the notion of effective dimension $\tilde{d}$ (Spokoiny and Panov (2021), Spokoiny (2025)) in recent high-dimensional studies of the Gaussian Bernstein–von Mises theorem and the Laplace approximation. However, unlike the bounds in these studies, the one reported in (17) decays to zero with n, up to a poly-log term, rather than with $\sqrt{n}$, for any given dimension.

REMARK 4.3. Similar to Remark 2.6 our proofs can be easily modified to show that the TV distance between the target posterior and the classical Gaussian Laplace approximation is, up to a poly-log term, of order $1/\sqrt{n}$. This upper bound is worse than those derived for the skew-modal approximation. Theorem D.6 in the Supplementary Material further refines this result by proving that, up to a poly-log term, this upper bound is sharp, whenever the posterior displays local asymmetries. More specifically, under condition (D.30) in the Supplementary Material, we prove that, on an event with high probability (approaching 1), the TV distance between the posterior and the classical Laplace approximation (GM) is bounded from below by $C_d/\sqrt{n} + O(M_n^{c_8} d^3/n)$, for some constant $C_d > 0$, possibly depending on d. Crucially, the proof of this lower bound implies that $\|\Pi_n(\cdot) - \hat{P}^n_{\mathrm{GM}}(\cdot)\|_{\mathrm{TV}} - \|\Pi_n(\cdot) - \hat{P}^n_{\mathrm{SKS}}(\cdot)\|_{\mathrm{TV}}$ is also bounded from below by $C_d/\sqrt{n} + O(M_n^{c_8} d^3/n)$. This result strengthens (15)–(17).

REMARK 4.4. Since the TV distance is invariant with respect to scale and location transformations, the above results can be stated also for the original parametrization θ of interest. Focusing, in particular, on the choice $F(\cdot) = \Phi(\cdot)$, this yields the density

$$\hat{p}^n_{\mathrm{SKS}}(\theta) = 2\phi_d(\theta; \hat{\theta}, J^{-1}_{\hat{\theta}})\Phi\big((\sqrt{2\pi}/12)\ell^{(3)}_{\hat{\theta},stl}(\theta-\hat{\theta})_s(\theta-\hat{\theta})_t(\theta-\hat{\theta})_l\big), \tag{18}$$

which coincides with that of the generalized skew-normal (GSN) sub-class (Ma and Genton (2004)) and is guaranteed to approximate the posterior for θ with the rate in Theorem 4.1.

Our novel skew-modal approximation provides, therefore, a similarly tractable, yet more accurate, alternative to the classical Gaussian from the Laplace method. This is because the closed-form skew-modal density can be evaluated at a similar computational cost as the one for the Gaussian, when d is not too large. Moreover, it admits a straightforward i.i.d. sampling scheme that facilitates Monte Carlo estimation of any functional of interest. Recalling Section 2, such a scheme simply relies on sign perturbations of samples from a d-variate Gaussian and, hence, can be directly implemented via standard R packages. Notice that, although the nonasymptotic bound in (17) can be also derived for the theoretical skew-symmetric approximations in Section 2, the focus on the skew-modal is motivated by the fact that such an approximation provides the solution implemented in practice. Section 4.2 derives and studies an even more scalable, yet similarly accurate, approximation for posterior marginals.

4.2. *Marginal skew-modal approximation and theoretical guarantees.* The skew-modal in Section 4.1 targets the joint posterior. In practice, the marginals of this posterior are often the main object of interest (Rue, Martino and Chopin (2009)). For studying these quantities, it is possible to simulate i.i.d. values from the joint skew-modal in (14), leveraging the strategy in Section 2, and then retain only samples from the marginals of interest. This requires, however, multiple evaluations of the cubic function in the skewness-inducing factor. Below we derive a skew-modal approximation for posterior marginals that addresses this issue.

To accomplish the above goal, denote with $\mathcal{C} \subseteq \{1, \dots, d\}$ the set containing the indexes for the elements of θ on which we are interested in. Let $d_{\mathcal{C}}$ be the cardinality of $\mathcal{C}$, and $\bar{\mathcal{C}} = \mathcal{C}^c$ the complement of $\mathcal{C}$. Finally, write $\hat{h} = (\hat{h}_{\mathcal{C}}, \hat{h}_{\bar{\mathcal{C}}})$. Accordingly, the matrix $\hat{\Omega} = (J_{\hat{\theta}}/n)^{-1}$ can be partitioned in two diagonal blocks $\hat{\Omega}_{\mathcal{C}\mathcal{C}}$, $\hat{\Omega}_{\bar{\mathcal{C}}\bar{\mathcal{C}}}$, and an off-diagonal one $\hat{\Omega}_{\bar{\mathcal{C}}\mathcal{C}} = \hat{\Omega}^{\intercal}_{\mathcal{C}\bar{\mathcal{C}}}$.

Under the regularity conditions stated in Sections 2.3 and 4.1, we can write, for $n \to \infty$,

$$\pi_n(\hat{\theta} + \hat{h}/\sqrt{n}) \propto \exp\big(-j_{\hat{\theta},st}\hat{h}_s\hat{h}_t/(2n) + (\ell^{(3)}_{\hat{\theta},stl}/n)\hat{h}_s\hat{h}_t\hat{h}_l/(6\sqrt{n})\big) + O_{P^n_0}(n^{-1}).$$

Notice that the second order term in the above expression is proportional to the kernel of a Gaussian and, therefore, can be decomposed as

$$\exp\big(-j_{\hat{\theta},st}\hat{h}_s\hat{h}_t/(2n)\big) \propto \phi_d(\hat{h}; 0, \hat{\Omega}) = \phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}})\phi_{d-d_{\mathcal{C}}}(\hat{h}_{\bar{\mathcal{C}}}; \Lambda_{\mathcal{C}}\hat{h}_{\mathcal{C}}, \bar{\Omega}),$$

where $\Lambda_{\mathcal{C}} = \hat{\Omega}_{\bar{\mathcal{C}}\mathcal{C}}\hat{\Omega}^{-1}_{\mathcal{C}\mathcal{C}}$ and $\bar{\Omega} = \hat{\Omega}_{\bar{\mathcal{C}}\bar{\mathcal{C}}} - \hat{\Omega}_{\bar{\mathcal{C}}\mathcal{C}}\hat{\Omega}^{-1}_{\mathcal{C}\mathcal{C}}\hat{\Omega}_{\mathcal{C}\bar{\mathcal{C}}}$.

To obtain a marginal skew-modal approximation, let us leverage again the fact that the third order term converges to zero in probability, and that $e^x = 1 + x + O(x^2)$, for $x \to 0$. With these results, an approximation for the posterior marginal of $\hat{h}_{\mathcal{C}}$ is, therefore, proportional to

$$\begin{aligned} &\int \phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}})\phi_{d-d_{\mathcal{C}}}(\hat{h}_{\bar{\mathcal{C}}}; \Lambda_{\mathcal{C}}\hat{h}_{\mathcal{C}}, \bar{\Omega})\big(1 + (\ell^{(3)}_{\hat{\theta},stl}/n)\hat{h}_s\hat{h}_t\hat{h}_l/(6\sqrt{n})\big)\, d\hat{h}_{\bar{\mathcal{C}}} \\ &\quad = \phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}})\big[1 + \{(1/n)/(6\sqrt{n})\}\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\ell^{(3)}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l)\big], \end{aligned} \tag{19}$$

where $\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\ell^{(3)}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l)$ denotes the expectation with respect to $\phi_{d-d_{\mathcal{C}}}(\hat{h}_{\bar{\mathcal{C}}}; \Lambda_{\mathcal{C}}\hat{h}_{\mathcal{C}}, \bar{\Omega})$.

Leveraging standard properties of the expected value, the above expectation can be further decomposed as

$$\begin{aligned}(20)\qquad &\ell^{(3)}_{\hat\theta,stl}\hat h_s\hat h_t\hat h_l + 3\ell^{(3)}_{\hat\theta,str}\hat h_s\hat h_t\mathbb{E}_{\hat h_{\bar{\mathcal{C}}}|\hat h_{\mathcal{C}}}(\hat h_r)\\ &\quad + 3\ell^{(3)}_{\hat\theta,srv}\hat h_s\mathbb{E}_{\hat h_{\bar{\mathcal{C}}}|\hat h_{\mathcal{C}}}(\hat h_r\hat h_v) + \ell^{(3)}_{\hat\theta,rvk}\mathbb{E}_{\hat h_{\bar{\mathcal{C}}}|\hat h_{\mathcal{C}}}(\hat h_r\hat h_v\hat h_k),\end{aligned}$$

with $s,t,l\in\mathcal{C}$ and $r,v,k\in\bar{\mathcal{C}}$. Therefore, the above expected values simply require the first three noncentral moments of the multivariate Gaussian having density $\phi_{d-d_{\mathcal{C}}}(\hat h_{\bar{\mathcal{C}}};\Lambda_{\mathcal{C}}\hat h_{\mathcal{C}},\bar\Omega)$. These are

$$\mathbb{E}_{\hat h_{\bar{\mathcal{C}}}|\hat h_{\mathcal{C}}}(\hat h_r) = \Lambda_{\mathcal{C},rl}\hat h_l,\qquad \mathbb{E}_{\hat h_{\bar{\mathcal{C}}}|\hat h_{\mathcal{C}}}(\hat h_r\hat h_v) = \bar\Omega_{rv}+\Lambda_{\mathcal{C},rt}\Lambda_{\mathcal{C},vl}\hat h_t\hat h_l,$$

$$\mathbb{E}_{\hat h_{\bar{\mathcal{C}}}|\hat h_{\mathcal{C}}}(\hat h_r\hat h_v\hat h_k) = \bar\Omega_{rv}\Lambda_{\mathcal{C},ks}\hat h_s+\bar\Omega_{rk}\Lambda_{\mathcal{C},vs}\hat h_s+\bar\Omega_{vk}\Lambda_{\mathcal{C},rs}\hat h_s+\Lambda_{\mathcal{C},rs}\Lambda_{\mathcal{C},vt}\Lambda_{\mathcal{C},kl}\hat h_s\hat h_t\hat h_l.$$

Hence, letting

$$\begin{aligned}(21)\qquad \nu^n_{1,s} &= 3\ell^{(3)}_{\hat\theta,srv}\bar\Omega_{rv}+3\ell^{(3)}_{\hat\theta,rvk}\bar\Omega_{rv}\Lambda_{\mathcal{C},ks},\\ \nu^n_{3,stl} &= \ell^{(3)}_{\hat\theta,stl}+3\ell^{(3)}_{\hat\theta,str}\Lambda_{\mathcal{C},rl}+3\ell^{(3)}_{\hat\theta,srv}\Lambda_{\mathcal{C},rt}\Lambda_{\mathcal{C},vl}+\ell^{(3)}_{\hat\theta,rvk}\Lambda_{\mathcal{C},rs}\Lambda_{\mathcal{C},vt}\Lambda_{\mathcal{C},kl},\end{aligned}$$

the summation in (20) can be written as $\nu^n_{1,s}\hat h_s+\nu^n_{3,stl}\hat h_s\hat h_t\hat h_l$, with $s,t,l\in\mathcal{C}$. Replacing this quantity in (19), yields $2\phi_{d_{\mathcal{C}}}(\hat h_{\mathcal{C}};0,\hat\Omega_{\mathcal{CC}})(1/2+\eta\alpha_{\eta,\mathcal{C}}(\hat h_{\mathcal{C}}))$, with

$$(22)\qquad \alpha_{\eta,\mathcal{C}}(\hat h_{\mathcal{C}}) = \{1/(12\eta\sqrt{n})\}(1/n)\big(\nu^n_{1,s}\hat h_s+\nu^n_{3,stl}\hat h_s\hat h_t\hat h_l\big).$$

Therefore, by leveraging the reasoning as in Section 2.1, we can write

$$2\phi_{d_{\mathcal{C}}}(\hat h_{\mathcal{C}};0,\hat\Omega_{\mathcal{CC}})\big(1/2+\eta\alpha_{\eta,\mathcal{C}}(\hat h_{\mathcal{C}})\big) = 2\phi_{d_{\mathcal{C}}}(\hat h_{\mathcal{C}};0,\hat\Omega_{\mathcal{CC}})F\big(\alpha_{\eta,\mathcal{C}}(\hat h_{\mathcal{C}})\big)+O_{P^n_0}(n^{-1}),$$

where $\eta\in\mathbb{R}$, and $F(\cdot):\mathbb{R}\to[0,1]$ is a univariate cdf satisfying $F(-x)=1-F(x)$, $F(0)=1/2$ and $F(x)=F(0)+\eta x+O(x^2)$. As a result, the posterior marginal density for the vector with indexes in $\mathcal{C}$ can be approximated by

$$(23)\qquad \hat p^n_{\mathrm{SKS},\mathcal{C}}(\hat h_{\mathcal{C}}) = 2\phi_{d_{\mathcal{C}}}(\hat h_{\mathcal{C}};0,\hat\Omega_{\mathcal{CC}})w_{\mathcal{C}}(\hat h_{\mathcal{C}}) = 2\phi_{d_{\mathcal{C}}}(\hat h_{\mathcal{C}};0,\hat\Omega_{\mathcal{CC}})F\big(\alpha_{\eta,\mathcal{C}}(\hat h_{\mathcal{C}})\big).$$

Note that $\alpha_{\eta,\mathcal{C}}(\hat h_{\mathcal{C}})$ in (22) is an odd polynomial of $\hat h_{\mathcal{C}}$, and that $\alpha_{\eta,\mathcal{C}}(\hat h_{\mathcal{C}})=\mathbb{E}_{\hat h_{\bar{\mathcal{C}}}|\hat h_{\mathcal{C}}}\{\hat\alpha_\eta(\hat h)\}$.

Equation (23) shows that, once the quantities defining $\hat p^n_{\mathrm{SKS},\mathcal{C}}(\hat h_{\mathcal{C}})$ are pre-computed, then the cost of inference under such an approximating density scales with $d_{\mathcal{C}}$, and no more with d. As a consequence, when the focus is on univariate marginals, that is, $d_{\mathcal{C}}=1$, the computational gains over the joint approximation in (14) can be substantial, and calculation of functionals can be readily performed via one-dimensional numerical integration methods.

Theorem 4.5 below clarifies that, besides being effective from a computational perspective, the above solution preserves the same theoretical accuracy guarantees in approximating the target marginal posterior density $\pi_{n,\mathcal{C}}(\hat h_{\mathcal{C}})=\int\pi_n(\hat h)\,d\hat h_{\bar{\mathcal{C}}}$.

THEOREM 4.5. *Let $\Pi_{n,\mathcal{C}}(S)=\int_S\pi_{n,\mathcal{C}}(\hat h_{\mathcal{C}})\,d\hat h_{\mathcal{C}}$ for $S\subset\mathbb{R}^{d_{\mathcal{C}}}$. Then, under the assumptions of Theorem 4.1, we have that*

$$(24)\qquad \big\|\Pi_{n,\mathcal{C}}(\cdot)-\hat P^n_{\mathrm{SKS},\mathcal{C}}(\cdot)\big\|_{\mathrm{TV}} = O_{P^n_0}(M_n^{c_9}/n),$$

for some $c_9>0$, where $\hat P^n_{\mathrm{SKS},\mathcal{C}}(S)=\int_S\hat p^n_{\mathrm{SKS},\mathcal{C}}(\hat h_{\mathcal{C}})\,d\hat h_{\mathcal{C}}$ with $\hat p^n_{\mathrm{SKS},\mathcal{C}}(\hat h_{\mathcal{C}})$ defined as in (23).

REMARK 4.6. As for Remark 4.4, Theorem 4.5 holds also in the original parametrization θ. Considering the GSN case with $F(\cdot) = \Phi(\cdot)$ and letting $J^{-1}_{\hat{\theta},\mathcal{C}\mathcal{C}} = (J^{-1}_{\hat{\theta}})_{\mathcal{C}\mathcal{C}}$, this implies that

$$\begin{aligned}&\hat{p}^n_{\mathrm{SKS},\mathcal{C}}(\theta_{\mathcal{C}})\\ &\quad = 2\phi_{d_{\mathcal{C}}}(\theta_{\mathcal{C}}; \hat{\theta}_{\mathcal{C}}, J^{-1}_{\hat{\theta},\mathcal{C}\mathcal{C}})\Phi\left(\frac{\sqrt{2\pi}}{12}\left\{\frac{\nu^n_{1,s}}{n}(\theta-\hat{\theta})_s + \nu^n_{3,stl}(\theta-\hat{\theta})_s(\theta-\hat{\theta})_t(\theta-\hat{\theta})_l\right\}\right),\end{aligned} \tag{25}$$

approximates $\pi_{n,\mathcal{C}}(\hat{h}_{\mathcal{C}})$ with rate as in Theorem 4.5, with $s, t, l \in \mathcal{C}$, and $\nu^n_{1,s}$, $\nu^n_{3,stl}$ as in (21).

5. Empirical analysis of skew-modal approximations. Sections 5.1–5.2 demonstrate on both synthetic datasets and real-data applications that the join and marginal skew-modal approximations (SKEW-M) in Section 4 achieve remarkable accuracy improvements relative to the Gaussian-modal counterpart (GM) from the Laplace method. These improvements are again in line with the rates we derived theoretically. Comparisons against other state-of-the-art approximations from VB and EP are also discussed (see Appendix E in the Supplementary Material). In the following, we focus, in particular, on assessing performance of the generalized skew-normal approximations in Remarks 4.4 and 4.6.

5.1. *Exponential model revisited.* Let us first replicate the simulation study in Section 3.1 with focus on the practical skew-modal approximation in Section 4.1, rather than its population version which assumes knowledge of θ_*. Consistent with this focus, the performance of the SKEW-M approximation in (18) is compared against the GM solution $\mathrm{N}(\hat{\theta}, J^{-1}_{\hat{\theta}})$ arising from the Laplace method (e.g., Gelman et al. ((2014), p. 318)). Note that the additional Assumptions 9–10 required by Theorem 4.1 and Remark 4.4 are satisfied. In fact, $\hat{\theta}$ is asymptotically equivalent to the maximum likelihood estimator which implies that Assumption 9 is fulfilled. Moreover, in view of the expressions for the first three log-likelihood derivatives within Appendix E.2 of the Supplementary Material, also 10 holds.

Table 2 reports the same summaries as in the second part of Table 1, but now with a focus on comparing the SKEW-M approximation in (18) and the GM solution. Results are in line with those in Section 3.1, and show, for example, that to achieve the same accuracy attained by the skew-modal with $n = 20$, the Gaussian from the Laplace method requires a sample size of $\bar{n} \approx 500$. These results are strengthened in Appendix E.4 of the Supplementary Material which confirms the findings of Section 3.1 and again clarifies that the theory in Theorem 4.1 closely matches the empirical behavior observed in practice (including in misspecified settings).

5.2. *Probit and logistic regression model.* We consider now a real-data application on the Cushings dataset (Venables and Ripley (2002)), openly available in the `R` library `Mass`. In this case the true data-generative model is not known and, therefore, this analysis is useful to evaluate again the performance in possibly misspecified contexts.

The data are obtained from a medical study on $n = 27$ individuals, aimed at investigating the relationship between four different sub-types of the Cushing's syndrome and two steroid

TABLE 2

For each n from $n = 10$ to $n = 50$, sample size $\bar{n}$ required by the classical Gaussian from the Laplace method (GM) *to obtain the same* TV *and* FMAE *achieved by our skew-modal solution* (SKEW-M) *with sample size n*

	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 50$
$\bar{n} : \mathrm{TV}^{\bar{n}}_{\mathrm{GM}} = \mathrm{TV}^{n}_{\mathrm{SKEW\text{-}M}}$	150	260	470	730	>2500
$\bar{n} : \mathrm{FMAE}^{\bar{n}}_{\mathrm{GM}} = \mathrm{FMAE}^{n}_{\mathrm{SKEW\text{-}M}}$	190	390	650	1030	>2500

TABLE 3

For the probit and logistic regression, comparison among the accuracy of the skew-modal approximation (SKEW-M) *and the classical Gaussian one from the Laplace method* (GM). *Performance is measured in terms of* (*i*) TV *distances from the target joint posterior and its marginals*, (*ii*) *absolute error* (ERR) *in approximating the posterior means and* (*iii*) *average absolute error* (AVE-PR) *in the approximation of the posterior probabilities of being affected by bilateral hyperplasia for each patient. Bold values indicate best performance for each measure*

	TV_{θ}	TV_{θ_0}	TV_{θ_1}	TV_{θ_2}	ERR_{θ_0}	ERR_{θ_1}	ERR_{θ_2}	AVE-PR
Probit								
SKEW-M	**0.10**	**0.02**	**0.04**	**0.05**	**0.003**	**0.002**	**0.020**	**0.005**
GM	0.18	0.09	0.07	0.11	0.097	0.008	0.054	0.022
Logit								
SKEW-M	**0.14**	**0.05**	**0.07**	**0.07**	**0.072**	**0.006**	**0.052**	**0.009**
GM	0.22	0.10	0.09	0.14	0.183	0.016	0.112	0.033

metabolites. To simplify the analysis, we consider here the binary response $X_i \in \{0; 1\}$ which takes value 1 if patient i is affected by bilateral hyperplasia, and 0 otherwise, for $i = 1 \dots, n$. The two observed covariates z_{i1} and z_{i2} measure the urinary excretion rate for the two steroid metabolites, respectively, of the ith patient. In the following, we focus on the two most popular regression models for binary data, namely, probit and logistic regression with mean functions $\Phi(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2})$ and $1/(1 + \exp[-(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2})])$, respectively.

Under both regression models, Bayesian inference proceeds via independent weakly informative Gaussian priors $\text{N}(0, 25)$ for the coefficients in $\theta = (\theta_0, \theta_1, \theta_2)^{\intercal}$. Such priors, combined with the likelihoods, yield posteriors for θ which we approximate under both the joint and marginal skew-modal approximations (SKEW-M). Table 3 compares, via different measures, the accuracy of these solutions relative to the one obtained under the classical Gaussian approximation from the Laplace method. Notice that all these approximations can be readily derived from the closed-form derivatives of the log-likelihood and log-prior for both the probit and logistic regression. Moreover, since the prior distribution is Gaussian, the MAP under both models coincides with the ridge-regression estimator and therefore can be computed via basic R functions.

Table 3 displays the Monte Carlo estimates of the TV distances from the posterior distribution and its marginals, along with errors in approximating the posterior means for the three regression parameters and the posterior probabilities of being affected by a bilateral hyperplasia. Under probit, the latter quantity is defined as $\text{AVE-PR} = \sum_{i=1}^{n} |\text{pr}_i - \hat{\text{pr}}_{\text{APP},i}|/n$ with $\text{pr}_i = \int \Phi(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2}) \pi_n(\theta)\, d\theta$ and $\hat{\text{pr}}_{\text{APP},i} = \int \Phi(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2}) \hat{p}^n_{\text{APP}}(\theta)\, d\theta$, for each $i = 1, \dots, n$, where $\hat{p}^n_{\text{APP}}(\theta)$ is any generic approximation for $\pi_n(\theta)$. The logistic case follows by replacing $\Phi(\cdot)$ with the logit link. These measures and all those in Table 3 are computed via Monte Carlo as detailed in the code at https://github.com/Francesco16p/SMA. Samples under the target posterior are obtained via Hamiltonian Monte Carlo.

As shown in Table 3, the proposed SKEW-M solutions generally yield remarkable accuracy improvements relative to GM, under both models. More specifically, SKEW-M almost halves, on average, the TV distance associated with GM, while providing a much more accurate approximation for the posterior means and posterior probabilities. This is an important accuracy gain provided that the ratio between the absolute error made by GM in posterior means approximation and the actual value of these posterior means is, on average, ≈ 0.25. These gains are observed also when comparing the approximated 95% highest posterior density intervals with those of the target posterior. Also in this case SKEW-M is twice more accurate than GM.

As discussed in the Supplementary Material, the SKEW-M outperforms also state-of-the-art VB (Consonni and Marin (2008), Durante and Rigon (2019), Fasano, Durante and Zanella

(2022)), and is competitive with EP (Chopin and Ridgway (2017)). The latter result is particularly remarkable since SKEW-M only leverages the local behavior of the posterior distribution within a neighborhood of its mode, while EP is known to provide an accurate global solution aimed at matching the first two moments of the target posterior. Appendix E.6 in the Supplementary Material clarifies that important empirical gains are achieved also by the marginal skew-modal approximation from Section 4.2, even when the focus is on a high-dimensional study with $n = 333$ and $d = 135$.

6. Discussion. Through a novel treatment of a third order version of the Laplace method, this article shows that it is possible to derive valid, closed-form and tractable SKS approximations of posterior distributions. Under general assumptions which account for both misspecified models and non-i.i.d. settings, such a novel family of approximations is shown to admit a Bernstein–von Mises type result that establishes remarkable improvements in convergence rates to the target posterior relative to those of the classical Gaussian limiting approximation. The specialization of this general theory to regular parametric models yields SKS approximations with direct methodological impact and immediate applicability under a novel skew-modal solution that is obtained by replacing the unknown θ_* entering the theoretical version with the MAP. The empirical studies on both simulated data and real applications confirm that the remarkable accuracy improvements encoded in our asymptotic and nonasymptotic theory are visible also in practice, even for small-sample regimes. This provides further support to the superior theoretical, methodological and practical performance of the proposed skewed approximations.

The above advancements open new avenues that stimulate research in the field of Bayesian inference based on skewed approximations. As shown in a number of contributions appearing after our article and referencing to our results, interesting directions include the introduction of skewness in other deterministic approximations, such as VB (e.g., Tan (2024)), and further refinements of the high-dimensional results implied by the nonasymptotic bounds we derive for the proposed skew-modal approximation. Katsevich (2024) provides an interesting contribution along this latter direction, which leverages a theoretical approach based on Hermite polynomial expansions to show that d can possibly grow faster than $n^{1/3}$, under suitable models. However, unlike for our results, the focus is on studying nonvalid skewed approximating densities. The notion of effective dimension $\tilde{d}$ introduced by Spokoiny and Panov (2021) and Spokoiny (2025) for the study of the Gaussian Laplace approximation in high dimensions is also worth further investigations under our skewed extension, since $\tilde{d}$ can be possibly $o(d)$.

Semiparametric settings (e.g., Bickel and Kleijn (2012), Castillo and Rousseau (2015)) are also of interest. Additionally, although the inclusion of skewness is arguably sufficient to yield an accurate approximation of intractable posterior distributions, accounting for kurtosis might provide additional improvements both in theory and in practice. To this end, a relevant research direction is to seek for an alternative to the Gaussian density in the symmetric part, possibly obtained from an extension to the fourth order of our novel treatment of the Laplace method. Our conjecture is that such a generalization would provide an additional order-of-magnitude improvement in the rates, while yielding an approximation still within the general skew-symmetric family.

Acknowledgments. We are grateful to the Editor, the Associate Editor and the referees for the constructive feedbacks, which helped us in improving the preliminary version of the article.

Funding. Botond Szabo is co-funded by the European Union (ERC grant, BigBayesUQ, project n. 101041064).

Francesco Pozza is funded by the European Union (ERC grant, PrSc-HDBayLe, project n. 101076564). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

SUPPLEMENTARY MATERIAL

Supplementary Material to "Skewed Bernstein–von Mises theorem and skew-modal approximations" (DOI: 10.1214/24-AOS2429SUPPA; .pdf). This supplement contains proofs, technical lemmas and further results.

Code for "Skewed Bernstein–von Mises theorem and skew-modal approximations" (DOI: 10.1214/24-AOS2429SUPPB; .zip). This directory contains code to implement the analyses in Section 5.2. For the most recent and updated version of the code, refer to https://github.com/Francesco16p/SMA.

REFERENCES

ANCESCHI, N., FASANO, A., DURANTE, D. and ZANELLA, G. (2023). Bayesian conjugacy in probit, tobit, multinomial probit and extensions: A review and new results. *J. Amer. Statist. Assoc.* **118** 1451–1469. MR4595508 https://doi.org/10.1080/01621459.2023.2169150

ARELLANO-VALLE, R. B. and AZZALINI, A. (2006). On the unification of families of skew-normal distributions. *Scand. J. Stat.* **33** 561–574. MR2298065 https://doi.org/10.1111/j.1467-9469.2006.00503.x

AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 367–389. MR1983753 https://doi.org/10.1111/1467-9868.00391

BERNSTEIN, S. (1917). Theory of Probability, Moskow.

BICKEL, P. J. and KLEIJN, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *Ann. Statist.* **40** 206–237. MR3013185 https://doi.org/10.1214/11-AOS921

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BOCHKINA, N. A. and GREEN, P. J. (2014). The Bernstein–von Mises theorem and nonregular models. *Ann. Statist.* **42** 1850–1878. MR3262470 https://doi.org/10.1214/14-AOS1239

BOUCHERON, S. and GASSIAT, E. (2009). A Bernstein–von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148. MR2471588 https://doi.org/10.1214/08-EJS262

CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473 https://doi.org/10.1214/14-AOS1246

CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

CHALLIS, E. and BARBER, D. (2012). Affine independent variational inference. *Adv. Neural Inf. Process. Syst.* **25** 1–9.

CHOPIN, N. and RIDGWAY, J. (2017). Leave Pima Indians alone: Binary regression as a benchmark for Bayesian computation. *Statist. Sci.* **32** 64–87. MR3634307 https://doi.org/10.1214/16-STS581

CONSONNI, G. and MARIN, J.-M. (2008). Mean-field variational approximate Bayesian inference for latent variable models. *Comput. Statist. Data Anal.* **52** 790–798. MR2418528 https://doi.org/10.1016/j.csda.2006.10.028

DEHAENE, G. and BARTHELMÉ, S. (2018). Expectation propagation in the large data limit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 199–217. MR3744718 https://doi.org/10.1111/rssb.12241

DURANTE, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika* **106** 765–779. MR4031198 https://doi.org/10.1093/biomet/asz034

DURANTE, D., POZZA, F. and SZABO, B. (2024). Supplement to "Skewed Bernstein–von Mises theorem and skew-modal approximations." https://doi.org/10.1214/24-AOS2429SUPPA, https://doi.org/10.1214/24-AOS2429SUPPB

DURANTE, D. and RIGON, T. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statist. Sci.* **34** 472–485. MR4017524 https://doi.org/10.1214/19-STS712

FASANO, A. and DURANTE, D. (2022). A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *J. Mach. Learn. Res.* **23** Paper No. [30], 26. MR4420755

FASANO, A., DURANTE, D. and ZANELLA, G. (2022). Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika* **109** 901–919. MR4519107 https://doi.org/10.1093/biomet/asac026

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Stat.* **41** 851–864. MR0263198 https://doi.org/10.1214/aoms/1177696963

KASPRZAK, M. J., GIORDANO, R. and BRODERICK, T. (2022). How good is your Gaussian approximation of the posterior? Finite-sample computable error bounds for a variety of useful divergences. Available at arXiv:2209.14992.

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics*: *Essays in Honor of George A. Barnard* 473–487.

KATSEVICH, A. (2024). The Laplace approximation accuracy in high dimensions: a refined analysis and new skew adjustment. Available at arXiv:2306.07262.

KATSEVICH, A. and RIGOLLET, P. (2024). On the approximation accuracy of Gaussian variational inference. *Ann. Statist.* **52** 1384–1409. MR4804813 https://doi.org/10.1214/24-aos2393

KLEIJN, B. J. K. and VAN DER VAART, A. W. (2012). The Bernstein–Von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381. MR2988412 https://doi.org/10.1214/12-EJS675

KOERS, G., SZABO, B. and VAN DER VAART, A. (2023). Misspecified Bernstein–Von Mises theorem for hierarchical models. Available at arXiv:2308.07803.

KOLASSA, J. E. and KUFFNER, T. A. (2020). On the validity of the formal Edgeworth expansion for posterior densities. *Ann. Statist.* **48** 1940–1958. MR4134781 https://doi.org/10.1214/19-AOS1871

LAPLACE, P. S. (1810). *Théorie Analytique des Probabilités*, 3rd ed. Courcier, Paris.

LE CAM, L. and YANG, G. L. (1990). *Asymptotics in Statistics*: *Some Basic Concepts*. *Springer Series in Statistics*. Springer, New York. MR1066869 https://doi.org/10.1007/978-1-4684-0377-0

LECAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Stat.* **1** 277–329. MR0054913

MA, Y. and GENTON, M. G. (2004). Flexible class of skew-symmetric distributions. *Scand. J. Stat.* **31** 459–468. MR2087837 https://doi.org/10.1111/j.1467-9469.2004.03_007.x

MCCULLAGH, P. (2018). *Tensor Methods in Statistics*, 2nd ed. Dover Publications, New York.

MINKA, T. P. (2001). Expectation propagation for approximate Bayesian inference. *Proc. Uncertainty Artif. Intell.* **17** 362–369.

OPPER, M. and ARCHAMBEAU, C. (2009). The variational Gaussian approximation revisited. *Neural Comput.* **21** 786–792. MR2478318 https://doi.org/10.1162/neco.2008.08-07-592

PACE, L. and SALVAN, A. (1997). *Principles of Statistical Inference*: *From a Neo-Fisherian Perspective*. *Advanced Series on Statistical Science & Applied Probability* **4**. World Scientific, River Edge, NJ. MR1476674

PANOV, M. and SPOKOINY, V. (2015). Finite sample Bernstein–von Mises theorem for semiparametric problems. *Bayesian Anal.* **10** 665–710. MR3420819 https://doi.org/10.1214/14-BA926

RAY, K. and SZABÓ, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *J. Amer. Statist. Assoc.* **117** 1270–1281. MR4480711 https://doi.org/10.1080/01621459.2020.1847121

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SPOKOINY, V. (2025). Inexact Laplace approximation and the use of posterior mean in Bayesian inference. *Bayesian Anal.* **20** 1303–1330. MR4832249 https://doi.org/10.1214/23-BA1391

SPOKOINY, V. and PANOV, M. (2021). Accuracy of Gaussian approximation for high-dimensional posterior distribution. *Bernoulli* (in print).

TAN, L. S. (2024). Variational inference based on a subclass of closed skew normals. *J. Comput. Graph. Statist.* (in print).

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VEHTARI, A., GELMAN, A., SIVULA, T., JYLÄNKI, P., TRAN, D., SAHAI, S., BLOMSTEDT, P., CUNNINGHAM, J. P., SCHIMINOVICH, D. et al. (2020). Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *J. Mach. Learn. Res.* **21** Paper No. 17, 53. MR4071200

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.

VON MISES, R. (1931). *Wahrscheinlichkeitsrechnung*. Springer, Berlin.

WANG, C. and BLEI, D. M. (2013). Variational inference in nonconjugate models. *J. Mach. Learn. Res.* **14** 1005–1031. MR3063617

WANG, J., BOYER, J. and GENTON, M. G. (2004). A skew-symmetric representation of multivariate distributions. *Statist. Sinica* **14** 1259–1270. MR2126352

WANG, Y. and BLEI, D. M. (2019). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* **114** 1147–1161. MR4011769 https://doi.org/10.1080/01621459.2018.1473776

WENG, R. C. (2010). A Bayesian Edgeworth expansion by Stein's identity. *Bayesian Anal.* **5** 741–763. MR2740155 https://doi.org/10.1214/10-BA526