Tight Dimension Dependence of the Laplace Approximation

Anya Katsevich
akatsevi@mit.edu

May 21, 2023

Abstract

In Bayesian inference, a common technique to approximately sample from and compute statistics of high-dimensional posteriors is to use the Laplace approximation — a Gaussian proxy for the true posterior. The Laplace approximation accuracy improves as sample size grows, but the question of how fast dimension d can grow with sample size n has not been fully resolved. Prior works have shown that $d^3 \ll n$ is a sufficient condition for accuracy of the approximation. But by deriving the leading order asymptotics of the TV distance between the two measures, we show that in fact $d^2 \ll n$ is sufficient, and we show for a logistic regression posterior that this growth condition is necessary. Furthermore, through another leading order asymptotic expansion, we derive a computable correction to the mode, which is the Laplace approximation to the mean. Incorporating this *skew correction* improves the mean approximation accuracy by two orders of magnitude.

1 Introduction

Consider the posterior distribution π of a parameter $x \in \mathbb{R}^d$ given i.i.d. data $y_i, i = 1 \ldots, n$, a prior ν on x, and a data-generating model $p(\cdot \mid x)$:

$$\pi(x) = \pi(x \mid \{y_i\}_{i=1}^n) \propto \nu(x) \prod_{i=1}^n p(y_i \mid x), \quad x \in \mathbb{R}^d. \tag{1.1}$$

In modern applications, the dimension d can be very large, so that computing summary statistics of π — given by high dimensional integrals — is the central challenge in Bayesian inference. A standard approach in Bayesian inference to approximate integrals with respect to π is to use the *Laplace approximation*: namely, to write $\pi \propto e^{-V}$ and replace V by its second order Taylor expansion around the minimizer $\hat{m}$ of V, which is the mode of π. This yields the Gaussian density

$$\hat{\gamma} = \mathcal{N}\left(\hat{m},\ \nabla^2 V(\hat{m})^{-1}\right), \quad \hat{m} = \arg\min_{x \in \mathbb{R}^d} V(x), \tag{1.2}$$

and we can now easily evaluate integrals with respect to $\hat{\gamma}$. In fact, this approximation automatically gives us an estimate for the mean and covariance of π, but we can also approximate expectations of other observables f (e.g. indicators of credible sets) as $\int f d\pi \approx \int f d\hat{\gamma}$. To understand the rationale behind the Laplace approximation, note that for $\pi \propto e^{-V}$ of the form (1.1), we can write $V(x) = nv(x)$, where

$$v(x) = -\frac{1}{n} \log \nu(x) - \frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid x). \tag{1.3}$$

Therefore,

$$\int f d\pi = \frac{\int_{\mathbb{R}^d} f e^{-nv} dx}{\int_{\mathbb{R}^d} e^{-nv} dx}, \tag{1.4}$$

and such integrals can be approximated using *Laplace's method* for large n (note however that v can depend on n, while in the classical theory v remains fixed). The idea of Laplace's method is that when n is large, the density $\pi \propto e^{-nv(x)}$ concentrates around $x = \hat{m}$, the global minimizer of v. But near $\hat{m}$, the function v is well-approximated by its second order Taylor expansion. The resulting density is precisely the Gaussian density $\hat{\gamma}$ defined in (1.2).

The central question addressed in this work is the accuracy of the Laplace approximation in the case when dimension d grows with sample size n. Specifically, how fast can d grow with n such that $\hat{\gamma}$ still yields an accurate approximation to π when $n \gg 1$? To address this question, we derive the leading order asymptotics of the observable expectation error $\int f d\pi - \int f d\hat{\gamma}$ for generic f. We use this expansion to conclude that the TV distance between π and $\hat{\gamma}$ is small provided $d^2 \ll n$, up to v-dependent constants (which we will discuss at length). Furthermore, the asymptotic expansion opens the door to *correcting* the Laplace approximation, since the leading order term of $\int f d\pi - \int f d\hat{\gamma}$ is explicit and computable. In particular, for $f(x) = x$ this amounts to a computable skew correction to the mode $\hat{m}$, which improves the approximation of the mean by two orders of magnitude.

Before describing our contributions in more detail, we explain the "fixed-sample" and model-agnostic perspective taken in this work.

The Laplace approximation as fixed-sample BvM. Due to the randomness of the data $\{y_i\}_{i=1}^n$, the posterior π is itself random. The Bernstein-von Mises (BvM) theorem considers π in the large n limit under the frequentist assumption that $\{y_i\}_{i=1}^n$ is drawn i.i.d. from the distribution $p(\cdot \mid x_0)$, where x_0 is the ground truth parameter. In this case, the BvM theorem states that, with high probability under the ground truth data-generating process, the posterior, when centered on an efficient estimator, converges in TV as $n \to \infty$ to a zero mean Gaussian with covariance matrix given by the inverse of the Fisher information at x_0.The classic BvM theorem is stated for fixed parameter dimension, but in the last several decades, BvM results in growing parameter dimension have been shown. We review these below.

In practice, the BvM has two drawbacks. First, it does not yield an implementable Gaussian approximation, since the covariance of the limiting Gaussian depends on the ground truth. Second, it does not allow us to quantify the distance between the limiting Gaussian and a *given* posterior π as a function of the *fixed samples* $\{y_i\}_{i=1}^n$. But the Laplace approximation *is* implementable, and recent works (discussed below) have obtained error bounds on this approximation which depend explicitly on the function v for which π is defined via $\pi \approx e^{-nv}$. Note that v incorporates both the log prior and the log-likelihood, which depends on the given samples $\{y_i\}_{i=1}^n$. For a distance δ between measures (or an f-divergence), such a bound takes the general form

$$\delta(\pi, \hat{\gamma}) \leq \epsilon(d, n, v). \tag{1.5}$$

Typically, $\epsilon(d, n, v)$ depends on v through its derivatives in a neighborhood of the mode $\hat{m}$. We, too, take this fixed-sample perspective. Our approach is also model agnostic: we do not require that π has the form of a posterior distribution at all. Rather, we simply assume $\pi \propto e^{-nv}$ for some function v. Note that v is defined on $\mathbb{R}^d$, and therefore inherently depends on dimension. We also allow v to depend on n, although our results our most useful in the case when this dependence is mild, as in (1.3).

Dimension dependence: prior work. A long line of work has established in various settings that the BvM holds when $d = d_n$ satisfies $d_n^3/n \to 0$ as $n \to \infty$, up to logarithmic and model-specific factors. [Ghosal, 1999, Ghosal, 2000] prove BvMs for posteriors arising from linear regression models and exponential families, respectively. [Boucheron and Gassiat, 2009] proves a BvM for the posterior of a discrete probability mass function truncated to its first d_n entries. [Spokoiny, 2013, Panov and Spokoiny, 2015] prove BvMs for growing parametric, and semiparametric statistical models, respectively. [Lu, 2017] proves a BvM for nonlinear Bayesian inverse problems. See all of the above works for further references on BvMs with growing parameter dimension.

We have summarized the above BvM results under the generic condition $d_n^3/n \to 0$ as $n \to \infty$. However, we emphasize that important model-specific growth conditions must also be satisfied. To give an example, [Lu, 2017] requires that $\sigma(d_n)^2 \log d_n \sqrt{d_n^3/n} \to 0$, where $\sigma(d_n)^{-1}$ is a lower bound on the smallest eigenvalue of the gradient of the forward operator. [Boucheron and Gassiat, 2009] requires that $d_n^3/(n \inf_{i \leq d_n} \theta_0(i)) \to 0$, where $\theta_0(i)$ is the probability of state $i = 1, 2, 3, \ldots$ under the ground truth probability mass function θ_0.

More recent "finite-sample" works, which bound the accuracy of the Laplace approximation (1.2), have conveyed these model specific growth conditions through the function v (where $\pi \propto e^{-nv}$). Namely, they obtain bounds in terms of a "universal" factor depending on d and n only, and a "model-specific" factor depending on v. Specifically, the works [Helin and Kretschmann, 2022, Spokoiny, 2022, Dehaene, 2019] have obtained error bounds on the Laplace approximation of the form

$$\mathrm{TV}(\pi, \hat{\gamma}) \lesssim c_3(v)\sqrt{d^3/n}, \quad \mathrm{KL}(\,\pi \,||\, \hat{\gamma}) \lesssim c_3(v)^2 d^3/n.$$

Here, $c_3(v)$ is a ratio of third and second order derivatives of v, whose definition varies slightly from paper to paper. We note that the TV and KL bounds due to [Spokoiny, 2022] are actually in terms of d_{eff}^3/n, where $d_{\text{eff}} \leq d$ is an *effective* dimension which depends on the strength of regularization by a Gaussian prior.

Remarkably, the above TV bound was recently tightened in [Kasprzak et al., 2022]. Without strengthening the assumptions of the above works, the authors show that in fact,

$$\text{TV}(\pi, \hat{\gamma}) \lesssim c_3(v)\sqrt{d^2/n}, \tag{1.6}$$

where $c_3(v)$ is an analogous constant to the above. Interestingly, dimension dependence was not the stated aim of this work, and the dependence of the bound on d was not made explicit in the paper. We review this work and the authors' proof method in more detail below.

Main Contributions. The result (1.6) by [Kasprzak et al., 2022] tightens the dimension dependence of previous bounds, showing that $d^2 \ll n$ is *sufficient* for Laplace approximation accuracy, up to model-dependent conditions. However, the question remains: what are the *necessary* conditions on d, n, and v to ensure Laplace approximation accuracy? This question can be resolved with an asymptotic expansion of the TV error. Early progress toward an asymptotic expansion in the high dimensional regime was made in [Shun and McCullagh, 1995].

Our first main contribution is to derive the leading order asymptotics of the TV error, bringing greater clarity to the question of Laplace approximation accuracy as a function of v, d, and n.

Theorem 1.1 (Informal)**.** *Let $V = nv \in C^4$ have a unique strict minimum $\hat{m}$. Let $\pi \propto e^{-V}$ and $\hat{\gamma} = \mathcal{N}\left(\hat{m},\, \nabla^2 V(\hat{m})^{-1}\right)$ be its Laplace approximation. If the third and fourth order derivatives of V grow at most polynomially and V grows at least polynomially (any power greater than zero) at infinity, then for an explicit quantity*

$$L = L(\nabla^2 V(\hat{m}), \nabla^3 V(\hat{m})),$$

we have the decomposition $\text{TV}\left(\pi,\, \hat{\gamma}\right) = L + R$. The leading order term L and the remainder R are bounded as

$$0 \leq L \leq c_3(v)\frac{d}{\sqrt{n}}, \qquad |R| \leq f(c_3(v), c_4(v))\left(\frac{d}{\sqrt{n}}\right)^2.$$

See Theorem V1 for the formal statement and explicit formulas for L and f, and Section 2.1 for the definitions of c_3 and c_4. In most cases, the upper bound on R is an order of magnitude smaller than the upper bound on L. Therefore, the first implication of the theorem is a tighter upper bound on the TV error thanks to tight control over the explicit term L.

More importantly, however, the term L determines the relationship between d, n, and v under which the Laplace approximation is accurate. We believe the upper bound $|L| \leq c_3(v)d/\sqrt{n}$ is generically tight. But even in cases where it is not tight, the term L still serves as the starting point for analysis. For example, if $L \sim (d/\sqrt{n})^2$, this implies one should compute the second order

term in the expansion (which our theory provides in semi-explicit form). If e.g. $L \sim \sqrt{d}/\sqrt{n}$, then this is a useful indication that $d/\sqrt{n}$ is not the fundamental unit of error.

As mentioned above, we prove Theorem 1.1 by deriving the leading order asymptotics for the observable expectation error $\int f d\pi - \int f d\hat{\gamma}$. This allows one to improve the accuracy of the Laplace approximation to an observable expectation of interest. Arguably the most important observable is the mean. Thus, **our second main contribution is to derive a skew correction to the mode to better approximate the mean**.

Theorem 1.2 (Informal)**.** *Consider the same setting and conditions as in Theorem 1.1, and additionally suppose $V \in C^5$, with $\nabla^5 V$ growing polynomially at infinity. Let m_π be the first moment of π, and note that $\hat{m}$ is the first moment of $\hat{\gamma}$. Then for an explicit quantity*

$$L = L(\nabla^2 V(\hat{m}), \nabla^3 V(\hat{m})),$$

we have the decomposition $\nabla^2 V(\hat{m})^{1/2}(m_\pi - \hat{m}) = L + R$, where

$$\|L\| \leq c_3(v)\left(\frac{d}{\sqrt{n}}\right), \qquad \|R\| \leq f\left(c_3(v), c_4(v), c_5(v)\right)\left(\frac{d}{\sqrt{n}}\right)^3.$$

The message of this theorem is that $\hat{L} := \nabla^2 V(\hat{m})^{-1/2} L$ is an explicit, computable *skew correction* term. Indeed, when π is skewed, the mode $\hat{m}$ is likely not to be a good approximation of the mean m_π. But the theorem shows that estimating m_π by $\hat{m} + \hat{L}$ rather than by $\hat{m}$ decreases the normalized error by a factor of d^2/n. See the below paragraph on computability, as well as Example 2.1, for a discussion of the cost of computing $\hat{L}$.

We note that [Kasprzak et al., 2022] also bounds the mean error (without the correction), obtaining the same order of magnitude, $d/\sqrt{n}$, as do we in the above theorem. The recent work [Durante et al., 2023] also derives a skew correction — not just to the mode $\hat{m}$, but to the Gaussian distribution $\hat{\gamma}$ itself. The resulting distribution, which belongs to the class of generalized skew-normal distributions, leads to a TV error of order $O(1/n)$, which is a factor of $1/\sqrt{n}$ better than the TV error for the original Laplace approximation. However, the TV error bound for the skew-normal approximation scales exponentially in dimension.

In addition to the above asymptotics of the mean error, we also derive an upper bound on the covariance error. The leading contribution to this error is already of order $(d/\sqrt{n})^2$ (compare to the leading mean error, which has order $d/\sqrt{n}$), so deriving the explicit correction term that would make the covariance error even smaller seems unnecessary.

A necessary price to pay for the generality of our results is that our bounds depend on the data and statistical model only indirectly, through the constants $c_3(v), c_4(v), c_5(v)$. Determining the scaling of these constants with dimension must be done on a case-by-case basis. We show how this can be done for logistic

regression, an important model used often in practice. Thus **our third main contribution is to prove high probability upper bounds on the Laplace approximation error for a logistic regression posterior with Gaussian design.** We show that if $d^2/n < 1$, then the coefficients $c_3(v)$, $f(c_3(v), c_4(v))$, $f(c_3(v), c_4(v), c_5(v))$ arising in the bounds on L and R in Theorems 1.1 and 1.2 are all bounded by an absolute constant with high probability. This leads to state of the art error bounds on the Laplace approximation for logistic regression, in terms of $d/\sqrt{n}$ alone. We also show numerically that the leading order terms L in the TV and mean error decompositions are bounded from below by $d/\sqrt{n}$, showing that in this example, $d^2/n \ll 1$ *is necessary and sufficient for accurate Laplace approximation.*

Furthermore, the logistic regression analysis paves the way to deriving bounds on the Laplace approximation error for other generalized linear models with Gaussian design.

Computability and comparison to Gaussian VI. Gaussian variational inference (VI) offers another Gaussian approximation to posteriors π. It is defined as

$$\hat{\gamma}^{\mathrm{VI}} = \operatorname*{argmin}_{p \in \mathcal{P}_{\mathrm{Gauss}}} \mathrm{KL}(\, p \,||\, \pi), \tag{1.7}$$

where $\mathcal{P}_{\mathrm{Gauss}}$ is the family of nondegenerate Gaussian distributions on $\mathbb{R}^d$. For a measure $\pi \propto e^{-nv}$ on $\mathbb{R}^d$, [Katsevich and Rigollet, 2023] bounds the mean and covariance error of Gaussian VI in terms of d and n. They found that the normalized mean approximation error is upper bounded by $(d^3/n)^{3/2}$, which in its n dependence significantly outperforms the Laplace mean approximation error. However, Theorem 1.2 shows that if we can compute the skew correction term $\hat{L}$, then the estimate $\hat{m} + \hat{L}$ is just as accurate as the Gaussian VI mean error — order $(d^2/n)^{3/2}$. Note that the Laplace and VI covariance error both have the same n scaling; see Figure 1 in [Katsevich and Rigollet, 2023].

We show in Example 2.1 below that for a generalized linear model, computing the mean correction term $\hat{L}$ is no more computationally expensive than computing the Laplace approximation itself, which requires 1) finding the mode $\hat{m}$, 2) computing the Hessian $\nabla^2 V(\hat{m})$, and 3) inverting the Hessian to obtain the covariance estimate. In fact, fully inverting the Hessian is not even required to compute $\hat{L}$; one need only solve the linear systems $\nabla^2 V(\hat{m}) a_i = x_i$ for a_i, $i = 1, \ldots, n$, where the x_i are the predictor variables. In particular, computing the third derivative tensor $\nabla^3 V(\hat{m})$ is straightforward for GLMs (and moreover, in order to evaluate $\hat{L}$ one need not first compute and store this tensor).

However, for other statistical models — particularly Bayesian inverse problems — computing the third derivative of V may be prohibitively expensive. In this case Gaussian VI may be more feasible, since it only requires the first derivative of V. See [Lambert et al., 2022, Diao et al., 2023] for algorithmic implementations of Gaussian VI.

Organization. The rest of the paper is organized as follows. In Section 2, we state our assumptions and main results on the Laplace approximation. In Section 3, we specialize these results to logistic regression. In Section 4, we present the most general version of our asymptotic expansion to arbitrary order. We outline the proof of the expansion, and show how the theorems in Section 2 follow from it.

Notation. For a measure π on $\mathbb{R}^d$ with finite first and second moments, we let m_π, Σ_π be the mean and covariance of π. For a function V with a unique minimum, we define H_V to be the Hessian of V at the minimum. We let γ denote the density of the standard normal distribution $\mathcal{N}(0, I_d)$ in d dimensions, and we write either $\int f d\gamma$ or $\mathbb{E}[f(Z)]$ or $\gamma(f)$ for the expectation of f under γ. We write $Z = (Z_1, \dots, Z_d)$ to denote a standard multivariate normal random variable $Z \sim \gamma$ in $\mathbb{R}^d$, and Z_1 to denote a standard normal in $\mathbb{R}$. For an observable $f : \mathbb{R}^d \to \mathbb{R}$ such that $\int |f|^p d\gamma < \infty$, we define

$$\|f\|_p = \left(\int |f|^p d\gamma \right)^{\frac{1}{p}}.$$

A tensor T of order k is an array $T = (T_{i_1 i_2 \dots i_k})_{i_1,\dots,i_k=1}^d$. For two order k tensors T and S we let $\langle T, S \rangle$ be the entrywise inner product. We say T is symmetric if $T_{i_1 \dots i_k} = T_{j_1 \dots j_k}$, for all permutations $j_1 \dots j_k$ of $i_1 \dots i_k$.

Let H be a symmetric positive definite matrix. For a vector $x \in \mathbb{R}^d$, we let $\|x\|_H$ denote $\|x\|_H = \sqrt{x^T H x}$. For an order k tensor T, we define the H-weighted operator norm of T to be

$$\|T\|_H := \sup_{\|x_1\|_H = \dots = \|x_k\|_H = 1} \langle T, x_1 \otimes \dots \otimes x_k \rangle. \tag{1.8}$$

When $H = I_d$, the norm $\|T\|_{I_d}$ is the regular operator norm, and in this case we omit the subscript. For a symmetric, order 3 tensor T and a symmetric matrix A, we let $\langle T, A \rangle \in \mathbb{R}^d$ be the vector with coordinates

$$\langle T, A \rangle_i = \sum_{j,k=1}^d T_{ijk} A_{jk}, \quad i = 1, \dots, d. \tag{1.9}$$

Note that $\|\langle T, A \rangle\| = \sup_{\|u\|=1} \langle T, A \otimes u \rangle$, and if $A = \sum_{i=1}^d \lambda_i v_i v_i^T$ is the eigendecomposition of A then

$$\|\langle T, A \rangle\| = \sup_{\|u\|=1} \langle T, A \otimes u \rangle \le \sum_{i=1}^d |\lambda_i| \, |\langle T, v_i \otimes v_i \otimes u \rangle| \le d \|A\| \|T\|. \tag{1.10}$$

2 Main Result

Let $\pi \propto e^{-nv}$ on $\mathbb{R}^d$. In this section, we first state our assumptions on v, n, and d. We then present our results on 1) the TV distance between π and its Laplace approximation, 2) the first moment error (with and without skew correction), and 3) the covariance error.

2.1 Assumptions on the potential

Here, we state and discuss our assumptions on v, d, and n.

Assumption A1. $v \in C^4$, with unique global minimizer $x = \hat{m}$, and $H_v = \nabla^2 v(\hat{m}) \succ 0$.

Assumption A2. There exist $c_3, c_4 > 0$ such that the following bounds hold on the operator norms of third and fourth order derivative tensors, at $\hat{m}$ and in a vanishing neighborhood of $\hat{m}$, respectively:

$$\begin{aligned} \|\nabla^3 v(\hat{m})\|_{H_v} &\le c_3, \\ \|\nabla^4 v(\hat{m}+x)\|_{H_v} &\le c_4, \quad \forall \|x\|_{H_v} \le 4\sqrt{d/n}. \end{aligned} \tag{2.1}$$

Assumption A3. For some $q > 0$ and the same c_4 as in Assumption A2, we have the following global bound on the growth of the fourth derivative of $t \mapsto v(\hat{m} + tu)$:

$$|\langle \nabla^4 v(\hat{m}+tu), u^{\otimes 4}\rangle| \le c_4 \max(1,t)^q, \quad \forall \|u\|_{H_v} = 1, t \ge 0. \tag{2.2}$$

Note that $\sup_{\|u\|_{H_v}=1} |\langle \nabla^4 v(\hat{m}+tu), u^{\otimes 4}\rangle| \le \sup_{\|u\|_{H_v}=1} \|\nabla^4 v(\hat{m}+tu)\|_{H_v}$. Therefore, Assumption A3 is implied by the following, stronger assumption, which may be easier to interpret: for some $q > 0, \tilde{c}_4 > 0$ we have

$$\|\nabla^4 v(\hat{m}+x)\|_{H_v} \le \begin{cases} \tilde{c}_4, & \|x\|_{H_v} \le 1, \\ \tilde{c}_4 \|x\|_{H_v}^q, & \|x\|_{H_v} \ge 1. \end{cases} \tag{2.3}$$

If (2.3) holds, then both the second bound in (2.1) and Assumption A3 hold with $c_4 = \tilde{c}_4$. Note that since $v \in C^4$, we can always find a finite constant $\tilde{c}_4$ such that $\sup_{\|x\|_{H_v} \le 1} \|\nabla^4 v(\hat{m}+x)\|_{H_v} \le \tilde{c}_4$. Therefore, (2.3) simply states that we can extend this uniform (constant) bound inside the unit ball $\{\|x\|_{H_v} \le 1\}$, to a polynomial growth bound outside of it. The actual assumption A3 is analogous but slightly weaker, since we only consider the action of $\nabla^4 v(\hat{m}+tu)$ in the direction of u itself. This only makes a difference for $t > 0$, since at $t = 0$ we have $\sup_{\|u\|_{H_v}=1} \langle \nabla^4 v(\hat{m}), u^{\otimes 4}\rangle = \|\nabla^4 v(\hat{m})\|_{H_v}$. Thus in particular, Assumption A3 implies that c_4 is no smaller than $\|\nabla^4 v(\hat{m})\|_{H_v}$ and in fact, no smaller than $\sup_{\|x\|_{H_v} \le \sqrt{d/n}} \|\nabla^4 v(\hat{m}+x)\|_{H_v}$ by Assumption A2.

Assumption A4. For some $c_0 > 0, \alpha > 0$, and $0 < r < 1$ satisfying

$$4c_3 r + c_4 r^2 \le 6, \tag{2.4}$$

the following lower bound holds on the growth of v at far-range:

$$v(\hat{m}+x) - v(\hat{m}) \ge c_0 \|x/r\|_{H_v}^{\alpha}, \quad \forall \|x\|_{H_v} \ge r \tag{2.5}$$

Remark 2.1. *It will be important for our main results in Section 2.2 that nc_0 satisfy some minimal growth condition. To verify this condition, it is convenient if we can take $c_0 = \inf_{\|x\|_{H_v}=r} [v(\hat{m}+x) - v(\hat{m})]$ in (2.5), meaning that*

$$\inf_{\|x\|_{H_v}\geq r} \frac{v(\hat{m}+x)-v(\hat{m})}{\|x/r\|_{H_v}^{\alpha}} = \inf_{\|x\|_{H_v}=r} v(\hat{m}+x)-v(\hat{m}) =: c_0 \tag{2.6}$$

The convenience of this c_0 is that it is easy to bound from below. Indeed, Lemma 4.3 below implies that

$$c_0 = \inf_{\|x\|_{H_v}=r} v(\hat{m}+x)-v(\hat{m}) \geq r^2/4. \tag{2.7}$$

The condition (2.6) is slightly stronger than Assumption A4, which only requires that the lefthand side of (2.6) is bounded below by some $c_0 > 0$. *However, if v is convex, for example, then it is straightforward to show that (2.6) holds for $\alpha = 1$ and any $r > 0$. Therefore, assuming (2.6) holds for some $0 < \alpha < 1$ and some $r > 0$ is weaker than convexity, though stronger than (2.5).*

Our theorem on the first moment approximation error will also require the following additional assumption.

Assumption A5. $v \in C^5$, and there exists a positive constant c_5 such that with the same q from Assumption A3, we have the following bound on the growth of the fifth derivative of $t \mapsto v(\hat{m}+tu)$:

$$|\langle \nabla^5 v(\hat{m}+tu), u^{\otimes 5}\rangle| \leq c_5 \max\left(1, \frac{t}{\sqrt{d/n}}\right)^q, \quad \forall \|u\|_{H_v} = 1, t \geq 0. \tag{2.8}$$

This assumption on the fifth derivative is analogous to but slightly weaker than Assumption A3 on the fourth derivative. Here, the transition from uniform boundedness to polynomial growth occurs at $\|x\|_{H_v} = \sqrt{d/n}$ rather than at $\|x\|_{H_v} = 1$ (where $x = tu$).

Remark 2.2. *Note that for all $\|u\|_{H_v} = 1$ and $t \geq 0$, we have*

$$|\langle \nabla^k v(\hat{m}+tu), u^{\otimes k}\rangle| \leq \|\nabla^k v(\hat{m}+tu)\|_{H_v} \leq \frac{\|\nabla^k v(\hat{m}+tu)\|}{\lambda_{\min}(H_v)^{k/2}} \tag{2.9}$$

(The lefthand inequality for $k = 4$ was already discussed following Assumption A3.) Therefore, Assumptions A2, A3, and A5 are implied by the following three inequalities, respectively:

$$\begin{aligned}
&\frac{\|\nabla^3 v(\hat{m})\|}{\lambda_{\min}(H_v)^{3/2}} \leq c_3, \\
&\frac{\|\nabla^4 v(\hat{m}+x)\|}{\lambda_{\min}(H_v)^2} \leq c_4 \max(1, \|x\|_{H_v})^q, \quad \forall x \in \mathbb{R}^d, \\
&\frac{\|\nabla^5 v(\hat{m}+x)\|}{\lambda_{\min}(H_v)^{5/2}} \leq c_5 \max\left(1, \left\|x/\sqrt{d/n}\right\|_{H_v}\right)^q, \quad \forall x \in \mathbb{R}^d.
\end{aligned} \tag{2.10}$$

In some cases, including the logistic regression posterior in Section 3, it is simpler to bound the lefthand quantities in (2.10) then it is to check the original assumptions.

Discussion. Let us comment on the dependence on d, n of the quantities $c_0, c_3, c_4, c_5, q, \alpha, r$, for a function v stemming from a posterior distribution, as in (1.3). Consider the asymptotic regime in which $n \to \infty$ and d grows with n. In this case, we have a sequence of priors $\nu = \nu_d$ on $\mathbb{R}^d$, and models $p(\cdot \mid x) = p_d(\cdot \mid x)$, $x \in \mathbb{R}^d$. Typically, the prior and model do not depend on n explicitly (only through d). Thus v depends on n only via the scaled-down prior $\frac{1}{n}\nu_d$, and via the average of the n functions $p(y_i \mid \cdot)$. If the data $\{y_i\}_{i=1}^n$ are distributed i.i.d. from some distribution Q (e.g. $Q = p(\cdot \mid x_0)$ in the well-specified case with ground truth x_0), then for large n we expect that $v(x) \approx \mathbb{E}_{Y \sim Q}[-\log p_d(Y \mid x)]$. This is only heuristic, since d grows with n, but it suggests that the above quantities depend only mildly on n itself.

On the other hand, these quantities may depend strongly on d. For example, recall from Remark 2.2 that c_3, c_4, c_5 can be obtained as upper bounds on $\|\nabla^k v\| \lambda_{\min}(H_v)^{-k/2}$ for $k = 3, 4, 5$. If $\lambda_{\min}(H_v)$ is going to zero with d, as in ill-posed inverse problems, then we can expect c_3, c_4, c_5 to grow with dimension. (Note that the function v in Bayesian inverse problems can be written in a similar form to (1.3); see [Lu, 2017].)

We will see in the theorem statements in the next section that our bounds on the Laplace accuracy are only small if $c_3 d/\sqrt{n}$ and $c_4 d^2/n$ are small. Therefore, if c_3, c_4 grow with dimension then the Laplace approximation is only valid if $n \gg d^{2+p}$ for some $p > 0$. This is natural; we cannot expect $n \gg d^2$ to be a universal condition that guarantees accuracy of the Laplace approximation in all cases.

Another implication of growing c_3, c_4 is that it forces us to take r small to satisfy (2.4), and hence c_0 will also be small. We will discuss this in the next section following our theorem statements. Finally, since q and α are powers of polynomial growth, it seems less common for these constants to also scale with dimension. For example, if v remains convex for all d, then we can always take $\alpha = 1$ (see Remark 2.1).

2.2 Main results on the Laplace approximation error

Recall that $\pi \propto e^{-V}$, where $V = nv$, and

$$\hat{m} = \arg\min_{x \in \mathbb{R}^d} V(x), \quad H_V = \nabla^2 V(\hat{m}).$$

The below theorems involve the following constants:

$$K_{p,\ell} = \exp\left(A(q)\bar{c}_3 d/\sqrt{n}\right)(1 + B_{p,\ell}), \tag{2.11}$$

where $A(q)$ is some constant depending only on q,

$$\begin{aligned} \bar{c}_3 &= c_3 + c_4 \frac{d}{\sqrt{n}}, \\ B_{p,\ell} &= \exp\left((p + \ell(4+q) + d) \log(r\sqrt{n} e^{1/\alpha}) - n c_0 \right). \end{aligned} \tag{2.12}$$

Here, p and ℓ are parameters arising in our general asymptotic expansion of $\int f d\pi - \int f d\hat{\gamma}$. Namely, ℓ is the order of the expansion and p is the power of polynomial growth assumed on f (see Proposition 4.1 for more details).

For simplicity, we assume in the below theorems that q and α, powers of polynomial growth of v, are absolute constants.

Theorem V1 (TV asymptotics and error bound)**.** *Let L be given by*

$$L = \frac{1}{12} \mathbb{E} \left| \langle \nabla^3 V(\hat{m}), (H_V^{-1/2} Z)^{\otimes 3} \rangle \right|, \tag{2.13}$$

where the expectation is with respect to $Z \sim \mathcal{N}(0, I_d)$. If v satisfies Assumptions A1-A4, then

$$\begin{aligned} &\mathrm{TV}\left(\pi,\, \mathcal{N}(\hat{m}, H_V^{-1})\right) = L + R, \\ &\qquad |L| \lesssim c_3 \frac{d}{\sqrt{n}}, \quad |R| \lesssim K_{0,2}(\bar{c}_3^2 + c_4) \left(\frac{d}{\sqrt{n}} \right)^2. \end{aligned} \tag{2.14}$$

Discussion: upper bound. Consider the leading order term L from (2.13). Let us explain why $L \lesssim c_3 d/\sqrt{n}$, since this bound is at the heart of our overall proof. Let $W(x) = V(\hat{m} + H_V^{-1/2} x)$. Then we can write L as $L = \mathbb{E}\,|\langle \nabla^3 W(0), Z^{\otimes 3} \rangle|$. Now, one can show that

$$\|\nabla^3 W(0)\| = \frac{1}{\sqrt{n}} \|\nabla^3 v(\hat{m})\|_{H_v} \leq \frac{c_3}{\sqrt{n}},$$

explaining why L scales with n as $1/\sqrt{n}$. This is also clear from the original definition (2.13) of L, since $H_V^{-1/2} \sim 1/\sqrt{n}$ and $\nabla^3 V \sim n$. However, it is not immediately clear why the upper bound on L scales with d as d^1. Indeed, the straightforward bound on L is the following:

$$L^2 \leq \mathbb{E}\left[\langle \nabla^3 W(0), Z^{\otimes 3} \rangle^2 \right] \leq \|\nabla^3 W(0)\|^2\, \mathbb{E}\left[\|Z\|^6 \right] \lesssim \left(c_3 \frac{d\sqrt{d}}{\sqrt{n}} \right)^2. \tag{2.15}$$

But in fact, we have

$$\begin{aligned} L^2 &\leq \mathbb{E}\left[\langle \nabla^3 W(0), Z^{\otimes 3} \rangle^2 \right] = 6 \|\nabla^3 W(0)\|_F^2 + 9 \|\langle \nabla^3 W(0), I_d \rangle\|^2 \\ &\leq 15 d^2 \|\nabla^3 W(0)\|^2 \leq 15 \left(c_3 \frac{d}{\sqrt{n}} \right)^2. \end{aligned} \tag{2.16}$$

The equality is proved in Lemma D.1. To prove that the remainder $\mathrm{TV}(\cdot) - L$ is bounded as $(d/\sqrt{n})^2$, we prove the more general fact that

$$\mathbb{E}\left[\langle T, Z^{\otimes 3}\rangle^{2k}\right] \lesssim (d\|T\|)^{2k} \tag{2.17}$$

for a symmetric order 3 tensor T. See Section 4.5 and Appendix D.

Lower bound. To show that $\mathrm{TV}\left(\pi,\, \mathcal{N}(\hat{m}, H_V^{-1})\right) \sim d/\sqrt{n}$, we need to show L is bounded both above and below by a multiple of $d/\sqrt{n}$. This is true if the inequalities in (2.16) are tight up to constants. It may be possible to show these inequalities are generically tight in some sense, but we do not investigate this further. Rather, we believe the value of Theorem V1 lies in its application to individual models, since one can exploit the problem-specific structure to derive tailored upper and lower bounds on L. For example, we will show in Section 3 that for a posterior stemming from logistic regression, L is indeed bounded *both* above and below by $d/\sqrt{n}$. To give an example of a non-generic case, suppose π is a product measure. Then one obtains that $\nabla^3 W(0)$ is a diagonal tensor, and it follows that the Frobenius norm is an order of magnitude smaller than d times the operator norm.

Theorem V2 (Mean asymptotics and error bound)**.** *Let L be given by*

$$L = -\frac{1}{2} H_V^{-1/2} \langle \nabla^3 V(\hat{m}), H_V^{-1} \rangle. \tag{2.18}$$

If v satisfies Assumptions A1-A4, then

$$\begin{aligned} H_V^{1/2}(m_\pi - \hat{m}) &= L + R, \\ \|L\| \lesssim c_3 \frac{d}{\sqrt{n}}, \quad \|R\| &\lesssim K_{1,2}(\bar{c}_3^2 + c_4)\left(\frac{d}{\sqrt{n}}\right)^2 \end{aligned} \tag{2.19}$$

If v also satisfies Assumption A5, then in fact

$$\|R\| \lesssim K_{1,3}(\bar{c}_3 c_4 + \bar{c}_3^3 + c_5 d^{-1/2})\left(\frac{d}{\sqrt{n}}\right)^3 + \left(\bar{c}_3 \frac{d}{\sqrt{n}}\right)^4 \tag{2.20}$$

Remark 2.3. *Note that the mode $\hat{m}$ is a poor estimate of the mean m_π if π is skewed, and we should think of $H_V^{-1/2} L$ as a correction to the mode which accounts for skew. Theorem V2 shows that if we can compute*

$$\hat{L} = H_V^{-1/2} L = -\frac{1}{2} H_V^{-1} \langle \nabla^3 V(\hat{m}), H_V^{-1} \rangle, \tag{2.21}$$

then the skew-corrected approximation $m_\pi \approx \hat{m} + \hat{L}$ will be much more accurate then the approximation $m_\pi \approx \hat{m}$ by the mode alone. This is evident in Figure 1, for a logistic regression example.

Example 2.1 (Skew correction for GLMs)**.** Consider a generalized linear model (GLM) with likelihood

$$p(\{y_i\}_{i=1}^n \mid \{x_i\}_{i=1}^n, \theta) \propto \exp\left(\sum_i y_i \theta^T x_i - \phi(\theta^T x_i)\right).$$

Here, the y_i are scalar, dependent variables, the $x_i \in \mathbb{R}^d$ are the independent variables, and θ is the parameter whose posterior distribution we are interested in. For example if $\phi(s) = \log(1 + e^s)$ then this is the likelihood for logistic regression. For a flat prior on $\theta \in \mathbb{R}^d$, the posterior of θ given the data takes the same form, i.e. $\pi(\theta) \propto e^{-V(\theta)}$ where

$$V(\theta) = \sum_i \left[-y_i \theta^T x_i + \phi(\theta^T x_i)\right].$$

It is then straightforward to compute that for $k \geq 2$, we have

$$\nabla^{(k)} V(\theta) = \sum_i \phi^{(k)}(\theta^T x_i) x_i^{\otimes k},$$

and hence the skew correction $\hat{L}$ defined in (2.21) is given by

$$\begin{aligned} \hat{L} &= -\frac{1}{2} \sum_{i=1}^n \phi'''(\hat{\theta}^T x_i)(x_i^T H_V^{-1} x_i) H_V^{-1} X_i, \\ \text{where} \quad H_V &= \sum_i \phi''(\hat{\theta}^T x_i) x_i x_i^T \end{aligned} \tag{2.22}$$

and $\hat{\theta}$ is the mode of π, in this case also the MLE. Note that computing $\hat{L}$ does not require inverting H_V, but rather only solving the linear systems $x_i = H_V a_i$ for a_i.

Theorem V3 (Covariance error bound)**.** *If v satisfies Assumptions A1-A4, then we have the upper bound*

$$\|H_V^{1/2}(\Sigma_\pi - H_V^{-1})H_V^{1/2}\| \lesssim K_{2,2}(c_4 + \bar{c}_3^2)\left(\frac{d}{\sqrt{n}}\right)^2 + \|H_V^{1/2}(m_\pi - \hat{m})\|^2.$$

The bound on $\|H_V^{1/2}(m_\pi - \hat{m})\|^2$ can be obtained by adding together the bounds on L and R from (2.19) in Theorem V2. Thus we se see that the total covariance error has order $(d/\sqrt{n})^2$ (holding c_3, c_4 fixed).

We note that [Kasprzak et al., 2022] also bounds the covariance, assuming $v \in C^3$. Our bound is tighter by a factor of $d/\sqrt{n}$ due to a symmetry-related cancellation that can be made when $v \in C^4$. [Spokoiny, 2022] also provides the ingredients to obtain a covariance bound but does not state a bound explicitly.

Remark 2.4. *Given a lower bound on $\lambda_{\min}(H_V)$, the bounds in Theorems V2 and V3 can be transformed into bounds on $\|m_\pi - \hat{m}\|$ and $\|\Sigma_\pi - H_V^{-1}\|$, respectively.*

Finally, we present our general result on the leading order correction to the Laplace approximation of an observable expectation.

Theorem V4. *Suppose* $|f(x)-f(\hat{m})| \le C_f \|x-\hat{m}\|_{H_V}^p$ *for some* $p, C_f \ge 0$, *and that* v *satisfies Assumptions A1-A4. Then*

$$\mathbb{E}_{X\sim\pi}[f(X)] = \mathbb{E}_{X\sim\hat{\gamma}}[f(X)] + L(f) + R(f),$$

where the leading term $L(f)$ *is given by*

$$L(f) = -\frac{1}{6}\mathbb{E}_{X\sim\hat{\gamma}}\left[f(X)\left\langle \nabla^3 V(\hat{m}),\, (X-\hat{m})^{\otimes 3}\right\rangle\right], \tag{2.23}$$

and $L(f), R(f)$ *are bounded as*

$$\begin{aligned} |L(f)| &\le \|f-\hat{\gamma}(f)\|_{L^2(\hat{\gamma})}\frac{c_3 d}{\sqrt{n}}, \\ |R(f)| &\lesssim K_{p,2}\left(C_f \vee \|f-\hat{\gamma}(f)\|_{L^4(\hat{\gamma})}\right)(\bar{c}_3^2 + c_4)\frac{d^2}{n}. \end{aligned} \tag{2.24}$$

Remark 2.5. *Note that* $\|x-\hat{m}\|_{H_V} = n^{p/2}\|x-\hat{m}\|_{H_v}$, *so if* $f \sim 1$ *grows polynomially then we expect* $C_f \sim n^{-p/2}$.

Remark 2.6. *Theorem V3 is proven by taking* $f(x) = (u^T x)^2$, $\|u\| = 1$ *in Theorem V4. Note that since* f *is even,* $L(f) = 0$, *which explains why the covariance error is of order* $(d/\sqrt{n})^2$ *rather than* $d/\sqrt{n}$. *In Theorem V2, the leading order term and the first of the two bounds on* R *also follow from Theorem V4, with* $f(x) = u^T x$. *To get the second bound on* R *we use the more general asymptotic expansion of* $\int f d\pi$ *shown in Section 4.3; see Proposition 4.1.*

Example 2.2 (Observable correction for GLMs)**.** Consider the same set-up as in Example 2.1, and suppose we want to compute $\mathbb{E}_{\theta\sim\pi}[f(\theta)]$, the posterior expectation of a polynomially bounded observable f. Then according to Theorem V4, we can improve the accuracy of the Laplace approximation $\mathbb{E}_{\theta\sim\pi}[f(\theta)] \approx \mathbb{E}_{\theta\sim\hat{\gamma}}[f(\theta)]$ by adding the correction term $L(f)$. For the GLM, this corrected approximation takes the form

$$\mathbb{E}_{\theta\sim\pi}[f(\theta)] \approx \mathbb{E}_{\theta\sim\hat{\gamma}}\left[f(\theta)\left(1 - \frac{1}{6}\sum_{i=1}^n \phi'''(x_i^T\hat{\theta})(x_i^T\theta - x_i^T\hat{\theta})^3\right)\right], \tag{2.25}$$

where $\hat{\theta}$ is the mode of π.

Consider the constants $K_{p,\ell}$ appearing in the above theorems. As we have noted in the discussion in Section 2.1, if c_3, c_4 are large then we are forced to take r small to satisfy (2.4), and hence c_0 will also be small. If as a result, the exponent of $B_{p,\ell}$ (defined in (2.12)) grows with d, then the prefactor $K_{p,\ell}$ in our bounds would blow up. However, the following lemma rules out this possibility under mild assumptions.

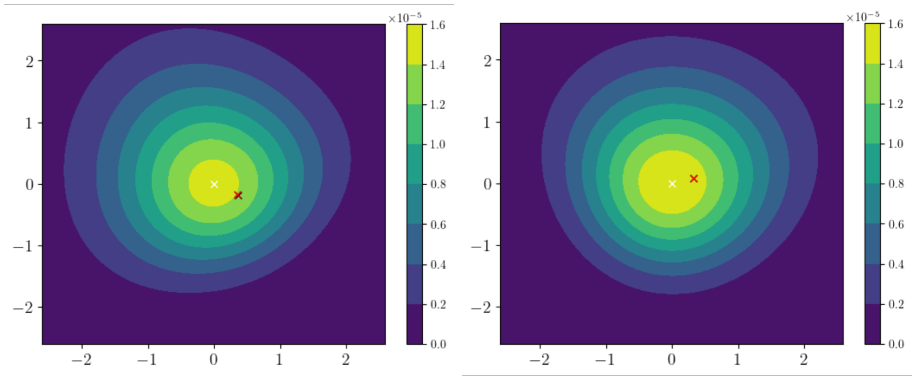

Figure 1: Contour plots of the densities $\rho_n := T_{\#}\pi$ with $n = 15$ (left), and $n = 50$ (right). Here, π is the logistic regression posterior from (3.1), and the rescaling map $T(x) = H_V^{1/2}(x - \hat{m})$ shifts the mode to zero, marked by a white X. The red X marks the location of the true mean (upon rescaling), and underneath the red X is a black X, marking the location of the skew adjustment L. On the scale depicted here, the red and black X's are indistinguishable.

Lemma 2.1. *Suppose* $c_3 d/\sqrt{n} \le 1$ *and* $c_4 d^2/n \le 1$, *and* $\log n \le \min(\sqrt{n/d}, \sqrt{d})$. *Suppose also that* (2.6) *is satisfied with* $r = \log n\sqrt{d/n}$ *and some* $\alpha > 0$. *Finally, suppose* $p + \ell(4+q) + d \le Cd$ *for some* $C > 1$, *and*

$$\max\left(C \log\log n, \frac{C}{2}\log d, C\alpha^{-1}\right) \le \frac{1}{12}\log^2 n. \tag{2.26}$$

Then Assumption A4 is satisfied with this choice of r *(including* (2.4)*), and* $B_{p,\ell} \le 1$.

Assuming $c_3 d/\sqrt{n} \le 1$ and $c_4 d^2/n \le 1$ is reasonable since our bounds in the above theorems are only small if $c_3 d/\sqrt{n}$ and $c_4 d^2/n$ are small. That n is bounded by $e^{\sqrt{d}}$ is also a very weak requirement for large d. Also, recall from Remark 2.1 that (2.6) is only a slight strengthening of (2.5) from Assumption A4. In particular, if v is convex, then (2.6) is automatically satisfied for any $r > 0$ and $\alpha = 1$. See Appendix A for the proof of this lemma.

3 Example: Logistic Regression

In this section, we consider a posterior arising from logistic regression with Gaussian design. We describe the setting in Section 3.1. In Section 3.2, we verify Assumptions A1-A5, deriving high probability upper bounds on the constants c_3, c_4, c_5. This leads to overall error bounds on the mean error, covariance error, and TV distance of the Laplace approximation. In Section 3.3 we show these bounds are tight in their dimension dependence: namely, we show the leading

order terms of the TV and mean asymptotics are bounded above and below by $d/\sqrt{n}$.

3.1 Setting

We observe n pairs (X_i, Y_i), $i = 1, \dots, n$, where $X_i \in \mathbb{R}^d$ are features and $Y_i \in \{0, 1\}$ are corresponding labels, distributed according to

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \quad Y_i \mid X_i \sim \text{Bernoulli}(s(\beta^T X_i)).$$

Here, s is the sigmoid $s(t) = (1 + e^{-t})^{-1}$ and β is the ground truth parameter. For simplicity we take $\beta = (1, 0, \dots, 0)$.

We let $b \in \mathbb{R}^d$ denote a generic parameter indexing the distributions $Y \mid x \sim \text{Bernoulli}(s(b^T x))$. Assuming a flat prior, the posterior distribution of b given the data is

$$\pi(b \mid (X_i, Y_i)_{i=1}^n) \propto \prod_{i=1}^n p(Y_i \mid X_i, b) = \prod_{i=1}^n s(b^T X_i)^{Y_i} (1 - s(b^T X_i))^{1-Y_i}. \quad (3.1)$$

Note that $\pi \propto e^{-V}$, where

$$V(b) = -\sum_{i=1}^n \left[Y_i \log s(b^T X_i) + (1 - Y_i) \log(1 - s(b^T X_i)) \right]. \quad (3.2)$$

The function V is convex, and it is well-known that if the data is not linearly separable then there exists a strict global minimum

$$\hat{\beta} = \arg\min_{b \in \mathbb{R}^d} V(b).$$

This point $\hat{\beta}$ is both the MAP and the MLE, since the posterior coincides with the likelihood. We record for future reference the second and third derivative of V:

$$\begin{aligned} \nabla^2 V(b) &= \sum_{i=1}^n s'(b^T X_i) X_i X_i^T, \\ \nabla^3 V(b) &= \sum_{i=1}^n s''(b^T X_i) X_i^{\otimes 3}. \end{aligned} \quad (3.3)$$

3.2 Upper bounds on the approximation error

In this section, we verify Assumptions A1-A5 for the logistic regression setting described above, and bound the constants $c_0, c_3, c_4, c_5, r, q, \alpha$ with high probability with respect to the randomness in the data. We then apply Theorems V1, V2, V3 to obtain state of the art bounds on the Laplace approximation error for logistic regression.

To check Assumption A1, we need to show the MLE $\hat{\beta}$ is finite, and bounded, with high probability. The works [Sur and Candès, 2019, Candès and Sur, 2020, Sur, 2019] study existence of the MLE for logistic regression in the high-dimensional regime $d/n \to \kappa \in (0,1)$. Convergence of the MLE to the ground truth (and therefore boundedness of the MLE) has also been studied for generalized linear models [Spokoiny, 2017, Spokoiny, 2013, Spokoiny, 2012, He and Shao, 2000] and exponential families [Portnoy, 1988] in limiting regimes in which $d/n \to 0$.

For convenience to the reader, we provide our own proof that the MLE is bounded with high probability, which is tailored to our specific setting. However, the proof method is inspired by Theorem 3.4 of [Spokoiny, 2017], and uses in a crucial way Lemma 7 of Chapter 3 of [Sur, 2019]. The following lemma proves boundedness of the MLE and at the same time, provides a lower bound on the eigenvalues of $H_v = \nabla^2 v(\hat{\beta})$.

Lemma 3.1. *There exist absolute constants $0 < A_0 < 1$ and $0 < A_1, A_2$ such that if $d/n < A_0$ then*

$$\mathbb{P}\left(\|\hat{\beta} - \beta\| \leq 1, \quad \lambda_{\min}(H_v) \geq A_2\right) \geq 1 - e^{-d/2} - 5e^{-A_1 n}.$$

See Appendix F for the proof. This verifies Assumption A1 (with high probability). Now, as mentioned in Remark 2.2, we can find the required constants c_3, c_4, c_5 by bounding ratios of the form $\|\nabla^k v(\hat{\beta} + tu)\|/\lambda_{\min}(H_v)^{k/2}$. We have already lower bounded $\lambda_{\min}(H_v)$, so it remains to upper bound the operator norms of $\nabla^k v$ for $k = 3, 4, 5$. We also bound the matrix norm of $\nabla^2 v$ for use in a later lemma.

Lemma 3.2. *Suppose $d \leq n \leq e^{\sqrt{d}}$. Then there exist absolute constants $B_0, B_1 > 0$ such that for $k = 2, 3, 4, 5$, we have*

$$\sup_{b \in \mathbb{R}^d} \|\nabla^k v(b)\| \leq B_0 \left(1 + \frac{d^{k/2}}{n}\right) \tag{3.4}$$

with probability at least $1 - 2\exp(-B_1\sqrt{nd}/\log(2n/d))$.

See Appendix F for the proof, which relies on a result of [Adamczak et al., 2010] bounding quantities of the form $\sup_{\|u\|=1} \frac{1}{n}\sum_{i=1}^n |X_i^T u|^k$ with high probability. Combining the last two lemmas with the inequality (2.9), we have the following corollary.

Corollary 3.1. *Assume $d/n < A_0$ from Lemma 3.1, and let $q = 0$. Then there exist absolute constants $C_0, C_1 > 0$ such that with probability at least $1 - 7\exp(-C_1\sqrt{nd}/\log(2n/d)) - e^{-d/2}$, Assumptions A2, A3, and A5 hold with constants $c_k \leq C_0(1 + d^{k/2}/n)$, $k = 3, 4, 5$.*

This corollary shows that if d^2/n is bounded by an absolute constant, then so are c_3 and c_4, but that c_5 scales as $\sqrt{d}$. Fortuitously, however, c_5 appears in our main theorems only through $c_5 d^{-1/2}$ (recall (2.20)).

It remains to find r, c_0, α such that Assumption A4 is satisfied. Since c_3 and c_4 are bounded, we can take $r > 0$ small enough but bounded from below by an absolute constant, to satisfy (2.4). Next, we find c_0 and α.

Lemma 3.3. *Assume $d/n < A_0$ from Lemma 3.1. Let c_3, c_4 be from the above corollary, and find $r \geq C$ such that (2.4) is satisfied. Then there exist absolute constants $C_0, C_1 > 0$ such that*

$$\begin{aligned}\mathbb{P}\bigg(v(\hat{\beta}+x) - v(\hat{\beta}) \geq C_0\|x/r\|_{H_v} \quad \forall \|x\|_{H_v} \geq r\bigg) \\ \geq 1 - 7\exp(-C_1\sqrt{nd}/\log(2n/d)) - e^{-d/2}.\end{aligned} \tag{3.5}$$

In other words, with high probability, Assumption A4 is satisfied with $\alpha = 1$ and $c_0 = C_0$.

See Appendix F for the proof. Having checked all the assumptions and bounded the constants appearing in them, we are nearly ready to apply Theorems V1-V3 to logistic regression. But first we bound the constant K appearing in these theorems, defined in (2.11). If d^2/n is bounded above by an absolute constant, then c_3, c_4 are bounded from above, and c_0 is bounded from below, by absolute constants. It follows that B is exponentially small in n, and hence K is bounded above by an absolute constant.

Corollary 3.2. *Consider the logistic regression setting described in Section 3.1. If d^2/n is small enough, then there exist absolute constants $C, C_1, C_2, C_3 > 0$ such that the following bounds on the TV error, mean error, and covariance error hold with probability at least $1 - C_1\exp(-C_2\sqrt{nd}/\log(2n/d)) - C_3e^{-d/2}$:*

$$\begin{aligned}\mathrm{TV}\left(\pi,\, \mathcal{N}\left(\hat{\beta}, \nabla^2 V(\hat{\beta})^{-1}\right)\right) &\leq C\frac{d}{\sqrt{n}}, \\ \sqrt{n}\left\|\mathbb{E}\left[b \mid \{X_i, Y_i\}_{i=1}^n\right] - \hat{\beta}\right\| &\leq C\frac{d}{\sqrt{n}}, \\ n\left\|\mathrm{Cov}\left(b \mid \{X_i, Y_i\}_{i=1}^n\right) - \nabla^2 V(\hat{\beta})^{-1}\right\| &\leq C\left(\frac{d}{\sqrt{n}}\right)^2.\end{aligned} \tag{3.6}$$

Moreover, let

$$\hat{L} = -\frac{1}{2}\sum_{i=1}^n s''(\hat{\beta}^T X_i)(X_i^T H_V^{-1} X_i) H_V^{-1} X_i.$$

When the MAP (mode) $\hat{\beta}$ is adjusted by this skew correction term, the mean error becomes

$$\sqrt{n}\left\|\mathbb{E}\left[b \mid \{X_i, Y_i\}_{i=1}^n\right] - (\hat{\beta} + \hat{L})\right\| \leq C\left(\frac{d}{\sqrt{n}}\right)^3.$$

Here, $\hat{L} = H_V^{-1/2} L$, where L is defined in (2.18). We have used formula (3.3) for $\nabla^3 V$ to derive the formula for $\hat{L}$. This corollary follows immediately from Theorems V1-V3, and the fact that $c_3, c_4, c_5 d^{-1/2} \lesssim C$ and $H_v = \nabla^2 v(\hat{\beta}) \succeq C I_d$ with high probability.

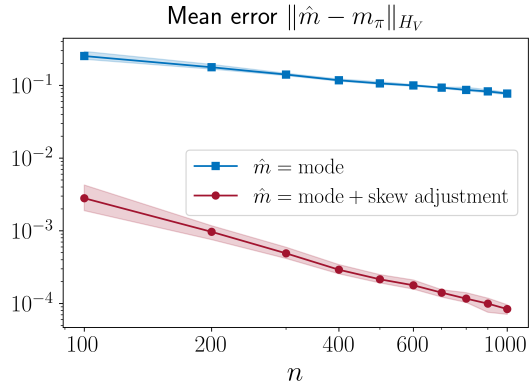

Figure 2: $\pi \propto e^{-V}$ is the likelihood of logistic regression given n observations in dimension $d = 2$. The slopes of the best-fit lines are -0.51 for the Laplace approximation and -1.52 for the adjusted approximation.

Remark 3.1. *[Kasprzak et al., 2022] also considers the Laplace approximation error for logistic regression. Although the "universal" contribution to their error bounds is $d/\sqrt{n}$, which is the right scaling with d, the authors obtain dimension dependent upper bounds on the v-dependent component of the error, e.g. on their analogue of c_3. This is because the authors bound the operator norm of $\nabla^3 v$ as $\|\nabla^3 v\| \lesssim \frac{1}{n}\sum_{i=1}^n \|X_i\|^3$, which for Gaussian design will scale with d as $d\sqrt{d}$. Thus Corollary 3.2 yields the tightest known bounds, since we show the error scales as $d/\sqrt{n}$ even when taking into account the v-dependent constants.*

Figure 2 displays the n scaling of the mean error on a log-log plot in dimension $d = 2$. The blue curve shows the error $\|H_V^{1/2}(m_\pi - \hat{m})\|$ of the standard Laplace mean approximation as a function of n, while the red curve shows the error $\|H_V^{1/2}(m_\pi - \hat{m} - \hat{L})\|$ of the skew-adjusted Laplace mean approximation. The slope of the best-fit lines to these two curves are -0.51 and -1.52, respectively, as predicted by Corollary 3.2.

See also Figure 1, which displays the contour plots of the densities $\rho_n := T_\# \pi$ for $n = 15$ and $n = 50$, in $d = 2$. Here, π is the logistic regression posterior from (3.1), and the rescaling map $T(x) = H_V^{1/2}(x - \hat{m})$ shifts the mode to zero, marked by a white X. The red X is the (rescaled) mean, and underneath the red X is a black X marking the location of the skew correction term L.

3.3 Dimension scaling of leading order terms L

We now show that the leading order terms L in the TV and mean error expansions (see (2.13) and (2.18)) are bounded from *below* by $d/\sqrt{n}$. This shows that $d^2 \ll n$ is *necessary* for the Laplace approximation to be accurate. We will show this numerically for the random, n-sample posterior $\pi \propto e^{-V}$, and rigorously for a measure $\pi \propto e^{-\bar{V}_\infty}$ inspired by the population log likelihood.

We start with this second example: let

$$\bar{V}_\infty(b) = n\mathbb{E}_{X,Y}[Y \log s(X^T b) + (1-Y)\log(1-s(X^T b))].$$

Then $\bar{V}_\infty = \mathbb{E}_X[V_\infty(b)]$, where

$$\begin{aligned} V_\infty(b) = \mathbb{E}_Y[V(b) \mid \{X_i\}_{i=1}^n] = \sum_{i=1}^n \mathbb{E}[Y_i \mid X_i] \log s(b^T X_i) \\ + (1 - \mathbb{E}[Y_i \mid X_i]) \log(1 - s(b^T X_i)) \end{aligned} \tag{3.7}$$

is the population log likelihood. Note that the logistic regression model treats the X_i as fixed, so V_∞ is the population log likelihood for each set of $\{X_i\}_{i=1}^n$, while $\bar{V}_\infty$ is the expectation of V_∞ over the X_i.

Since the ground truth $b = \beta$ minimizes the population log likelihood V_∞ for each fixed $\{X_i\}_{i=1}^n$, it follows that β is also the minimizer of $\bar{V}_\infty$. Furthermore, we compute

$$\nabla^k \bar{V}_\infty(\beta) = n\mathbb{E}[s^{(k-1)}(X_1) X^{\otimes k}], \quad k = 2, 3. \tag{3.8}$$

(Recall that $\beta = (1, 0, \ldots, 0)$.) In Appendix F, we prove the following lemmas.

Lemma 3.4 (Leading order TV)**.** *Let L be as in (2.13) for $V = \bar{V}_\infty$, and define $a_{k,p} = \mathbb{E}[s^{(k)}(X_1) X_1^p]$, where $X_1 \sim \mathcal{N}(0,1)$. Then*

$$L \geq \frac{2}{a_{1,2}^{1/2}\sqrt{n}} \left((d-1)\frac{|a_{2,1}|}{a_{1,0}} - \frac{2|a_{2,3}|}{a_{1,2}} \right).$$

Lemma 3.5 (Leading order mean error)**.** *Let L be as in (2.18) for $V = \bar{V}_\infty$, and let $a_{k,p}$ be as in Lemma 3.4. Then*

$$\|L\| = \frac{1}{2a_{1,2}^{1/2}\sqrt{n}} \left| (d-1)\frac{a_{2,1}}{a_{1,0}} + \frac{a_{2,3}}{a_{1,2}} \right|.$$

We conclude that the leading order terms L are bounded from below by $d/\sqrt{n}$ for the distribution $\pi \propto e^{-\bar{V}_\infty}$. We now show numerically the same property of the n-sample random posterior for logistic regression: namely, that the leading order terms L do not go to zero if $d/\sqrt{n}$ remains bounded from below. We take an increasing sequence of dimensions d, and let n be either $n = 2d^2$ (lefthand plot of Figure 3) or $n = d^{2.5}$ (righthand plot). In the former regime, we see that L remains bounded below as d increases, while in the latter regime, L goes to zero. In both plots, the solid lines represent the average over 20 n-sample posteriors, and the shaded region depicts the 25%-75% quantile.

4 Proof overview

In this section, we first apply a scale-removing change of variables to V, and restate Theorems V1-V4 in terms of the new function W. In Section 4.3, we

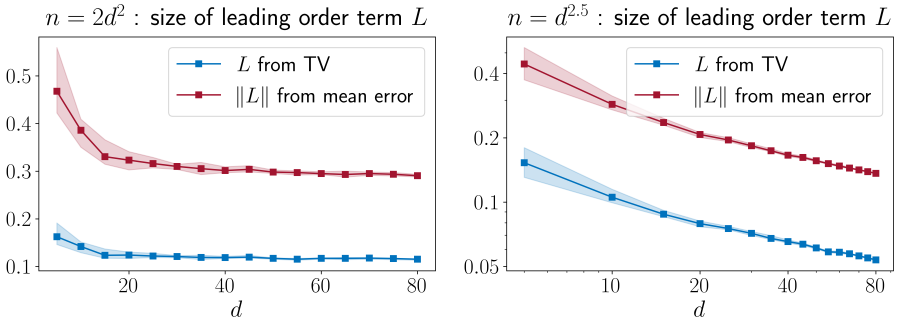

Figure 3: The size of the leading order term L from the TV asymptotics (2.13)-(2.14) and the mean error asymptotics (2.18)-(2.19) for the logistic regression posterior $\pi \propto e^{-V}$, with V given by (3.2). In the lefthand plot, we take $n = 2d^2$ for an increasing sequence of dimensions d. We see that L does not go to zero with d, which confirms that $d^2/n \to 0$ is *necessary* for an accurate Laplace approximation. In the righthand plot we take $n = d^{2.5}$. The plot (on a log-log scale) shows that the leading order term goes to zero in this regime, as predicted by our upper bounds in Corollary 3.2.

state our general asymptotic expansion and go over the main features of the proof. Section 4.4 then finishes the proof of Theorems V1-V4 using the general result from Section 4.3. Finally, Section 4.5 goes over the proof of the main lemma which shows why $d/\sqrt{n}$, rather than $d\sqrt{d}/\sqrt{n}$, is the fundamental unit of error. We also review the key ingredient in the proof of [Kasprzak et al., 2022] which allows the authors to obtain the rate $d/\sqrt{n}$.

4.1 Simplifying Coordinate Transformation

We begin by rewriting the quantities from Theorems V1-V4 in terms of a function W obtained from V by changing variables. Namely, let $\hat{\gamma} = \mathcal{N}(\hat{m}, H_V^{-1})$ be the Laplace approximation to $\pi \propto e^{-V}$, and define the linear map $T : \mathbb{R}^d \to \mathbb{R}^d$ by $T(x) = H_V^{1/2}(x - \hat{m})$. Then we let

$$\rho = T_{\#}\pi \propto e^{-W}, \quad \gamma = T_{\#}\hat{\gamma} = \mathcal{N}(0, I_d),$$

where

$$W(x) = V(\hat{m} + H_V^{-1/2}x) = nv(\hat{m} + H_v^{-1/2}x/\sqrt{n}). \tag{4.1}$$

Note that W is minimized at zero, with Hessian $\nabla^2 W(0) = I_d$, so that $\gamma = \mathcal{N}(0, I_d)$ is the Laplace approximation to ρ. In other words, the Laplace approximation is affine invariant in the following sense: if $\hat{\gamma}_\mu$ is defined to be the Laplace approximation to μ, then we have seen that $\hat{\gamma}_{T_{\#}\pi} = T_{\#}\hat{\gamma}_\pi$ for all affine maps T.

Lemma 4.1. *Let W be given by (4.1) and $\rho \propto e^{-W}$. Then*

1. $\mathrm{TV}\left(\pi,\, \mathcal{N}\left(\hat{m}, H_V^{-1}\right)\right) = \mathrm{TV}\left(\rho,\, \mathcal{N}(0, I_d)\right)$, *and the leading order term* L *from Theorem V1 can be expressed in terms of* W *as follows:*

$$L = \frac{1}{12}\mathbb{E}\left|\langle \nabla^3 W(0), Z^{\otimes 3}\rangle\right| \tag{4.2}$$

Therefore, Theorem V1 is equivalent to

$$\begin{aligned} &\mathrm{TV}\left(\rho,\, \mathcal{N}(0, I_d)\right) = L + R, \\ &L \le c_3 \frac{d}{\sqrt{n}}, \quad |R| \le K(\bar{c}_3^2 + c_4)\left(\frac{d}{\sqrt{n}}\right)^2. \end{aligned} \tag{W1}$$

2. *We have* $H_V^{1/2}(m_\pi - \hat{m}) = m_\rho$, *and the leading order term* L *from Theorem V2 can be expressed in terms of* W *as follows:*

$$L = -\frac{1}{2}\langle \nabla^3 W(0), I_d\rangle. \tag{4.3}$$

Therefore, Theorem V2 is equivalent to showing

$$\begin{aligned} &m_\rho = L + R, \\ &\|L\| \lesssim c_3 \frac{d}{\sqrt{n}}, \quad \|R\| \lesssim K(\bar{c}_3^2 + c_4)\left(\frac{d}{\sqrt{n}}\right)^2, \\ &\|R\| \overset{A5}{\lesssim} K(\bar{c}_3 c_4 + \bar{c}_3^3 + c_5 d^{-1/2})\left(\frac{d}{\sqrt{n}}\right)^3 + \left(\bar{c}_3 \frac{d}{\sqrt{n}}\right)^4 \end{aligned} \tag{W2}$$

The last line is the bound on $\|R\|$ *that holds under Assumption A5 in addition to Assumptions A1-A4.*

3. *We have* $\|H_V^{1/2}(\Sigma_\pi - H_V^{-1})H_V^{1/2}\| = \|\Sigma_\rho - I_d\|$. *Therefore, Theorem V3 is equivalent to showing*

$$\|\Sigma_\rho - I_d\| \lesssim K(c_4 + \bar{c}_3^2)\left(\frac{d}{\sqrt{n}}\right)^2 + \|m_\rho\|^2. \tag{W3}$$

4. *Let* $h = f \circ T^{-1}$. *Then* $|f(x) - f(\hat{m})| \le C_f \|x - \hat{m}\|_{H_V}^p$ *if and only if* $|h(x) - h(0)| \le C_f \|x\|^p$. *Also,* $\pi(f) = \rho(h)$, $\hat{\gamma}(f) = \gamma(h)$, *and the leading order term* $L(f)$ *from (2.23) can be expressed in terms of* W *and* h *as follows:*

$$L(f) = \tilde{L}(h) = -\frac{1}{6}\mathbb{E}\left[h(Z)\langle \nabla^3 W(0), Z^{\otimes 3}\rangle\right]. \tag{4.4}$$

Furthermore, $\|f\|_{L^p(\hat{\gamma})} = \|h\|_p$. *Therefore, Theorem V4 is equivalent to showing that if* $|h(x) - h(0)| \le C_h \|x\|^p$ *then*

$$\begin{aligned} &\int h d\rho = \int h d\gamma + \tilde{L}(h) + R(h), \\ &|\tilde{L}(h)| \le c_3 \|h - \gamma(h)\|_2 \frac{d}{\sqrt{n}}, \\ &|R(h)| \lesssim K_{p,2}\left(C_h \vee \|h - \gamma(h)\|_4\right)(\bar{c}_3^2 + c_4)\left(\frac{d}{\sqrt{n}}\right)^2. \end{aligned} \tag{W4}$$

See Appendix B.1 for the proof. This lemma shows that comparing π to $\hat{\gamma}$ in the sense of Theorems V1-V4 is equivalent to comparing $\rho \propto e^{-W}$ to $\gamma \propto e^{-\|x\|^2/2}$, the standard normal distribution. The next lemma shows that W's derivatives of order higher than two are small, implying that $W(x) \approx \|x\|^2/2 + \text{const.}$ and hence $\rho \approx \gamma$.

Lemma 4.2. *Let W be given by (4.1), and suppose v satisfies Assumptions A1-A4 with constants $c_0, c_3, c_4, \alpha, q, r$. Then $W \in C^4$ and has unique global minimizer $x = 0$, with $\nabla^2 W(0) = I_d$. Furthermore, we have*

$$\|\nabla^3 W(0)\| \le \frac{c_3}{\sqrt{n}}, \tag{4.5}$$

$$\|\nabla^4 W(x)\| \le \frac{c_4}{n}, \quad \forall \|x\| \le 4\sqrt{d}, \tag{4.6}$$

$$|\langle \nabla^4 W(tu), u^{\otimes 4}\rangle| \le \frac{c_4}{n} \max\left(1, \frac{t}{\sqrt{n}}\right)^q, \quad \forall \|u\| = 1, t \ge 0, \tag{4.7}$$

$$W(x) - W(0) \ge n c_0 \left\| \frac{x}{r\sqrt{n}} \right\|^\alpha, \quad \forall \|x\| \ge r\sqrt{n}. \tag{4.8}$$

If v satisfies Assumption A5, then we also have

$$|\langle \nabla^5 W(tu), u^{\otimes 5}\rangle| \le \frac{c_5}{n\sqrt{n}} \max\left(1, \frac{t}{\sqrt{d}}\right)^q, \quad \forall \|u\| = 1, t \ge 0. \tag{4.9}$$

See Appendix B.1 for the proof. We record one more property: namely, the condition (2.4) relating c_3, c_4, r ensures that W is lower bounded by a quadratic in an $O(\sqrt{n})$ ball around the origin.

Lemma 4.3. *If v satisfies Assumptions A1-A4, then W satisfies*

$$W(x) - W(0) \ge \frac{\|x\|^2}{4}, \quad \forall \|x\| \le r\sqrt{n}.$$

See Appendix B.1 for the proof. In preparation to bound the TV distance and observable expectation errors between ρ and γ, we define a function r such that $d\rho \propto e^{-r} d\gamma$. Namely, we let $r(x) = W(x) - \|x\|^2/2 + \text{const.}$, where the constant is chosen so that $\gamma(r) = 0$. We will write r in a more enlightening form by considering the Taylor expansion of W:

$$W(x) = W(0) + \frac{1}{2}\|x\|^2 + \frac{1}{3!}\langle \nabla^3 W(0), x^{\otimes 3}\rangle + r_4(x), \tag{4.10}$$

for a remainder r_4.

Definition 4.1. Let

$$r(x) = \frac{1}{3!}\langle \nabla^3 W(0), x^{\otimes 3}\rangle + r_4(x) - \gamma(r_4), \tag{4.11}$$

where r_4 is the remainder in the Taylor expansion (4.10) of W. Then r satisfies $\gamma(r) = 0$ and $d\rho \propto e^{-r} d\gamma$.

4.2 TV upper bound: [Kasprzak et al., 2022] proof outline

The main insight behind our bounds on the leading order TV error and remainder is that the straightforward operator norm inequality $|\langle \nabla^3 W(0), x^{\otimes 3}\rangle| \leq \|\nabla^3 W(0)\|\|x\|^3$ leads to the bound

$$\mathbb{E}\left[\langle \nabla^3 W(0), Z^{\otimes 3}\rangle^{2k}\right] \leq \|\nabla^3 W(0)\|^{2k}\mathbb{E}\left[\|Z\|^{6k}\right] \lesssim_k (c_3 d\sqrt{d}/\sqrt{n})^{2k},$$

but a more careful argument shows that

$$\mathbb{E}\left[\langle \nabla^3 W(0), Z^{\otimes 3}\rangle^{2k}\right] \lesssim_k (c_3 d/\sqrt{n})^{2k}.$$

(See more on this below, as well as the discussion following Theorem V1). However, if one is only interested in an upper bound on the TV error, then the work [Kasprzak et al., 2022] shows we can avoid bounding $\mathbb{E}\left[\langle \nabla^3 W(0), Z^{\otimes 3}\rangle^{2k}\right]$ in the first place. Since this work does not focus on dimension dependence, it is enlightening to review the authors' proof method from the perspective of dimension.

We wish to bound $\mathrm{TV}(\pi, \hat{\gamma}) = \mathrm{TV}(\rho, \gamma)$, where $\rho = T_{\#}\pi \propto e^{-W}$ is the distribution obtained by removing the scale from π, and γ is the standard normal distribution (the Laplace approximation to ρ). First, suppose ρ is strongly log-concave. By Pinsker's inequality and the log Sobolev inequality (LSI), we then have

$$\mathrm{TV}(\rho, \gamma) \leq \sqrt{\mathrm{KL}(\,\gamma \,||\, \rho)} \lesssim \mathbb{E}_\gamma[\|\nabla \log(\gamma/\rho)\|^2] = \mathbb{E}\left[\|\nabla r(Z)\|^2\right]. \tag{4.12}$$

Here, $\lesssim$ hides the constant of strong log-concavity, and r is the function from Definition 4.1 satisfying $d\rho \propto e^{-r} d\gamma$. But now note that the leading order term of ∇r is

$$\nabla r(x) \approx \frac{1}{2}\langle \nabla^3 W(0), x^{\otimes 2}\rangle,$$

and therefore in a neighborhood of the origin we have

$$\|\nabla r(x)\| \lesssim \|\nabla^3 W(0)\|\|x\|^2. \tag{4.13}$$

Therefore, $\mathbb{E}\left[\|\nabla r(Z)\|^2\right] \lesssim \|\nabla^3 W(0)\|^2 \mathbb{E}\left[\|Z\|^4\right] \lesssim c_3^2 d^2/n$. The key point is that applying an LSI allows us to bound the TV error in terms of the *gradient* of r, whose leading order term is now a *second* order polynomial. In other words, we brought the order of the polynomial in x down from 3 to 2. To get the right dependence on d, one can now bound the Gaussian expectation of $\|\nabla r\|$ using the straightforward operator norm inequality (4.13).

This is the essence of the proof. Note that [Kasprzak et al., 2022] does not actually assume $\rho \propto e^{-W}$ is strongly log concave but rather applies an LSI locally, in a neighborhood of zero. In this neighborhood, we have $W(x) \approx W(0) + \|x\|^2/2$, so there is some absolute constant of strong log-concavity. One must then also deal with tail integrals, since this whole argument holds only locally.

4.3 Asymptotics for an observable expectation

We have shown that comparing π to $\hat{\gamma}$ is equivalent to comparing $\rho \propto e^{-W}$ to $\gamma \propto e^{-\|x\|^2/2}$, and Lemma 4.2 already suggests that $W(x) \approx W(0) + \|x\|^2/2$ and hence $\rho \approx \gamma$. In this section, we make this precise. Note that $\mathrm{TV}(\rho, \gamma)$, $\|m_\rho\|$, and $\|\Sigma_\rho - I_d\|$ are all given by $\sup_{f \in \mathcal{F}} \int f d\rho - \int f d\gamma$ for different classes $\mathcal{F}$. Therefore, we start by decomposing $\int f d\rho - \int f d\gamma$ into leading and remainder terms, for a polynomially bounded function f.

To state our preliminary decomposition of $\int f d\rho - \int f d\gamma$, we introduce some notation.

Definition 4.2. Let

$$\begin{aligned} L_k(f) &= \int g \sum_{j=0}^{k} \frac{(-1)^j}{j!} r^j d\gamma, \\ E_{k+1}(f) &= \int g e^{-r} d\gamma - L_k(f) \end{aligned} \tag{4.14}$$

("L" for leading, "E" for error), where $g = f$ if f is a constant function, and $g = f - \gamma(f)$ otherwise. For example,

$$\begin{aligned} L_2(f) &= \int (f - \gamma(f)) \left(1 - r + \frac{r^2}{2}\right) d\gamma = \int (f - \gamma(f)) \left(-r + \frac{r^2}{2}\right) d\gamma, \\ E_3(f) &= \int (f - \gamma(f)) \left(e^{-r} - 1 + r - \frac{r^2}{2}\right) d\gamma = \int (f - \gamma(f)) \left(-\frac{r^3}{3!} + -\frac{r^4}{4!} - ...\right) d\gamma, \\ L_1(1) &= \int (1 - r) d\gamma = 1, \quad E_2(1) = \int (e^{-r} - 1 + r) d\gamma = \end{aligned} \tag{4.15}$$

With this notation, and using that $d\rho \propto e^{-r} d\gamma$, we have for any $m \geq 1$ that

$$\int f d\rho - \int f d\gamma = \frac{\int (f - \gamma(f)) e^{-r} d\gamma}{\int e^{-r} d\gamma} = \frac{L_m(f) + E_{m+1}(f)}{L_{m-1}(1) + E_m(1)}. \tag{4.16}$$

As this formula suggests, we can write $\int f d\rho - \int f d\gamma$ as $L_m(f)/L_{m-1}(1)$ plus a remainder. Indeed, a straightforward computation in Appendix B.2 shows that

$$\left| \int f d\rho - \int f d\gamma - \frac{L_m(f)}{L_{m-1}(1)} \right| \leq |E_{m+1}(f)| + |E_m(1)| \left| \frac{L_m(f)}{L_{m-1}(1)} \right| \tag{4.17}$$

Now, from the formula (4.11) for r we see that $r = o(1)$, since derivatives of W of order higher than 2 are $o(1)$. Therefore, the L_m's and E_m's are dominated by the contribution from the lowest power of r. The lowest power of r contributing to $L_k(f)$ is r^1 for all $k \geq 1$ (note that the r^0 term always drops out, as seen in the first line of (4.15)), while the lowest power of r contributing to $E_k(f), E_k(1)$ is r^k (meaning the first power of r appearing in the tail of the Taylor expansion of the exponential; see e.g. the second expression for $E_3(f)$ in (4.15)). Therefore

the second error term in (4.17) involving $E_m(1)L_m(f)$ is on the same order as the first error term E_{m+1}, in terms of total powers of r. This explains why we split the denominator of (4.16) as $L_{m-1} + E_m$ rather than $L_m + E_{m+1}$.

To be a bit more precise, we will essentially show that the righthand side of (4.17) is dominated by $\|r\|_{2m+2}^{m+1}$. The following key lemma shows that $\|r\|_{2m+2} \sim d/\sqrt{n}$. This is the reason why $d/\sqrt{n}$, rather than $d\sqrt{d}/\sqrt{n}$, is the fundamental unit of error.

Lemma 4.4. *We have*

$$\mathbb{E}\left[\left\langle \nabla^3 W(0), Z^{\otimes 3}\right\rangle^{2m}\right] \lesssim_m \left(d\|\nabla^3 W(0)\|\right)^{2m} \leq \left(c_3 \frac{d}{\sqrt{n}}\right)^{2m}, \tag{4.18}$$

$$\mathbb{E}\left[\left(r_4(Z) - \gamma(r_4)\right)^{2m}\right] \lesssim_m \left(c_4 \frac{d^2}{n}\right)^{2m}, \tag{4.19}$$

and therefore

$$\|r\|_{2m} \lesssim_m \bar{c}_3 d/\sqrt{n}.$$

Since this lemma is central, we outline its proof in Section 4.5 below. Using the key lemma 4.4 as well as a few additional technical results, we obtain the following bound on the remainder in (4.17), which holds for any polynomially bounded f.

Proposition 4.1. *Let* $|f(x) - f(0)| \leq C_f \|x\|^p$ *and* $\epsilon = d/\sqrt{n}$, *and define* $K_{p,m+1} = e^{A(q)\bar{c}_3\epsilon}(1 + B_{p,m+1})$, *where*

$$B_{p,m+1} = \exp\left((p + (m+1)(q+4) + d)\log(r\sqrt{n}e^{1/\alpha}) - nc_0\right).$$

Then

$$\begin{aligned}
&\left|\int f d\rho - \int f d\gamma - L_1(f)\right| \lesssim_p K_{p,2}\left(C_f \vee \|f - \gamma(f)\|_4\right)\left(\bar{c}_3\epsilon\right)^2, \\
&\left|\int f d\rho - \int f d\gamma - L_2(f)\right| \lesssim_p K_{p,3}\left(C_f \vee \|f - \gamma(f)\|_4\right)\left(\bar{c}_3\epsilon\right)^3\left(1 + \bar{c}_3\epsilon\right).
\end{aligned} \tag{4.20}$$

If $m \geq 3$, *then there exists a constant* $C(m) < 1$ *such that if* $\bar{c}_3\epsilon \leq C(m)$ *then*

$$\begin{aligned}
&\left|\int f d\rho - \int f d\gamma - \frac{L_m(f)}{L_{m-1}(1)}\right| \\
&\qquad \lesssim_{p,m} K_{p,m+1}\left(C_f \vee \|f - \gamma(f)\|_4\right)\left(\bar{c}_3\epsilon\right)^{m+1}.
\end{aligned} \tag{4.21}$$

When $m \geq 3$, the denominator $L_{m-1}(1)$ no longer equals one exactly, and we show in Lemma B.3 that $|L_{m-1}(1) - 1| \lesssim_m \bar{c}_3\epsilon + (\bar{c}_3\epsilon)^{m-1}$. This upper bound must be small enough to ensure that $L_{m-1}(1) > 0$, explaining why there is a smallness condition on $\bar{c}_3\epsilon$ when $m \geq 3$.

Proposition 4.1 is of independent interest, since it can be used to approximate expectations under ρ by expectations under the standard normal, for

arbitrary polynomially bounded f. Of course, there remains the difficulty of writing $L_m(f)/L_{m-1}(1)$ in a tractable way. The idea is to Taylor expand r to sufficiently high order, so that $L_m(f)/L_{m-1}(1)$ is given by a ratio of the form $\int f q_m d\gamma / \int q_m d\gamma$ plus a negligible remainder, where q_m is a polynomial whose coefficients are functions of the derivatives of W at the origin (recall that $r = W - \|\cdot\|^2/2 + \text{const.}$ so r and W have the same derivatives at the origin).

For small values of m, this is relatively straightforward. For example, when $m = 1$ we have $L_1(f) = -\int f r d\gamma$, using that $\int r d\gamma = 0$. The above proposition then gives a bound on the deviation $|\int f d\rho - \int f d\gamma + \int (f - \gamma(f)) r d\gamma|$. Now we substitute $r = \frac{1}{3!}\langle \nabla^3 W(0), x^{\otimes 3}\rangle + (r_4 - \gamma(r_4))$ and treat $\int (f - \gamma(f))(r_4 - \gamma(r_4)) d\gamma$ as a remainder term. Using the bound on $\|r_4\|_2$ provided in Lemma 4.4, we immediately obtain the following corollary of Proposition 4.1:

Corollary 4.1. *Let* $|f(x) - f(0)| \le C_f \|x\|^p$, *and*

$$L(f) = -\frac{1}{6}\int f(x)\langle \nabla^3 W(0), x^{\otimes 3}\rangle d\gamma(x). \tag{4.22}$$

Then

$$\left|\int f d\rho - \int f d\gamma - L(f)\right| \lesssim_p K_{p,2}\,(C_f \vee \|f - \gamma(f)\|_4)\,(\bar{c}_3^2 + c_4)\frac{d^2}{n} \tag{4.23}$$

Although we have called $L_1(f)$ a "leading order" term, we note that $L(f)$ is itself the leading order contribution to $L_1(f) = -\int f r d\gamma$, which is obtained by dropping $r_4 - \gamma(r_4)$ from r.

In Section 4.4, we apply this corollary to get the leading order asymptotics of the TV error, mean error, and covariance. For the mean, we will also apply Proposition 4.1 with $m = 2$ to get a stronger result. Similarly, one can apply Corollary 4.1 with $m = 1$ for all 1-Lipschitz f to derive the leading order term of the Wasserstein-1 distance. Note that the bound depends on $\|f - \gamma(f)\|_4$ rather than on $\|f\|_4$. This is important since Lipshitz functions of a standard normal are sub-Gaussian *when centered on their mean.*

We end this section by giving a few details about the proof of Proposition 4.1. We start with the upper bound (4.17), and use the following preliminary decomposition of the error E_k.

Lemma 4.5. *Let* $\mathcal{U} = \{\|x\| \le 4\sqrt{d}\}$. *Then*

$$E_k(f) \le 2\|g\|_4\,(1 + \|e^{-r}\chi_{\mathcal{U}}\|_4)\,\|r\|_{2k}^k + \int_{\mathcal{U}^c} |g||r|^k e^{-r} d\gamma, \tag{4.24}$$

where $g = f$ *if* f *is constant, and* $g = f - \gamma(f)$ *otherwise.*

See Appendix B.2 for the proof of this decomposition. The main contribution to this upper bound is the first summand, while the second is exponentially small. Lemma 4.4 gives us the $(\bar{c}_3 d/\sqrt{n})^k$ scaling of the first term, but we also need to show $\|e^{-r}\chi_{\mathcal{U}}\|_4$ is an $O(1)$ correction. Recall again formula (4.11) for r. On $\{\|x\| \le 4\sqrt{d}\}$, the function r is dominated by the cubic $\langle \nabla^3 W(0), x^{\otimes 3}\rangle$,

which can reach values up to $d\sqrt{d}/\sqrt{n}$. This seems like an obstacle to bounding $\|e^{-r}\chi_{\mathcal{U}}\|_4$. However, one can show that ∇r is L-Lipschitz, with a Lipschitz constant $L \sim d/\sqrt{n}$, in this region. Therefore, we can apply Herbst's argument on the exponential integrability of Lipschitz functions with respect to the Gaussian (or any measure satisfying a log Sobolev inequality). See Lemma B.2 in Appendix B.2 for more details.

See Appendix B.2 for the full proof of Proposition 4.1, in which we upper bound E_k using (4.24), and we also upper bound $L_m(f)$ and lower bound $L_{m-1}(1)$.

4.4 Finishing the proof of Theorem V1, V2, V3, V4

Recall that it is equivalent to prove (W1), (W2), (W3), (W4). First, we use Corollary 4.1 with all $\|f\|_\infty \le 1/2$ (so that $p = 0$, $C_f = 1$), to get the TV asymptotics for ρ. We have

$$\begin{aligned}
|R| &= \left| \mathrm{TV}(\rho,\gamma) - \frac{1}{12}\mathbb{E}\left|\langle \nabla^3 W(0), Z^{\otimes 3}\rangle\right| \right| \\
&= \left| \sup_{\|f\|_\infty \le 1/2} \left(\int f d\rho - \int f d\gamma \right) - \frac{1}{6} \sup_{\|f\|_\infty \le 1/2} \mathbb{E}\left[f(Z)\langle \nabla^3 W(0), Z^{\otimes 3}\rangle \right] \right| \\
&\le \sup_{\|f\|_\infty \le 1/2} \left| \int f d\rho - \int f d\gamma - \frac{1}{6}\mathbb{E}\left[f(Z)\langle \nabla^3 W(0), Z^{\otimes 3}\rangle \right] \right| \\
&\lesssim K_{0,2}(\bar{c}_3^2 + c_4)\frac{d^2}{n}.
\end{aligned} \tag{4.25}$$

To finish the proof of (W1), it remains to bound $|L| = \frac{1}{12}\mathbb{E}\left|\langle \nabla^3 W(0), Z^{\otimes 3}\rangle\right|$. The fact that $|L| \lesssim c_3 d/\sqrt{n}$ is shown in Corollary D.1; see also Section 4.5 for a discussion of this result.

Next we turn to the proof of (W2). We apply Corollary 4.1 with functions $f_u(x) = u^T x$, with $\|u\| = 1$ (so that $p = 1$, $C_f = 1$). First we compute the leading order term $L(f_u)$ defined in (4.22). We have

$$\begin{aligned}
L(f_u) &= -\frac{1}{6}\int u^T x \langle \nabla^3 W(0), x^{\otimes 3}\rangle d\gamma(x) = -\frac{1}{2}u^T \int \langle \nabla^3 W(0), x^{\otimes 2}\rangle d\gamma(x) \\
&= -\frac{1}{2}u^T \langle \nabla^3 W(0), I_d\rangle = u^T L.
\end{aligned} \tag{4.26}$$

Here, L is as in (4.3), and we used Gaussian integration by parts to get the

second equality. It follows by Corollary 4.1 that

$$\begin{aligned}\|R\| = \|m_\rho - L\| &= \sup_{\|u\|=1} |u^T m_\rho - u^T L| \\ &= \sup_{\|u\|=1} \left| \int u^T x d\rho(x) - \int u^T x d\gamma(x) - L(f_u) \right| \\ &\lesssim K_{1,2}(\bar{c}_3^2 + c_4)\frac{d^2}{n},\end{aligned} \tag{4.27}$$

noting that $\|f_u\|_4 = \mathbb{E}\,[(u^T Z)^4]^{1/4} \leq 2$ for all $\|u\| = 1$. This proves the first bound on R stated in (W2). To prove the second bound, we use Proposition 4.1 with $m = 2$. We show that the leading order contribution to $L_2(f_u)$ is $L(f_u)$ itself. Hence, the leading order term does not change, and the remainder turns out to be an order of magnitude smaller. See Appendix B.3 for this proof of the second bound on R. The upper bound on $\|L\|$ stated in (W2) follows from (1.10) and the fact that $\|\nabla^3 W(0)\| \leq c_3/\sqrt{n}$ by Lemma 4.2.

To prove (W3), we apply Corollary 4.1 with functions $f_u(x) = (u^T x)^2$, for $\|u\| = 1$ (so that $p = 2$, $C_f = 1$). First we note that $L(f_u) = 0$ since $(u^T x)^2 \langle \nabla^3 W(0), x^{\otimes 3}\rangle$ is an odd order polynomial. Thus we have

$$\begin{aligned}\|R\| = \|\Sigma_\rho - I_d\| &\leq \sup_{\|u=1} u^T \mathbb{E}_{X\sim\rho}[XX^T - I_d]u + \|m_\rho\|^2 \\ &= \sup_{\|u\|=1} \left| \int (u^T x)^2 d\rho(x) - \int (u^T x)^2 d\gamma(x) - L(f_u) \right| + \|m_\rho\|^2 \\ &\lesssim K_{2,2}(\bar{c}_3^2 + c_4)\frac{d^2}{n} + \|m_\rho\|^2,\end{aligned} \tag{4.28}$$

noting that $\|f_u\|_4 = \mathbb{E}\,[(u^T Z)^8]^{1/4} \leq 4$ for all $\|u\| = 1$.

Finally, note that Corollary 4.1 nearly finishes the proof of (W4), and it remains only to bound $L(f)$. This bound follows from Cauchy-Schwarz and Corollary D.1.

4.5 Proof of Key Lemma 4.4

We start by discussing the bound (4.19) on $\mathbb{E}\,[(r_4(Z) - \gamma(r_4))^{2m}]$. Recall that r_4 is the fourth order remainder in the Taylor expansion (4.10) of W. By the Taylor remainder theorem, r_4 can be written as

$$r_4(x) = \frac{1}{4!}\langle \nabla^4 W(tx), x^{\otimes 4}\rangle$$

for some $t \in [0, 1]$ depending on x. Recall from Lemma 4.2 that $\nabla^4 W$ scales as $1/n$, so Cauchy-Schwarz essentially gives that

$$\mathbb{E}\,[|r_4(Z)|^p] \lesssim (1/n)^p \mathbb{E}\,[\|Z\|^{4p}] \lesssim_p (d^2/n)^p,$$

since $\mathbb{E}\,[\|Z\|^k] \sim d^{k/2}$. Therefore, it is completely straightforward to show that $\|r_4\|_p \lesssim d^2/n$. Lemma C.1 provides the rigorous proof, showing that

$$\|r_4\|_{2m} \lesssim_m c_4 d^2/n, \tag{4.29}$$

and hence $\mathbb{E}\left[(r_4(Z)-\gamma(r_4))^{2m}\right] \lesssim_m (c_4 d^2/n)^{2m}$. Next, we turn to the bound (4.18) on $\mathbb{E}\left[\langle \nabla^3 W(0), Z^{\otimes 3}\rangle^{2m}\right]$. We wish to show this scales as $(d/\sqrt{n})^{2m}$. Recall that $\|\nabla^3 W(0)\| \le c_3/\sqrt{n}$, so a straightforward bound would give

$$\mathbb{E}\left[\langle \nabla^3 W(0), Z^{\otimes 3}\rangle^{2m}\right] \le (c_3/\sqrt{n})^{2m}\mathbb{E}\left[\|Z\|^{6m}\right] \lesssim (c_3 d\sqrt{d}/\sqrt{n})^{3m}$$

rather than the desired bound $(d/\sqrt{n})^{2m}$. To get the right scaling with d, we first show that it is sufficient to prove $\mathbb{E}\left[\langle \nabla^3 W(0), H_3(Z)\rangle^{2m}\right] \lesssim (d/\sqrt{n})^{2m}$, where $H_3(Z)$ is the tensor of third order Hermite polynomials of Z. To prove this modified result, we first compute the base case $2m = 2$ explicitly. Namely, we show that

$$\mathbb{E}\left[\langle \nabla^3 W(0), H_3(Z)\rangle^2\right] = 3!\|\nabla^3 W(0)\|_F^2 \le 3! d^2 \|\nabla^3 W(0)\|^2 \le 3! c_3 (d/\sqrt{n})^2, \tag{4.30}$$

which is the right dependence on d for the case $2m = 2$. We then use hypercontractivity of the Ornstein-Uhlenbeck semigroup, for which the Hermite polynomials are eigenfunctions, to relate the $2m$ norm to the 2 norm:

$$\mathbb{E}\left[\langle \nabla^3 W(0), H_3(Z)\rangle^{2m}\right] \lesssim_m \mathbb{E}\left[\langle \nabla^3 W(0), H_3(Z)\rangle^2\right]^m \lesssim_m (c_3 d/\sqrt{n})^{2m}.$$

The final result, shown in Corollary D.1, is that

$$\mathbb{E}\left[\left\langle \nabla^3 W(0), Z^{\otimes 3}\right\rangle^{2m}\right] \lesssim_m (c_3 d/\sqrt{n})^{2m}. \tag{4.31}$$

See Section D leading up to this corollary for the full proof. This finishes the proof of Lemma 4.4.

A Proof of Lemma 2.1

Proof of Lemma 2.1. Using that $c_3 d/\sqrt{n} \le 1$, $c_4 d^2/n \le 1$, $r = \log n\sqrt{d/n}$, and $\log n \le \sqrt{d}$, we get

$$\begin{aligned} 4c_3 r + c_4 r^2 &\le 4\frac{\sqrt{n}}{d}(\log n\sqrt{d/n}) + \frac{n}{d^2}(\log n\sqrt{d/n})^2 \\ &= 4\frac{\log n}{\sqrt{d}} + \frac{\log^2 n}{d} < 6, \end{aligned} \tag{A.1}$$

verifying (2.4). Since (2.6) holds with this choice of r, it follows that Assumption A4 is satisfied. Now, by Lemma 4.3, we have that

$$\frac{\|y\|^2}{4} \le W(y) - W(0) = nv(\hat{m} + H_v^{-1/2} y/\sqrt{n}) - nv(\hat{m}), \quad \forall \|y\| \le r\sqrt{n}.$$

Let $x = H_v^{-1/2} y/\sqrt{n}$. Then $\|y\| = \sqrt{n}\|x\|_{H_v}$, so in terms of x this inequality takes the form

$$\frac{n}{4}\|x\|_{H_v}^2 \le nv(\hat{m} + x) - nv(\hat{m}), \quad \forall \|x\|_{H_v} \le r.$$

Hence

$$nc_0 := \inf_{\|x\|_{H_v}=r} nv(\hat{m}+x) - nv(\hat{m}) \geq \frac{nr^2}{4} = n\frac{\log^2 n}{4}(d/n) = \frac{\log^2 n}{4}d.$$

But then, using that $p + (4+q)\ell + d \leq Cd$, we have

$$\begin{aligned}\log B_{p,\ell} &= (p+(4+q)\ell+d)\left(\log(r\sqrt{n}) + \alpha^{-1}\right) - nc_0 \\ &\leq Cd(\log(\sqrt{d}\log n) + \alpha^{-1}) - \frac{\log^2 n}{4}d \\ &= \left(C\log\log n + \frac{C}{2}\log d + C\alpha^{-1} - \frac{\log^2 n}{4}\right)d \leq 0,\end{aligned} \tag{A.2}$$

so $B_{p,\ell} \leq 1$. To get the final inequality we used (2.26). □

B Proofs from Section 4

B.1 Proofs from Section 4.1 relating V to W

For the proof of Lemma 4.1, we use the following identity.

Lemma B.1. *Let $W(x) = V(Ax+b)$ for a symmetric matrix A. Then*

$$\langle \nabla^3 W(0), I_d\rangle = A\langle \nabla^3 V(b), A^2\rangle. \tag{B.1}$$

Proof. Let A_j denote the jth column of A. Then

$$\begin{aligned}\langle \nabla^3 W(0), I_d\rangle_i = \langle \nabla^3 W(0), I_d \otimes e_i\rangle &= \sum_{j=1}^d \langle \nabla^3 W(0), e_j\otimes e_j\otimes e_i\rangle \\ &= \sum_{j=1}^d \langle \nabla^3 V(b),\ A_j\otimes A_j\otimes A_i\rangle \\ &= \langle \nabla^3 V(b),\ A^2\otimes A_i\rangle = A_i^T\langle \nabla^3 V(b), A^2\rangle \\ &= (A^T\langle \nabla^3 V(b), A^2\rangle)_i = (A\langle \nabla^3 V(b), A^2\rangle)_i.\end{aligned} \tag{B.2}$$

To get the third line we noted that $\sum_{j=1}^d A_j\otimes A_j = AA^T = A^2$. We have shown that the ith coordinate of $\langle \nabla^3 W(0), I_d\rangle$ equals the ith coordinate of $A\langle \nabla^3 V(b), A^2\rangle$, so $\langle \nabla^3 W(0), I_d\rangle = A\langle \nabla^3 V(b), A^2\rangle$. □

Proof of Lemma 4.1. The fact that the TV distances are equal follows from the data processing inequality and the fact that T is a bijection. The formula (4.2) for L in terms of W is immediate using the relationship between V and W. To prove point 2, note that T is linear, so $m_\rho = T(m_\pi) = H_V^{1/2}(m_\pi - \hat{m})$. The formula (4.3) for L in terms of W follows from Lemma B.1 with $A = H_V^{-1/2}$ and $b = \hat{m}$. Point 3 follows from the fact that $\Sigma_\rho = H_V^{1/2}\Sigma_\pi H_V^{1/2}$. Finally, point 4

follows using the definition of the H_V weighted norms and standard change of variables.

□

Proof of Lemma 4.2. We first relate tensor norms and inner products involving W to the corresponding quantities for v. In the next two equations, let $y(x) = H_v^{-1/2}x/\sqrt{n}$, so that $\|y(x)\|_{H_v} = \|x\|/\sqrt{n}$. First, we have

$$\begin{aligned}
\|\nabla^k W(x)\| &= \sup_{\|u\|=1} \langle \nabla^k W(x), u^{\otimes k}\rangle \\
&= \sup_{\|u\|=1} \left\langle n\nabla^k v(\hat{m} + H_v^{-1/2}x/\sqrt{n}),\ (H_v^{-1/2}u/\sqrt{n})^{\otimes k}\right\rangle \\
&= n^{1-k/2} \sup_{\|u\|=1} \left\langle \nabla^k v(\hat{m} + y(x)),\ H_v^{-1/2}u^{\otimes k}\right\rangle \\
&= n^{1-k/2} \sup_{\|w\|_{H_v}=1} \left\langle \nabla^k v(\hat{m} + y(x)),\ w^{\otimes k}\right\rangle \\
&= n^{1-k/2} \|\nabla^k v(\hat{m} + y(x))\|_{H_v},
\end{aligned} \tag{B.3}$$

where the last line follows by (G.1) below (using the fact that $\nabla^k v(\hat{m} + y(x))$ is a symmetric tensor). Second,

$$\begin{aligned}
\langle \nabla^k W(tu), u^{\otimes k}\rangle &= \left\langle n\nabla^k v(\hat{m} + H_v^{-1/2}tu/\sqrt{n}),\ (H_v^{-1/2}u/\sqrt{n})^{\otimes k}\right\rangle \\
&= n\left\langle \nabla^k v(\hat{m} + ty(u)),\ y(u)^{\otimes k}\right\rangle
\end{aligned} \tag{B.4}$$

Using (B.3), the relationship between $\|x\|$ and $\|y(x)\|_{H_v}$, and Assumption A2, we see that $\|\nabla^3 W(0)\| = n^{-1/2}\|\nabla^3 v(\hat{m})\|_{H_v} \le c_3/\sqrt{n}$, and

$$\sup_{\|x\|\le 4\sqrt{d}} \|\nabla^4 W(x)\| = n^{-1} \sup_{\|y\|_{H_v}\le 4\sqrt{d/n}} \|\nabla^4 v(\hat{m} + y)\| \le c_4/n.$$

This proves (4.5) and (4.6). Now, fix $\|u\| = 1$, so that $\|\sqrt{n}y(u)\|_{H_v} = 1$. Using (B.4) and Assumption A3, we then have

$$\begin{aligned}
\langle \nabla^4 W(tu), u^{\otimes 4}\rangle &= n\left\langle \nabla^4 v(\hat{m} + ty(u)), y(u)^{\otimes 4}\right\rangle \\
&= \frac{1}{n}\left\langle \nabla^4 v\left(\hat{m} + \frac{t}{\sqrt{n}}\sqrt{n}y(u)\right), (\sqrt{n}y(u))^{\otimes 4}\right\rangle \\
&\le \frac{c_4}{n}\max\left(1, \frac{t}{\sqrt{n}}\right)^q,
\end{aligned} \tag{B.5}$$

proving (4.7). Next, fix $\|x\| \ge r\sqrt{n}$, so that $\|y(x)\|_{H_v} \ge r$. Then by Assumption A4, we have

$$\begin{aligned}
W(x) - W(0) &= nv(\hat{m} + H_v^{-1/2}x/\sqrt{n}) - nv(\hat{m}) \\
&= nv(\hat{m} + y(x)) - nv(\hat{m}) \ge nc_0\|y(x)/r\|_{H_v}^{\alpha} \\
&= nc_0\|x/(r\sqrt{n})\|^{\alpha}.
\end{aligned} \tag{B.6}$$

This proves (4.8). Finally, (4.9) follows from Assumption A5 analogously to how (4.7) follows from Assumption A3. □

Proof of Lemma 4.3. Recall from Assumption A4 that $r \le 1$. Fix $\|x\| \le r\sqrt{n} \le \sqrt{n}$ and write $x = \|x\|u$ for some $\|u\| = 1$. Then for some $t \in [0, \|x\|] \subset [0, \sqrt{n}]$ we have by Taylor's theorem

$$\begin{aligned} W(x) - W(0) &= \frac{\|x\|^2}{2} + \frac{1}{6}\langle \nabla^3 W(0), x^{\otimes 3}\rangle + \frac{1}{24}\langle \nabla^4 W(tu), (\|x\|u)^{\otimes 4}\rangle \\ &\ge \frac{\|x\|^2}{2} - \frac{c_3\|x\|^3}{6\sqrt{n}} - \frac{c_4\|x\|^4}{24n} \\ &= \frac{\|x\|^2}{2}\left(1 - \frac{c_3\|x\|}{3\sqrt{n}} - \frac{c_4\|x\|^2}{12n}\right). \end{aligned} \tag{B.7}$$

using that $\|\nabla^3 W(0)\| \le c_3/\sqrt{n}$ by (4.5), and $\sup_{|t|\le\sqrt{n}}\langle \nabla^4 W(tu), u^{\otimes 4}\rangle \le c_4/n$ for all $\|u\| = 1$, which follows by (4.7). Upper bounding $\|x\|$ by $r\sqrt{n}$, we get the further lower bound

$$W(x) - W(0) \ge \frac{\|x\|^2}{2}\left(1 - \frac{c_3 r}{3} - \frac{c_4 r^2}{12}\right) \ge \|x\|^2/4, \tag{B.8}$$

using (2.4) to get the last inequality. □

B.2 Proof of Proposition 4.1

The proof of Proposition 4.1 will follow from the lemmas stated in Section 4.3 and a number of additional lemmas proved here.

Proof of (4.17). We compute

$$\begin{aligned} \left|\int f d\rho - \int f d\gamma - \frac{L_m(f)}{L_{m-1}(1)}\right| &= \left|\frac{L_m(f) + E_{m+1}(f)}{L_{m-1}(1) + E_m(1)} - \frac{L_m(f)}{L_{m-1}(1)}\right| \\ &= \left|\frac{E_{m+1}(f)L_{m-1}(1) - L_m(f)E_m(1)}{(L_{m-1}(1) + E_m(1))L_{m-1}(1)}\right| \\ &\le \left|\frac{E_{m+1}(f)L_{m-1}(1) - L_m(f)E_m(1)}{L_{m-1}(1)}\right| \\ &\le |E_{m+1}(f)| + |E_m(1)|\left|\frac{L_m(f)}{L_{m-1}(1)}\right|, \end{aligned} \tag{B.9}$$

where in the third line we used that $L_{m-1}(1) + E_m(1) = \int e^{-r} d\gamma \ge e^{\int -r d\gamma} = 1$, since $\int r d\gamma = 0$. □

Proof of Lemma 4.5. The Taylor remainder of e^{-r} of order k is given by

$$e^{-r} - \left(\sum_{j=0}^{k-1}(-1)^j r^j/j!\right) = \frac{r^k}{k!}e^{\xi}$$

for some ξ between 0 and $-r$. But then $e^{\xi} \le 1 + e^{-r}$, so that

$$\left|e^{-r} - \left(\sum_{j=0}^{k-1}(-1)^j r^j/j!\right)\right| \le |r|^k + |r|^k e^{-r}.$$

Now, we have

$$
\begin{aligned}
|E_k(f)| &= \left| \int g\Big(e^{-r} - \big(\sum_{j=0}^{k-1}(-1)^j r^j/j!\big)\Big)d\gamma \right| \le \int |g||r|^k d\gamma + \int |g||r|^k e^{-r} d\gamma \\
&= \int |g| r^k d\gamma + \int |g| r^k (e^{-r}\chi_{\mathcal{U}}) d\gamma + \int_{\mathcal{U}^c} |g||r|^k e^{-r} d\gamma \\
&\le \|g\|_2 \|r\|_{2k}^k + \|g\|_4 \|r\|_{2k}^k \|e^{-r}\chi_{\mathcal{U}}\|_4 + \int_{\mathcal{U}^c} |g||r|^k e^{-r} d\gamma \\
&\le \|g\|_4 \|r\|_{2k}^k (1 + \|e^{-r}\chi_{\mathcal{U}}\|_4) + \int_{\mathcal{U}^c} |g||r|^k e^{-r} d\gamma.
\end{aligned}
\tag{B.10}
$$

To get the second to last line, we applied Cauchy-Schwarz to $\int |g| r^2 d\gamma$ and generalized Hölder to $\int |g||r|^k (e^{-r}\chi_{\mathcal{U}}) d\gamma$ with powers $4, 2, 4$ to the factors $|g|$, $|r|^k$, and $e^{-r}\chi_{\mathcal{U}}$, respectively. □

Next, we bound $\|e^{-r}\chi_{\mathcal{U}}\|_4$.

Lemma B.2. *Let $\epsilon = d/\sqrt{n}$. Then $\|e^{-r}\chi_{\mathcal{U}}\|_4 \le e^{A(q)\bar{c}_3\epsilon}$.*

Proof. We first bound $\|\nabla r(x)\|$ for $x \in \mathcal{U} = \{\|x\| \le 4\sqrt{d}\}$. To do so, we Taylor expand ∇r:

$$
\begin{aligned}
\nabla r(x) &= \nabla(W(x) - \|x\|^2/2) = \nabla W(x) - \nabla W(0) - x \\
&= \nabla W(x) - \nabla W(0) - \langle \nabla^2 W(0), x\rangle \\
&= \frac{1}{2}\langle \nabla^3 W(0), x^{\otimes 2}\rangle + \frac{1}{6}\langle \nabla^4 W(tx), x^{\otimes 3}\rangle,
\end{aligned}
\tag{B.11}
$$

for some $t \in [0,1]$. Therefore,

$$
\begin{aligned}
\sup_{\|x\|\le 4\sqrt{d}} \|\nabla r(x)\| &\le \frac{1}{2} 16 d \|\nabla^3 W(0)\| + \frac{1}{6} 64 d\sqrt{d} \sup_{\|x\|\le 4\sqrt{d}} \|\nabla^4 W(x)\| \\
&\le 8\frac{c_3 d}{\sqrt{n}} + 12\frac{c_4 d\sqrt{d}}{n} \le 12(c_3\epsilon + c_4\epsilon^2) = 12\bar{c}_3\epsilon =: L
\end{aligned}
\tag{B.12}
$$

To get the third inequality we used point 1 of Lemma 4.2. Hence, $-4r$ is $4L$-Lipschitz in $\mathcal{U}$. Let $\gamma_{\mathcal{U}}$ be the Gaussian measure restricted to $\mathcal{U}$. Using Herbst's argument on the exponential integrability of Lipschitz functions with respect to a measure satisfying a log Sobolev inequality (for $\gamma_{\mathcal{U}}$, with constant 1) we have

$$
\begin{aligned}
\left(\int_{\mathcal{U}} e^{-4r} d\gamma\right)^{1/4} &\le \left(\int e^{-4r} d\gamma_{\mathcal{U}}\right)^{1/4} \le \left(e^{\int(-4r)d\gamma_{\mathcal{U}}} e^{(4L)^2/2}\right)^{1/4} \\
&= \exp\left(-\int r d\gamma_{\mathcal{U}} + 2L^2\right) \le \exp\left(\int_{\mathcal{U}^c} |r| d\gamma + 2L^2\right),
\end{aligned}
\tag{B.13}
$$

using that $\int r d\gamma = 0$ on $\mathbb{R}^d$. See Proposition 5.4.1 of [Bakry et al., 2014] for a reference on Herbst's argument. Now, using Lemma C.2, we have on the set $\mathcal{U}^c$ the following upper bound:

$$|r(x)| \lesssim_q \bar{c}_3\epsilon\|x\|^{4+q}. \tag{B.14}$$

Hence

$$\int_{\mathcal{U}^c} |r|d\gamma \lesssim_q \bar{c}_3\epsilon\mathbb{E}\left[\|Z\|^{4+q}\{\|Z\| \geq 4\sqrt{d}\}\right] \lesssim_q \bar{c}_3\epsilon, \tag{B.15}$$

and

$$\int_{\mathcal{U}^c} |r|d\gamma + 2L^2 \lesssim_q \bar{c}_3\epsilon,$$

recalling that $L = 12\bar{c}_3\epsilon$. Therefore,

$$\|e^{-r}\chi_{\mathcal{U}}\|_4 \leq \exp\left(A(q)\bar{c}_3\epsilon\right)$$

for some constant $A(q)$, as desired. □

Using the above bounds, the key Lemma 4.4 bounding $\|r\|$, and the tail bound in Lemma E.1 below, we get the following further upper bound on $E_k(f)$.

Corollary B.1. *Suppose $|f(x) - f(0)| \leq C_f\|x\|^p$. Then*

$$|E_k(f)| \lesssim_{p,k} K_{p,k}(C_f \vee \|g\|_4)(\bar{c}_3\epsilon)^k, \tag{B.16}$$

where as usual the meaning of $E_k(f)$ and g depends on whether or not f is constant.

Proof. First, note that if $|f(x) - f(0)| \leq C_f\|x\|^p$, then $|\gamma(f - f(0))| \lesssim_p C_f d^{p/2}$, and therefore on $\{\|x\| \geq 4\sqrt{d}\}$, we have $|g(x)| = |f(x) - \gamma(f)| \lesssim_p C_f\|x\|^p$. Alternatively, if f is a constant, then $g = f = C_f$ satisfies $|g| \leq C_f\|x\|^p$ with $p = 0$. Substituting this inequality, as well as the bound from Lemma B.2 into (4.24), we get

$$\begin{aligned} |E_k(f)| &\lesssim_p e^{A(q)\bar{c}_3\epsilon}\|g\|_4\|r\|_k^k + C_f\int_{\mathcal{U}^c} \|x\|^p|r|^k e^{-r}d\gamma \\ &\lesssim_{p,k} e^{A(q)\bar{c}_3\epsilon}\|g\|_4(\bar{c}_3\epsilon)^k + C_f K_{p,k}(\bar{c}_3\epsilon)^k \\ &\lesssim K_{p,k}(C_f \vee \|g\|_4)(\bar{c}_3\epsilon)^k, \end{aligned} \tag{B.17}$$

as desired. We used Lemma 4.4 and Lemma E.1 to get the second line. □

Next, we bound $|L_m(f)|$ and $|L_{m-1}(1) - 1|$.

Lemma B.3. *We have $L_0(1) = L_1(1) = 1$ and*

$$\begin{aligned} |L_m(f)| &\lesssim_m \|f - \gamma(f)\|_2(\bar{c}_3\epsilon + (\bar{c}_3\epsilon)^m), \\ |L_m(1) - 1| &\lesssim_m \bar{c}_3\epsilon + (\bar{c}_3\epsilon)^m \end{aligned} \tag{B.18}$$

Proof. $L_0(1) = 1$ by definition, $L_1(1) = \int (1-r)d\gamma = 1$ since $\int r d\gamma = 0$, and $L_1(f) = -\int (f-\gamma(f))r d\gamma$, so $|L_1(f)| \le \|f-\gamma(f)\|_2 \|r\|_2 \le \|f-\gamma(f)\|_2 (\bar{c}_3\epsilon)$ by Lemma 4.4. Next take $m \ge 2$. Note that

$$L_m(f) = \int (f-\gamma(f))(1-r+...)d\gamma = \int (f-\gamma(f))(-r+...)d\gamma,$$

i.e. the term with r^0 goes away. Hence

$$\begin{aligned} |L_m(f)| &\le \sum_{k=1}^m \int |f-\gamma(f)| \frac{|r|^k}{k!} d\gamma \\ &\le \|f-\gamma(f)\|_2 \sum_{k=1}^m \|r\|_{2k}^k \lesssim_m \|f-\gamma(f)\|_2 \sum_{k=1}^m (\bar{c}_3\epsilon)^k \\ &\lesssim_m \|f-\gamma(f)\|_2 (\bar{c}_3\epsilon + (\bar{c}_3\epsilon)^m) \end{aligned} \tag{B.19}$$

Similarly, note that $L_m(1) - 1 = \int (-r+...)d\gamma$, and hence

$$\begin{aligned} |L_m(1)-1| &\le \sum_{k=1}^m \int \frac{|r|^k}{k!} d\gamma \\ &\le \sum_{k=1}^m \|r\|_k^k \lesssim_m \sum_{k=1}^m (\bar{c}_3\epsilon)^k \\ &\lesssim_m \bar{c}_3\epsilon + (\bar{c}_3\epsilon)^m. \end{aligned} \tag{B.20}$$

□

Using the above lemmas, we can now prove Proposition 4.1.

Proof of Proposition 4.1. Using (4.17), Corollary B.1, and Lemma B.3, we have

$$\begin{aligned} &\left| \int f d\rho - \int f d\gamma - \frac{L_m(f)}{L_{m-1}(1)} \right| \\ &\lesssim_{m,p} K_{p,m+1}(C_f \vee \|f-\gamma(f)\|_4)(\bar{c}_3\epsilon)^{m+1} + K_{p,m}(\bar{c}_3\epsilon)^m \|f-\gamma(f)\|_2 \frac{(\bar{c}_3\epsilon) + (\bar{c}_3\epsilon)^m}{L_{m-1}(1)} \\ &\lesssim_{m,p} K_{p,m+1}(C_f \vee \|f-\gamma(f)\|_4)(\bar{c}_3\epsilon)^{m+1} \left(1 + \frac{1 + (\bar{c}_3\epsilon)^{m-1}}{L_{m-1}(1)}\right) \end{aligned} \tag{B.21}$$

Now we distinguish between three cases: $m = 1$, $m = 2$, $m \ge 3$. If $m = 1$ then the expression in parentheses equals 2, an absolute constant, and we are done. If $m = 2$ then $L_{m-1}(1) = 1$ and the expression in parentheses is bounded by $1 + \bar{c}_3\epsilon$. Finally, consider $m \ge 3$ and suppose $\bar{c}_3\epsilon \le 1$. Then by Lemma B.3 we have $|L_{m-1}(1) - 1| \lesssim_m \bar{c}_3\epsilon$, so if $\bar{c}_3\epsilon \le C(m)$ for a small enough constant $C(m)$, then $L_{m-1} > 1/2$, and hence the expression in parentheses is bounded by an absolute constant. □

B.3 Proofs from Section 4.4

Proof of second bound on R from (W2). Let $f_u(x) = u^T x$, and recall the notation $L_k(f)$ from Section 4.3. We have

$$L_2(f_u) = -\mathbb{E}\left[u^T Z r(Z)\right] + \frac{1}{2}\mathbb{E}\left[u^T Z r(Z)^2\right].$$

We will use the following two representations of $r(x)$, which stem from Taylor expanding W either to third or fourth order, and recalling that r is defined to have expectation zero:

$$\begin{aligned} r(x) &= p_3(x) + [r_4(x) - \gamma(r_4)], \\ &= p_3(x) + [p_4(x) - \gamma(p_4)] + [r_5(x) - \gamma(r_5)], \end{aligned} \tag{B.22}$$

where

$$p_k(x) = \frac{1}{k!}\langle \nabla^k W(0), k^{\otimes 4}\rangle, \quad r_k(x) = \frac{1}{k!}\langle \nabla^k W(tx), x^{\otimes k}\rangle,$$

for some $t \in [0,1]$, and the second representation in (B.22) holds under Assumption A5. Thus

$$\begin{aligned} L_2(f_u) =& -\mathbb{E}\left[u^T Z\left(p_3(Z) + p_4(Z) - \gamma(p_4) + r_5(Z) - \gamma(r_5)\right)\right] \\ &+ \frac{1}{2}\mathbb{E}\left[u^T Z(p_3(Z) + r_4(Z) - \gamma(r_4))^2\right] \\ =& u^T L - \mathbb{E}\left[u^T Z r_5(Z)\right] \\ &+ \mathbb{E}\left[u^T Z p_3(Z)(r_4(Z) - \gamma(r_4))\right] + \frac{1}{2}\mathbb{E}\left[u^T Z(r_4(Z) - \gamma(r_4))^2\right]. \end{aligned} \tag{B.23}$$

We used the second representation of r to write $-\mathbb{E}\left[u^T Z r(Z)\right]$ in the first line of (B.23), and the first representation of r to write $\frac{1}{2}\mathbb{E}\left[u^T Z r(Z)^2\right]$ in the second line of (B.23). In the third line of (B.23), we used that $-\mathbb{E}\left[u^T Z p_3(Z)\right]$ is precisely $L(f_u) = u^T L$ (as shown in (4.26)). We also used that $\mathbb{E}\left[u^T Z(p_4(Z) - \gamma(p_4))\right] = 0$, since this observable is an odd order polynomial of Z. Finally, in the fourth line we expanded the square (from the second line), and used that $\mathbb{E}\left[u^T Z p_3(Z)^2\right] = 0$ since the observable is an odd order polynomial. Now, let

$$R_{2,u} = -\mathbb{E}\left[u^T Z r_5(Z)\right] + \mathbb{E}\left[u^T Z p_3(Z)(r_4(Z) - \gamma(r_4))\right] + \frac{1}{2}\mathbb{E}\left[u^T Z(r_4(Z) - \gamma(r_4))^2\right],$$

so that $L_2(f_u) = u^T L + R_{2,u}$. We have the bound

$$\begin{aligned} |R_{2,u}| &\lesssim \|f_u\|_2 \|r_5\|_2 + \|f_u\|_2 \|p_3\|_4 \|r_4\|_4 + \|f_u\|_2 \|r_4\|_4^2 \\ &\lesssim c_5 d^{-1/2}\epsilon^3 + c_3 c_4 \epsilon^3 + c_4^2 \epsilon^4 = (\bar{c}_3 c_4 + c_5 d^{-1/2})\epsilon^3, \end{aligned} \tag{B.24}$$

using Corollary D.1 and Lemma C.1.

Finally, we apply Proposition 4.1 with the functions $f_u = u^T x$, $\|u\| = 1$, and $m = 2$. We have

$$\begin{aligned}
\|R\| = \|m_\rho - L\| &= \sup_{\|u\|=1} \int u^T x d\gamma(x) - u^T L \\
&= \sup_{\|u\|=1} \left| \int f_u d\rho - \int f_u d\gamma - u^T L \right| \\
&\leq \sup_{\|u\|=1} \left| \int f_u d\rho - \int f_u d\gamma - L_2(f_u) \right| + \sup_{\|u\|=1} R_{2,u} \\
&\lesssim K_{1,3}((\bar{c}_3\epsilon)^3 + (\bar{c}_3\epsilon)^4) + (\bar{c}_3 c_4 + c_5 d^{-1/2})\epsilon^3 \\
&\lesssim K_{1,3}(\bar{c}_3 c_4 + \bar{c}_3^3 + c_5 d^{-1/2})\epsilon^3 + \bar{c}_3^4 \epsilon^4
\end{aligned} \tag{B.25}$$

□

C Auxiliary Results

Lemma C.1. *Let r_4 and r_5 be the fourth and fifth order remainders in the Taylor expansion of W about zero, i.e.*

$$r_k(x) = W(x) - \sum_{j=0}^{k-1} \frac{1}{j!} \langle \nabla^j W(0), x^{\otimes j} \rangle, \quad k = 4, 5. \tag{C.1}$$

Then

$$\|r_4\|_p \lesssim_{p,q} c_4 \epsilon^2, \quad \forall p \geq 1. \tag{C.2}$$

If Assumption A5 also holds, then

$$\|r_5\|_p \lesssim_{p,q} c_5 d^{-1/2} \epsilon^3, \quad \forall p \geq 1. \tag{C.3}$$

Proof. Fix $x \in \mathbb{R}^d$ and let $u = x/\|x\|$. Then $r_k(x)$ is the kth order Taylor remainder of the function $W_u : t \mapsto W(tu)$ about $t = 0$, evaluated at $t = \|x\|$. Therefore, using (4.7), we have

$$|r_4(x)| \leq \sup_{s \in [0, \|x\|]} |W_u^{(4)}(s)| \frac{\|x\|^4}{4!} \leq \frac{c_4}{24n} \|x\|^4 (1 + \|x/\sqrt{n}\|^q). \tag{C.4}$$

Hence

$$\begin{aligned}
\mathbb{E}\left[|r_4(Z)|^p\right] &\lesssim_p \left(\frac{c_4}{n}\right)^p \left(\mathbb{E}\left[\|Z\|^{4p}\right] + \frac{\mathbb{E}\left[\|Z\|^{(4+q)p}\right]}{(\sqrt{n})^{qp}} \right) \\
&\lesssim_{p,q} \left(\frac{c_4}{n}\right)^p d^{2p} = (c_4 \epsilon^2)^p,
\end{aligned} \tag{C.5}$$

using that $d \leq n$. Taking the pth root gives the desired bound. When Assumption A5 holds we can use (4.9) to show, similarly to (C.4), that

$$|r_5(x)| \leq \sup_{s \in [0, \|x\|]} |W_u^{(5)}(s)| \frac{\|x\|^5}{5!} \leq \frac{c_5}{120n} \|x\|^5 (1 + \|x/\sqrt{d}\|^q). \tag{C.6}$$

From here, an analogous calculation gives

$$\|r_5\|_p \lesssim_{p,q} \frac{c_5 d^{2.5}}{n\sqrt{n}} = c_5 d^{-1/2}\epsilon^3,$$

as desired. □

Lemma C.2. *We have*

$$|r(x)| \lesssim \bar{c}_3\epsilon\|x\|^{4+q} \tag{C.7}$$

for all $x \in \mathcal{U}^c = \{x \,:\, \|x\| \geq 4\sqrt{d}\}$.

Proof. Using the Taylor expansion (4.11) of r, Lemma 4.2, Lemma C.1, and (C.4) from the proof of Lemma C.1, we get

$$\begin{aligned} |r(x)| &\leq \frac{\|\nabla^3 W(0)}{3!}\|x\|^3 + |r_4(x)| + |\gamma(r_4)| \\ &\lesssim \frac{c_3}{\sqrt{n}}\|x\|^3 + \frac{c_4}{n}\|x\|^4(1 + \|x/\sqrt{n}\|^q) + c_4\epsilon^2 \\ &\leq c_3\epsilon\|x\|^3 + c_4\epsilon^2\|x\|^4(1 + \|x/\sqrt{n}\|^q) + c_4\epsilon^2. \end{aligned} \tag{C.8}$$

Now note that $1 \leq \|x\|^3 \leq \|x\|^4 \leq \|x\|^{4+q}$ on $\mathcal{U}^c$, and $\|x/\sqrt{n}\| \leq \|x\|$. Then the last line above is bounded by $|r(x)| \lesssim (c_3\epsilon + c_4\epsilon^2)\|x\|^{4+q} = \bar{c}_3\epsilon\|x\|^{4+q}$, as desired. □

D Hermite-Related Proofs

D.1 Very Brief Hermite Primer

Let $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}^d_{\geq 0}$. We let $|\gamma| = \gamma_1 + \cdots + \gamma_d$, and $\gamma! = \gamma_1! \ldots \gamma_d!$. Then

$$H_\gamma(x_1, \ldots, x_d) = \prod_{i=1}^d H_{\gamma_i}(x_i),$$

where $H_k(x)$ is the order k univariate Hermite polynomial. We have $H_0(x) = 1, H_1(x) = x, H_2(x) = x^2 - 1, H_3(x) = x^3 - 3x$. We have

$$\mathbb{E}\left[H_\gamma(Z)H_{\gamma'}(Z)\right] = \delta_{\gamma,\gamma'}\gamma!.$$

Given $i, j, k \in [d]$, let $\gamma(ijk) = (\gamma_1, \ldots, \gamma_d)$ be given by

$$\gamma_\ell = \delta_{i\ell} + \delta_{j\ell} + \delta_{k\ell}, \quad \ell = 1, \ldots, d.$$

In other words γ_ℓ is the number of times index $\ell \in [d]$ repeats within the string ijk. For example

$$\gamma(111) = (3, 0, \ldots, 0), \quad \gamma(113) = (2, 0, 1, 0, \ldots, 0).$$

We define $\mathbf{H}_3(x)$ as the $d \times d \times d$ tensor, with entries

$$\mathbf{H}_3^{ijk}(x_1, \ldots, x_d) = H_{\gamma(ijk)}(x_1, \ldots, x_d).$$

One can show that $\mathbf{H}_3(x) = x^{\otimes 3} - 3\mathrm{Sym}(x \otimes I_d)$.

D.2 Proof of main L^p bound

First, we prove

Lemma D.1. *If T is a symmetric $d \times d \times d$ tensor, then*

$$\mathbb{E}\left[\langle T, \mathbf{H}_3(Z)\rangle^2\right] = 3!\|T\|_F^2.$$

Proof. Note that given a γ with $|\gamma| = 3$, there are $3!/\gamma!$ tuples $(i,j,k) \in [d]^3$ for which $\gamma(ijk) = \gamma$. We let T_γ denote T_{ijk} for any ijk for which $\gamma(ijk) = \gamma$. This is well-defined since T is symmetric. Now, since both T and $\mathbf{H}_3$ are symmetric tensors, we can write the inner product between T and $\mathbf{H}_3$ by grouping together equal terms. In other words, for all $3!/\gamma!$ tuples i,j,k such that $\gamma(ijk) = \gamma$, we have $T_{ijk}\mathbf{H}_3^{ijk} = T_\gamma H_\gamma$. Therefore,

$$\langle T, \mathbf{H}_3(Z)\rangle = \sum_{|\gamma|=3} \frac{3!}{\gamma!} T_\gamma H_\gamma(Z).$$

Using this formula, we get

$$\begin{aligned}
\mathbb{E}\left[\langle T, \mathbf{H}_3(Z)\rangle^2\right] &= \sum_{|\gamma|=3,|\gamma'|=3}^{d} \frac{3!}{\gamma!}\frac{3!}{\gamma'!} T_\gamma T_{\gamma'} \mathbb{E}\left[H_\gamma(Z) H_{\gamma'}(Z)\right] \\
&= \sum_{|\gamma|=3}^{d} \frac{3!}{\gamma!}\frac{3!}{\gamma!} T_\gamma^2 \gamma! = 3! \sum_{|\gamma|=3}^{d} \frac{3!}{\gamma!} T_\gamma^2 \qquad \text{(D.1)} \\
&= 3! \sum_{i,j,k=1}^{d} T_{ijk}^2 = 3!\|T\|_F^2.
\end{aligned}$$

□

Lemma D.2. *If T is a symmetric $d \times d \times d$ tensor, then*

$$\|\langle T, \mathbf{H}_3\rangle\|_{2k} \le \sqrt{6}(2k-1)^{3/2} d \|T\|.$$

Proof. Let $\mathcal{L}$ be the generator for the d-dimensional Ornstein-Uhlenbeck process. Then it is known that $(\mathcal{L}H_\gamma)(x) = -|\gamma| H_\gamma(x)$, i.e. the H_γ are eigenfunctions of $\mathcal{L}$ with corresponding eigenvalues $-|\gamma|$. Hence, $P_t\langle T, \mathbf{H}_3\rangle = e^{-3t}\langle T, \mathbf{H}_3\rangle$, where $P_t = e^{t\mathcal{L}}$. Now, by hypercontractivity (see e.g. Chapter 5.2.2 of [Bakry et al., 2014]), we have

$$e^{-3t}\|\langle T, H_3\rangle\|_{q(t)} = \|P_t\langle T, \mathbf{H}_3\rangle\|_{q(t)} \le \|\langle T, \mathbf{H}_3\rangle\|_2,$$

where $q(t) = 1 + e^{2t}$. Setting $2k = q(t)$ we get $e^{3t} = (2k-1)^{3/2}$, so that

$$\|\langle T, H_3\rangle\|_{2k} \le (2k-1)^{3/2}\|\langle T, \mathbf{H}_3\rangle\|_2 \le (2k-1)^{3/2}\sqrt{6}\|T\|_F, \qquad \text{(D.2)}$$

where the last inequality is by Lemma D.1. Finally, we use that $\|T\|_F \le d\|T\|$. □

Corollary D.1. *If T is a symmetric $d \times d \times d$ tensor, then*

$$\|\langle T, x^{\otimes 3}\rangle\|_{2k} \le 6(2k-1)^{3/2} d\|T\|.$$

In particular, if $p_3(x) = \frac{1}{3!}\langle \nabla^3 W(0), x^{\otimes 3}\rangle$ then

$$\|p_3\|_{2k} \le (2k-1)^{3/2} c_3 \epsilon.$$

Proof. The second statement follows from the first by recalling $\|\nabla^3 W(0)\| \lesssim c_3/\sqrt{n}$ by Lemma 4.2. To prove the first statement, we use that since $\mathbf{H}_3(x) = x^{\otimes 3} - 3\mathrm{Sym}(x \otimes I_d)$ and T is symmetric, we have

$$\langle T, \mathbf{H}_3(x)\rangle = \langle T, x^{\otimes 3} - 3x \otimes I_d\rangle = \langle T, x^{\otimes 3}\rangle - 3\langle T, x \otimes I_d\rangle = \langle T, x^{\otimes 3}\rangle - 3y^T x,$$

where $y = \langle T, I_d\rangle$. Note that $\|y\| \le d\|T\|$. Therefore, $\langle T, x^{\otimes 3}\rangle = \langle T, \mathbf{H}_3(x)\rangle + 3y^T x$, so that

$$\begin{aligned} \|\langle T, x^{\otimes 3}\rangle\|_{2k} &\le \|\langle T, \mathbf{H}_3\rangle\|_{2k} + 3\|y^T x\|_{2k} \\ &\le \sqrt{6}(2k-1)^{3/2} d\|T\| + 3((2k-1)!!)^{1/2k} d\|T\| \\ &\le 6(2k-1)^{3/2} d\|T\|, \end{aligned} \tag{D.3}$$

since $(2k-1)!!^{1/2k} \le (2k-1)^{3/2}$ and $\sqrt{6} + 3 \le 6$. □

E Negligible Tail

Lemma E.1. *We have*

$$\begin{aligned} &\int_{\mathcal{U}^c} \|x\|^p |r(x)|^{k+1} e^{-r(x)} d\gamma(x) \\ &\qquad \lesssim_{p,k} e^{A(q)\bar{c}_3\epsilon}(1 + B_{p,k+1})\,(\bar{c}_3\epsilon)^{k+1} = K_{p,k+1}(\bar{c}_3\epsilon)^{k+1}. \end{aligned} \tag{E.1}$$

We start with two supplementary lemmas. For their proof, note that by comparing the Taylor expansion (4.10) of W and the second line of the Taylor expansion (4.11) of r, we have that

$$r(x) = W(x) - W(0) - \frac{\|x\|^2}{2} - \gamma(r_4),$$

and hence

$$\begin{aligned} e^{-r(x)}\gamma(x) &= (2\pi)^{-d/2} \exp\left(-r(x) - \|x\|^2/2\right) \\ &= (2\pi)^{-d/2} \exp\left(\gamma(r_4) + W(0) - W(x)\right) \\ &\le (2\pi)^{-d/2} e^{A(q)c_4\epsilon^2} \exp\left(W(0) - W(x)\right), \end{aligned} \tag{E.2}$$

where the last line uses Lemma C.1 to bound $|\gamma(r)| \le \|r_4\|_2 \le A(q)c_4\epsilon^2$.

Lemma E.2. *We have*

$$I := \int_{4\sqrt{d}\le\|x\|\le r\sqrt{n}} \|x\|^p e^{-r(x)}\gamma(x)dx \lesssim_p e^{A(q)c_4\epsilon^2}(3/4)^{d/2}. \tag{E.3}$$

Proof. When $\|x\| \le r\sqrt{n}$, we have by Lemma 4.3 that $W(x) - W(0) \ge \|x\|^2/4$. Therefore using (E.2) we have that

$$e^{-r(x)}\gamma(x) \le (2\pi)^{-d/2} e^{A(q)c_4\epsilon^2} e^{-\|x\|^2/4}, \quad \|x\| \le r\sqrt{n}.$$

Substituting into the integral I, we get

$$I \le e^{A(q)c_4\epsilon^2}(2\pi)^{-d/2} \int_{4\sqrt{d}\le\|x\|} \|x\|^p e^{-\|x\|^2/4}dx, \tag{E.4}$$

and

$$\begin{aligned}
&(2\pi)^{-d/2} \int_{4\sqrt{d}\le\|x\|} \|x\|^p e^{-\|x\|^2/4}dx \\
&\qquad = \sqrt{2}^{d+p}\mathbb{E}\left[\|Z\|^p\{\|Z\| \ge 2\sqrt{2}\sqrt{d}\}\right] \\
&\qquad \lesssim_p \sqrt{2}^d \sqrt{d}^p e^{-\frac{d}{4}(2\sqrt{2}-1)^2} \\
&\qquad \lesssim_p d^{p/2}\left(\sqrt{2}e^{-\frac{1}{4}(2\sqrt{2}-1)^2}\right)^d \\
&\qquad \le d^{p/2}(3/4)^d \lesssim_p (3/4)^{d/2}.
\end{aligned} \tag{E.5}$$

Here we used that $d^{p/2}(3/4)^d = \left[d^{p/2}(3/4)^{d/2}\right](3/4)^{d/2}$ and the expression in square brackets is bounded above by a constant $C(p)$ for all $d \in \mathbb{N}$. Substituting (E.5) into (E.4) finishes the proof. □

Lemma E.3. *We have*

$$\int_{\|x\|\ge r\sqrt{n}} \|x\|^p e^{-r(x)}\gamma(x)dx \lesssim e^{A(q)c_4\epsilon^2} \exp\left((p+d)\log(r\sqrt{n}e^{\frac{1}{\alpha}}) - nc_0\right). \tag{E.6}$$

Proof. Recall from point 3 of Lemma 4.2 that

$$W(x) - W(0) \ge nc_0\|x/(r\sqrt{n})\|^\alpha, \quad \forall\|x\| \ge r\sqrt{n}$$

and hence using (E.2) we have

$$\begin{aligned}
e^{-r(x)}\gamma(x) &\le \frac{e^{A(q)c_4\epsilon^2}}{(2\pi)^{d/2}} \exp\left(-nc_0\|x/(r\sqrt{n})\|^\alpha\right) \\
&= \frac{e^{A(q)c_4\epsilon^2}}{(2\pi)^{d/2}} \exp\left(-\|Mx/N\|^\alpha\right),
\end{aligned} \tag{E.7}$$

where $M = (nc_0)^{1/\alpha}$ and $N = r\sqrt{n}$. Therefore,

$$\begin{aligned}
&\int_{\|x\|\geq r\sqrt{n}} \|x\|^p e^{-r(x)}\gamma(x)dx \\
&\qquad \leq \frac{e^{A(q)c_4\epsilon^2}}{(2\pi)^{d/2}} \int_{\|x\|\geq r\sqrt{n}} \|x\|^p \exp\left(-\|Mx/N\|^\alpha\right) dx \qquad \text{(E.8)} \\
&\qquad = \frac{S_{d-1}e^{A(q)c_4\epsilon^2}}{(2\pi)^{d/2}} \int_N^\infty u^{p+d-1} e^{-(\frac{M}{N}u)^\alpha} du,
\end{aligned}$$

where S_{d-1} is the surface area of the unit d-sphere. Now we apply Lemma E.4 with $a = N = r\sqrt{n}, b = M/N$ (where $M = (nc_0)^{1/\alpha}$) and $\alpha = \alpha$ to get that

$$\begin{aligned}
\int_{\|x\|\geq r\sqrt{n}} \|x\|^p e^{-r(x)}\gamma(x)dx &\lesssim_\alpha e^{A(q)c_4\epsilon^2}(r\sqrt{n})^{p+d} e^{\frac{p+d}{\alpha} - nc_0} \\
&= e^{A(q)c_4\epsilon^2} \exp\left((p+d)\log(r\sqrt{n}e^{\frac{1}{\alpha}}) - nc_0\right), \qquad \text{(E.9)}
\end{aligned}$$

as desired. □

Combining the above two lemmas, we immediately get

Corollary E.1. *We have*

$$\int_{\mathcal{U}^c} \|x\|^p e^{-r(x)}\gamma(x)dx \lesssim_p e^{A(q)c_4\epsilon^2}\left(1 + \exp\left((p+d)\log(r\sqrt{n}e^{\frac{1}{\alpha}}) - nc_0\right)\right).$$

We can now prove Lemma E.1.

Proof of Lemma E.1. Fix $\|x\| \geq 4\sqrt{d} \geq 1$. Using Lemma C.2, we have $|r(x)| \lesssim_q \bar{c}_3\epsilon\|x\|^{4+q}$ on the set $\mathcal{U}^c$, and therefore

$$\|x\|^p |r(x)|^{k+1} \lesssim_q (\bar{c}_3\epsilon)^{k+1}\|x\|^{p+(k+1)(4+q)}$$

on this set. Therefore using Corollary E.1 we have

$$\begin{aligned}
\int_{\mathcal{U}^c} \|x\|^p |r|^{k+1} e^{-r} d\gamma &\lesssim (\bar{c}_3\epsilon)^{k+1} \int_{\mathcal{U}^c} \|x\|^{p+(k+1)(4+q)} e^{-r} d\gamma \\
&\lesssim e^{A(q)c_4\epsilon^2}\left(1 + \exp\left((p+(k+1)(4+q)+d)\log(r\sqrt{n}e^{\frac{1}{\alpha}}) - nc_0\right)\right)(\bar{c}_3\epsilon)^{k+1} \\
&\leq e^{A(q)\bar{c}_3\epsilon}(1 + B_{p,k+1})(\bar{c}_3\epsilon)^{k+1} = K_{p,k+1}(\bar{c}_3\epsilon)^{k+1}, \qquad \text{(E.10)}
\end{aligned}$$

as desired. □

E.1 Auxiliary results for proof of Lemma E.1

Lemma E.4. *Assume* $(ab)^\alpha > (p+d)/\alpha$. *Then*

$$I := \frac{S_{d-1}}{(2\pi)^{d/2}} \int_a^\infty u^{p+d-1} e^{-(bu)^\alpha} du \lesssim \frac{a^{p+d}}{\alpha \Gamma(d/2) 2^{d/2}} e^{\frac{p+d}{\alpha} - (ab)^\alpha}. \tag{E.11}$$

Proof. First, let $u = s/b$ so that $u^{p+d-1}du = b^{-p-d}s^{p+d-1}ds$. Hence

$$I = \frac{S_{d-1} b^{-p-d}}{(2\pi)^{d/2}} \int_{ab}^\infty s^{p+d-1} e^{-s^\alpha} ds. \tag{E.12}$$

Next, let $s = t^{1/\alpha}$, so that

$$s^{p+d-1}ds = t^{(p+d-1)/\alpha} \frac{1}{\alpha} t^{\frac{1}{\alpha}-1} ds = \frac{1}{\alpha} t^{\frac{p+d}{\alpha}-1} dt.$$

Hence

$$\begin{aligned} I &= \frac{S_{d-1} b^{-p-d}}{\alpha (2\pi)^{d/2}} \int_{(ab)^\alpha}^\infty t^{\frac{p+d}{\alpha}-1} e^{-t} dt \\ &= \frac{S_{d-1} b^{-p-d} \Gamma((p+d)/\alpha)}{\alpha (2\pi)^{d/2}} \mathbb{P}(X \geq (ab)^\alpha), \end{aligned} \tag{E.13}$$

where $X \sim \Gamma((p+d)/\alpha, 1)$. Now, we show in Lemma E.5 that if $\lambda > c$ and $X \sim \Gamma(c,1)$, then $\Gamma(c)\mathbb{P}(X \geq \lambda) \leq e^{c-\lambda}\lambda^c$. Applying this result with $\lambda = (ab)^\alpha$ and $c = (p+d)/\alpha$, we get

$$I \lesssim \frac{S_{d-1} b^{-p-d} (ab)^{p+d}}{\alpha (2\pi)^{d/2}} e^{\frac{p+d}{\alpha} - (ab)^\alpha} = \frac{S_{d-1} a^{p+d}}{\alpha (2\pi)^{d/2}} e^{\frac{p+d}{\alpha} - (ab)^\alpha}. \tag{E.14}$$

To conclude, we substitute the formula $S_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$. □

Lemma E.5. *Let* $X \sim \Gamma(c,1)$. *Then for* $\lambda > c$, *we have*

$$\Gamma(c)\mathbb{P}(X \geq \lambda) \leq e^{c-\lambda}\lambda^c.$$

Proof. The mgf of $\Gamma(c,1)$ is $\mathbb{E}[e^{Xt}] = (1-t)^{-c}$, defined for $t < 1$. Hence for all $t \in (0,1)$ we have

$$\mathbb{P}(X \geq \lambda) \leq e^{-\lambda t}(1-t)^{-c} = f(t). \tag{E.15}$$

Now,

$$f'(t) = -\lambda f(t) + \frac{c}{1-t} f(t) = f(t)\left(\frac{c}{1-t} - \lambda\right),$$

and we find that $t = 1 - \frac{c}{\lambda}$ is the minimizer of f. Substituting this value of t into (E.15) gives

$$\mathbb{P}(X \geq \lambda) \leq e^{c-\lambda}(\lambda/c)^c.$$

Multiplying both sides by $\Gamma(c)$ and using that $\Gamma(c) \leq c^c$ gives the desired bound. □

F Proofs from Section 3: Logistic Regression

Proof of Lemma 3.1. First note that we can write v as

$$v(b) = \left(\frac{1}{n}\sum_{i=1}^{n} Y_i X_i\right)^T b - \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{b^T X_i}).$$

Also, let $v_\infty(b) = \mathbb{E}\left[v(b) \mid \{X_i\}_{i=1}^n\right]$, which is given by

$$v_\infty(b) = \left(\frac{1}{n}\sum_{i=1}^{n} s(\beta^T X_i) X_i\right)^T b - \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{b^T X_i}), \tag{F.1}$$

recalling that $\mathbb{E}\left[Y_i \mid X_i\right] = s(\beta^T X_i)$. Note also that $\nabla v_\infty(\beta) = 0$, and that

$$\nabla v(\beta) = \nabla v(\beta) - \nabla v_\infty(\beta) = \frac{1}{n}\sum_{i=1}^{n} (Y_i - s(\beta^T X_i)) X_i. \tag{F.2}$$

.

We will show that if d/n is small enough then

$$v(b) > v(\beta), \quad \forall \|b - \beta\| = 1 \tag{F.3}$$

with high probability. It then follows that $v(b') > v(\beta)$ for all $\|b' - \beta\| \geq 1$. Indeed, fix such a b' and let b be the point on the segment connecting β to b' that is distance 1 away from β. Then by convexity of v, we have

$$v(b') - v(\beta) \geq \frac{v(b) - v(\beta)}{\|b' - \beta\|} > 0.$$

This implies that the minimizer $\hat{\beta}$ of v lies inside the unit ball around β, i.e. $\|\hat{\beta} - \beta\| < 1$. To prove (F.3), Taylor expand v around β, and evaluate at a point b such that $\|b - \beta\| = 1$:

$$\begin{aligned} v(b) - v(\beta) =& \nabla v(\beta)^T (b - \beta) + \frac{1}{2}(b - \beta)^T \nabla^2 v(\xi)(b - \beta) \\ &\geq \left(\frac{1}{n}\sum_{i=1}^{n} (Y_i - s(\beta^T X_i)) X_i\right)^T (b - \beta) \\ &\quad + \frac{\|b - \beta\|^2}{2} \inf_{\|x - \beta\| \leq 1} \lambda_{\min}(\nabla^2 v(x)) \\ &\geq -\left\|\frac{1}{n}\sum_{i=1}^{n} (Y_i - s(\beta^T X_i)) X_i\right\| + \frac{1}{2} \inf_{\|x\| \leq 2} \lambda_{\min}(\nabla^2 v(x)). \end{aligned} \tag{F.4}$$

In the first line of (F.4), ξ is a point on the interval between b and β. In the second line, we used equation (F.2) for $\nabla v(\beta)$. In the third line, we used that $\|b - \beta\| = 1$, and that $\|\beta\| = 1$, so $\{\|x - \beta\| \leq 1\} \subset \{\|x\| \leq 2\}$.

Now, we apply Lemma 7 of Chapter 3 of the PhD thesis [Sur, 2019]. (This result also appears in Lemma 4 of the paper [Sur et al., 2019], but the form in which it appears in the thesis is closer to our setting.) The lemma states that if $d/n < C < 1$ for some absolute constant C, then there exist $C_1, C_2, C_3 > 0$ such that the event

$$E_1 = \left\{ \inf_{\|b\| \le r} \lambda_{\min}(\nabla^2 v(b)) \ge C_1 s'\left(C_2 r\right) \quad \forall r \ge 0. \right\}$$

has probability at least $1 - 4e^{-C_3 n}$. (In fact, the lemma proves an even stronger statement than this). Furthermore, if $d/n < 1/(4\log 3)$ then by Lemma F.1

$$E_2 = \left\{ \left\| \frac{1}{n} \sum_{i=1}^n (Y_i - s(\beta^T X_i)) X_i \right\| \le 4\sqrt{d/n} \right\} \tag{F.5}$$

has probability at least $1 - e^{-d/2} - e^{-n/4}$. Therefore,

$$\mathbb{P}(E_1 \cap E_2) \ge 1 - e^{-d/2} - 5e^{-(C_3 \wedge 0.25)n} = 1 - e^{-d/2} - 5e^{-A_1 n}$$

(where $A_1 = C_3 \wedge 0.25$) and on $E_1 \cap E_2$ we have that

$$\begin{aligned} v(b) - v(\beta) &\ge -\left\| \frac{1}{n} \sum_{i=1}^n (Y_i - s(\beta^T X_i)) X_i \right\| + \frac{1}{2} \inf_{\|x\| \le 2} \lambda_{\min}(\nabla^2 v(x)) \\ &\ge -4\sqrt{d/n} + \frac{C_1}{2} s'(2C_2) \end{aligned} \tag{F.6}$$

If $d/n < \min\left(C, (4\log 3)^{-1}, \left(\frac{C_1}{8} s'(2C_2)\right)^2\right) =: A_0$ then this lower bound on $v(b) - v(\beta)$ is positive. (Recall that $d/n < C, d/n < 1/(4\log 3)$ is necessary to apply the aforementioned lemmas). We conclude that $\|\hat{\beta} - \beta\| \le 1$ on $E_1 \cap E_2$. But then we also have on $E_1 \cap E_2$ that

$$\lambda_{\min}(\nabla^2 v(\hat{\beta})) \ge \inf_{\|b\| \le 2} \lambda_{\min}(\nabla^2 v(b)) \ge C_1 s'\left(2C_2\right) =: A_2.$$

We conclude that

$$E_1 \cap E_2 \subseteq \{\|\hat{\beta} - \beta\| \le 1, \quad \lambda_{\min}(\nabla^2 v(\hat{\beta})) \ge A_2\},$$

and hence the righthand event has probability at least $1 - e^{-d/2} - 5e^{-A_1 n}$ as well. This concludes the proof. □

Proof of Lemma 3.2. We have

$$\nabla^k v(b) = \frac{1}{n} \sum_{i=1}^n s^{(k-1)}(b^T x_i) x_i^{\otimes k}. \tag{F.7}$$

This is a symmetric tensor, so Theorem 2.1 of [Zhang et al., 2012] implies that $\|\nabla^k v(b)\| = \sup_{\|u\|=1} \langle \nabla^k v(b), u^{\otimes k} \rangle$, i.e. it suffices to consider the action of

$\nabla^k v(b)$ on the product of k copies of the same unit vector, rather than k arbitrary unit norm vectors. Now, using that $|s^{(k-1)}|$ is bounded uniformly over $\mathbb{R}$, we have

$$\|\nabla^k v(b)\| \leq \|s^{(k-1)}\|_\infty \sup_{\|u\|=1} \frac{1}{n} \sum_{i=1}^n |X_i^T u|^k. \tag{F.8}$$

We now apply Proposition 4.4 of [Adamczak et al., 2010], which implies that there are absolute constants B_0', B_1' such that the events

$$F_k = \left\{ \sup_{\|u\|=1} \frac{1}{n} \sum_{i=1}^n |X_i^T u|^k \leq B_0' \left(1 + \frac{d^{k/2}}{n}\right) \right\} \tag{F.9}$$

satisfy $\mathbb{P}(F_k) \geq 1 - 2\exp(-B_1'\sqrt{nd}/\log(2n/d))$ for all $k = 2, 3, 4, 5$. This result and (F.8) imply that (3.4) is satisfied with $B_0 = \max_{k=2,3,4,5} \|s^{(k-1)}\|_\infty B_0'$ and $B_1 = B_1'$. □

For the next proof, we return to the notation of the proof of Lemma 3.1. We note that that proof implies

$$\begin{aligned} E_1 \cap E_2 &\subseteq E_1 \cap \{\|\hat{\beta} - \beta\| \leq 1\} \\ &= \{\|\hat{\beta} - \beta\| \leq 1, \ \inf_{\|b\| \leq r} \lambda_{\min}(\nabla^2 v(b)) \geq A(r) \ \forall r \geq 0\} =: E_3, \end{aligned} \tag{F.10}$$

where $A(r) = C_1 s'(C_2 r)$. In particular, $\lambda_{\min}(H_v) = \lambda_{\min}(\nabla^2 v(\hat{\beta})) \geq A(2)$ on E_3. (The lower bound $A(2)$ was denoted A_2 in the above proof.) Note also that $\mathbb{P}(E_3) \geq \mathbb{P}(E_1 \cap E_2) \geq 1 - e^{-d/2} - 5e^{-A_1 n}$.

Proof of Lemma 3.3. In this proof we write H for H_v, for brevity. Define

$$f(y) = v(\hat{\beta} + H^{-1/2} y) - v(\hat{\beta}),$$

and let $k = \inf_{\|y\|=r} f(y)$. First, we will show that $f(y) \geq \kappa \|y/r\|$ for all $\|y\| \geq r$. Since v is convex, so is f. Now, fix y such that $\|y\| \geq r$, and let $u = ry/\|y\|$ be the point on the line segment connecting 0 to u which has norm r. By convexity of f, it follows that

$$\begin{aligned} f(y) = f(y) - f(0) &\geq \frac{\|y\|}{r}(f(u) - f(0)) = \|y/r\| f(u) \\ &\geq \|y/r\| \left[\inf_{\|u\|=r} f(u)\right] = \kappa \|y/r\|, \end{aligned} \tag{F.11}$$

as desired. Now, fix x such that $\|x\|_H \geq r$. Let $x = H^{-1/2} y$ and note that $\|y\| = \|x\|_H \geq r$. Using the definition of f, we have

$$v(\hat{\beta} + x) - v(\hat{\beta}) = v(\hat{\beta} + H^{-1/2} y) - v(\hat{\beta}) = f(y) \geq \kappa \|y/r\| = \kappa \|x/r\|_H.$$

To finish the proof, we show that on a subset of the event E_3 defined in (F.10), there exists an absolute constant C_0 such that $\kappa \geq C_0$. Recall from the remarks following the definition of E_3 that $\lambda_{\min}(H) \geq A(2)$ on E_3.

Fix y such that $\|y\| = r$ and Taylor expand f around zero. Then for some $t \in [0,1]$, we have on the event E_3 that

$$\begin{aligned} f(y) = f(y) - f(0) &= \frac{1}{2} y^T H^{-1/2} \nabla^2 v \left(\hat{\beta} + t H^{-1/2} y\right) H^{-1/2} y \\ &\geq \frac{1}{2} \inf_{\|y\|=r, t\in[0,1]} \lambda_{\min}\left(\nabla^2 v(\hat{\beta} + t H^{-1/2} y)\right) \|H^{-1/2} y\|^2 \\ &\geq \frac{1}{2} \inf_{\|u\| \leq \|\hat{\beta}\| + \|H^{-1/2}\| r} \lambda_{\min}\left(\nabla^2 v(u)\right) \frac{r^2}{\|H\|} \\ &\geq \frac{1}{2} \inf_{\|u\| \leq 2 + r/\sqrt{A(2)}} \lambda_{\min}\left(\nabla^2 v(u)\right) \frac{r^2}{\|H\|} \\ &\geq \frac{1}{2} A(2 + r/\sqrt{A(2)}) \frac{r^2}{\|H\|}. \end{aligned} \tag{F.12}$$

Therefore,

$$\kappa = \inf_{\|y\|=r} f(y) \geq \frac{1}{2} A(2 + r/\sqrt{A(2)}) \frac{r^2}{\|H\|} \tag{F.13}$$

on E_3. To conclude the proof, we recall from Lemma 3.2 that

$$E_4 = \{\sup_{b \in \mathbb{R}^d} \|\nabla^2 v(b)\| \leq B_0(1 + d/n) \leq 2B_0\}$$

has probability at least $1 - e^{-B_1\sqrt{nd}/\log(2n/d)}$. Thus

$$\begin{aligned} \mathbb{P}(E_3 \cap E_4) &\geq 1 - 2\exp(-B_1\sqrt{nd}/\log(2n/d)) - e^{-d/2} - 5e^{-A_1 n} \\ &\geq 1 - 7\exp(-C_1\sqrt{nd}/\log(2n/d)) - e^{-d/2} \end{aligned} \tag{F.14}$$

for some C_1. On $E_3 \cap E_4$ we have

$$\kappa \geq \frac{1}{2} A(2 + r/\sqrt{A(2)}) \frac{r^2}{\|H\|} \geq \frac{1}{2} A(2 + r/\sqrt{A(2)}) \frac{r^2}{2B_0} =: C_0.$$

This concludes the proof. □

For the next two proofs, recall that $a_{k,p} = \mathbb{E}[s^{(k)}(Z_1) Z_1^p]$ for $Z_1 \sim \mathcal{N}(0,1)$. Also, we use V to denote $\bar{V}_\infty$ for brevity.

Proof of Lemma 3.4. Using (3.8), we have

$$\begin{aligned} H_V = \nabla^2 V(\beta) &= \mathbb{E}[s'(Z_1) Z Z^T] \\ &= n\mathrm{diag}(a_{1,2}, a_{1,0} \ldots, a_{1,0}), \end{aligned} \tag{F.15}$$

and

$$\nabla^3 V(\beta) = n\mathbb{E}[s''(Z_1) Z^{\otimes 3}].$$

Now, for a fixed vector $b \in \mathbb{R}^d$, we compute

$$\begin{aligned}\langle \nabla^3 V(\beta), b^{\otimes 3}\rangle &= n\mathbb{E}\left[s''(Z_1)(b^T Z)^3\right] \\ &= n\mathbb{E}\left[s''(Z_1)(b_1 Z_1)^3\right] + 3n\mathbb{E}\left[s''(Z_1)Z_1\right]\mathbb{E}\left[(b_{2:d}^T Z_{2:d})^2\right] \\ &= n\left(a_{2,3}b_1^3 + 3a_{2,1}b_1\|b_{2:d}\|^2\right).\end{aligned} \tag{F.16}$$

Hence

$$|\langle \nabla^3 V(\beta), b^{\otimes 3}\rangle| \geq n\left(3|a_{2,1}||b_1|\|b_{2:d}\|^2 - |a_{2,3}||b_1|^3\right). \tag{F.17}$$

Now substitute

$$b = H_V^{-1/2} Z = n^{-1/2}\left(a_{1,2}^{-1/2} Z_1, a_{1,0}^{-1/2} Z_2, \ldots, a_{1,0}^{-1/2} Z_d\right)$$

into (F.17) and take expectations on both sides:

$$\begin{aligned}L &= \mathbb{E}\left|\langle \nabla^3 V(\beta), (H_V^{-1/2} Z)^{\otimes 3}\rangle\right| \\ &\geq 3\frac{(d-1)}{\sqrt{n}}\frac{|a_{2,1}|}{a_{1,2}^{1/2} a_{1,0}}\mathbb{E}\left[|Z_1||Z_2|^2\right] - \frac{|a_{2,3}|}{a_{1,2}^{3/2}}\frac{\mathbb{E}\left[|Z_1|^3\right]}{\sqrt{n}} \\ &\geq \frac{2}{a_{1,2}^{1/2}\sqrt{n}}\left(\frac{|a_{2,1}|}{a_{1,0}}(d-1) - \frac{2|a_{2,3}|}{a_{1,2}}\right),\end{aligned} \tag{F.18}$$

as desired. □

Proof of Lemma 3.5. Using the formulas for H_V and $\nabla^3 V(\beta)$ from the proof of Lemma 3.4, we have

$$\begin{aligned}L &= -\frac{1}{2}H_V^{-1/2}\langle \nabla^3 V(\beta), H_V^{-1}\rangle \\ &= -\frac{n}{2}\mathbb{E}\left[s''(Z_1) Z^T H_V^{-1} Z H_V^{-1/2} Z\right] \\ &= -\frac{1}{2}\mathbb{E}\left[s''(Z_1)\left(a_{1,2}^{-1}Z_1^2 + \sum_{j=2}^d a_{1,0}^{-1}Z_j^2\right) H_V^{-1/2} Z\right] \\ &= -\frac{1}{2a_{1,2}}\mathbb{E}\left[s''(Z_1) Z_1^2 H_V^{-1/2} Z\right] - \frac{1}{2a_{1,0}}\sum_{j=2}^d \mathbb{E}\left[s''(Z_1) Z_j^2 H_{V_\infty}^{-1/2} Z\right] \\ &= -\frac{a_{2,3}}{2a_{1,2}^{3/2}\sqrt{n}}e_1 - \frac{(d-1)a_{2,1}}{2a_{1,0}a_{1,2}^{1/2}\sqrt{n}}e_1 \\ &= -\frac{1}{2\sqrt{n}a_{1,2}^{1/2}}\left(\frac{a_{2,3}}{a_{1,2}} + (d-1)\frac{a_{2,1}}{a_{1,0}}\right)e_1.\end{aligned} \tag{F.19}$$

Taking the norm of this vector gives the result. □

Lemma F.1. *Let* $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, $i = 1, \dots, n$, *and* $Y_i \mid X_i \sim \text{Bernoulli}(s(X_{i1}))$, *where* $s(t) = (1 + e^{-t})^{-1}$ *is the sigmoid and* X_{i1} *is the first coordinate of* X_i. *If* $d/n < 1/(4 \log 3)$ *then*

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} (Y_i - s(X_{i1}))X_i \right\| \geq 4\sqrt{d/n} \right) \leq e^{-n/4} + e^{-d/2}. \tag{F.20}$$

Proof. Let $A = \frac{1}{n} \sum_{i=1}^{n} (Y_i - s(X_{i1}))X_i$, and let $\mathcal{N}$ be a 1/2-net of the sphere S^{d-1}. Then

$$\|A\| \leq 2 \sup_{u \in \mathcal{N}} u^T A,$$

and hence

$$\mathbb{P}(\|A\| \geq t) \leq \mathbb{P}(\sup_{u \in \mathcal{N}} u^T A \geq t/2) \leq 3^d \sup_{\|u\|=1} \mathbb{P}(u^T A \geq t/2),$$

where we have used a union bound and the fact that $|\mathcal{N}| \leq 3^d$. Now, we have

$$\begin{aligned}
\mathbb{P}(u^T A \geq t/2) &= \mathbb{E}\left[\mathbb{P}\left(u^T A \geq t/2 \mid \{X_i\}_{i=1}^n \right) \right] \\
&= \mathbb{E}\left[\mathbb{P}\left(\sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i \mid X_i]) u^T X_i \geq nt/2 \mid \{X_i\}_{i=1}^n \right) \right] \\
&\leq \mathbb{E}\left[\exp\left(-\frac{2n^2(t/2)^2}{\sum_{i=1}^{n} (u^T X_i)^2} \right) \right] = \mathbb{E}\left[\exp\left(-\frac{n^2 t^2}{2\|Z_n\|^2} \right) \right],
\end{aligned} \tag{F.21}$$

where $Z_n \sim \mathcal{N}(0, I_n)$. We have used Hoeffding's inequality in the third line, and the fact that $(u^T X_i)_{i=1}^n$ is a standard normal on $\mathbb{R}^n$. Now, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-\frac{n^2 t^2}{2\|Z_n\|^2} \right) \right] &\leq \mathbb{P}(\|Z_n\| \geq 2\sqrt{n}) + \exp\left(-\frac{n^2 t^2}{2(2\sqrt{n})^2} \right) \\
&\leq e^{-n/2} + e^{-nt^2/8}.
\end{aligned} \tag{F.22}$$

Taking $t = 4\sqrt{d/n}$ and combining all of the preceding bounds, we conclude that

$$\begin{aligned}
\mathbb{P}(\|A\| \geq 4\sqrt{d/n}) &\leq e^{d \log 3 - n/2} + e^{d(\log 3 - 2)} \\
&\leq e^{-n/4} + e^{-d/2},
\end{aligned} \tag{F.23}$$

using that $d/n < 1/4 \log 3$ to get the last line. □

G Matrix-weighted operator norm of a tensor

In Theorem 2.1 of [Zhang et al., 2012] it is shown that

$$\sup_{\|x_1\| = \cdots = \|x_m\| = 1} \langle T, x_1 \otimes \cdots \otimes x_m \rangle = \sup_{\|x\|=1} \langle T, x^{\otimes m} \rangle, \quad \forall T \in \text{Sym}^m(\mathbb{R}^n), \tag{G.1}$$

where $\mathrm{Sym}^m(\mathbb{R}^n)$ is the set of order m, dimension n symmetric tensors. We now show the same is true for the H-weighted tensor operator norm, i.e. we claim that

$$\|T\|_H := \sup_{\|x_1\|_H=\cdots=\|x_m\|_H=1} \langle T, x_1 \otimes \cdots \otimes x_m \rangle = \sup_{\|x\|_H=1} \langle T, x^{\otimes m} \rangle. \tag{G.2}$$

Here, $\|x\|_H = \sqrt{x^T H x}$ for a symmetric matrix H. To prove (G.2), we start by noting the following identity.

Lemma G.1. *Given a symmetric order m, dimension d tensor T and a $d \times d$ matrix A, define the tensor $S(T, A)$ by*

$$S(T, A)_{i_1 \ldots i_m} = \sum_{j_1,\ldots,j_m=1}^{d} T_{j_1 \ldots j_m} A_{i_1 j_1} \ldots A_{i_m j_m}. \tag{G.3}$$

Then

$$\langle T, (Ay_1) \otimes \cdots \otimes (Ay_m) \rangle = \langle S(T, A), y_1 \otimes \cdots \otimes y_m \rangle \tag{G.4}$$

for all $y_1, \ldots, y_m \in \mathbb{R}^d$.

Proof. We have

$$\begin{aligned}
\langle T, (Ay_1) \otimes \cdots \otimes (Ay_m) \rangle &= \sum_{j_1,\ldots,j_m=1}^{d} T_{j_1 \ldots j_m} (Ay_1)_{j_1} \ldots (Ay_m)_{j_m} \\
&= \sum_{i_1,\ldots,i_m=1}^{d} \sum_{j_1,\ldots,j_m=1}^{d} T_{j_1 \ldots j_m} A_{j_1 i_1} (y_1)_{i_1} \ldots A_{j_m i_m} (y_m)_{i_m} \\
&= \sum_{i_1,\ldots,i_m=1}^{d} (y_1)_{i_1} \ldots (y_m)_{i_m} \sum_{j_1,\ldots,j_m=1}^{d} T_{j_1 \ldots j_m} A_{j_1 i_1} \ldots A_{j_m i_m} \\
&= \sum_{i_1,\ldots,i_m=1}^{d} (y_1)_{i_1} \ldots (y_m)_{i_m} S_{i_1 \ldots i_m} \\
&= \langle S(T, A), y_1 \otimes \cdots \otimes y_m \rangle
\end{aligned} \tag{G.5}$$

□

We now prove (G.2). First let $S(T, H^{-1/2})$ be defined as in Lemma G.1, and

note that this is a symmetric tensor, by the symmetry of T. Therefore,

$$\begin{aligned}
\|T\|_H &= \sup_{\|x_1\|_H=\cdots=\|x_m\|_H=1} \langle T, x_1 \otimes \cdots \otimes x_m \rangle \\
&= \sup_{\|y_1\|=\cdots=\|y_m\|=1} \langle T, (H^{-1/2} y_1) \otimes \cdots \otimes (H^{-1/2} y_m) \rangle \\
&= \sup_{\|y_1\|=\cdots=\|y_m\|=1} \langle S(T, H^{-1/2}), y_1 \otimes \cdots \otimes y_m \rangle \\
&= \sup_{\|y\|=1} \langle S(T, H^{-1/2}), y^{\otimes m} \rangle \\
&= \sup_{\|y\|=1} \langle T, (H^{-1/2} y)^{\otimes m} \rangle = \sup_{\|x\|_H=1} \langle T, x^{\otimes m} \rangle,
\end{aligned} \tag{G.6}$$

as desired. The third line used Lemma G.1, the fourth line uses (G.1) (since $S(T, H^{-1/2})$ is symmetric), and the fifth line again uses Lemma G.1.

References

[Adamczak et al., 2010] Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2010). Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561.

[Bakry et al., 2014] Bakry, D., Gentil, I., Ledoux, M., et al. (2014). *Analysis and geometry of Markov diffusion operators*, volume 103. Springer.

[Boucheron and Gassiat, 2009] Boucheron, S. and Gassiat, E. (2009). A Bernstein-Von Mises Theorem for discrete probability distributions. *Electronic Journal of Statistics*, 3(none):114 – 148.

[Candès and Sur, 2020] Candès, E. J. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27 – 42.

[Canonne, 2019] Canonne, C. (2019). A short note on poisson tail bounds. http://www.cs.columbia.edu/ccanonne/files/misc/2017-poissonconcentration.pdf.

[Dehaene, 2019] Dehaene, G. P. (2019). A deterministic and computable bernstein-von mises theorem. *arXiv preprint arXiv:1904.02505*.

[Diao et al., 2023] Diao, M., Balasubramanian, K., Chewi, S., and Salim, A. (2023). Forward-backward gaussian variational inference via jko in the bures-wasserstein space. *arXiv preprint arXiv:2304.05398*.

[Durante et al., 2023] Durante, D., Pozza, F., and Szabo, B. (2023). Skewed bernstein-von mises theorem and skew-modal approximations. *arXiv preprint arXiv:2301.03038*.

[Ghosal, 1999] Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, pages 315–331.

[Ghosal, 2000] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis*, 74(1):49–68.

[He and Shao, 2000] He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135.

[Helin and Kretschmann, 2022] Helin, T. and Kretschmann, R. (2022). Non-asymptotic error estimates for the laplace approximation in bayesian inverse problems. *Numerische Mathematik*, 150(2):521–549.

[Kasprzak et al., 2022] 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. *arXiv preprint arXiv:2209.14992*.

[Katsevich and Rigollet, 2023] Katsevich, A. and Rigollet, P. (2023). On the approximation accuracy of gaussian variational inference. *arXiv preprint arXiv:2301.02168*.

[Lambert et al., 2022] Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2022). Variational inference via wasserstein gradient flows. *arXiv preprint arXiv:2205.15902*.

[Lu, 2017] Lu, Y. (2017). On the bernstein-von mises theorem for high dimensional nonlinear bayesian inverse problems. *arXiv preprint arXiv:1706.00289*.

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

[Portnoy, 1988] Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, pages 356–366.

[Schudy and Sviridenko, 2012] Schudy, W. and Sviridenko, M. (2012). Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. SIAM.

[Shun and McCullagh, 1995] Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):749–760.

[Spokoiny, 2012] Spokoiny, V. (2012). Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909.

[Spokoiny, 2013] Spokoiny, V. (2013). Bernstein-von mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*.

[Spokoiny, 2017] Spokoiny, V. (2017). Penalized maximum likelihood estimation and effective dimension. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):389 – 429.

[Spokoiny, 2022] Spokoiny, V. (2022). Dimension free non-asymptotic bounds on the accuracy of high dimensional laplace approximation. *arXiv preprint arXiv:2204.11038*.

[Sur, 2019] Sur, P. (2019). *A modern maximum-likelihood theory for high-dimensional logistic regression*. PhD thesis, Stanford University.

[Sur and Candès, 2019] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

[Sur et al., 2019] Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175:487–558.

[Zhang et al., 2012] Zhang, X., Ling, C., and Qi, L. (2012). The best rank-1 approximation of a symmetric tensor and related spherical optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 33(3):806–821.