Journal of Machine Learning Research 23 (2022) 1-56 Submitted 6/21; Revised 9/22; Published 10/22

Variational Inference in high-dimensional linear regression

Sumit Mukherjee SM3949@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Subhabrata Sen SUBHABRATASEN@FAS.HARVARD.EDU
Department of Statistics
Harvard University
Cambridge, MA 02138, USA

Editor: Pierre Alquier

Abstract

We study high-dimensional bayesian linear regression with product priors. Using the nascent theory of *non-linear large deviations* (Chatterjee and Dembo, 2016), we derive sufficient conditions for the leading-order correctness of the naive mean-field approximation to the log-normalizing constant of the posterior distribution. Subsequently, assuming a true linear model for the observed data, we derive a limiting infinite dimensional variational formula for the log normalizing constant for the posterior. Furthermore, we establish that under an additional "separation" condition, the variational problem has a unique optimizer, and this optimizer governs the probabilistic properties of the posterior distribution. We provide intuitive sufficient conditions for the validity of this "separation" condition. Finally, we illustrate our results on concrete examples with specific design matrices.

Keywords: Variational Inference, Linear regression, Naive Mean-Field Approximation

1. Introduction

In this age of big-data, statisticians routinely analyze large, high-dimensional datasets arising from applications in genomics, finance, public policy etc., with the goal of discovering relationships between the response variable, and the observed features. The linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

is arguably the most common framework for this task when the response is continuous. Under a Bayesian formalism, the statistician posits a prior distribution π for the coefficient vector $\boldsymbol{\beta}$, and constructs the corresponding posterior. Subsequent inference is based solely on this posterior distribution. Two questions are naturally relevant in this setting:

1. What are the statistical properties of Bayesian procedures, particularly in high-dimensions?

2. Are these procedures computationally tractable for large datasets with high-dimensional features?

The theoretical performance of high-dimensional Bayesian methods have been examined extensively in recent times. In Bayesian asymptotic theory, given data $(y_i, \mathbf{x}_i)_{i=1}^n$, $x_i \in \mathbb{R}^p$, one assumes the correctness of a frequentist model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon$, and studies frequentist inference of the coefficient vector β_0 under Bayesian procedures. Ghosal (1999) established a version of the traditional Bernstein-Von-Mises theorem as long as $p^4 \log p/n \to 0$. A lot of recent attention has been focused on the high-dimensional regime $p \gg n$ with an additional assumption on the sparsity of $\boldsymbol{\beta}_0$. In this context, spike-and-slab based approaches (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005) have been established to exhibit optimal frequentist properties (Castillo et al., 2015). In particular, these posterior distributions "contract" to the underlying truth $\boldsymbol{\beta}_0$ at the minimax optimal estimation rate (we refer the interested reader to Banerjee et al. (2021) for a formal definition of posterior contraction, and a survey of recent breakthroughs in this area). Despite the superior theoretical properties of this approach, the sharp spike-and-slab based approaches suffer from a computational bottleneck. In the special case of linear regression, continuous shrinkage priors (see e.g. Carvalho et al. (2010)) provide a comparatively more tractable alternative from the computational perspective (Bhattacharya et al., 2016). Contraction properties of the corresponding posteriors were characterized in Song and Liang (2017). Unfortunately, this strategy is specific to the linear model, and does not generalize beyond this setting.

In general, computationally tractable Bayesian inference for high-dimensional models presents significant challenges. MCMC based strategies have been explored extensively for this purpose. Despite rapid progress in MCMC methodology and supporting theory, these methods are still slower than competing frequentist methodology e.g. those based on convex optimization. Variational methods (Wainwright and Jordan, 2008) provide an attractive general option to the Bayesian statistician. The simplest form of variational inference approximates the true posterior distribution using a product distribution—this version is often referred to as *naive mean-field Variational Bayes* (nVB). Computing the best approximating product distribution is computationally fast, and thus these methods provide a practical option for modern datasets. We refer the interested reader to Blei et al. (2017) for an introduction to variational Bayes methods in statistics and machine learning.

Although variational methods provide a computationally feasible strategy, supporting theoretical evidence has been relatively scarce. In Wang and Titterington (2006); Wang and Blei (2019b), the authors established the correctness of this approach in parametric models in the classical fixed p setting. Subsequently Wang and Blei (2019a) study variational Bayes in misspecified models. In the context of the linear model, early work by Neville et al. (2014); Ormerod et al. (2017) focused on variational inference in the low dimensional linear model, while Carbonetto and Stephens (2012) provides an early variational approximation for variable selection.

The past two years have witnessed rapid progress in the analysis of variational methods for high-dimensional models (Alquier et al., 2016; Alquier and Ridgway, 2020; Han and Yang, 2019; Ray and Szabó, 2021; Ray et al., 2020; Yang et al., 2020). These results focus on high-dimensional models with a sparse underlying truth $\boldsymbol{\beta}_0$, and study the contraction properties of the variational posterior. In particular, they derive sufficient conditions for the variational posterior to contract at the minimax optimal rate. Correctness of variational methods have also been established for community detection (Bickel et al., 2013; Zhang and

Zhou, 2020), a poisson mixed model (Hall et al., 2011), frequentist models (Westling and McCormick, 2019) and mixture models (Chérief-Abdellatif and Alquier, 2018).

In sharp contrast, nVB methods can fail in truly high-dimensional settings, e.g. in versions of topic modeling (Ghorbani et al., 2019). In this specific situation, the correct variational approximation is provided by the TAP approximation from spin-glass theory (Mézard et al., 1987). Fan et al. (2018) establishes rigorous guarantees regarding the validity of the TAP approach in the context of the simpler $\mathbb{Z}_2$ synchronization problem. Evidence regarding the inadequacy of the nVB approximation also arises in the work of Fasano et al. (2019). The authors analyze the nVB approximation for probit regression in the regime n fixed and $p \to \infty$. In this regime, the variational posterior exhibits an *over-shrinking* phenomenon; to remedy this issue, the authors propose an alternative, locally factored approximation, which resolves this over-shrinking problem in this setting.

There is thus an immediate need to understand general properties of statistical problems which ensure the correctness of the nVB approximation. In this paper, we study the accuracy of this approximation in Bayesian linear regression with product priors. We provide easily verifiable conditions which ensure the asymptotic accuracy of the nVB approximation. Further, we illustrate that the approximation can yield detailed information regarding the statistical properties of this model, which might be unavailable using other techniques. We elaborate on our specific contributions below.

1.1 Contributions

Our main contributions in this paper are as follows:

(i) In Theorem 7, we provide sufficient conditions for the asymptotic tightness of the nVB approximation. The conditions are easily verifiable, and can be explicitly checked in specific applications. This provides rigorous theoretical support for widely used mean-field approximation based methodology.

(ii) Assuming a true frequentist linear model for the data, we derive a limiting variational formula for the log normalizing constant in Theorem 17. We emphasize that in contrast to existing results, we do not assume any sparsity on the true regression coefficients. We refer the interested reader to Lee et al. (2020) for a motivating discussion on the importance of such scenarios in scientific applications.

(iii) Under an additional "separation" condition (11), we establish that the limiting variational problem has a unique optimizer (see Theorem 20 for a precise statement). Under this separation condition, empirical averages of samples drawn from the posterior distribution concentrate around explicit deterministic limits; in addition, these limits are characterized by the optimizer (Corollary 22). We also provide interpretable sufficient conditions which enforce this separation condition (Lemma 25).

We emphasize that in existing analyses of high-dimensional Bayesian linear regression, properties of the posterior are directly established (Castillo et al., 2015), and the variational posterior is analyzed independently (Ray and Szabó, 2021). Our approach is inherently different—we first establish the correctness of the mean-field approximation, and then study the posterior through the lens of the mean-field approximation formula.

(iv) We further illustrate our general results by applying them to three specific examples—a two factor ANOVA model, a gaussian design setting with spiked covariance, and a sparse bernoulli design. In each case, we identify the specific limiting functional which determines the limiting log normalizing constant.

(v) From a theoretical perspective, our results crucially utilize recent advances in the theory of *non-linear large deviations*. Initiated in the seminal paper Chatterjee and Dembo (2016), the theory of non-linear large deviations was originally conceived to answer some deep questions concerning large deviations of sub-graph counts in sparse random graphs. In Basak and Mukherjee (2017), one of the authors successfully utilized this framework to establish the tightness of the naive mean field approximation for the log-partition function in a family of Potts models. To the best of our knowledge, these developments have not been utilized previously for other statistical models. In this paper, we demonstrate the usefulness of these tools in the context of high dimensional statistics; we hope that this spurs an in-depth study of their applicability for other high dimensional problems. We consider this to be a key conceptual contribution of this paper.

(vi) Our results are intimately related to *local asymptotic* scaling regimes from classical statistics (Van der Vaart, 2000). In fact, for high-dimensional linear regression with iid gaussian design, our results can be viewed as natural extensions of classical local-asymptotic results under the iterated limit $p \to \infty$ following $n \to \infty$. We elaborate on this connection in Section 3.

1.2 Non-linear large deviations and related results

In a breakthrough paper, Chatterjee and Dembo (Chatterjee and Dembo, 2016) introduced the theory of *non-linear large deviations* with the goal of studying large deviations for non-linear functions of bernoulli random variables. As an application of this general machinery, they characterized sharp deviation probabilities for sub-graph counts in sparse Erdős-Rényi random graphs. Subsequent extensions by Eldan (Eldan, 2018) and Augeri (Augeri, 2019) allow one to track a wider regime of sparsity for the binary variables. In a different direction, Yan (Yan, 2020) extended the Chatterjee-Dembo framework to general bounded Banach-space valued variables. See also Austin (2019) for related decompositions of general Gibbs measures using information theoretic ideas. These results have galvanized the study of large deviations for sub-graph counts on sparse random graphs, and the past three years have witnessed rapid progress in this direction. We refer the interested reader to Cook et al. (2021) and references therein for a survey of recent progress in this area.

At the heart of the Chatterjee-Dembo framework lies a tight approximation bound for the log-normalizing constant for general Gibbs type distributions in terms of the naive mean-field approximation formula. This framework was utilized by one of the authors (Basak and Mukherjee, 2017) to derive asymptotic limits for the log normalizing constant of Potts models on several sequences of graphs. Similar results were independently derived by (Jain et al., 2018, 2019) using different techniques.

In this paper, we study Bayesian linear regression through the lens of non-linear large deviations, and uncover precise statistical properties of these models by analyzing the mean-field variational problem.

1.3 Setup

We observe $\{(y_i, x_i) : 1 \le i \le n\}$, $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$. We set $\mathbf{y} = (y_i) \in \mathbb{R}^n$ and $\mathbf{X}^{\mathrm{T}} = [x_1, \cdots, x_n]$. Throughout, we work in an asymptotic setting where p and $n = n(p)$ are both going to ∞. A natural Bayesian model for such a dataset assumes

$$\beta_1, \cdots, \beta_p \sim^{iid} \pi, \quad \boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \cdots, \beta_p), \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \ \ \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \tag{1}$$

In the display above, π is a probability distribution supported on $[-1, 1]$, and we consider an iid prior on the regression coefficients. For our subsequent discussion, the precise interval $[-1, 1]$ for the prior support is not crucial— our arguments go through unchanged, as long as the prior has bounded support. Note that in the context of Bayesian inference for linear regression, the regression parameters are often drawn from a parametric family, and the parameters specifying the prior are, in turn, sampled from a hyper-prior. In our discussions, we will restrict ourselves to the simpler setting where the prior π is fixed—however, we do not make any assumptions on π apart from the bounded support assumption. Throughout, we assume that the noise variance $\sigma^2 > 0$ is fixed and known to the statistician.

Given the Bayesian model (1), one naturally constructs the posterior distribution

$$\frac{\mathrm{d}\mu_{\mathbf{y},\mathbf{X}}}{\mathrm{d}\pi^{\otimes p}}(\boldsymbol{\beta}) \propto \exp\Big(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\Big) \propto \exp\Big(-\frac{1}{2\sigma^2}\Big[\boldsymbol{\beta}^{\mathrm{T}} A_p \boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}} D_p \boldsymbol{\beta} - 2\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}\Big]\Big),$$

where $\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$, and D_p, A_p are the diagonal and off-diagonal matrices obtained from $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. More precisely,

$$\begin{aligned} A_p(i,i) &:= 0, & D_p(i,i) &:= \Big(\mathbf{X}^{\mathrm{T}}\mathbf{X}\Big)_{ii}, \\ A_p(i,j) &:= \Big(\mathbf{X}^{\mathrm{T}}\mathbf{X}\Big)_{ij}, & D_p(i,j) &:= 0. \end{aligned}$$

Since the term $\boldsymbol{\beta}^{\mathrm{T}} D\boldsymbol{\beta}$ is additive in the components of $\boldsymbol{\beta}$, we can absorb the term $\boldsymbol{\beta}^{\mathrm{T}} D\boldsymbol{\beta}$ in the base measure $\pi^{\otimes p}$, via the following definition:

Definition 1 (Exponential Family) *For any $\gamma := (\gamma_1, \gamma_2) \in \mathbb{R}^2$ define a probability measure π_γ on $[-1, 1]$ as*

$$\frac{\mathrm{d}\pi_\gamma}{\mathrm{d}\pi}(z) := \exp\Big(\gamma_1 z - \frac{\gamma_2}{2}z^2 - c(\gamma)\Big), \quad c(\gamma) := \log \int_{[-1,1]} \exp\Big(\gamma_1 z - \frac{\gamma_2}{2}z^2\Big)\,\mathrm{d}\pi(z).$$

Using this definition we can write the posterior distribution μ as

$$\mu_{\mathbf{y},\mathbf{X}}(\mathrm{d}\boldsymbol{\beta}) \propto \exp\Big(-\frac{1}{2\sigma^2}\Big[\boldsymbol{\beta}^{\mathrm{T}} A\boldsymbol{\beta} - 2\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}\Big]\Big)\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i), \quad \text{where } \pi_i := \pi_{(0,d_i)},\ d_i := \frac{D_{ii}}{\sigma^2}.$$

A central object in the theoretical study of these posterior distributions is the normalizing constant, also referred to as the "partition function" in statistical physics parlance. Formally, we define

$$Z_p(\mathbf{y}, \mathbf{X}) = \int_{[-1,1]^p} \exp\Big(- \frac{1}{2\sigma^2}\Big[\boldsymbol{\beta}^{\mathrm{T}} A \boldsymbol{\beta} - 2\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}\Big]\Big) \prod_{i=1}^{p} \pi_i(\mathrm{d}\beta_i). \tag{2}$$

The partition function is intractable for most priors, unless special conjugacy properties are satisfied between the prior and the likelihood. Henceforth, we suppress the dependence of μ, Z_p on $\mathbf{y}$, $\mathbf{X}$ whenever it is clear from the context. The classical Gibbs variational principle characterizes the partition function as

$$\log Z_p = \sup_{Q} \Big(\mathbb{E}_Q\Big[- \frac{1}{2\sigma^2}\Big[\boldsymbol{\beta}^{\mathrm{T}} A \boldsymbol{\beta} - 2\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}\Big]\Big] - \mathrm{D}_{\mathrm{KL}}(Q \| \prod_{i=1}^{p} \pi_i)\Big),$$

where the supremum ranges over probability distributions Q on $[-1,1]^p$ (see e.g. Wainwright and Jordan (2008)). In fact, the supremum in the above variational problem is attained by $Q = \mu$. The naive mean-field approximation to Z_p restricts the supremum to product distributions, and thus obtains a universal lower bound. Formally, we have,

$$\log Z_p \geq \sup_{Q = \prod_{i=1}^{p} Q_i} \Big(\mathbb{E}_Q\Big[- \frac{1}{2\sigma^2}\Big[\boldsymbol{\beta}^{\mathrm{T}} A \boldsymbol{\beta} - 2\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}\Big]\Big] - \mathrm{D}_{\mathrm{KL}}(Q \| \prod_{i=1}^{p} \pi_i)\Big).$$

The optimizer in the display above provides the best approximation to μ among product distributions under KL divergence. This paper focuses on the tightness of the naive mean-field lower bound in the context of linear regression. For an in-depth survey of variational inference, we refer the interested reader to Bishop (2006); Blei et al. (2017); Wainwright and Jordan (2008).

Outline: The rest of the paper is structured as follows. We collect our results in Section 2. We discuss some directions for future enquiry in Section 3. The main results are established in Section 4. We defer some proofs to the Appendix.

2. Results

We collect our main results in this section. To this end, we first discuss some elementary facts regarding exponential families. The next result collects some analytic properties of the cumulant generating function $c(\cdot)$, which will be relevant for our subsequent discussion. For the sake of completeness, we provide a proof in the Appendix.

Lemma 2 *Let $c(\cdot) : \mathbb{R}^2 \mapsto \mathbb{R}$ be as in Definition 1. Assume that both $\{-1, 1\}$ belong to the support of π. Then the following conclusions hold.*

(i) $\dot{c}(\gamma_1, \gamma_2) := \frac{\partial c(\gamma_1, \gamma_2)}{\partial \gamma_1}$ is strictly increasing in γ_1, with $\lim_{\gamma_1 \to \pm\infty} \dot{c}(\gamma_1, \gamma_2) = \pm 1$ for every $\gamma_2 \in \mathbb{R}$.

(ii) For any $t \in (-1, 1)$, there exists a unique $h(t, \gamma_2)$ such that $\dot{c}(h(t, \gamma_2), \gamma_2) = t$. Further, $\lim_{t \to \pm 1} h(t, \gamma_2) = \pm\infty$ for every $\gamma_2 \in \mathbb{R}$.

Armed with these basic facts, we can formally state the naive-mean field approximation to the log-normalizing constant (2).

Definition 3 *Define a possibly extended real valued function G on $[-1,1]\times\mathbb{R}$ by setting*

$$\begin{aligned} G(u,d) :=& uh(u,d) - c(h(u,d),d) + c(0,d) && \text{if } u\in(-1,1), d\in\mathbb{R},\\ :=& \mathrm{D}_{\mathrm{KL}}(\pi_\infty\|\pi_{(0,d)}) && \text{if } u=1, d\in\mathbb{R},\\ :=& \mathrm{D}_{\mathrm{KL}}(\pi_{-\infty}\|\pi_{(0,d)}) && \text{if } u=-1, d\in\mathbb{R}, \end{aligned}$$

where π_∞ and $\pi_{-\infty}$ are degenerate distributions which puts mass 1 at 1 and -1 respectively.

We will need the following facts about the derivatives of G. We defer the proof of Lemma 4 to the Appendix.

Lemma 4 *We have, for $u\in(-1,1)$ and $d\in\mathbb{R}$,*

$$\frac{\partial G}{\partial u}(u,d) = h(u,d),\ \frac{\partial G}{\partial d} = \frac{1}{2}\int_{-1}^{1} z^2 \mathrm{d}\pi_{(h(u,d),d)}(z) - \frac{1}{2}\int_{-1}^{1} z^2\mathrm{d}\pi_{(0,d)}(z).$$
$$\frac{\partial^2 G}{\partial^2 u}(u,d) = \frac{1}{\ddot{c}(h(u,d),d)} > 0.$$

Consequently, we have, $\sup_{u\in(-1,1),d\in\mathbb{R}} |\frac{\partial G}{\partial d}(u,d)| \le \frac{1}{2}$.

Definition 5 *Define $M_p : [-1,1]^p \to \mathbb{R}$ as*

$$M_p(\mathbf{u}) := \Big\{ -\frac{1}{2\sigma^2}\Big[\mathbf{u}^{\mathrm{T}} A\mathbf{u} - 2\mathbf{z}^{\mathrm{T}}\mathbf{u}\Big] - \sum_{i=1}^{p} G(u_i,d_i)\Big\}. \tag{3}$$

Lemma 6 *With $M_p(\cdot)$ as in (3), we have,*

$$\sup_{Q=\prod_{i=1}^p Q_i} \Big(\mathbb{E}_Q\Big[-\frac{1}{2\sigma^2}\Big[\boldsymbol{\beta}^{\mathrm{T}} A\boldsymbol{\beta} - 2\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}\Big]\Big] - \mathrm{D}_{\mathrm{KL}}(Q\|\prod_{i=1}^p \pi_i)\Big) = \sup_{\mathbf{u}\in[-1,1]^p} M_p(\mathbf{u}). \tag{4}$$

This result follows immediately from elementary facts about exponential families. We refer the interested reader to (Wainwright and Jordan, 2008, Section 5.3).

2.1 Validity of the naive mean-field approximation

Throughout the paper, we use the usual Landau notation $O(\cdot)$ and $o(\cdot)$ for deterministic sequences dependent on p.

Theorem 7 *Assume that the matrix A_p satisfies the two conditions*

$$\mathrm{tr}(A_p^2) = o(p), \tag{5}$$

$$\sup_{\mathbf{u}\in[-1,1]^p} \sum_{i=1}^{p} \Big|\sum_{j=1}^{p} A_p(i,j)u_j\Big| = O(p). \tag{6}$$

(i) We have, setting $R_p := \sup_{\mathbf{u}\in[-1,1]^p} M_p(\mathbf{u})$*, as* $p \to \infty$*,*

$$\log Z_p - R_p = o(p). \tag{7}$$

(ii) If $b_i := \mathbb{E}_\mu(\beta_i|\beta_k, k \neq i)$*, then the vector* $\mathbf{b} := (b_1, \ldots, b_p)$ *satisfies*

$$M_p(\mathbf{b}) - R_p = o(p).$$

(iii) Suppose there exists $\hat{\mathbf{u}} \in [-1,1]^p$ *which satisfies that for every* $\eta > 0$ *we have*

$$\limsup_{p\to\infty} \frac{1}{p}\Big[\sup_{\mathbf{u}\in[-1,1]^p:\|\mathbf{u}-\hat{\mathbf{u}}\|_2^2\geq p\eta} M_p(\mathbf{u}) - R_p\Big] < 0. \tag{8}$$

Assume further that the empirical measure $\frac{1}{p}\sum_i \delta_{D_p(i,i)}$ *is uniformly integrable. Then, for any* $\varepsilon > 0$ *and for any continuous function* $\zeta : [-1,1] \times [0,1] \to \mathbb{R}$ *we have*

$$\mu_{\mathbf{y},\mathbf{X}}\Big(\Big|\frac{1}{p}\sum_{i=1}^p \zeta\Big(\beta_i, \frac{i}{p}\Big) - \frac{1}{p}\sum_{i=1}^p \int_{[-1,1]} \zeta\Big(z, \frac{i}{p}\Big)\pi_{(h(\hat{u}_i,d_i),d_i)}(dz)\Big| > \varepsilon\Big) \to 0.$$

Remark 8 *The careful reader would have already noticed that Theorem 7 is stated for* deterministic *data* $(\mathbf{y}, \mathbf{X})$*. In practical applications, it is often more natural to assume that the data* $(\mathbf{y}, \mathbf{X})$ *is, in turn, sampled from some underlying distribution* $\mathcal{P}$*. The conclusions of Theorem 7 continue to hold as long as the sufficient conditions hold asymptotically with high probability under* $\mathcal{P}$*.*

The condition (6) is extremely mild, and specifies that the log normalizing constant is asymptotically of order p. This condition is satisfied, for example, whenever the matrix $A_p = \mathbf{X}^{\mathrm{T}}\mathbf{X} - D_p$ has spectral norm $O(1)$. The assumption (5) is the non-trivial assumption in the statement above, and effectively guarantees the accuracy of the naive mean field approximation to leading exponential order. Note that if $\lambda_1 \geq \cdots \geq \lambda_p$ denote the eigenvalues of A_p, then $\mathrm{tr}(A_p^2) = \sum_{i=1}^p \lambda_i^2$. Thus the requirement $\mathrm{tr}(A_p^2) = o(p)$ can be qualitatively interpreted to mean that the eigenstructure of A is "dominated" by a few top eigenvalues. Similar results were derived in the context of naive mean-field approximation for Potts models by one of the authors in Basak and Mukherjee (2017). Covariance matrices with an approximately low rank are ubiquitous in modern datasets, and we believe these conditions are satisfied in diverse applications of practical interest. Our result provides formal evidence to the correctness of widely used mean-field approximations in these settings. We prove Theorem 7 in Section 4.1.

The third part of our theorem provides further insights into the posterior distribution, assuming that the naive mean field approximation is "dominated" by a unique factorized distribution. In the language of Statistical Physics, these distributions are referred to be in a "pure phase".

We now demonstrate a few applications of Theorem 7 to concrete examples, which cover both deterministic and random design matrices. We defer the proofs of these corollaries to Appendix C.

Corollary 9 *Let* $\mathbf{X}$ *be any sequence of deterministic design matrices with* $\mathbf{X}^{\mathrm{T}}\mathbf{X} = A_p + D_p$, *where* A_p *and* D_p *represent the off diagonal and diagonal parts of* $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ *as before. Suppose* A_p *satisfies* (5) *and* (6), *and the empirical measure* $\frac{1}{p}\sum_{i=1}^{p} D_p(i,i)$ *is uniformly integrable. Then the conclusion of Theorem 7 holds.*

Corollary 10 *Suppose that the* i^{th} *row of the design matrix* $\mathbf{X}$ *equals* $n^{-1/2}\mathbf{x}_i$, *where* $\{\mathbf{x}_i\}_{1\le i\le n} \overset{i.i.d.}{\sim} N(0,\Gamma_p)$. *Assume that* $p = o(n)$, *and the following conditions hold:*

(a) The off diagonal part $\Gamma_{p,\mathrm{off}}$ *of the covariance matrix* Γ_p *satisfies* $\mathrm{tr}(\Gamma_{p,\mathrm{off}}^2) = o(p)$.

(b) $\|\Gamma_p\|_2 = O(1)$.

Then the conclusion of Theorem 7 holds.

Sparse design matrices arise routinely in coding theory (Gallager, 1962; MacKay, 1999) and genomics (Wu et al., 2011; Tang et al., 2014). Our next corollary discusses the accuracy of the nVB approximation in the context of a linear regression problem with a sparse bernoulli design. We assume that the entries of the design are independent, but not necessarily identical.

Corollary 11 *Suppose that the* $(i,j)^{th}$ *entry of the design matrix* $\mathbf{X}$ *equals* $\sqrt{\frac{p}{n}}B(i,j)$, *where* $\{B(i,j)\}_{1\le i\le n, 1\le j\le p}$ *are mutually independent Bernoullis with* $\mathbb{P}(B(i,j)=1) \le \frac{\lambda}{p}$, *for some* $\lambda > 0$ *free of* p. *If* $p = o(n)$, *the conclusion of Theorem 7 applies.*

2.2 Scaling limit for the log-normalizing constant

Under the asymptotic validity of the naive mean-field approximation, we derive an asymptotic scaling limit for the log-normalizing constant. The limiting description for the log-normalizing constant is in terms of an infinite dimensional variational problem. We will subsequently illustrate that under an additional separation condition, the NMF variational problem at finite n, p has a unique near optimizer, which can be approximated in terms of the optimizer of the associated infinite dimensional limiting problem. As a concrete takeaway, the optimizer of the limiting problem characterizes the typical behavior of empirical averages under the posterior. To this end, we will require some notation.

The theory of dense graph limits was developed in Borgs et al. (2008, 2012); Lovász and Szegedy (2007), and has received tremendous attention over the last decade in Probability, Combinatorics, Computer Science and Statistics. We refer the interested reader to Lovász (2012) for an in-depth survey of this area. Follow up work of Borgs et. al. (Borgs et al., 2019, 2018) has extended this theory significantly beyond the regime of dense graphs—the resulting L^p-convergence theory can handle sparse graphs and weighted matrices. Utilizing this set up, our next result will show that under the assumption that the off diagonal matrix A_p obtained from $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ converges in cut norm to a suitable graphon (need not be bounded), the corresponding log normalizing constant converges in probability to a deterministic optimization problem. We note that the use of cut-norms on matrices precedes the development of graph limit theory (see e.g. Frieze and Kannan (1999) and references therein).

Definition 12 *A function $W : [0,1]^2 \mapsto \mathbb{R}$ is called symmetric if $W(x,y) = W(y,x)$ for all $x, y \in [0,1]$. Any symmetric function $W : [0,1]^2 \mapsto \mathbb{R}$ which is L^1 integrable, i.e. $\|W\|_1 := \int_{[0,1]^2} |W(x,y)|dxdy < \infty$ is called a graphon. Let $\mathcal{W}$ denote the space of all graphons.*

The cut norm of a graphon W is given by

$$\|W\|_\square = \Big| \sup_{S,T \subset [0,1]} \int_{S \times T} W(x,y)dxdy \Big|.$$

The cut norm is equivalent to the $L^\infty \mapsto L^1$ operator norm defined by

$$\|W\|_{\infty \mapsto 1} := \sup_{f,g:\|f\|_\infty, \|g\|_\infty \leq 1} \Big| \int_{[0,1]^2} W(x,y)f(x)g(x)dxdy \Big|. \tag{9}$$

More precisely, we have $\|W\|_\square \leq \|W\|_{\infty \mapsto 1} \leq 4\|W\|_\square$. It also follows from (9) that the cut norm is weaker than the L^1 norm, i.e. convergence in L^1 implies convergence in cut norm.

Definition 13 *Given a symmetric $p \times p$ matrix B with real entries, define a piecewise constant function on $[0,1]^2$ by dividing $[0,1]^2$ into p^2 smaller squares each of length $1/p$, and set*

$$\begin{aligned} W_B(x,y) :=& B(i,j) \text{ if } \lceil px \rceil = i, \lceil py \rceil = j \text{ with } i \neq j, \\ =& 0 \text{ otherwise.} \end{aligned}$$

We will also need the following notion for embedding vectors into functions.

Definition 14 (Vector to function) *Given $\mathbf{t} := (t_1, \cdots, t_p) \in \mathbb{R}^p$, define the piecewise constant function $w_{\mathbf{t},p}$ on $[0,1]$ by dividing $[0,1]$ into p intervals $\cup_{i=1}^p \frac{1}{p}(i-1, i]$ of equal length $1/p$, and setting $w_{\mathbf{t}}(x) := t_i$ if x is in the i^{th} interval, i.e. $\lceil px \rceil = i$.*

To derive the scaling limit, we will assume an underlying model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. We say that a random variable $f := f(\mathbf{y}, \mathbf{X}) \overset{P|\mathbf{X}}{\longrightarrow} 0$ if for any $\varepsilon > 0$, $\mathbb{P}\Big[|f| > \varepsilon | \mathbf{X}\Big] \to 0$ as $p \to \infty$. Thus this convergence is conditional on the sequence of design matrices. In our subsequent analysis, the following representation lemma will be crucial.

Lemma 15 *Suppose we are in the setting of Theorem 7. Assume further that*

$$\sum_{i=1}^p D_p(i,i) = O(p).$$

Then there exists $(\xi_1, \cdots \xi_p) \overset{i.i.d.}{\sim} N(0,1)$ such that

$$\frac{1}{p} \sup_{\mathbf{u} \in [-1,1]^p} \Big| M_p(\mathbf{u}) - \widetilde{M}_p(\mathbf{u}) \Big| \overset{P|\mathbf{X}}{\to} 0,$$

where

$$\widetilde{M}_p(\mathbf{u}) := -\frac{1}{2\sigma^2}\Big[\mathbf{u}'A\mathbf{u} - 2\sum_{i=1}^p u_i \sqrt{D_p(i,i)}\xi_i - 2\boldsymbol{\beta}_0'\mathbf{X}'\mathbf{X}\mathbf{u}\Big] - \sum_{i=1}^p G(u_i, d_i).$$

We will derive a limiting formula for the log-normalizing constant (2) in terms of a variational problem on a space of probability distributions. This requires the following definition.

Definition 16 *Let $\mathscr{F}$ denote the space of all bounded measurable functions from $[0,1]\times\mathbb{R}$ to $[-1,1]$. Fixing $W\in\mathcal{W}, g,\psi\in L^1[0,1]$, define a functional $\mathcal{G}_{W,g,\psi}(\cdot):\mathscr{F}\to\mathbb{R}$ by setting*

$$\begin{aligned}\mathcal{G}_{W,g,\psi}(F) :=& -\frac{1}{2\sigma^2}\mathbb{E}[W(X,X')F(X,Z')F(X',Z')] + \frac{1}{\sigma^2}\mathbb{E}[g(X)F(X,Z)]\\ &+\frac{1}{\sigma^2}\mathbb{E}[\sqrt{\psi(X)}F(X,Z)Z] - \mathbb{E}\Big[G\Big(F(X,Z),\frac{\psi(X)}{\sigma^2}\Big)\Big],\end{aligned}$$

where $(X,X')\overset{i.i.d.}{\sim}U[0,1]$ and $(Z,Z')\overset{i.i.d.}{\sim}N(0,1)$ are mutually independent.

Theorem 17 *Suppose $\mathbf{y}\sim N\Big(\mathbf{X}\boldsymbol{\beta}_0,\sigma^2\mathbf{I}\Big)$. Writing $\mathbf{X}^{\mathrm{T}}\mathbf{X}=A_p+D_p$, set $\mathbf{D}$ to denote the vector in $\mathbb{R}^p$ containing the diagonal entries of D_p. Assume the following:*

(i) $d_\square(W_{pA_p},W)\to 0$ *for some* $W\in\mathcal{W}$.

(ii) $w_{\boldsymbol{\beta}_0}\overset{L^1}{\to}\phi(\cdot)$, *and* $w_{\mathbf{D}}\overset{L^1}{\to}\psi(\cdot)$.

(iii) $\mathrm{tr}(A_p^2)=o(p)$.

Define a function $g\in L^1[0,1]$ by

$$g(x)=\int_{[0,1]}W(x,y)\phi(y)\mathrm{d}y+\psi(x)\phi(x).$$

Then we have,

$$\frac{1}{p}\sup_{\mathbf{u}\in[-1,1]^p}M_p(\mathbf{u})\overset{P|\mathbf{X}}{\longrightarrow}\sup_{F\in\mathscr{F}}\mathcal{G}_{W,g,\psi}(F). \tag{10}$$

Remark 18 *For convenience of the reader, we point out the analogues between the discrete objects (vectors, matrices) and their continuum versions (functions) used in the above theorem, in the list below:*

$$pA_p\leftrightarrow W,\quad \boldsymbol{\beta}_0\leftrightarrow\phi,\quad \mathbf{D}\leftrightarrow\psi$$

Lemma 15 and Theorem 17 are established in Section 4.2.

2.3 Uniqueness of the optimizer

Theorem 7 identifies general conditions for the asymptotic tightness of the naive mean-field lower bound to the log-normalizing constant. The Gibbs Variational Principle (Wainwright and Jordan, 2008) establishes that under these settings, the distribution μ can be approximated, to the leading order, by a product distribution. However, this does not specify whether the "best" approximation is unique. Indeed, the ferromagnetic Ising model on the complete graph, henceforth referred to as the Curie-Weiss model, provides a classical example where the mean-field lower bound is tight, but without an external magnetization, the model has two distinct optimizers at "low temperature".

In this section, we identify some conditions which guarantee the uniqueness of the minimizers. From a statistical perspective, these conditions allow us to conclude that the posterior distribution roughly behaves like a product distribution. To this end, our next lemma identifies a set of sufficient conditions.

Definition 19 *Let $(\xi_1, \ldots, \xi_p) \overset{i.i.d.}{\sim} N(0,1)$ be as constructed in Lemma 15. For any $\mathbf{u} \in [-1,1]^p$, define a random empirical measure $L_p^{(\mathbf{u})}$ on $\mathbb{R}^3$ by setting*

$$L_p^{(\mathbf{u})} := \frac{1}{p}\sum_{i=1}^{p} \delta_{(\frac{i}{p},\xi_i,u_i)}.$$

Theorem 20 *Suppose all assumptions of Theorem 17 hold. Assume further that there exists $\mathbf{u}_p^* \in [-1,1]^p$ such that for all $\varepsilon > 0$ there exists $\delta' > 0$ such that*

$$\mathbb{P}\Big[\sup_{\mathbf{u}:\|\mathbf{u}-\mathbf{u}_p^*\|_2^2 > p\varepsilon} \frac{1}{p}\{M_p(\mathbf{u}) - M_p(\mathbf{u}_p^*)\} < -\delta' | \mathbf{X}\Big] = 1 - o(1). \tag{11}$$

Then the following conclusions hold.

(i) The limiting variational problem (10) has a unique optimizer $F^* \in \mathscr{F}$.

(ii) $L_p^{(\mathbf{u}_p^*)} \xrightarrow{P} \mu^*$, where μ^* is the law of $(X, Z, F^*(X,Z))$ with $X \sim U[0,1]$ and $Z \sim N(0,1)$ mutually independent.

(iii) F^* satisfies the fixed point equation:

$$F^*(x,z) \overset{a.s.}{=} \dot{c}\Big(\frac{1}{\sigma^2}\Big(-\mathbb{E}[W(x,X)F^*(X,Z)] + g(x) + \sqrt{\psi(x)}z\Big), \frac{\psi(x)}{\sigma^2}\Big), \tag{12}$$

where $X \sim U[0,1]$ and $Z \sim N(0,1)$ are mutually independent.

Remark 21 *Similar to remark 18, we point out the new discrete and continuum analogues appearing in Theorem 20.*

$$\mathbf{u}_p^* \leftrightarrow F^*, \quad L_p(\mathbf{u}_p^*) \leftrightarrow \mu^*.$$

Combining Theorem 7 and Theorem 20, we get the following corollary, which deduces a Law of Large numbers under the posterior distribution.

Corollary 22 *Suppose* (6), (11), *and the three conditions (i), (ii), (iii) of Theorem 17 hold. Then the optimization in Theorem 17 has a unique solution $F^* \in \mathscr{F}$ that satisfies*

$$F^*(x,z) \overset{a.s.}{=} \dot{c}\Big(\frac{1}{\sigma^2}\Big(-\mathbb{E}[W(x,X)F^*(X,Z)] + g(x) + \sqrt{\psi(x)}z\Big), \frac{\psi(x)}{\sigma^2}\Big),$$

Further, for any continuous function $\zeta : [-1,1]^2 \mapsto \mathbb{R}$ we have

$$\mu_{\mathbf{y},\mathbf{X}}\Big(\Big|\frac{1}{p}\sum_{i=1}^{p}\zeta(\beta_i,\beta_{0,i}) - \mathbb{E}\Big[\int_{[-1,1]}\zeta(w,\phi(X))d\pi_{\left(h(F^*(X,Z),\frac{\psi(X)}{\sigma^2}),\frac{\psi(X)}{\sigma^2}\right)}(w)\Big]\Big| > \varepsilon\Big) \overset{P|\mathbf{X}}{\to} 0,$$

where $X \sim U([0,1])$ and $Z \sim N(0,1)$ are independent.

Corollary 22 illustrates how one can extract properties of the high-dimensional posterior using the NMF approximation, and is one of the main takeaways of this paper. The result has two parts, and should be interpreted as follows: first, assume a non-trivial scaling of the problem (6), and that the finite dimensional NMF variational problem has a unique approximate optimizer (11). In this setting, provided the problem parameters W, $\boldsymbol{\beta}_0$, $\mathbf{D}$ have appropriate continuum limits and the problem is approximately NMF, the infinite dimensional variational problem in Theorem 17 has a unique optimizer F^*. In turn, this optimizer F^* satisfies a functional fixed point equation. Note that in principle, this fixed point equation can still have multiple solutions—we simply claim that the global optimizer is one of the solutions to this fixed point relation. This optimizer governs the Law of Large Numbers (LLN) behavior under the posterior distribution—this is the main takeaway from the second part of the corollary above. Formally, if $\boldsymbol{\beta}$ is a sample from the posterior $\mu_{\mathbf{y},\mathbf{X}}$ and ζ is any continuous test function, the empirical average $\frac{1}{p}\sum_{i=1}^{p}\zeta(\beta_i,\beta_{0,i})$ concentrates around an explicit, deterministic limit characterized in terms of the optimizer F^*. Note that there is an additional technical subtlety—due to the randomness in the distribution of $\mathbf{y}$ given $\mathbf{X}$, the posterior distribution $\mu_{\mathbf{y},\mathbf{X}}(\cdot)$ is actually random! Thus the aforementioned behavior of the empirical averages (under the posterior) is only valid with high probability under the distribution $P|\mathbf{X}$. To further elucidate the usefulness of this corollary, we describe two specific applications in the next remark.

Remark 23 *We highlight two specific applications of Corollary 22. First, setting $\zeta(x,y) = (x-y)^2$, we have,*

$$\frac{1}{p}\sum_{i=1}^{p}(\beta_i - \beta_{0,i})^2 \overset{P|\mathbf{X}}{\to} \mathbb{E}\Big[\int_{[-1,1]}(w-\phi(X))^2 d\pi_{\left(h(F^*(X,Z),\frac{\psi(X)}{\sigma^2}),\frac{\psi(X)}{\sigma^2}\right)}(w)\Big].$$

Thus although we do not have posterior contraction in this regime, the normalized L^2 distance between the true vector $\boldsymbol{\beta}_0$ and a sample $\boldsymbol{\beta}$ from the posterior $\mu_{\mathbf{y},\mathbf{X}}$ converges to the explicit value characterized above. In other words, the posterior is "concentrated" on a thin shell of a fixed radius around the true vector $\boldsymbol{\beta}_0$.

As a second application, we consider a prior π such that zero is an isolated point in its support. In this case, $\zeta(x,y) = \mathbf{1}(x=0)$ is almost surely continuous under π, and thus

$$\frac{1}{p}\sum_{i=1}^{p}\mathbf{1}(\beta_i = 0) \overset{P|\mathbf{X}}{\to} \mathbb{E}\Big[\pi_{\left(h(F^*(X,Z),\frac{\psi(X)}{\sigma^2}),\frac{\psi(X)}{\sigma^2}\right)}(\{0\})\Big].$$

In this case, the proportion of zero coordinates in a sample from the posterior distribution can be tracked using our framework. These results are particularly useful for "spike and slab" type priors π.

Remark 24 *To use Corollary 22 in concrete examples, one needs to compute the functions (W, g, ψ), and verify the conditions* (5), (6) *and* (11). *Of these, the separation condition* (11) *is somewhat implicit, while the other two conditions are relatively easy to verify directly. The following lemma provides two sufficient conditions to this end. In particular, it shows that* (11) *holds in the so called* high temperate regime*, or if π has a density (with respect to Lebesgue measure) which is log concave.*

Lemma 25 *(i) Suppose there exists $\lambda > 0$ and $\mathbf{u}_p^* \in [-1,1]^p$ such that for all $\mathbf{u} \in [-1,1]^p$ we have*

$$\mathbb{P}\Big[M_p(\mathbf{u}^*) - M_p(\mathbf{u}) \geq \lambda \|\mathbf{u} - \mathbf{u}^*\|_2^2 \Big| \mathbf{X}\Big] = 1 - o(1), \tag{13}$$

for some $\lambda > 0$, free of $\mathbf{u}$ and p. Then the condition (11) *in Theorem 20 holds.*

(ii) In particular (13) *holds under either of the following conditions:*

(a)

$$\limsup_{p\to\infty} \sup_{1\leq i\leq p} \sum_{j\neq i} |A_p(i,j)| < \sigma^2.$$

(b) Let the prior π have a density with respect to Lebesgue measure on $[-1,1]$,

$$\frac{\mathrm{d}\pi}{\mathrm{d}x} = \frac{1}{Z}\exp(-V(x)),$$

where V is even, $V : [0,1] \to \mathbb{R}$ is increasing, and V' is convex on $[0,1)$. Further, we assume that $\liminf_{p\to\infty} \lambda_{\min}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) > 0$.

We collect the proofs of Theorem 20 and Lemma 25 in Section 4.3.

2.4 Applications

To illustrate the utility of Theorem 17 and Corollary 22, we apply our results to specific examples in this section. The proofs are deferred to Appendix C.

Spiked Covariance Matrix: We consider linear regression with mean-zero gaussian features. We assume a spike covariance structure (Johnstone and Lu, 2009) on the features.

Corollary 26 *Suppose that the i^{th} row of the design matrix $\mathbf{X}$ equals $n^{-1/2}\mathbf{x}_i$, where $\{\mathbf{x}_i\}_{1\leq i\leq n} \overset{i.i.d.}{\sim} N(0,\Gamma_p)$, where $\Gamma_p = \mathbf{I} + \mathbf{v}\mathbf{v}'$, with $v_i := \frac{1}{\sqrt{p}} G(i/p)$, and $G : [0,1] \mapsto \mathbb{R}$ is continuous almost surely. Further assume that $p \ll n$, and the true regression coefficient $\boldsymbol{\beta}_0$ satisfies $w_{\boldsymbol{\beta}_0} \xrightarrow{L^1} \phi$, for some $\phi \in L^1[0,1]$.*

(a) Then for any $\varepsilon > 0$,

$$\frac{1}{p}\log Z_p(\mathbf{y}, \mathbf{X}) \xrightarrow{P|\mathbf{X}} \sup_{F\in\mathcal{F}} \mathcal{G}_{W,g,\psi}(F),$$

where

$$W(x,y) = G(x)G(y) \quad \psi(x) = 1, \quad g(x) = G(x)\int_{[0,1]} G(y)\phi(y)dy + \phi(x).$$

(b) Consider the following two cases: either

(i) $\lambda := \max_{1\le i\le n}\sum_{j\neq i}\Gamma_p(i,j) < \sigma^2$, or

(ii) the conditions of Lemma 25 Part (b) (ii) hold.

Then the conclusions of Corollary 22 hold.

Remark 27 *We now consider a very special case to illustrate our results. Suppose $\mathbf{X}$ has IID entries distributed as $N(0, \frac{1}{n})$, and $\sigma = 1$. In this case the assumptions of Corollary 26 are satisfied with*

$$W \equiv 0, \quad G \equiv 0, \quad g = \phi,$$

and so the optimizing $F^*(x,z)$ in Corollary 22 simplifies to

$$F^*(x,z) = \dot{c}(\phi(x) + z, 1).$$

The next example re-visits the sparse bernoulli design setting introduced in Corollary 11.

Corollary 28 *Suppose that the $(i,j)^{th}$ entry of the design matrix $\mathbf{X}$ equals $\sqrt{\frac{p}{n}}B(i,j)$, where $\{B(i,j)\}_{1\le i\le n, 1\le j\le p}$ are mutually independent Bernoulli random variables, with $\mathbb{P}(B(i,j)=1) = \frac{1}{p}G\Big(i/n, j/p\Big)$, where G is a function on $[0,1]^2$ which is continuous almost surely. Further assume that $p \ll \sqrt{n}$, and the true regression coefficient $\boldsymbol{\beta}_0$ satisfies $w_{\boldsymbol{\beta}_0} \xrightarrow{L^1} \phi$, for some function $\phi \in L^1[0,1]$.*

(a) Then

$$\frac{1}{p}\log Z_p(\mathbf{y}, \mathbf{X}) \xrightarrow{P} \sup_{F\in\mathcal{F}} \mathcal{G}_{W,g,\psi}(F),$$

where

$$W(x,y) = \int_{[0,1]} G(t,x)G(t,y)dt, \quad \psi(x) = \int_{[0,1]} G(t,x)dt,$$
$$g(x) = \int_{[0,1]} W(x,y)\phi(y)dy + \phi(x)\psi(x).$$

(b) Consider the following two cases:

(i) Set

$$S(x) := \int_{[0,1]} W(x,y)dy = \int_{[0,1^2]} G(t,x)G(t,y)dtdy.$$

Suppose $\sigma^2 > \operatorname{ess\,sup} S(X)$, where $X \sim U[0,1]$;

(ii) the conditions of Lemma 25 Part (b) (ii) hold.

Then the conclusions of Corollary 22 hold.

As our last example, we consider a sequence of design matrices arising in the study of two-way ANOVA designs.

Corollary 29 *Suppose that we have a two factor ANOVA model of the form*

$$y_{ij} = \frac{1}{\sqrt{p}}(\tau_i + \gamma_j) + \xi_{ij}, 1 \leq i, j \leq \tilde{p}$$

Here $(\tau_1, \ldots, \tau_{\tilde{p}}) \in [-1,1]^{\tilde{p}}$ *are the levels of the first factor, and* $(\gamma_1, \ldots, \gamma_{\tilde{p}}) \in [-1,1]^{\tilde{p}}$ *are the levels of the second factor, and* $\xi_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$. *Setting* $n = \tilde{p}^2$ *and* $p = 2\tilde{p}$, *let* $\mathbf{y} \in \mathbb{R}^n$ *denote the vector obtained by linearizing the matrix* $((y_{ij}))$ *row-wise, and* $\mathbf{X}$ *denote the corresponding* $n \times p$ *design matrix, and let* $\boldsymbol{\beta} := (\tau, \gamma) \in [-1,1]^p$ *be the unknown parameter.*

Assume that the true regression coefficient $\boldsymbol{\beta}_0 \in [-1,1]^p$ *satisfies* $w_{\boldsymbol{\beta}_0} \overset{L^1}{\to} \phi$, *for some function* $\phi \in L^1[0,1]$.

(a) Then

$$\frac{1}{p} \log Z_p(\mathbf{y}, \mathbf{X}) \overset{P}{\to} \sup_{F \in \mathcal{F}} \mathcal{G}_{W,g,\psi}(F),$$

where $\psi = \frac{1}{2}$,

$$\begin{aligned} W(x,y) :=& 0 \text{ if } (x,y) \in [0,.5)^2 \cup (.5,1]^2, \\ =& 1 \text{ if } (x,y) \in [0,.5) \times (.5,1] \cup (.5,1] \times [0,.5). \end{aligned}$$

and $g(x) = \int_{[0,1]} W(x,y)\phi(y)dy + \frac{1}{2}\phi(x)$.

(b) Consider the following two cases:

(i) $\sigma^2 > \frac{1}{2}$.

(ii) the conditions of Lemma 25 Part (b) (ii) hold.

Then the conclusions of Corollary 22 hold.

3. Discussions

We discuss some limitations of our current results, and collect some questions for future enquiry.

(i) **The bounded support assumption on the prior**—A vital technical assumption in our analysis concerns the bounded support assumption on the prior. We note that for a general prior with unbounded support, the posterior $\mu_{\mathbf{y},\mathbf{X}}$ might not even be a proper probability distribution. One intuitively expects that under appropriate "tail-decay" conditions on the prior, the results in this paper should generalize. Going beyond the bounded support assumption requires extending the theory of non-linear large deviations to probability measures on unbounded spaces, and is thus beyond the scope of this paper.

(ii) **Connection with classical asymptotics**—The curious reader might wonder about the connections between the results in this paper, and classical low-dimensional bayesian asymptotic theory. To explicate this connection, recall the model (1) with the covariates being drawn iid from a $N(0, I_p)$ distribution. Consider the classical asymptotic regime with fixed dimension p and $n \to \infty$. Let $\boldsymbol{\beta}_0/\sqrt{n} \in \mathbb{R}^p$ be a sequence of coefficients on the *local* scale. Using the theory of equivalence of statistical experiments (Van der Vaart, 2000), the likelihood is equivalent to an observation $\tilde{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}_0, (\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n})^{-1})$. Further, the Fisher information matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n)^{-1}$ is concentrated around the identity matrix, and thus the data is asymptotically equivalent to $\tilde{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}_0, I_p)$. In turn, the posterior, constructed from an iid product prior, is equivalent to

$$\mathrm{d}\mu(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}) \propto \prod_{i=1}^{p} \exp\Big(-\frac{1}{2}(\beta_i - \beta_{0,i})^2\Big)\mathrm{d}\pi(\beta_i).$$

Finally, consider an additional limit $p \to \infty$. Assuming $w_{\boldsymbol{\beta}_0} \xrightarrow{L^1} \phi$, by the weak law of large numbers, for any continuous function $\zeta : [-1, 1]^2 \to \mathbb{R}$,

$$\frac{1}{p}\sum_{i=1}^{p} \zeta(\beta_i, \beta_{0,i}) \xrightarrow{P} \mathbb{E}\Big[\int_{[-1,1]} \zeta(w, \phi(X)) d\pi_{(\phi(X)+Z,1)}(w)\Big],$$

where $X \sim U([0, 1])$ and $Z \sim N(0, 1)$ are independent. We emphasize that this is an iterated limit, and p is not a function of n in this discussion.

Note that this behavior is consistent with Corollary 22 upon setting $W = 0$ and $\psi = 1$. Thus our results are naturally compatible with the *local asymptotics* regime—in fact, for gaussian designs, our results establish the validity of this behavior as long as $p = o(n)$. Of course, our framework is significantly more general, and can handle situations where the design distribution changes with n, p, e.g. sparse Bernoulli designs.

(iii) **Extensions to Gibbs posteriors and fractional posteriors**—A careful study of our proof reveals that our main results do not depend strongly on the correct model specification. As a result, we expect similar techniques to be broadly useful in the study of Gibbs and fractional posteriors (Alquier et al., 2016; Yang et al., 2020).

(iv) **Extensions to other GLMs**—Another natural question of interest concerns the applicability of these ideas to more general models, e.g. logistic regression. We consider this to be an extremely interesting question, and plan to explore this in the future.

(v) **Extensions to models with latent characteristics**—Variational methods are ubiquitous in applications with latent characteristics e.g. topic modeling, community detection etc. In contrast, the relevant variables are all observed in the linear model framework. It will be interesting to explore the applicability of our techniques to models with latent features.

(vi) **Connections with benign overfitting**—A recent series of papers have explored the *benign overfitting* phenomenon in the context of high-dimensional linear regression (Bartlett et al., 2020; Hastie et al., 2022; Tsigler and Bartlett, 2020). The main takeaway from this line of research is that if the population covariance matrix has a long and fat tail, the min-norm interpolant continues to have excellent prediction performance in high dimensions. There are some high-level conceptual similarities between this phenomenon, and the results reported in this paper. Indeed, both phenomena are driven by the eigenvalues of the covariates—in our case, the condition $\mathrm{Tr}(\mathbf{A}^2) = o(p)$ essentially demands that the matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ has only $o(p)$ many eigenvalues which are $O(1)$. Broadly, this highlights the importance of covariate geometry in governing the practical performance of ML methods in high-dimensions.

(vii) **Connections with Bayesian deep learning**—Some recent papers in Bayesian deep learning (Farquhar et al., 2020; Foong et al., 2020) experimentally establish that the nVB approximation might not be accurate in Bayesian neural networks, and that increasing the depth might help ameliorate some of the issues. Our results, on the other hand, derive sufficient conditions for the accuracy of the nVB approximation under the simple case of the linear model with product priors. While one should definitely not expect the nVB approximation to be uniformly accurate (especially in high-dimensions), it would be interesting to identify conditions on the covariates which guarantee its accuracy in Bayesian Neural Networks.

4. Proofs

We prove the main results in this section. Theorem 7 is established in Section 4.1, Theorem 17 is established in Section 4.2, while Theorem 20 is proved in Section 4.3.

4.1 Proof of Theorem 7

Our first lemma collects some basic facts about the exponential family π_γ. The proof is deferred to the Appendix.

Lemma 30 *In the setting of Lemma 2, we have the following conclusions.*

(i) Set $\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^2$. The derivatives of $c(\gamma_1, \gamma_2)$ with respect to γ_1 are given by

$$\dot{c}(\gamma) := \frac{\partial c(\gamma_1,\gamma_2)}{\partial \gamma_1} = \int_{[-1,1]} z \mathrm{d}\pi_\gamma(z), \quad \ddot{c}(\gamma) := \frac{\partial^2 c(\gamma_1,\gamma_2)}{\partial \gamma_1^2} = \int_{[-1,1]} (z - \dot{c}(\gamma))^2 \mathrm{d}\pi_\gamma(z).$$

(ii) We have, for $\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^2$,

$$\mathrm{D}_{\mathrm{KL}}(\pi_\gamma \| \pi_{(0,\gamma_2)}) = \gamma_1 \dot{c}(\gamma) - c(\gamma) + c(0, \gamma_2).$$

(iii) Consider a sequence $\gamma_k = (\gamma_{1,k}, \gamma_{2,k}) \in \mathbb{R}^2$ such that $\gamma_{1,k} \to \pm\infty$ and $\limsup |\gamma_{2,k}| < \infty$. Then $\pi_{\gamma_k} \xrightarrow{w} \pi_{\pm\infty}$, where $\pi_{\pm\infty}$ is the degenerate distribution which puts mass 1 at ± 1.

(iv) For any $d > 0$, the function $G(\cdot, d) : [-1, 1] \to \mathbb{R}^+$ is lower semicontinuous.

(v) For $x \in [-1, 1]$, $y \in \mathbb{R}$ and $r \in \mathbb{N}$, define

$$H(x, y) = \int_{[-1,1]} z^r \mathrm{d}\pi_{(h(x,y),y)}(z) \tag{14}$$

Then we have,

$$\sup_{x\in(-1,1),y\in\mathbb{R}} \Big|\frac{\partial H(x, y)}{\partial y}\Big| \le 2. \tag{15}$$

We now turn to the proof of Theorem 7. To this end, set

$$m_i(\boldsymbol{\beta}) := \sum_{j=1}^{p} A_{ij}\beta_j, \quad \theta_i := \frac{z_i - m_i(\boldsymbol{\beta})}{\sigma^2}, \tag{16}$$

where we recall that σ^2 denotes the noise variance in our linear regression model (1), and $\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$. Observe that $A_{ii} = 0$ for all $1 \le i \le p$ implies that $\mu(\cdot|(\beta_j)_{j\neq i}) = \pi_{(\theta_i,d_i)}$, and thus $\mathbb{E}_\mu[\beta_i|(\beta_j)_{j\neq i}] = \dot{c}(\theta_i, d_i)$. We define

$$\mathbf{b} = (\dot{c}(\theta_i, d_i))_{1\le i\le p}. \tag{17}$$

We will prove that certain statistics under the posterior distribution μ can be "well-approximated" by the vector of conditional means. To this end, we establish the following results.

Lemma 31 *Under the conditions of Theorem 7, setting $f(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}A\boldsymbol{\beta} + \frac{1}{\sigma^2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$, we have,*

$$\log Z_p = \log\Big[\int_{[-1,1]^p} e^{f(\boldsymbol{\beta})}\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i)\Big] = \sup_{\mathbf{u}\in[-1,1]^p} M_p(\mathbf{u}) + o(p), \tag{18}$$

$$\mathbb{E}_\mu\Big[\boldsymbol{\beta}^{\mathrm{T}}A\boldsymbol{\beta} - \mathbf{b}^{\mathrm{T}}A\mathbf{b}\Big]^2 = o(p^2), \tag{19}$$

$$\mathbb{E}_\mu\Big[\sum_{i=1}^{p} m_i(\boldsymbol{\beta})(\beta_i - b_i)\Big]^2 = o(p^2), \tag{20}$$

$$\mathbb{E}_\mu\Big[f(\boldsymbol{\beta}) - f(\mathbf{b}) - \sum_{i=1}^{p}\theta_i(\beta_i - b_i)\Big]^2 = o(p^2). \tag{21}$$

In the above displays, the random vectors $\boldsymbol{\theta} := (\theta_1, \cdots, \theta_p)^{\mathrm{T}}$ and $\mathbf{b}$ are as defined in (16) and (17) respectively.

Lemma 32 *Suppose $\phi : [-1, 1] \mapsto \mathbb{R}$ is a bounded measurable function, and A_p satisfies (5) and (6). Then for any $\mathbf{c} \in [-1, 1]^p$ we have*

$$\limsup_{p\to\infty}\frac{1}{p}\mathbb{E}_\mu\left[\sum_{i=1}^{p} c_i\Big\{\phi(\beta_i) - \mathbb{E}_\mu[\phi(\beta_i)|\beta_k, k \neq i]\Big\}\right]^2 < \infty.$$

We defer the proofs of these lemmas to the end of this section, and establish Theorem 7, given these lemmas.

Proof of Theorem 7

Proof of Part (i). This is exactly (18) of Lemma 31.

Proof of Part (ii). With $f(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}} A\boldsymbol{\beta} + \frac{1}{\sigma^2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$ as in Lemma 31, define

$$C_p(\varepsilon) = \Big\{\mathbf{u} \in [-1,1]^p : f(\mathbf{u}) - \sum_{i=1}^{p}\Big(u_i h(u_i, d_i) - c(h(u_i, d_i), d_i)\Big) < \sup_{\mathbf{u}\in[-1,1]^p} M_p(\mathbf{u}) - p\varepsilon\Big\}.$$

Note that it suffices to establish that for all $\varepsilon > 0$, $\mu(\mathbf{b} \in C_p(\varepsilon)) \to 0$ as $p \to \infty$. To this end, define the event

$$E_p(\varepsilon/2) = \Big\{\boldsymbol{\beta} \in [-1,1]^p : |f(\boldsymbol{\beta}) - f(\mathbf{b}) - \sum_{i=1}^{p}\theta_i(\beta_i - b_i)| \le \frac{p\varepsilon}{2}\Big\}.$$

Note that (21), in combination with Chebychev inequality, implies that $\mu(E_p(\varepsilon/2)^c) \to 0$ as $p \to \infty$. Therefore,

$$\mu(C_p(\varepsilon)) \le \mu(C_p(\varepsilon) \cap E_p(\varepsilon/2)) + \mu(E_p(\varepsilon/2)^c) = \mu(C_p(\varepsilon) \cap E_p(\varepsilon/2)) + o(1). \tag{22}$$

Now, we have,

$$\begin{aligned}
&\mu(C_p(\varepsilon) \cap E_p(\varepsilon/2)) = \frac{1}{Z_p}\int_{C_p(\varepsilon)\cap E_p(\varepsilon/2)} \exp(f(\boldsymbol{\beta}))\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i)\\
&\le \frac{\exp(p\varepsilon/2)}{Z_p}\int_{C_p(\varepsilon)\cap E_p(\varepsilon/2)} \exp\Big[f(\mathbf{b}) + \sum_{i=1}^{p}\theta_i(\beta_i - b_i)\Big]\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i)\\
&\le \frac{\exp(-p\varepsilon/2 + \sup_{\mathbf{u}\in[-1,1]^p} M_p(\mathbf{u}))}{Z_p}\int_{C_p(\varepsilon)\cap E_p(\varepsilon/2)} \exp\Big[\sum_{i=1}^{p}(\theta_i\beta_i - c(\theta_i, d_i))\Big]\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i),
\end{aligned}$$

where the first inequality follows using the definition of $E_p(\varepsilon/2)$, while the last inequality follows from the definition of $C_p(\varepsilon)$. Using the display above in combination with (7), it suffices to establish that

$$\log\int_{C_p(\varepsilon)\cap E_p(\varepsilon/2)} \exp\Big[\sum_{i=1}^{p}(\theta_i\beta_i - c(\theta_i, d_i))\Big]\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i) = o(p).$$

This argument would be relatively straight-forward if the θ_i were fixed constants independent of $\boldsymbol{\beta}$, as

$$\begin{aligned}
&\int_{C_p(\varepsilon)\cap E_p(\varepsilon/2)} \exp\Big[\sum_{i=1}^{p}(\theta_i\beta_i - c(\theta_i, d_i))\Big]\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i)\\
&\le \int_{[-1,1]^p} \exp\Big[\sum_{i=1}^{p}(\theta_i\beta_i - c(\theta_i, d_i))\Big]\prod_{i=1}^{p}\pi_i(\mathrm{d}\beta_i) = 1.
\end{aligned}$$

We caution the reader that this is not the case, and the θ_i are themselves functions of $\boldsymbol{\beta}$, as specified in (16).

To overcome this issue, we proceed as follows. Fix $\delta > 0$, and let $\mathscr{D}_p(\delta)$ be a $\sqrt{p}\delta$ net of the set $\{A\boldsymbol{\beta}, \boldsymbol{\beta} \in [-1,1]^p\}$ in the euclidean metric, satisfying $\lim_{p\to\infty} \frac{1}{p}\log|\mathscr{D}_p(\delta)| = 0$. Under the assumption $\mathrm{tr}(A^2) = o(p)$, such a net was constructed in (Basak and Mukherjee, 2017, Lemma 3.4). Thus for every $\boldsymbol{\beta} \in [-1,1]^p$, there exists $\mathbf{p} \in \mathscr{D}_p(\delta)$ such that $\|A\boldsymbol{\beta} - \mathbf{p}\|_2 \le \delta\sqrt{p}$. For any $\mathbf{p} \in \mathscr{D}_p(\delta)$, we set

$$\mathcal{P}(\mathbf{p}) = \{\boldsymbol{\beta} \in [-1,1]^p : \|A\boldsymbol{\beta} - \mathbf{p}\|_2 \le \delta\sqrt{p}\}.$$

For $\boldsymbol{\beta} \in \mathcal{P}(\mathbf{p})$, setting $\theta_i^{\mathbf{p}} = \frac{z_i - p_i}{\sigma^2}$, we have,

$$\sum_{i=1}^p (\theta_i - \theta_i^{\mathbf{p}})^2 = \frac{1}{\sigma^4}\sum_{i=1}^p (m_i(\boldsymbol{\beta}) - p_i)^2 = \frac{1}{\sigma^4}\|A\boldsymbol{\beta} - \mathbf{p}\|_2^2 \le \frac{p\delta}{\sigma^4},$$

where the last inequality uses the definition of $\mathcal{P}(\mathbf{p})$. This gives,

$$\Big|\sum_{i=1}^p (\theta_i - \theta_i^{\mathbf{p}})\beta_i\Big| \le \frac{p\sqrt{\delta}}{\sigma^2}, \quad \Big|\sum_{i=1}^p \Big(c(\theta_i, d_i) - c(\theta_i^{\mathbf{p}}, d_i)\Big)\Big| \le \sum_{i=1}^p |\theta_i - \theta_i^{\mathbf{p}}| \le \frac{p\sqrt{\delta}}{\sigma^2}.$$

where the first inequality follows from Cauchy-Schwarz, and the second inequality follows from the smoothness of $c(\cdot,\cdot)$. Thus we have,

$$\begin{aligned}
&\int_{C_p(\varepsilon)\cap E_p(\varepsilon/2)} \exp\Big[\sum_{i=1}^p (\theta_i\beta_i - c(\theta_i, d_i))\Big] \prod_{i=1}^p \pi_i(\mathrm{d}\beta_i) \\
&\le \sum_{\mathbf{p}\in\mathscr{D}_p(\delta)} \int_{\mathcal{P}(\mathbf{p})} \exp\Big[\sum_{i=1}^p (\theta_i\beta_i - c(\theta_i, d_i))\Big] \prod_{i=1}^p \pi_i(\mathrm{d}\beta_i). \\
&\le \exp\Big(\frac{2p\sqrt{\delta}}{\sigma^2}\Big) \sum_{\mathbf{p}\in\mathscr{D}_p(\delta)} \int_{[-1,1]^p} \exp\Big[\sum_{i=1}^p (\theta_i^{\mathbf{p}}\beta_i - c(\theta_i^{\mathbf{p}}, d_i))\Big] \prod_{i=1}^p \pi_i(\mathrm{d}\beta_i) = \exp\Big(\frac{2p\sqrt{\delta}}{\sigma^2}\Big).
\end{aligned}$$

This concludes the proof, as $\delta > 0$ is arbitrary.

Proof of Part (iii). First, note that it suffices to show the result with $\zeta(x,y) = x^r y^s$ for any $r, s \in \mathbb{N}$. Thus we fix $r, s \in \mathbb{N}$ in the rest of the proof. Using Lemma 32 with $\phi(x) = x^r$ and $c_i = (\frac{i}{p})^s$ it follows that for any $\varepsilon > 0$,

$$\mu\Big(\Big|\frac{1}{p}\sum_{i=1}^p \beta_i^r \Big(\frac{i}{p}\Big)^s - \frac{1}{p}\sum_{i=1}^p H(b_i, d_i)\Big(\frac{i}{p}\Big)^s\Big| > \varepsilon\Big) \to 0,$$

where H is defined as in (14).To complete the proof, it thus suffices to show that

$$\mu\Big(\Big|\frac{1}{p}\sum_{i=1}^p H(b_i, d_i)\Big(\frac{i}{p}\Big)^s - \frac{1}{p}\sum_{i=1}^p H(\hat{u}_i, d_i)\Big(\frac{i}{p}\Big)^s\Big| > \varepsilon\Big) \to 0. \tag{23}$$

Fixing $K < \infty$ and setting $d_K(i) := d_i 1\{|d_i| \le K\}$, using Lemma 30 Part (v) we have

$$\frac{1}{p}\left|\sum_{i=1}^{p} H(\hat{u}_i, d_i)\Big(\frac{i}{p}\Big)^s - \sum_{i=1}^{p} H(\hat{u}_i, d_K(i))\Big(\frac{i}{p}\Big)^s\right| \le \frac{2}{p}\sum_{i=1}^{p} d_i 1\{|d_i| > K\}. \tag{24}$$

which goes to 0 as $p \to \infty$ followed by $K \to \infty$, using the uniform integrability assumption on $\frac{1}{p}\sum_{i=1}^{p}\delta_{d_i}$. Also, with $\delta > 0$ and setting

$$\begin{aligned}\hat{u}_\delta(i) :=& \hat{u}_i \text{ if } |\hat{u}_i| \le 1-\delta,\\ :=& 1-\delta \text{ if } \hat{u}_i > 1-\delta,\\ :=& -1+\delta \text{ if } \hat{u}_i < -1+\delta,\end{aligned}$$

we have

$$\begin{aligned}&\frac{1}{p}\left|\sum_{i=1}^{p} H(\hat{u}_i, d_K(i))\Big(\frac{i}{p}\Big)^s - \sum_{i=1}^{p} H(\hat{u}_\delta(i), d_K(i))\Big(\frac{i}{p}\Big)^s\right|\\ \le& \sup_{x\in[1-\delta,1],|y|\le K}|H(x,y) - H(1-\delta,y)| + \sup_{x\in[-1,1+\delta],|y|\le K}|H(x,y) - H(-1+\delta,y)|. \end{aligned}\tag{25}$$

Fixing $K > 0$, we now claim that the RHS of (25) converges to 0 as $\delta \to 0$. Given this claim, it follows from (24) and (25) that

$$\limsup_{K\to\infty}\limsup_{\delta\to 0}\limsup_{p\to\infty}\Big|\frac{1}{p}\sum_{i=1}^{p} H(\hat{u}_i, d_i)\Big(\frac{i}{p}\Big)^s - \frac{1}{p}\sum_{i=1}^{p} H(\hat{u}_\delta(i), d_K(i))\Big(\frac{i}{p}\Big)^s\Big| = 0.$$

A similar argument with $\hat{u}_i$ replaced by b_i gives

$$\limsup_{K\to\infty}\limsup_{\delta\to 0}\limsup_{p\to\infty}\Big|\frac{1}{p}\sum_{i=1}^{p} H(b_i, d_i)\Big(\frac{i}{p}\Big)^s - \frac{1}{p}\sum_{i=1}^{p} H(b_\delta(i), d_K(i))\Big(\frac{i}{p}\Big)^s\Big| = 0,$$

where $b_\delta(i) := b_i 1\{|b_i| \le 1-\delta\}$.

It suffices to show that for every δ, K fixed we have

$$\mu\Big(\Big|\frac{1}{p}\sum_{i=1}^{p} H(\hat{u}_\delta(i), d_K(i))\Big(\frac{i}{p}\Big)^s - \frac{1}{p}\sum_{i=1}^{p} H(b_\delta(i), d_K(i))\Big(\frac{i}{p}\Big)^s\Big| > \varepsilon\Big) \to 0.$$

But this follows on noting that

$$\sup_{|x|\le 1-\delta,|y|\le K}\Big|\frac{\partial H(x,y)}{\partial x}\Big| =: M < \infty,$$

and so

$$\begin{aligned}&\frac{1}{p}\left|\sum_{i=1}^{p} H(\hat{u}_\delta(i), d_K(i))\Big(\frac{i}{p}\Big)^s - \sum_{i=1}^{p} H(b_\delta(i), d_K(i))\Big(\frac{i}{p}\Big)^s\right|\\ &\qquad\le \frac{M}{p}\sum_{i=1}^{p}|\hat{u}_\delta(i) - b_\delta(i)| \le M\sqrt{\frac{1}{p}\sum_{i=1}^{p}\Big[\hat{u}_i - b_i\Big]^2},\end{aligned}$$

which converges to zero in probability under the posterior distribution $\mu_{\mathbf{y},\mathbf{X}}(\cdot)$. Here the last estimate uses part (ii) of this Theorem.

To complete the argument, it thus remains to verify the claim involving (25), for which it suffices to verify continuity of the function $(x, y) \mapsto H(x, y)$ at $x = \pm 1$, uniformly for $y \in [-K, K]$. By symmetry, it suffices to show that

$$\lim_{\delta \to 0} \sup_{x \in [1-\delta, 1], |y| \leq K} |H(x, y) - 1| \to 0.$$

Suppose this is not true. Then there exists sequences $\{x_k\}_{k \geq 1}, \{y_k\}_{k \geq 1}$ with $\lim_{k \to \infty} x_k = 1$, and $|y_k| \leq K$, such that $|H(x_k, y_k) - 1| > \varepsilon$ for all k, for some $\varepsilon > 0$. Without loss of generality, by passing to a subsequence, we can assume that y_k converges to $y \in [-K, K]$. Using Lemma 30 Part (iii) we have that $\pi_{(h(x_k,y_k),y_k)}$ converges weakly to δ_1, the point mass at 1. Consequently, using DCT we have

$$H(x_k, y_k) = \int_{[-1,1]} x^r d\pi_{(h(x_k,y_k),y_k)}(z) \to 1,$$

a contradiction. This verifies the claim, and hence completes the proof of part (iii).

■

Next we turn to the proof of Lemmas 31 and 32.

Proof of Lemma 31 Recall the notation

$$f(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2} \boldsymbol{\beta}^T A \boldsymbol{\beta} + \frac{1}{\sigma^2} \mathbf{z}^T \boldsymbol{\beta} = -\frac{1}{2\sigma^2} \sum_{i,j=1}^{p} A_p(i, j) \beta_i \beta_j + \frac{1}{\sigma^2} \sum_{i=1}^{p} z_i \beta_i$$

from the proof of Theorem 7 part (ii). Set

$$\widetilde{f}(\boldsymbol{\beta}) := -\frac{1}{2\sigma^2} \boldsymbol{\beta}^T A \boldsymbol{\beta} = -\frac{1}{2\sigma^2} \sum_{i,j=1}^{p} A_p(i, j) \beta_i \beta_j,$$

and note that we are in the setting of proof of (Yan, 2020, Thm 4), via the following connections:

$$X \leftrightarrow \boldsymbol{\beta},\ \hat{X} \leftrightarrow \mathbf{b},\ N \leftrightarrow 1,\ J_{N \times N} \leftrightarrow -\frac{1}{2\sigma^2},\ n \leftrightarrow p,\ A_n \leftrightarrow A_p,\ \widetilde{f}_i(x) \leftrightarrow m_i(\boldsymbol{\beta}),\ h \leftrightarrow \mathbf{z}.$$

The only disparity in the above connection is that Yan (2020) works with h which is free of i, whereas in our setup we have $h_i = z_i$ which varies with i. However, the proof of (Yan, 2020, Theorem 4) goes through verbatim under this more general choice of the linear term, as long as we have the mean field assumption $\mathrm{tr}(A^2) = o(p)$. Thus, (Yan, 2020, Theorem 4) gives

$$\log Z_p - \sup_{Q = \prod_{i=1}^{p} Q_i} \Big[-\frac{1}{2\sigma^2} (\mathbb{E}_Q[\mathbf{X}])^{\mathrm{T}} A (\mathbb{E}_Q[\mathbf{X}]) + \frac{1}{\sigma^2} \mathbf{z}^{\mathrm{T}} \mathbb{E}_Q[\mathbf{X}] - \sum_{i=1}^{p} D(Q_i \| \pi_i) \Big] = o(p),$$

where $\mathbb{E}_Q[\mathbf{X}]$ is the mean vector of $\mathbf{X} \sim Q = \prod_{i=1}^p Q_i$. The first conclusion (18) then follows upon using Lemma 6.
Also (19) and (20) follow from (Yan, 2020, (4.40)) and (Yan, 2020, (4.41)) respectively, by using the connections mentioned above.
It thus suffices to prove (21). But this follows upon observing that

$$\Big|f(\boldsymbol{\beta}) - f(\mathbf{b}) - \sum_{i=1}^p \theta_i(\beta_i - b_i)\Big| \le \frac{1}{\sigma^2}\Big|\sum_{i=1}^p m_i(\boldsymbol{\beta})(\beta_i - b_i)\Big| + \frac{1}{2\sigma^2}|\boldsymbol{\beta}^{\mathrm{T}} A\boldsymbol{\beta} - \mathbf{b}^{\mathrm{T}} A\mathbf{b}|,$$

and using (19) and (20).

■

Proof of Lemma 32

For any $i, j \in [p]$ with $i \neq j$, setting

$$t_i := \mathbb{E}[\phi(\beta_i)|\beta_k, k \neq i] = \int_{[-1,1]} \phi(z) d\pi_{(\theta_i, d_i)}(z), \quad \theta_i := \frac{z_i - \sum_{k=1}^p A_{ik}\beta_k}{\sigma^2}$$

we have

$$\mathbb{E}_\mu\Big[\phi(\beta_i) - t_i\Big]\Big[\phi(\beta_j) - t_j\Big] = \mathbb{E}_\mu\Big[\phi(\beta_i) - t_i\Big]\Big[\phi(\beta_j) - t_j^{(i)}\Big] + \mathbb{E}_\mu\Big[\phi(\beta_i) - t_i\Big]\Big[t_j^i - t_j\Big] \quad (26)$$

where

$$t_j^{(i)} := \int_{[-1,1]} \phi(z) d\pi_{(\theta_j^i, d_j)}(z), \quad \theta_j^{(i)} := \frac{z_j - \sum_{k\neq i} A_{jk}\beta_k}{\sigma^2}.$$

Since $t_j^{(i)}$ does not depend on β_i, we have

$$\mathbb{E}_\mu\Big[\phi(\beta_i) - t_i\Big]\Big[\phi(\beta_j) - t_j^{(i)}\Big] = \mathbb{E}_\mu\Big(\Big[\phi(\beta_j) - t_j^{(i)}\Big]\mathbb{E}_\mu\Big[\phi(\beta_i) - t_i\Big|\beta_k, k \neq i\Big]\Big) = 0.$$

For estimating the second term in the RHS of (26), setting

$$\ell(x) := \int_{[-1,1]} \phi(z) d\pi_{(x, d_i)}(z)$$

a Taylor's series expansion gives

$$t_j^{(i)} - t_j = \ell(\theta_j^{(i)}) - \ell(\theta_j) = -\frac{A_{ij}}{\sigma^2}\beta_i \ell'(\theta_j) + \frac{A_{ij}^2}{2\sigma^4}\beta_i^2 \ell''(\xi_j^{(i)}).$$

Using the last two displays and summing over i, j give

$$\begin{aligned}
&\Big|\sum_{i\neq j} c_i c_j \mathbb{E}_\mu\Big[\phi(\beta_i) - t_i\Big]\Big[\phi(\beta_j) - t_j\Big]\Big| \\
\le&\mathbb{E}_\mu\left[\left|\sum_{i\neq j} c_i c_j \frac{A_{ij}}{\sigma^2}\beta_i \ell'(\theta_j)\Big[\phi(\beta_i) - t_i\Big]\right|\right] + \mathbb{E}_\mu\left[\left|\sum_{i,j=1}^p c_i c_j \frac{A_{ij}^2}{2\sigma^4}\beta_i^2 \ell''(\xi_j^{(i)})\right|\right] \\
\le&\frac{2\|\phi\|_\infty}{\sigma^2} \sup_{\mathbf{u}\in[-1,1]^p} \sum_{i=1}^p \Big|\sum_{j=1}^p A_{ij}u_j\Big| + \frac{1}{2\sigma^4}\sum_{i,j=1}^p A_{ij}^2,
\end{aligned}$$

where the last estimate uses the fact that $\|\ell^{(r)}\|_\infty \leq 1$ for $r = 1, 2$. The desired conclusion follows upon using the given hypothesis on A_p.

■

4.2 Proof of Theorem 17

We start with a proof of Lemma 15.

Proof of Lemma 15 Without loss of generality assume that the samples $(Z_1, \cdots, Z_p)$ are arranged in increasing order of variance, i.e. increasing order of $\{D_p(i,i)\}_{i=1}^p$. Let $\{\delta_p : p \geq 1\}$ be a positive sequence converging to 0, such that

$$\frac{1}{\delta_p^4} \sum_{i,j=1}^p A_p(i,j)^2 = o(p).$$

The existence of such a sequence follows from the assumption (5). Let

$$q := \arg\min\{i \in [p] : D_p(i,i) \geq \delta_p\}, \quad r := \arg\max\{i \in [p] : D_p(i,i) \leq \frac{1}{\delta_p}\}.$$

Recall that if $Z \sim \mathcal{N}(0,1)$, $\mathbb{E}|Z| = \sqrt{\frac{2}{\pi}}$. This implies

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{u}\in[-1,1]^p} |\sum_{i=1}^{q-1} Z_i u_i| \leq \mathbb{E} \sum_{i=1}^{q-1} |Z_i| = \sqrt{\frac{2}{\pi}} \sum_{i=1}^{q-1} \sqrt{D_p(i,i)} \leq \sqrt{\frac{2\delta_p}{\pi}} p = o(p), \\
&\mathbb{E} \sup_{\mathbf{u}\in[-1,1]^p} |\sum_{i=r+1}^{p} Z_i u_i| \leq \mathbb{E} \sum_{i=r+1}^{p} |Z_i| \leq \sqrt{\frac{2}{\pi}} \sum_{i=1}^{p} \sqrt{D_p(i,i)} \mathbf{1}\Big(D_p(i,i) \geq \frac{1}{\delta_p}\Big) \qquad (27) \\
&\leq \sqrt{\frac{2\delta_p}{\pi}} \sum_{i=1}^{p} D_p(i,i) = o(p),
\end{aligned}$$

where we use $\sum_{i=1}^p D_p(i,i) = O(p)$. For $q \leq i \leq r$, setting $Y_i := \frac{Z_i}{\sqrt{D_p(i,i)}}$, let Γ_p denote the $(r-q+1 \times r-q+1)$ dimensional covariance matrix of the vector $\mathbf{Y} := (Y_q, \cdots, Y_r)^\top$. Define $\mathcal{E} = (\xi_q, \cdots, \xi_r)^\top := \Gamma_p^{-1/2}\mathbf{Y}$, and note that $(\xi_q, \cdots, \xi_r) \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Setting $v_i := u_i\sqrt{D_p(i,i)}$ and $\mathbf{v} := (v_q, \cdots, v_r)$, we have

$$\sum_{i=q}^r u_i Z_i = \sum_{i=q}^r v_i Y_i = \mathbf{v}^\top \Gamma^{1/2} \mathcal{E} = \mathbf{v}^\top \mathcal{E} + \mathbf{v}^\top B\mathcal{E} = \sum_{i=q}^r \xi_i \sqrt{D_p(i,i)} u_i + \mathbf{v}^\top B\mathcal{E}, \qquad (28)$$

where $B := \Gamma_p^{1/2} - \mathbf{I}_{r-q+1}$. We now claim that

$$\frac{1}{p} \sup_{\mathbf{v}\in\frac{1}{\delta_p}[-1,1]^{r-q+1}} \mathbf{v}^\top B\mathcal{E} \overset{P}{\to} 0. \qquad (29)$$

Given (29), it follows from (28) that

$$\frac{1}{p}\sup_{\mathbf{u}\in[-1,1]^{r-q+1}}\Big|\sum_{i=q}^{r}u_iZ_i-\sum_{i=q}^{r}u_i\sqrt{D_p(i,i)}\xi_i\Big|\xrightarrow{P}0. \tag{30}$$

Finally with $\{\xi_i\}_{[p]\setminus[q,r]}\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$ independent of $\{\xi_i\}_{i=q}^r$, using arguments similar to the derivation of (27) we get

$$\begin{aligned}\mathbb{E}\sup_{\mathbf{u}\in[-1,1]^p}\Big|\sum_{i=1}^{q-1}u_i\sqrt{D(i,i)}\xi_i\Big|&\le\sqrt{\frac{2\delta_p}{\pi}}p=o(p)\\ \mathbb{E}\sup_{\mathbf{u}\in[-1,1]^p}\Big|\sum_{i=r+1}^{p}u_i\sqrt{D(i,i)}\xi_i\Big|&\le\sqrt{\frac{2\delta_p}{\pi}}\sum_{i=1}^{p}D_p(i,i)=o(p).\end{aligned} \tag{31}$$

Combining (27), (30) and (31) the desired conclusion follows.

It thus remains to verify (29). To this effect, note that $\Gamma_p(i,i)=1$, and $\Gamma_p(i,j)=\frac{A_p(i,j)}{\sqrt{D_p(i,i)D_p(j,j)}}$. Setting $C_\delta(i,j):=\Gamma_p(i,j)1\{i\neq j\}$, and using Spectral Theorem, we have,

$$C_\delta=\sum_{\ell=q}^{r}\lambda_\ell\psi_\ell\psi_\ell^\top=\Psi\Lambda\Psi^\top.$$

Observe that Γ_p is positive semidefinite, and has eigenvalues $1+\lambda_\ell$ with $\lambda_\ell\ge-1$. Then we have

$$\sum_{\ell=q}^{r}\lambda_\ell^2=\sum_{i,j=1}^{p}C_\delta(i,j)^2\le\frac{1}{\delta_p^2}\sum_{i,j=1}^{p}A_p(i,j)^2. \tag{32}$$

Finally, the matrix B can expressed as

$$B=(\mathbf{I}_{r-q+1}+C_\delta)^{1/2}-\mathbf{I}_{r-q+1}=\sum_{\ell=q}^{r}\mu_\ell\psi_\ell\psi_\ell^\top=:\Psi\widetilde{\Lambda}\Psi^\top,$$

where $\mu_\ell:=\sqrt{1+\lambda_\ell}-1$, and $\widetilde{\Lambda}$ is a diagonal matrix with entries $\{\mu_\ell\}_{\ell=q}^r$. Consequently,

$$\mathbb{E}\sup_{\mathbf{v}\in\frac{1}{\delta_p}[-1,1]^{r-q+1}}|\mathbf{v}^\top B\mathcal{E}|=\frac{1}{\delta_p}\mathbb{E}\|B\mathcal{E}\|_1\le\frac{\sqrt{p}}{\delta_p}\mathbb{E}\|B\mathcal{E}\|_2=\frac{\sqrt{p}}{\delta_p}\mathbb{E}\|\Psi\widetilde{\Lambda}\mathcal{F}\|_2=\frac{\sqrt{p}}{\delta_p}\mathbb{E}\sqrt{\sum_{\ell=q}^{r}\mu_\ell^2\mathcal{F}_\ell^2},$$

where $\mathcal{F}_\ell:=\psi_\ell^\top\mathcal{E}$ are i.i.d. $\mathcal{N}(0,1)$ for $q\le\ell\le r$. By Jensen's inequality, the last expression is bounded by

$$\frac{\sqrt{p}}{\delta_p}\sqrt{\sum_{\ell=q}^{r}\mu_\ell^2}\le C\frac{\sqrt{p}}{\delta_p}\sqrt{\sum_{\ell=q}^{r}\lambda_\ell^2}=C\frac{\sqrt{p}}{\delta_p}\sqrt{\sum_{i,j=1}^{p}C_\delta(i,j)^2}\le C\frac{\sqrt{p}}{\delta_p^2}\sqrt{\sum_{i,j=1}^{p}A_p(i,j)^2}$$

where $C := \sup_{x \geq -1} \left| \frac{\sqrt{1+x}-1}{x} \right| < \infty$, and the last bound uses (32). The last term above is $o(p)$ by the choice of δ_p, and so we have verified (29). This completes the proof of the Lemma. ■

To establish Theorem 17, we need the following definitions.

Definition 33 *Let $\xi_1, \cdots, \xi_p \sim \mathcal{N}(0,1)$ be iid random variables obtained from Lemma 15. Let $X \sim U([0,1])$ be independent of the ξ_i variables. Given $\mathbf{u} \in [-1,1]^p$ set $Z = \xi_i$ and $U = u_i$ if $X \in [\frac{(i-1)}{p}, \frac{i}{p}]$. Define $\tilde{L}_p^{(\mathbf{u})}$ to be the joint distribution of (X, Z, U).*

Fix $W \in \mathcal{W}$, and ϕ, ψ are L^1 functions on $[0,1]$. Let $\tilde{\mathscr{F}}_{2,4}$ denote the space of all joint distributions $(X, Z, U) \sim \nu$ such that $X \sim U([0,1])$, $\mathbb{E}_\nu[Z^2] \leq 4$, $|U| \leq 1$. Define the functional $\tilde{\mathcal{G}}_{W,\phi,\psi} : \tilde{\mathscr{F}}_{2,4} \to \mathbb{R} \cup \{-\infty\}$ such that

$$\tilde{\mathcal{G}}_{W,\phi,\psi}(\nu) = \frac{1}{\sigma^2}\Big[-\frac{1}{2}\mathbb{E}[W(X_1,X_2)U_1U_2] + \mathbb{E}[\phi(X)U] + \mathbb{E}[\sqrt{\psi(X)}UZ]\Big] - \mathbb{E}\Big[G\Big(U, \frac{\psi(X)}{\sigma^2}\Big)\Big], \quad (33)$$

where $(X_1, Z_1, U_1), (X_2, Z_2, U_2)$ are iid copies from ν. Finally, let $\tilde{\mathscr{F}}$ denote the space of all probability measures ν on $\mathbb{R}^3$, such that if $(X, Z, U) \sim \nu$, then we have

$$X \sim U([0,1]), Z \sim \mathcal{N}(0,1), X \perp\!\!\!\perp Z, |U| \leq 1 \text{ a.s.} \quad (34)$$

We note that $\tilde{\mathscr{F}} \subset \tilde{\mathscr{F}}_{2,4}$, an observation that will be helpful in our subsequent analysis. The following stability estimates will be crucial in our proof of Theorem 17. The proof is deferred to the Appendix.

Lemma 34 *(i) We have, for $W, W' \in \mathcal{W}$, $\phi, \psi, \phi', \psi' \in L^1([0,1])$,*

$$\sup_{\nu \in \tilde{\mathscr{F}}_{2,4}} |\tilde{\mathcal{G}}_{W,\phi,\psi}(\nu) - \tilde{\mathcal{G}}_{W',\phi',\psi'}(\nu)| \lesssim \|W - W'\|_\square + \|\phi - \phi'\|_1 + \|\psi - \psi'\|_1.$$

(ii) Suppose the following assumptions hold:

(a) $W_k, W \in \mathcal{W}$ is such that $d_\square(W_k, W) \to 0$.
(b) $\phi_k, \phi \in \mathcal{L}$ is such that $\int_{[0,1]} |\phi_k(x) - \phi(x)|dx \to 0$.
(c) $\psi_k, \psi \in L_1[0,1]$ is such that $\int_{[0,1]} |\psi_k(x) - \psi(x)|dx \to 0$.

Then we have

$$\sup_{\nu \in \tilde{\mathscr{F}}_{2,4}} |\tilde{\mathcal{G}}_{W_k, W_k\cdot\phi_k + \phi_k\psi_k, \psi_k}(\nu) - \tilde{\mathcal{G}}_{W, W\cdot\phi+\phi\psi,\psi}(\nu)| \to 0 \quad (35)$$

as $k \to \infty$.

Lemma 35 *We have, for any $W \in \mathcal{W}$, $g, \psi \in L^1([0,1])$,*

$$\sup_{\nu \in \tilde{\mathscr{F}}} \tilde{\mathcal{G}}_{W,g,\psi}(\nu) = \sup_{F \in \mathscr{F}} \mathcal{G}_{W,g,\psi}(F).$$

Further, both the suprema are attained.

Remark 36 *Depending on whether $D_{\mathrm{KL}}(\pi_\infty\|\pi)$ and $D_{\mathrm{KL}}(\pi_{-\infty}\|\pi)$ are infinity or not, there are four possible cases. In our subsequent proofs, we consider the case $D_{\mathrm{KL}}(\pi_\infty\|\pi) < \infty$ and $D_{\mathrm{KL}}(\pi_{-\infty}\|\pi) = \infty$, noting that other cases follow by natural modifications.*

Lemma 37 *Let $d(\cdot,\cdot)$ be the 2-Wasserstein distance on $\Pr([0,1]\times\mathbb{R}\times[-1,1])$. For any $\delta > 0$, set*

$$\mathscr{B}_d(\tilde{\mathscr{F}},\delta) = \{\nu \in \Pr([0,1]\times\mathbb{R}\times[-1,1]) : \inf_{\nu'\in\tilde{\mathscr{F}}} d(\nu,\nu') < \delta\}.$$

Then there exists a sequence $\delta_p \to 0$ as $p\to\infty$ such that setting

$$F_p = \{\forall \mathbf{u}\in[-1,1]^p,\ \ \tilde{L}_p^{(\mathbf{u})} \in \mathscr{B}_d(\tilde{\mathscr{F}},\delta_p)\}$$

we have, $\mathbb{P}(F_p|\mathbf{X}) = 1 - o(1)$.

The proof of Lemma 37 is straightforward, and is thus omitted.

Lemma 38 *Let $\{\mathbf{u}_p : p\ge 1\} \in [-1,1]^p$ be such that $\tilde{L}_p^{(\mathbf{u}_p)}$ converges weakly to $\nu_0 \in \tilde{\mathscr{F}}$. Then for any $W\in\mathcal{W}$, $g,\psi\in L^1$,*

$$\limsup_{p\to\infty} \tilde{\mathcal{G}}_{W,g,\psi}(\tilde{L}_p^{(\mathbf{u}_p)}) \le \tilde{\mathcal{G}}_{W,g,\psi}(\nu_0).$$

Lemma 39 *Suppose we are in the setting of Theorem 17. Fix $F\in\mathscr{F}$ and $\delta, K > 0$ and $p\ge 1$. Let $V_i \sim U([(i-1)/p, i/p])$ be independent of ξ_i arising in Lemma 15. Set $d_i^K = d_i\mathbf{1}(|d_i|\le K)$ and define*

$$\tilde{u}_i = \begin{cases} F(V_i,\xi_i) & \text{if } F(V_i,\xi_i) \ge -1+\delta, \\ \dot{c}(0,d_i^K) & o.w. \end{cases} \tag{36}$$

Then we have, setting $\mathbf{u} = (\tilde{u}_i)$, for any $\varepsilon > 0$

$$\limsup_{K\to\infty}\limsup_{\delta\to 0}\limsup_{p\to\infty} \mathbb{P}\Big[\Big|\frac{1}{p}\widetilde{M}_p(\mathbf{u}) - \mathcal{G}_{W,g,\psi}(F)\Big| > \varepsilon|\mathbf{X}\Big] = 0.$$

We prove Theorem 17 assuming Lemma 35, 38 and 39. The corresponding proofs are deferred to the end of the section.

Proof of Theorem 17 Using Lemma 15, it suffices to prove that

$$\frac{1}{p}\sup_{\mathbf{u}\in[-1,1]^p} \widetilde{M}_p(\mathbf{u}) \overset{P|\mathbf{X}}{\to} \sup_{F\in\mathscr{F}} \mathcal{G}_{W,g,\psi}(F).$$

Upper bound: Note that

$$\begin{aligned}\frac{1}{p}\widetilde{M}_p(\mathbf{u}) &\overset{(b)}{=} \tilde{\mathcal{G}}_{W_{pA_p},W_{pA_p}\cdot w_{\boldsymbol{\beta}_0}+w_{\boldsymbol{\beta}_0}w_{\mathbf{D}},w_{\sigma^2\mathbf{d}}}(\tilde{L}_p^{(\mathbf{u})}) \\ &\overset{(c)}{=} \tilde{\mathcal{G}}_{W,g,\psi}(\tilde{L}_p^{(\mathbf{u})}) + \mathcal{E}_p^{(2)}(\mathbf{u}),\end{aligned}$$

where $\sup_{\mathbf{u}\in[-1,1]^p} |\mathcal{E}_p^{(2)}(\mathbf{u})| \overset{P|\mathbf{X}}{\to} 0$ as $p \to \infty$. Here, (b) uses the definition (33), and (c) uses Lemma 34 part (ii). To invoke Lemma 34, we use the fact that

$$\mathbb{P}[\forall \mathbf{u} \in [-1,1]^p, \tilde{L}_p^{(\mathbf{u})} \in \tilde{\mathscr{F}}_{2,4}] = \mathbb{P}\Big[\sum_{i=1}^p \xi_i^2 \le 4p\Big] \to 1 \tag{37}$$

as $p \to \infty$.

To complete the upper bound, invoking Lemma 35, it suffices to establish that for all $\varepsilon > 0$

$$\mathbb{P}\Big[\sup_{\mathbf{u}\in[-1,1]^p} \tilde{\mathcal{G}}_{W,g,\psi}(\tilde{L}_p^{(\mathbf{u})}) < \sup_{\nu\in\tilde{\mathscr{F}}} \tilde{\mathcal{G}}_{W,g,\psi}(\nu) + \varepsilon|\mathbf{X}\Big] = 1 - o(1). \tag{38}$$

We first turn to the proof of (38) by contradiction. Suppose there exists $\varepsilon > 0$, such that setting

$$E_p = \Big\{ \sup_{\mathbf{u}\in[-1,1]^p} \tilde{\mathcal{G}}_{W,g,\psi}(\tilde{L}_p^{(\mathbf{u})}) > \sup_{\nu\in\tilde{\mathscr{F}}} \tilde{\mathcal{G}}_{W,g,\psi}(\nu) + \varepsilon \Big\}.$$

we have $\limsup \mathbb{P}[E_p|\mathbf{X}] > 0$. Using Lemma 37, we have, $\limsup \mathbb{P}[F_p \cap E_p] > 0$. Therefore, there exists a subsequence $p_k \to \infty$ along which $E_{p_k} \cap F_{p_k} \neq \emptyset$.

Consequently, there exists $\boldsymbol{\xi}_{p_k} := (\xi_{1,p_k}, \cdots, \xi_{p_k,p_k}) \in \mathbb{R}^{p_k}$, $\boldsymbol{\xi}_{p_k} \in E_{p_k} \cap F_{p_k}$ and $\mathbf{u}_{p_k} \in [-1,1]^{p_k}$ such that

$$\tilde{\mathcal{G}}_{W,g,\psi}(\tilde{L}_p^{(\mathbf{u}_{p_k})}) > \sup_{\nu\in\tilde{\mathscr{F}}} \tilde{\mathcal{G}}_{W,g,\psi}(\nu) + \varepsilon, \quad \tilde{L}_{p_k}^{(\mathbf{u}_{p_k})} \in \mathscr{B}_d(\tilde{\mathscr{F}}, \delta_p).$$

This in turn implies that the sequence $\{\tilde{L}_{p_k}^{(\mathbf{u}_{p_k})} : k \ge 1\}$ is tight, and hence has a subsequence converging weakly to $\nu_0 \in \tilde{\mathscr{F}}$. Assuming without loss of generality that $\tilde{L}_{p_k}^{(\mathbf{u}_{p_k})} \overset{d}{\to} \nu_0$, the desired contradiction follows once we show that

$$\limsup_{k\to\infty} \tilde{\mathcal{G}}_{W,g,\psi}(\tilde{L}_p^{(\mathbf{u}_{p_k})}) \le \tilde{\mathcal{G}}_{W,g,\psi}(\nu_0). \tag{39}$$

But this follows from Lemma 38.

Lower Bound: It suffices to show that for all $\varepsilon > 0$,

$$\mathbb{P}\Big[\frac{1}{p}\sup_{\mathbf{u}\in[-1,1]^p} \widetilde{M}_p(\mathbf{u}) > \sup_{F\in\mathscr{F}} \mathcal{G}_{W,g,\psi}(F) - \varepsilon|\mathbf{X}\Big] = 1 - o(1).$$

To this end, let $F \in \mathscr{F}$ satisfy

$$\mathcal{G}_{W,g,\psi}(F) > \sup_{F\in\mathscr{F}} \mathcal{G}_{W,g,\psi}(F) - \frac{\varepsilon}{2}.$$

Note that if $\mathbb{E}[G(F(X,Z),\psi(X)] = \infty$, then the lower bound is trivial. Thus we assume, without loss of generality, that $\mathbb{E}[G(F(X,Z),\psi(X)] < \infty$. Next, fixing $K > 0$, we have,

$$\mathcal{G}_{W,g,\psi_K}(F) = \mathcal{G}_{W,g,\psi}(F) + o_K(1), \tag{40}$$

where $\psi_K(x) := \psi(x)\mathbf{1}(\psi(x) \le K)$, and we have used Lemma 34 Part (i). Further, setting $d_i^K = \min\{d_i, K\}$, a similar argument gives

$$\begin{aligned}&\tilde{\mathcal{G}}_{W_{pA_p}, W_{pA_p}\cdot w_{\boldsymbol{\beta}_0} + w_{\boldsymbol{\beta}_0} w_{\mathbf{D}}, w_{\sigma^2\mathbf{d}}}(\tilde{L}_p^{(\mathbf{u})}) \\ &= \tilde{\mathcal{G}}_{W_{pA_p}, W_{pA_p}\cdot w_{\boldsymbol{\beta}_0} + w_{\boldsymbol{\beta}_0} w_{\mathbf{D}}, w_{\sigma^2\mathbf{d}^K}}(\tilde{L}_p^{(\mathbf{u})}) + o_K(1). \end{aligned} \tag{41}$$

For any $p \ge 1$, let $\xi_1, \cdots, \xi_p$ denote the iid $\mathcal{N}(0,1)$ random variables arising in the optimization problem M_p. For each $i \in [p]$, generate $V_i \sim U([(i-1)/p, i/p])$ independent of each other, and of the ξ_i variables. Fixing $\delta > 0$, we set

$$\tilde{u}_i = \begin{cases} F(V_i, \xi_i) & \text{if } F(V_i, \xi_i) \ge -1 + \delta, \\ \dot{c}(0, d_i^K) & \text{o.w.} \end{cases} \tag{42}$$

Using Lemma 39, for any $\varepsilon > 0$, there exists $\delta, K > 0$ (depending on ε), such that with probability at least $1 - \varepsilon$,

$$\frac{1}{p} \sup_{\mathbf{u} \in [-1,1]^p} \widetilde{M}_p(\mathbf{u}) \ge \frac{1}{p} \widetilde{M}_p(\tilde{\mathbf{u}}) \ge \mathcal{G}_{W,g,\psi}(F) - \varepsilon \ge \sup_{F \in \mathscr{F}} \mathcal{G}_{W,g,\psi}(F) - 2\varepsilon.$$

As $\varepsilon > 0$ is arbitrary, this completes the proof of the lower bound. ∎

It remains to prove Lemma 35, 38 and 39. We establish each result in turn.

Proof of Lemma 35 To begin, note that for $F \in \mathscr{F}$, $(X, Z, F(X,Z)) \sim \nu \in \tilde{\mathscr{F}}$. Therefore,

$$\sup_{\nu \in \tilde{\mathscr{F}}} \tilde{\mathcal{G}}_{W,g,\psi}(\nu) \ge \sup_{F \in \mathscr{F}} \mathcal{G}_{W,g,\psi}(F).$$

We now turn to the reverse inequality. Fix $(X, Z, U) \sim \nu \in \tilde{\mathscr{F}}$. Setting $V := \Phi(Z)$ note that X, V are $U([0,1])$ iid. Define $F(X,V) = \mathbb{E}[U|X,V]$. Note that

$$\begin{aligned} \mathbb{E}[W(X,X')UU'] &= \mathbb{E}[W(X,X')F(X,V)F(X',V')] \\ \mathbb{E}[g(X)U] &= \mathbb{E}[g(X)F(X,V)], \\ \mathbb{E}\Big[G\Big(U, \frac{\psi(X)}{\sigma^2}\Big)\Big] &\ge \mathbb{E}\Big[G\Big(F(X,V), \frac{\psi(X)}{\sigma^2}\Big)\Big], \end{aligned}$$

where the final step follows from Jensen's inequality and the observation that $\frac{\partial^2}{\partial u^2} G(u,d) \ge 0$ from Lemma 4. This implies

$$\begin{aligned} \tilde{\mathcal{G}}_{W,g,\psi}(\nu) \le &- \frac{1}{2\sigma^2}\mathbb{E}[W(X,X')F(X,V)F(X,V')] + \frac{1}{\sigma^2}\mathbb{E}[g(X)F(X,V)] \\ &+ \frac{1}{\sigma^2}\mathbb{E}[\sqrt{\psi(X)}F(X,V)\Phi^{-1}(V)]\Big] - \mathbb{E}\Big[G\Big(F(X,V), \frac{\psi(X)}{\sigma^2}\Big)\Big]. \end{aligned}$$

Finally, noting that Φ^{-1} is strictly increasing, we have,

$$\mathbb{E}[\sqrt{\psi(X)}F(X,V)\Phi^{-1}(V)] \le \mathbb{E}[\sqrt{\psi(X)}F^*(X,V)\Phi^{-1}(V)],$$

where for each $x \in [0,1]$, $F^*(x,\cdot)$ is the monotone rearrangment of $F(x,\cdot)$. This last step follows from Hardy-Littlewood inequality. Observe that as the Uniform distribution is invariant under measure preserving transformations, the other terms in the functional above are unchanged by this rearrangement. Thus

$$\tilde{\mathcal{G}}_{W,g,\psi}(\nu) \leq \mathcal{G}_{W,g,\psi}(F^*).$$

Since $F^* \in \mathscr{F}$, the proof is complete.

Finally, we show that both suprema are attained. To this end, observe that by Lemma 30 Part (iv), $\tilde{\mathcal{G}}_{W,g,\psi}(\cdot)$ is upper semi-continuous. Moreover, $\tilde{\mathscr{F}}$ is compact under weak topology—this follows as the collection $\tilde{\mathscr{F}}$ is tight, and thus pre-compact by Prokhorov's theorem. Thus a maximum exists. To see that $\mathcal{G}_{W,\phi,\psi}(\cdot)$ attains it's maximum, start with a maximizer ν_0 of $\tilde{\mathcal{G}}_{W,\phi,\psi}(\cdot)$, and take F_0^* as obtained in the proof above. ■

Proof of Lemma 38 Observe that we can approximate W in L^1 by a continuous function $W^{(\ell)}$. This implies

$$\begin{aligned}\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[W(X_1,X_2)U_1U_2] &= \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[W^{(\ell)}(X_1,X_2)U_1U_2] + o_\ell(1) \\ &\overset{p\to\infty}{\to} \mathbb{E}_{\nu_0}[W^{(\ell)}(X_1,X_2)U_1U_2] + o_\ell(1) = \mathbb{E}_{\nu_0}[W(X_1,X_2)U_1U_2] + o_\ell(1).\end{aligned}$$

Upon sending $\ell \to \infty$, we conclude that $\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[W(X_1,X_2)U_1U_2] \to \mathbb{E}_{\nu_0}[W(X_1,X_2)U_1U_2]$. A similar argument shows that

$$\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[g(X)U] \overset{p\to\infty}{\to} \mathbb{E}_{\nu_0}[g(X)U].$$

Next, we note that we can approximate ψ in L^1 by a sequence of continuous functions $\{\psi^{(\ell)} : \ell \geq 1\}$, and thus, by Cauchy-Schwarz

$$\begin{aligned}|\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[\sqrt{\psi(X)}UZ] - \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[\sqrt{\psi^{(\ell)}(X)}UZ]| &\leq \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[|\sqrt{\psi}^{(\ell)}(X) - \sqrt{\psi(X)}| \cdot |Z|] \\ &\leq \sqrt{\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[(\sqrt{\psi^{(\ell)}(X)} - \sqrt{\psi(X)})^2]\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[Z^2]} \\ &\leq \sqrt{2\int_0^1 |\psi^{(\ell)}(x) - \psi(x)|\mathrm{d}x} = o_\ell(1).\end{aligned}$$

The last inequality uses that $\tilde{L}_p^{(\mathbf{u}_p)} \in \tilde{\mathscr{F}}_{2,4}$. This implies

$$\begin{aligned}\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[\sqrt{\psi(X)}UZ] &= \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[\sqrt{\psi^{(\ell)}(X)}UZ] + o_\ell(1) \\ &\overset{p\to\infty}{\to} \mathbb{E}_{\nu_0}[\sqrt{\psi^{(\ell)}(X)}UZ] + o_\ell(1) \\ &= \mathbb{E}_{\nu_0}[\sqrt{\psi(X)}UZ] + o_\ell(1).\end{aligned}$$

Sending $\ell \to \infty$, we have $\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[\sqrt{\psi(X)}UZ] \to \mathbb{E}_{\nu_0}[\sqrt{\psi(X)}UZ]$ as $p \to \infty$.

The desired conclusion follows once we establish that

$$\liminf_{p\to\infty} \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}\Big[G\Big(U, \frac{\psi(X)}{\sigma^2}\Big)\Big] \geq \mathbb{E}_{\nu_0}\Big[G\Big(U, \frac{\psi(X)}{\sigma^2}\Big)\Big].$$

But this follows on using Fatou's lemma and the lower semicontinuity of $G\Big(\cdot, \frac{\psi(X)}{\sigma^2}\Big)$ (Lemma 30 part (iv)). ■

Proof of Lemma 39 The definition of $\mathbf{u}$ in (36) implies

$$G(u_i, d_i^K) = \begin{cases} G(F(V_i,\xi_i), d_i^K) & \text{if } F(V_i,\xi_i) \geq -1+\delta, \\ 0 & \text{o.w.} \end{cases}$$

Observe that

$$\frac{1}{p}\sum_{i=1}^{p} G(u_i, d_i^K) = \frac{1}{p}\sum_{i=1}^{p} G(F(V_i,\xi_i), d_i^K)\mathbf{1}(F(V_i,\xi_i) \geq -1+\delta) \tag{43}$$

Now, the function $G : [-1+\delta, 1] \times [0, K] \to \mathbb{R}^+$ is bounded, and thus by Chebychev inequality,

$$\begin{aligned} &\frac{1}{p}\sum_{i=1}^{p} G(F(V_i,\xi_i), d_i^K)\mathbf{1}(F(V_i,\xi_i) \geq -1+\delta) \\ &-\frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big[G(F(V_i,\xi_i), d_i^K)\mathbf{1}(F(V_i,\xi_i) \geq -1+\delta)\Big] \overset{P|\mathbf{X}}{\to} 0. \end{aligned} \tag{44}$$

Also, note that

$$\begin{aligned} &\frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big[G(F(V_i,\xi_i), d_i^K)\mathbf{1}(F(V_i,\xi_i) \geq -1+\delta)\Big] \\ &= \mathbb{E}[G(F(X,Z), w_{\mathbf{d}^K,p}(X))\mathbf{1}(F(X,Z) \geq -1+\delta)], \\ &= \mathbb{E}\Big[G\Big(F(X,Z), \frac{\psi_K(X)}{\sigma^2}\Big)\mathbf{1}(F(X,Z) \geq -1+\delta)\Big] + o(1) \end{aligned} \tag{45}$$

where $X \sim U([0,1])$ and $Z \sim \mathcal{N}(0,1)$ are independent. The last display uses Lemma 4 along with the fact that $w_{\mathbf{d}^K,p} \to \psi_K/\sigma^2$ in L^1 for almost every $K > 0$. Now, the assumptions $\mathbb{E}[G(F(X,Z), \psi(X)/\sigma^2)] < \infty$ and $D_{\mathrm{KL}}(\pi_{-\infty}\|\pi) = \infty$ together imply $\mathbb{P}[F(X,Z) = -1] = 0$. In combination with Dominated Convergence Theorem, this implies

$$\lim_{\delta\to 0} \mathbb{E}\Big[G\Big(F(X,Z), \frac{\psi_K(X)}{\sigma^2}\Big)\mathbf{1}(F(X,Z) \geq -1+\delta)\Big] = \mathbb{E}\Big[G\Big(F(X,Z), \frac{\psi_K(X)}{\sigma^2}\Big)\Big]. \tag{46}$$

Combining (43), (44), (45) and (46), for any $\varepsilon > 0$, we have,

$$\limsup_{\delta\to 0}\limsup_{p\to\infty} \mathbb{P}\Big[\Big|\frac{1}{p}\sum_{i=1}^{p} G(u_i, d_i^K) - \mathbb{E}\Big[G\Big(F(X,Z), \frac{\psi_K(X)}{\sigma^2}\Big)\Big]\Big| > \varepsilon|\mathbf{X}\Big] = 0. \tag{47}$$

We now claim that for any $\varepsilon > 0$,

$$\limsup_{\delta\to 0}\limsup_{p\to\infty}\mathbb{P}\Big[\Big|\frac{1}{p}\sum_{i,j=1}^{p}A_p(i,j)u_iu_j - \mathbb{E}[W(X_1,X_2)F(X_1,Z_1)F(X_2,Z_2)]\Big| > \varepsilon|\mathbf{X}\Big] = 0.$$
$$\limsup_{\delta\to 0}\limsup_{p\to\infty}\mathbb{P}\Big[\Big|\mathbf{u}^{\mathrm{T}}A_p\boldsymbol{\beta}_0 + \mathbf{u}^{\mathrm{T}}D_p\boldsymbol{\beta}_0 - \mathbb{E}[g(X)F(X,Z)]\Big| > \varepsilon|\mathbf{X}\Big] = 0. \tag{48}$$
$$\limsup_{\delta\to 0}\limsup_{p\to\infty}\mathbb{P}\Big[\Big|\sum_{i=1}^{p}u_i\sqrt{D_p(i,i)}\xi_i - \mathbb{E}[\sqrt{\psi(X)}UZ]\Big| > \varepsilon|\mathbf{X}\Big] = 0.$$

We first turn to the quadratic form. We have,

$$\begin{aligned}\frac{1}{p}\sum_{ij=1}^{p}A_p(i,j)u_iu_j &= \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[W_{pA_p}(X_1,X_2)U_1U_2]\\ &\overset{(a)}{=} \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[W(X_1,X_2)U_1U_2] + o(1)\\ &\overset{(b)}{=} \mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[W^{(K)}(X_1,X_2)U_1U_2] + o_K(1).\end{aligned} \tag{49}$$

Here, the step (a) uses Assumption (a), and step (b) uses $W^{(K)} := W\mathbf{1}(|W| \le K) \overset{L^1}{\to} W$. We have, setting $\mathcal{M}_p(i,j) := \int_{[(i-1)/p,i/p]}\int_{[(j-1)/p,j/p]}W^{(K)}(x,y)\mathrm{d}x\mathrm{d}y$,

$$\mathbb{E}_{\tilde{L}_p^{(\mathbf{u})}}[W^{(K)}(X_1,X_2)U_1U_2] = \sum_{i,j=1}^{p}\mathcal{M}_p(i,j)u_iu_j = \sum_{i,j=1}^{p}\mathcal{M}_p(i,j)\mathbb{E}[u_i]\mathbb{E}[u_j] + o_P(1) \tag{50}$$

by the weak law of large numbers, and the observation $|\mathcal{M}_p(i,j)| \le K/p^2$. Here, the $o_P(1)$ term converges to zero in probability under the joint distribution of $\{(V_i,\xi_i) : 1 \le i \le p\}$. Further, we have,

$$\begin{aligned}&\Big|\sum_{i,j=1}^{p}\mathcal{M}_p(i,j)\mathbb{E}[u_i]\mathbb{E}[u_j] - \sum_{i,j=1}^{p}\mathcal{M}_p(i,j)\mathbb{E}[F(V_i,Z)]\mathbb{E}[F(V_j,Z)]\Big|\\ &\le 2K\mathbb{P}[F(X,Z) \le -1+\delta],\end{aligned} \tag{51}$$

where $X \sim U([0,1])$ and $Z \sim \mathcal{N}(0,1)$ are independent. Finally,

$$\begin{aligned}\sum_{i,j=1}^{p}\mathcal{M}_p(i,j)\mathbb{E}[F(V_i,Z)]\mathbb{E}[F(V_j,Z)] &= \mathbb{E}[W_{\mathcal{M}_p}(X,X')F(X,Z)F(X',Z')]\\ &\overset{(a)}{=} \mathbb{E}[W^{(K)}(X,X')F(X,Z)F(X',Z')] + o(1)\\ &\overset{(b)}{=} \mathbb{E}[W(X,X')F(X,Z)F(X',Z')] + o_K(1)\end{aligned} \tag{52}$$

where (a) uses $\|W_{\mathcal{M}_p} - W^{(K)}\|_1 \to 0$ as $p \to \infty$ and (b) uses $\|W - W^{(K)}\|_1 \to 0$ as $K \to \infty$. Combining (49), (50), (51), (52), the first conclusion of (48) follows upon sending $p \to \infty$, $\delta \to 0$ and $K \to \infty$.

The other conclusions of (48) follow using analogous arguments, and are thus omitted. ■

4.3 Proof of Theorem 20 and related results

The following continuity statement will be critical for the proof of Theorem 20. The proof is deferred to the Appendix.

Lemma 40 *Define a map* $m : \Pr\Big([0,1] \times \mathbb{R} \times [-1,1]\Big) \times \mathcal{W} \times [0,1] \to \mathbb{R}$, $(\mu, W, x) \mapsto \mathbb{E}_{(X,Z,U)\sim\mu}[W(x, X')U']$.

(i) Fix $x \in [0,1]$, $\mu \in \Pr\Big([0,1] \times \mathbb{R} \times [-1,1]\Big)$, *and* $W, W' \in \mathcal{W}$. *For any* $\varepsilon > 0$, *set* $S(\varepsilon) = \{x : |m(\mu, W, x) - m(\mu, W', x)| > \varepsilon\}$. *For* $X \sim U([0,1])$, *we have,*

$$\mathbb{P}[X \in S(\varepsilon)] \le \frac{2}{\varepsilon}\|W - W'\|_{\square}.$$

(ii) For any $W \in \mathcal{W}$, *if* $\nu_p \xrightarrow{w} \nu$, *then*

$$\sup_{x\in[0,1]} \Big|m(\nu_p, W, x) - m(\nu, W, x)\Big| \to 0$$

as $p \to \infty$.

Proof of Theorem 20 We break the proof into several steps. Note that using Lemma 35, both variational problems attain their suprema.

(i) We establish uniqueness of the maximizer assuming (11). If possible, let F_1 and F_2 be two distinct optimizers of $\mathcal{G}_{W,g,\psi}$ in $\mathscr{F}$. Therefore, using Theorem 17,

$$\frac{1}{p}\sup_{\mathbf{u}\in[-1,1]^p} M_p(\mathbf{u}) \xrightarrow{P|\mathbf{X}} \sup_{F\in\mathscr{F}} \mathcal{G}_{W,g,\psi}(F) = \mathcal{G}_{W,g,\psi}(F_1) = \mathcal{G}_{W,g,\psi}(F_2).$$

Let $\xi_1, \cdots, \xi_p$ be iid $\mathcal{N}(0,1)$ random variables arising in the functional $\widetilde{M}_p(\cdot)$. Further, let $V_i \sim U([(i-1)/p, i/p])$ be independent of the ξ_i variables. Fixing $\delta > 0$, following the recipe in (36), we define two sequences $\mathbf{u}^{(1)}_{p,\delta,K}$, $\mathbf{u}^{(2)}_{p,\delta,K}$ in $[-1,1]^p$ corresponding to F_1, F_2 respectively. Using Lemma 39, for $i = 1, 2$, for $\delta, K > 0$ we have

$$\mathbb{P}\Big[\frac{1}{p}M_p(\mathbf{u}^{(i)}_{p,\delta,K}) > \mathcal{G}_{W,g,\psi}(F_i) - \frac{\delta'}{4}\Big|\mathbf{X}\Big] \ge 1 - \mathcal{E}(\delta, K),$$

where $\limsup_{K\to\infty}\limsup_{\delta\to 0}\mathcal{E}(\delta, K) = 0$.

Moreover, using Theorem 17 and the assumption that F_1 and F_2 are optimizers of the limiting variational problem, we have, for $i = 1, 2$,

$$\mathbb{P}\Big[\frac{1}{p}M_p(\mathbf{u}^*_p) < \mathcal{G}_{W,g,\psi}(F_i) + \frac{\delta'}{4}\Big|\mathbf{X}\Big] = 1 - o(1).$$

On the event $\{\|\mathbf{u}_{p,\delta,K}^{(i)}-\mathbf{u}_p^*\|_2^2 > p\varepsilon, i=1,2\}\cap\{\frac{1}{p}M_p(\mathbf{u}_p^*) < \mathcal{G}_{W,g,\psi}(F_i)+\frac{\delta'}{4}\}$, for $i=1,2$ we have,

$$\frac{1}{p}M_p(\mathbf{u}_{p,\delta,K}^{(i)}) \le \frac{1}{p}M_p(\mathbf{u}_p^*) - \delta' \le \mathcal{G}_{W,g,\psi}(F_i) + \frac{\delta'}{4} - \delta' < \mathcal{G}_{W,g,\psi}(F_i) - \frac{\delta'}{4}.$$

Using (11), this implies, for all p sufficiently large, $\mathbb{P}[\|\mathbf{u}_{p,\delta,K}^{(i)}-\mathbf{u}_p^*\|_2^2 > p\varepsilon, i=1,2|\mathbf{X}] \le 3\mathcal{E}(\delta,K)$. By triangle inequality, $\mathbb{P}[\|\mathbf{u}_{p,\delta,K}^{(1)} - \mathbf{u}_{p,\delta,K}^{(2)}\|_2^2 > 2p\varepsilon|\mathbf{X}] \le 3\mathcal{E}(\delta,K)$. As $\mathbf{u}_{p,\delta,K}^{(i)} \in [-1,1]^p$, we have,

$$\frac{1}{p}\mathbb{E}[\|\mathbf{u}_{p,\delta,K}^{(1)} - \mathbf{u}_{p,\delta,K}^{(2)}\|_2^2] \le 2\varepsilon + 12\mathcal{E}(\delta,K).$$

Thus, with $X\sim U([0,1])$, $Z\sim\mathcal{N}(0,1)$ independent, we have,

$$\begin{aligned}
&\mathbb{E}[(F_1(X,Z)-F_2(X,Z))^2]\\
=&\frac{1}{p}\sum_{i=1}^p \mathbb{E}[(F_1(V_i,\xi_i)-F_2(V_i,\xi_i))^2]\\
\le&\frac{1}{p}\mathbb{E}[\|\mathbf{u}_{p,\delta,K}^{(1)} - \mathbf{u}_{p,\delta,K}^{(2)}\|_2^2] + \sum_{a=1,2}\mathbb{P}[F_a(X,Z)\le -1+\delta]\\
\le&2\varepsilon + 12\mathcal{E}(\delta,K) + \sum_{a=1,2}\mathbb{P}[F_a(X,Z)\le -1+\delta].
\end{aligned}$$

Setting $\delta\to 0$ followed by $K\to\infty$, we obtain that $\mathbb{E}[(F_1(X,Z)-F_2(X,Z))^2]\le 2\varepsilon$. Here we use that F_1, F_2 are optimizers; thus without loss of generality we can assume $\mathbb{E}[G(F_a(X,Z),\psi(X)] < \infty$ for $a=1,2$, which along with Remark 36 implies that $\mathbb{P}[F_a(X,Z)=-1]=0$.

Finally, note that $\varepsilon>0$ is arbitrary, and thus $F_1=F_2$ a.s.. This establishes the desired uniqueness.

(ii) As $L_p^{(\mathbf{u}_p^*)}$ and $\tilde{L}_p^{(\mathbf{u}_p^*)}$ are close in weak topology, it suffices to work with $\tilde{L}_p^{(\mathbf{u}_p^*)}$. With $\tilde{\mathbf{u}}_{p,\delta,K}$ as in (36), for any $\varepsilon>0$, Lemma 39 and (11) imply

$$\limsup_{K\to\infty}\limsup_{\delta\to 0}\limsup_{p\to\infty}\mathbb{P}[\|\tilde{\mathbf{u}}_{p,\delta,K}-\mathbf{u}_p^*\|_2^2 > p\varepsilon|\mathbf{X}] = 0.$$

Recalling the 2-Wasserstein distance $d(\cdot,\cdot)$ from Lemma 37, we have,

$$d(\tilde{L}_p^{(\mathbf{u}_p^*)}, \tilde{L}_p^{(\mathbf{u}_{p,\delta,K})}) \le \frac{1}{p}\|\tilde{\mathbf{u}}_{p,\delta,K}-\mathbf{u}_p^*\|_2^2,$$

and so it suffices to look at $\tilde{L}_p^{(\mathbf{u}_{p,\delta,K})}$. Let $\eta:\mathbb{R}^3\to\mathbb{R}$ be bounded continuous. Then

$$\mathbb{E}_{\tilde{L}_p^{(\mathbf{u}_{p,\delta,K})}}[\eta(X,Z,U)] = \frac{1}{p}\sum_{i=1}^p \eta(V_i,\xi_i,\tilde{u}_{p,\delta,K}(i)) = \frac{1}{p}\sum_{i=1}^p \mathbb{E}[\eta(V_i,\xi_i,\tilde{u}_{p,\delta,K}(i))] + o_P(1),$$

where the last equality follows from the law of large numbers and the boundedness of η. Here, the $o_P(1)$ term converges to zero in probability under the joint distribution of $\{(V_i, \xi_i) : 1 \le i \le p\}$. Finally,

$$\Big|\frac{1}{p}\sum_{i=1}^{p}\mathbb{E}[\eta(V_i,\xi_i,\tilde{u}_{p,\delta,K}(i))] - \mathbb{E}[\eta(X,Z,F(X,Z)\mathbf{1}(F(X,Z)>-1+\delta))]\Big| \\ \le \mathbb{P}[F(X,Z)\le -1+\delta].$$

This gives, for any $\varepsilon > 0$,

$$\limsup_{K\to\infty}\limsup_{\delta\to 0}\limsup_{p\to\infty}\mathbb{P}\Big[\Big|\mathbb{E}_{\tilde{L}_p^{(\mathbf{u}_{p,\delta,K})}}[\eta(X,Z,U)] - \mathbb{E}[\eta(X,Z,F(X,Z))]\Big| > \varepsilon|\mathbf{X}\Big] = 0,$$

where we again use that $\mathbb{P}[F(X,Z) = -1] = 0$. This completes the proof of Part (ii).

(iii) For any $p \ge 1$, the function $\widetilde{M}_p(\cdot)$ is upper semicontinuous on $[-1,1]^p$, and thus attains its maximum. Let $\{\hat{\mathbf{u}}_p : p \ge 1\}$ denote any sequence of global maximizers of $\widetilde{M}_p(\cdot)$. Lemma 15 and the separation condition (11) together imply that $\frac{1}{p}\|\mathbf{u}_p^* - \hat{\mathbf{u}}_p\|_2^2 \overset{P|\mathbf{X}}{\to} 0$. In turn, this implies

$$d(\tilde{L}_p^{(\mathbf{u}_p^*)}, \tilde{L}_p^{(\hat{\mathbf{u}}_p)}) \overset{P|\mathbf{X}}{\to} 0, \tag{53}$$

where $d(\cdot,\cdot)$ denotes the 2-Wasserstein distance. Thus $\tilde{L}_p^{(\hat{\mathbf{u}}_p)}$ converges weakly in probability to $\nu^* := (X, Z, F^*(X,Z))$.

Next, differentiating $\widetilde{M}_p(\cdot)$ at $\mathbf{u} \in (-1,1)^p$, we obtain

$$\nabla\widetilde{M}_p(\mathbf{u}) = \Big[\frac{1}{\sigma^2}\Big(- A_p\mathbf{u} + A_p\boldsymbol{\beta}_0 + D_p\boldsymbol{\beta}_0 + \mathrm{diag}(\sqrt{D_p(i,i)})\xi\Big) - h(\mathbf{u},\mathbf{d})\Big],$$

where $\xi = (\xi_1, \cdots, \xi_p)^{\mathrm{T}}$, and $h(\cdot,\cdot)$ is the vector obtained upon applying h coordinate-wise to the two vectors. Note that if $u_i \to +1$, then $h(u_i, d_i) \to \infty$, and thus $\frac{\partial}{\partial u_i}\widetilde{M}_p(\mathbf{u},\mathbf{d}) \to -\infty$. Consequently, $\hat{u}_i < 1$. A similar argument implies $\hat{u}_i > -1$, and thus $\hat{\mathbf{u}}_p \in (-1,1)^p$. This immediately implies that $\hat{\mathbf{u}}_p$ is a critical point of $\widetilde{M}_p$, and satisfies the fixed point equation

$$\hat{\mathbf{u}}_p = \dot{c}\Big(\frac{1}{\sigma^2}\Big(- A_p\hat{\mathbf{u}}_p + A_p\boldsymbol{\beta}_0 + D_p\boldsymbol{\beta}_0 + \mathrm{diag}(\sqrt{D_p(i,i)})\xi\Big), \mathbf{d}\Big). \tag{54}$$

Let $f_1 : [0,1] \to \mathbb{R}$ and $f_2 : \mathbb{R} \to \mathbb{R}$ be bounded continuous functions. Recalling the function m from Lemma 40 and setting $g_p = W_{pA_p}\cdot w_{\boldsymbol{\beta}_0} + w_{\boldsymbol{\beta}_0}w_{\mathbf{D}}$, we have,

$$\begin{aligned}
&\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}[f_1(X)f_2(Z)U] \\
=&\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W_{pA_p}, X) + g_p(X) + \sqrt{w_{\mathbf{D}}(X)}Z\Big), w_{\mathbf{d}}(X)\Big)\Big] \\
=&\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W_{pA_p}, X) + g(X) + \sqrt{\psi(X)}Z\Big), w_{\mathbf{d}}(X)\Big)\Big] + \mathcal{E}_p,
\end{aligned} \tag{55}$$

where we use $\ddot{c}(\theta, d) \leq 1$, and observe that

$$|\mathcal{E}_p| \lesssim \mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}[|(\sqrt{w_{\mathbf{D}}(X)} - \sqrt{\psi(X)})Z|] + \|g_p - g\|_1 \lesssim \|w_{\mathbf{D}} - \psi\|_1 + \|g_p - g\|_1 = o(1). \quad (56)$$

Define the good event

$$\mathscr{E}(\varepsilon) = \{|m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W_{pA_p}, X) - m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W, X)| \leq \varepsilon\} \cap \{|w_{\mathbf{D}}(X) - \psi(X)| \leq \varepsilon\}.$$

Then, upon observing that $\sup_{\gamma_1,\gamma_2} |\frac{\partial^2}{\partial\gamma_1\partial\gamma_2} c(\gamma_1, \gamma_2)| \lesssim 1$ and $\sup_{\gamma_1,\gamma_2} |\frac{\partial^2}{\partial\gamma_1^2} c(\gamma_1, \gamma_2)| \lesssim 1$, we have, on the event $\mathscr{E}(\varepsilon)$,

$$\begin{aligned}
&\Big|\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W_{pA_p}, X) + g(X) + \mathbb{E}[\sqrt{\psi(X)}Z]\Big), w_{\mathbf{d}}(X)\Big) - \\
&-\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W, X) + g(X) + \mathbb{E}[\sqrt{\psi(X)}Z]\Big), \frac{\psi(X)}{\sigma^2}\Big)\Big| \lesssim \varepsilon.
\end{aligned}$$

Thus we have,

$$\begin{aligned}
&\Big|\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W_{pA_p}, X) + g(X) + \sqrt{\psi(X)}Z\Big), w_{\mathbf{d}}(X)\Big)\Big] \\
&- \mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W, X) + g(X) + \sqrt{\psi(X)}Z\Big), \frac{\psi(X)}{\sigma^2}\Big)\Big]\Big| \\
&\lesssim \varepsilon + \mathbb{P}[\mathscr{E}(\varepsilon)^c] \\
&\lesssim \varepsilon + \frac{2}{\varepsilon}\|W - W_{pA_p}\|_\square + \mathbb{P}\Big[|w_{\mathbf{d}}(X) - \frac{\psi(X)}{\sigma^2}| > \varepsilon\Big] = O(\varepsilon) + o(1),
\end{aligned} \quad (57)$$

where the last inequality uses Lemma 40 part (i).

Finally, Lemma 40 part (ii) implies that

$$\begin{aligned}
&\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\tilde{L}_p^{(\hat{\mathbf{u}}_p)}, W, X) + g(X) + \sqrt{\psi(X)}Z\Big), \frac{\psi(X)}{\sigma^2}\Big)\Big] \\
=&\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\nu^*, W, X) + g(X) + \sqrt{\psi(X)}Z\Big), \frac{\psi(X)}{\sigma^2}\Big)\Big] + o(1).
\end{aligned} \quad (58)$$

Combining (55), (56), (57), (58), we have,

$$\begin{aligned}
&\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}[f_1(X)f_2(Z)U] \\
=&\mathbb{E}_{\tilde{L}_p^{(\hat{\mathbf{u}}_p)}}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\nu^*, W, X) + g(X) + \sqrt{\psi(X)}Z\Big), \frac{\psi(X)}{\sigma^2}\Big)\Big] + o(1) + O(\varepsilon).
\end{aligned}$$

Sending $p \to \infty$, and then $\varepsilon \to 0$, we obtain

$$\begin{aligned}
&\mathbb{E}_{\nu^*}[f_1(X)f_2(Z)U] \\
=&\mathbb{E}_{\nu^*}\Big[f_1(X)f_2(Z)\dot{c}\Big(\frac{1}{\sigma^2}\Big(- m(\nu^*, W, X) + g(X) + \sqrt{\psi(X)}Z\Big), \frac{\psi(X)}{\sigma^2}\Big)\Big].
\end{aligned}$$

The desired conclusion follows upon recalling that $(X, Z, F^*(X, Z)) \sim \nu^*$.

■

We next turn to the proof of Lemma 25. To this end, we require the following auxiliary lemma. The proof is deferred to the Appendix.

Lemma 41 *Suppose $r_p : [-1,1]^p \to \mathbb{R} \cup \{-\infty\}$ is upper semi-continuous on $[-1,1]^p$, finite and differentiable on $(-1,1)^p$, and*

$$\sup_{\mathbf{u} \in (-1,1)^p} \lambda_{\min}(H_p(\mathbf{x})) \leq -\eta,$$

where $H_p(\cdot)$ is the Hessian of $r_p(\cdot)$. Then there exists a unique global maximizer $\mathbf{x}_0 \in [-1,1]^p$, and further for any $\mathbf{x} \in [-1,1]^p$ we have

$$r_p(\mathbf{x}_0) - r_p(\mathbf{x}) \geq \frac{\eta}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2.$$

Proof of Lemma 25

Proof of (i) The equation (13) implies that the event

$$\sup_{\mathbf{u} \in [-1,1]^p : \|\mathbf{u} - \mathbf{u}_p^*\|_2^2 > p\varepsilon} \frac{1}{p}\{M_p(\mathbf{u}) - M_p(\mathbf{u}_p^*)\} < -\varepsilon\lambda$$

occurs with probability $1 - o(1)$.

Proof of (ii)(a) Differentiating $M_p(\cdot)$, twice, the Hessian is given by

$$H_p = -\frac{1}{\sigma^2}A_p - \Delta_p, \quad \Delta_p(i,i) = \frac{1}{\ddot{c}(h(u_i,d_i),d_i)}. \tag{59}$$

Note that $\ddot{c}(h(u,d),d)$ is the variance of a random variable supported on $[-1,1]$, and thus $\ddot{c}(h(u,d),d) \leq 1$. Under the assumption of (i), there exists $\delta > 0$ such that for all p large enough and all $\mathbf{x} \in (-1,1)^p$,

$$\mathbf{x}^{\mathrm{T}} H_p(u)\mathbf{x} = -\frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}} A_p \mathbf{x} - \mathbf{x}\Delta_p\mathbf{x} \leq -\delta\|\mathbf{x}\|_2^2. \tag{60}$$

It then follows from Lemma 41 that there exists $\mathbf{u}^* \in [-1,1]^p$ which is a unique optimizer of $M_p(\cdot)$, and further, for any $\mathbf{u} \in [-1,1]^p$ we have

$$M_p(\mathbf{u}^*) - M_p(\mathbf{u}) \geq \frac{\delta}{2}\|\mathbf{u} - \mathbf{u}^*\|_2^2.$$

This verifies (13).

Proof of (ii)(b) To begin, note that the prior is absolutely continuous with respect to the Lebesgue measure on $[-1,1]$ and thus $D_{KL}(\pi_\infty\|\pi) = D_{KL}(\pi_{-\infty}\|\pi) = \infty$. Thus to maximize $M_p(\cdot)$, it suffices to restrict to $(-1,1)^p$. As in (60), it suffices to bound the lower eigenvalue of the Hessian, uniformly in $(-1,1)^p$, for which using (59) and the fact that $\lambda_{\min}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ is bounded away from 0 it suffices to show that $\Delta_p(i,i) \geq d_i$. By GHS inequality (Ellis et al.,

1976), $\ddot{c}(h(u,d_i),d_i) \leq \ddot{c}(0,d_i)$, and thus the desired inequality follows, once we establish that

$$\sup_{d\geq 0} d\ddot{c}(0,d) \leq 1. \tag{61}$$

Consider now a scale family of parametric distributions $\{\mathcal{P}_\theta : \theta \geq 0\}$ with

$$\frac{\mathrm{d}\mathcal{P}_\theta}{\mathrm{d}x} \propto \exp(-\theta V(x)) \exp(-dx^2/2).$$

Since $V(\cdot)$ is even, it follows that the first moment under $\mathcal{P}_\theta$ is 0 for all θ. Also, since $V(\cdot)$ is an increasing, this family has monotone likelihood ratio in $T(x) = |x|$, and thus the second moment under the law $\mathcal{P}_\theta$ is decreasing in θ for $\theta > 0$, and so

$$\ddot{c}(0,d) = \mathsf{Var}_{\theta=1}(Z) \leq \mathsf{Var}_{\theta=0}(Z).$$

Proceeding to bound the RHS of the above display, for $d > 0$ we have

$$d\mathsf{Var}_{\theta=0}(Z) = \frac{\int_{-1}^{1} dz^2 \exp(-dz^2/2)\,\mathrm{d}z}{\int_{-1}^{1} \exp(-dz^2/2)\,\mathrm{d}z}.$$

We substitute $\sqrt{d}z = t$, so that

$$d\mathsf{Var}_{\theta=0}(Z) = \frac{\int_{-\sqrt{d}}^{\sqrt{d}} t^2 \exp(-t^2/2)\,\mathrm{d}t}{\int_{-\sqrt{d}}^{\sqrt{d}} \exp(-t^2/2)\,\mathrm{d}t}.$$

This is exactly the variance of a truncated standard Gaussian distribution, truncated to the interval $[-\sqrt{d}, \sqrt{d}]$. Indeed, the truncated variance is $1 - \frac{2\sqrt{d}\phi(\sqrt{d})}{2\Phi(\sqrt{d})-1} < 1$ (Johnson and Kotz, 1971), where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the pdf and cdf of the standard Gaussian distribution. ■

Acknowledgments

The authors thank Pragya Sur for discussions on high-dimensional regression. SM gratefully thanks NSF (DMS 1712037) for support during this research. The authors gratefully thank the Associate Editor and two anonymous referees for their helpful comments, which significantly improved the exposition.

Appendix A. Some technical lemmas

We collect some basic results about exponential families in the first subsection. We prove Lemma 41 in the next subsection.

A.1 Results on exponential families

We prove Lemmas 2, 4 and 30 in this section.

Proof of Lemma 30

(i) This follows by direct computation (see e.g. Lehmann and Casella (2006)).

(ii) This follows by direct calculation.

(iii) Recall from Definition 1 that

$$\frac{\mathrm{d}\pi_\gamma}{\mathrm{d}\pi}(z) = \exp\Big(\gamma_1 z - \frac{\gamma_2}{2}z^2 - c(\gamma)\Big).$$

Without loss of generality, we consider the case $\gamma_{1,k} \to \infty$. The case $\gamma_{1,k} \to -\infty$ follows using the same argument, with obvious modifications. Observe that as $\gamma_{1,k} \to \infty$ and $\limsup_k |\gamma_{2,k}| < \infty$, for k sufficiently large, the function $\gamma_{1,k}z - \frac{\gamma_{2,k}}{2}z^2$ is increasing on $[-1,1]$. This implies, for any $\varepsilon > 0$,

$$\int_{-1}^{1-\varepsilon} \exp\Big(\gamma_{1,k}z - \frac{\gamma_{2,k}}{2}z^2\Big)\mathrm{d}\pi(z) \le \exp\Big(\gamma_{1,k}(1-\varepsilon) - \frac{\gamma_{2,k}}{2}(1-\varepsilon)^2\Big)\pi([-1,1-\varepsilon]).$$

On the other hand, $\exp(c(\gamma_k))$ can be lower bounded by

$$\int_{1-\frac{\varepsilon}{2}}^{1} \exp\Big(\gamma_{1,k}z - \frac{\gamma_{2,k}}{2}z^2\Big)\mathrm{d}\pi(z) \ge \exp\Big(\gamma_{1,k}(1-\frac{\varepsilon}{2}) - \frac{\gamma_{2,k}}{2}(1-\frac{\varepsilon}{2})^2\Big)\pi([1-\frac{\varepsilon}{2},1]).$$

On taking ratios of the above two displays we get

$$\pi_{\gamma_k}([-1,1-\varepsilon]) \le \exp\Big(-\frac{\varepsilon}{2}\gamma_{1,k} + \frac{1}{2}\gamma_{2,k}^2\Big)\frac{\pi([-1,1-\varepsilon])}{\pi([1-\frac{\varepsilon}{2},1])}$$

Also note that the ratio of probabilities under π in the RHS is positive, as ± 1 is in the support of π. The desired conclusion now follows upon letting $k \to \infty$, upon noting that $\gamma_{1,k} \to \infty$, and $\gamma_{2,k}$ stays bounded.

(iv) The proof follows directly from the lower semicontinuity of KL divergence under weak convergence (Posner, 1975, Theorem 1).

(v) By definition,

$$H(x,y) = \int_{[-1,1]} z^r \exp\Big(h(x,y)z - \frac{y}{2}z^2 - c(h(x,y),y)\Big)\mathrm{d}\pi(z).$$

For $x \in \{-1,1\}$, $H(x,\cdot)$ is independent of y, and thus $\frac{\partial H(x,y)}{\partial y} = 0$. We assume henceforth that $x \in (-1,1)$. Differentiating, we obtain,

$$\begin{aligned}\frac{\partial H(x,y)}{\partial y} &= \int_{[-1,1]} z^r\Big[z\frac{\partial h(x,y)}{\partial y} - \frac{z^2}{2} - \Big(\dot{c}(h(x,y),y)\frac{\partial h(x,y)}{\partial y} + \frac{\partial c}{\partial \gamma_2}\Big|_{(h(x,y),y)}\Big)\Big]\mathrm{d}\pi_{(h(x,y),y)}(z)\\ &= \int_{[-1,1]} z^r\Big[(z-x)\frac{\partial h(x,y)}{\partial y} - \frac{z^2}{2} - \frac{\partial c}{\partial \gamma_2}\Big|_{(h(x,y),y)}\Big]\mathrm{d}\pi_{(h(x,y),y)}(z),\end{aligned}$$

where the last equality uses $\dot{c}(h(x,y),y)) = x$. Therefore,

$$\left|\frac{\partial H(x,y)}{\partial y}\right| \leq \left|\frac{\partial h(x,y)}{\partial y}\right| \mathbb{E}_{\pi_{(h(x,y),y)}}[|Z-x|] + \frac{1}{2} + \frac{1}{2} \leq \left|\frac{\partial h(x,y)}{\partial y}\right| + 1, \tag{62}$$

where we use $\mathbb{E}_{\pi_{(h(x,y),y)}}[|Z|^r] \leq 1$ for all $r \geq 1$. Now, differentiating $\dot{c}(h(x,y),y) = x$ in y, we have,

$$\begin{aligned}
\frac{\partial h(x,y)}{\partial y} =& -\frac{\frac{\partial^2 c}{\partial \gamma_1 \partial \gamma_2}(h(x,y),y)}{\ddot{c}(h(x,y),y)} \\
=& \frac{1}{2}\frac{\mathrm{Cov}_{\pi_{(h(x,y),y)}}(Z,Z^2)}{\mathrm{Var}_{\pi_{(h(x,y),y)}}(Z)} \\
=& \frac{1}{2}\frac{\mathbb{E}_{(h(x,y),y)}[(Z-x)(Z^2-x^2)]}{\mathbb{E}_{(h(x,y),y)}[(Z-x)^2]} \\
=& \frac{1}{2}\frac{\mathbb{E}_{(h(x,y),y)}[(Z-x)^2(Z+x)]}{\mathbb{E}_{(h(x,y),y)}[(Z-x)^2]}.
\end{aligned}$$

This implies $\sup_{x\in(-1,1),y\in\mathbb{R}} \left|\frac{\partial h(x,y)}{\partial y}\right| \leq 1$, where we use the trivial bound $|Z+x| \leq 2$. The desired conclusion follows by plugging this back into (62).

■

Proof of Lemma 2

(a) Lemma 30 Part (a) implies that $\ddot{c}(\gamma_1,\gamma_2) = \mathrm{Var}_{\pi_\gamma}(Z) > 0$ for all γ_1,γ_2. Thus $\dot{c}(\gamma_1,\gamma_2)$ is strictly increasing in γ_1. For every $\gamma_2 \in \mathbb{R}$, as $\gamma_1 \to \pm\infty$, $\pi_\gamma \xrightarrow{w} \pi_{\pm\infty}$ using Lemma 30 Part (c). The desired conclusion follows on noting that

$$\dot{c}(\gamma_1,\gamma_2) = \mathbb{E}_{\pi_\gamma}[Z] \overset{\gamma_1\to\pm\infty}{\to} \pm 1$$

by the Dominated Convergence Theorem.

(b) Using Part (a), we know that $\dot{c}(\gamma_1,\gamma_2)$ is strictly increasing in γ_1, and continuous. Further, if $\gamma_1 \to \pm\infty$, $\dot{c}(\gamma_1,\gamma_2) \to \pm 1$. Thus for any $t \in (-1,1)$, the existence of $h(t,\gamma_2)$ follows by the Intermediate Value Theorem. Finally, we argue that for any fixed $\gamma_2 \in \mathbb{R}$, as $t \to 1$, $h(t,\gamma_2) \to \infty$. The case $t \to -1$ is similar, and thus omitted.

We complete the proof by contradiction. Suppose, if possible, that

$$M := \lim_{t\to 1} h(t,\gamma_2) < \infty.$$

In this case, as -1 is in the support of π, $\pi([-1,0]) > 0$. This implies that

$$\begin{aligned} t = \dot{c}(h(t,\gamma_2),\gamma_2) &= \int_{[-1,1]} z \exp\Big(h(t,\gamma_2)z - \frac{\gamma_2}{2}z^2 - c(h(t,\gamma_2),\gamma_2)\Big)\mathrm{d}\pi(z) \\ &= \int_{[-1,0]} z \exp\Big(h(t,\gamma_2)z - \frac{\gamma_2}{2}z^2 - c(h(t,\gamma_2),\gamma_2)\Big)\mathrm{d}\pi(z) + \\ &\int_{(0,1]} z \exp\Big(h(t,\gamma_2)z - \frac{\gamma_2}{2}z^2 - c(h(t,\gamma_2),\gamma_2)\Big)\mathrm{d}\pi(z) \\ &\leq \pi_{(h(t,\gamma_2),\gamma_2)}((0,1]). \end{aligned}$$

To complete the argument, it suffices to show that $\liminf_{t\to 1}\pi_{(h(t,\gamma_2),\gamma_2)}([-1,0]) > 0$. But this follows on noting that the function $t \to c(h(t,\gamma_2),\gamma_2)$ is non-decreasing on $[0,1]$, and thus

$$\pi_{(h(t,\gamma_2),\gamma_2)}([-1,0]) \geq \exp(-M - \frac{\gamma_2}{2} - c(M,\gamma_2))\pi([-1,0]).$$

■

Proof of Lemma 4 The lemma follows by direct computation. First, note that

$$\frac{\partial G(u,d)}{\partial u} = h(u,d) + u\frac{\partial h}{\partial u}(u,d) - \dot{c}(h(u,d),d)\frac{\partial h}{\partial u}(u,d) = h(u,d),$$

where the last equality follows upon noting that $\dot{c}(h(u,d),d) = u$. For the second derivative, note that

$$\begin{aligned} \frac{\partial G(u,d)}{\partial d} =& u\frac{\partial h(u,d)}{\partial d} - \dot{c}(h(u,d),d)\frac{\partial h(u,d)}{\partial d} - \frac{\partial}{\partial\gamma_2}c(\gamma_1,\gamma_2)\Big|_{(h(u,d),d)} + \frac{\partial}{\partial\gamma_2}c(\gamma_1,\gamma_2)\Big|_{(0,d)} \\ =& -\frac{\partial}{\partial\gamma_2}c(\gamma_1,\gamma_2)\Big|_{(h(u,d),d)} + \frac{\partial}{\partial\gamma_2}c(\gamma_1,\gamma_2)\Big|_{(0,d)} \\ =& \frac{1}{2}\int_{[-1,1]} z^2 d\pi_{(h(u,d),d)}(z)dz - \frac{1}{2}\int_{[-1,1]} z^2 d\pi_{(0,d)}(z)dz. \end{aligned}$$

Finally, note that

$$\frac{\partial^2 G(u,d)}{\partial^2 u} = \frac{\partial h(u,d)}{\partial u} = \frac{1}{\ddot{c}(h(u,d),d)} > 0$$

where the last equality follows from differentiating the equation $\dot{c}(h(u,d),d) = u$. ■

A.2 Proof of Lemma 41

Proof Since $r_p(\cdot)$ is upper semi-continuous there exists a global maximizer in $[-1,1]^p$, say $\widetilde{\mathbf{x}}_0$. Fixing any $\mathbf{x}$ in $(-1,1)^p$, consider the function $f(t) := r_p((1-t)\widetilde{\mathbf{x}}_0 + t\mathbf{x})$. Then f is twice differentiable on $(0,1)$, and

$$f''(t) = (\widetilde{\mathbf{x}}_0 - \mathbf{x})^{\mathrm{T}} H_p((1-t)\widetilde{\mathbf{x}}_0 + t\mathbf{x})(\widetilde{\mathbf{x}}_0 - \mathbf{x}) \leq -\eta\|\widetilde{\mathbf{x}}_0 - \mathbf{x}\|_2^2.$$

Consequently, for any $\varepsilon \in (0,1)$ Taylor's expansion implies

$$f(1) \le f(\varepsilon) + (1-\varepsilon)f'(\varepsilon) - \frac{1}{2}\eta(1-\varepsilon)^2\|\widetilde{\mathbf{x}}_0 - \mathbf{x}\|_2^2.$$

Taking limits as $\varepsilon \to 0$ we get

$$f(1) \le f(0) + f'(0+) - \frac{1}{2}\eta\|\widetilde{\mathbf{x}}_0 - \mathbf{x}\|_2^2 \le f(0) - \frac{1}{2}\eta\|\widetilde{\mathbf{x}}_0 - \mathbf{x}\|_2^2,$$

where the last inequality uses the fact that $f'(0+) \le 0$, as $t = 0$ is a global maximizer of f. The last display is equivalent to

$$r_p(\mathbf{x}) \le r_p(\widetilde{\mathbf{x}}_0) - \frac{1}{2}\eta\|\widetilde{\mathbf{x}}_0 - \mathbf{x}\|_2^2.$$

This inequality then extends to $x \in [-1,1]^p$ by upper semi-continuity of r_p. ■

Appendix B. Stability Estimates for functionals

We prove Lemma 34 and 40 in this section.

B.1 Proof of Lemma 34

We start with the proof of Part (i). Define the functional $T_{W,\phi} : \tilde{\mathscr{F}}_{2,4} \to \mathbb{R}$

$$T_{W,\phi,\psi}(\nu) := -\frac{1}{2}\mathbb{E}[W(X_1,X_2)U_1U_2] + \mathbb{E}[U_1\phi(X_1)] + \mathbb{E}[\sqrt{\psi(X_1)}U_1Z_1].$$

where $(X_1,Z_1,U_1),(X_2,Z_2,U_2) \sim \nu$ are iid. Similarly define $I_\psi : \tilde{\mathscr{F}}_{2,4} \to \mathbb{R}\cup\{\infty\}$

$$I_\psi(\nu) = \mathbb{E}\Big[G\Big(U, \frac{\psi(X)}{\sigma^2}\Big)\Big].$$

We observe that $\tilde{\mathcal{G}}_{W,\phi,\psi}(\nu) = \frac{1}{\sigma^2}T_{W,\phi,\psi}(\nu) - I_\psi(\nu)$. Further,

$$\sup_{\nu\in\tilde{\mathscr{F}}_{2,4}} |T_{W,\phi,\psi}(\nu) - T_{\tilde{W},\tilde{\phi},\psi}(\nu)| \le \frac{1}{2}\|W-\tilde{W}\|_\square + \|\phi-\tilde{\phi}\|_1, \tag{63}$$

$$\sup_{\nu\in\tilde{\mathscr{F}}_{2,4}} |T_{W,\phi,\psi}(\nu) - T_{W,\phi,\tilde{\psi}}(\nu)| \le 2\|\psi-\tilde{\psi}\|_1, \tag{64}$$

$$\sup_{\nu\in\tilde{\mathscr{F}}_{2,4}} |I_\psi(\nu) - I_{\tilde{\psi}}(\nu)| \le \frac{1}{2\sigma^2}\|\psi-\tilde{\psi}\|_1. \tag{65}$$

Indeed, (63) is immediate from the definition of $T_{W,\phi,\psi}(\cdot)$, and (65) follows from Lemma 4 on noting that $\sup_{u\in(-1,1),d\in\mathbb{R}}|\frac{\partial G}{\partial d}(u,d)| \le \frac{1}{2}$. Finally, (64) follows on noting that by Cauchy-Schwarz inequality,

$$\mathbb{E}_\nu\Big[|\sqrt{\psi(X)}Z - \sqrt{\tilde{\psi}(X)}Z|\Big] \le \sqrt{\mathbb{E}_\nu\Big[Z^2\Big]\mathbb{E}_\nu\Big[\Big(\sqrt{\psi(X)} - \sqrt{\tilde{\psi}(X)}\Big)^2\Big]} \le 2\|\psi-\tilde{\psi}\|_1,$$

where the last inequality uses $(\sqrt{a}-\sqrt{b})^2 \le |a-b|$ and $\mathbb{E}_\nu[Z^2] \le 4$. This completes the proof of Part (i).

Next, we turn to the proof of Part (ii). Using the definition of cut norm (as in Definition 12) and Part (i), we have,

$$\sup_{\nu \in \tilde{\mathscr{F}}_{2,4}} |\tilde{\mathcal{G}}_{W_k, W_k.\phi_k+\phi_k\psi_k, \psi_k}(\nu) - \tilde{\mathcal{G}}_{W, W.\phi_k+\phi_k\psi, \psi}(\nu)| \lesssim \|W_k - W\|_\square + \|\psi_k - \psi\|_1 \to 0.$$

Using Part (i) again, it suffices to show that

$$W.\phi_k + \phi_k\psi \xrightarrow{L^1} W.\phi + \phi\psi.$$

To this end, note that $|W.\phi_k + \phi_k(x)\psi(x) - W.\phi - \phi(x)\psi(x)|$ converges to 0 in measure, and

$$|W.\phi_k(x) + \phi_k(x)\psi(x) - W.\phi(x) - \phi(x)\psi(x)| \le 2\int_{[0,1]} |W(x,y)|dy + 2\psi,$$

which is an integrable function. This completes the argument using DCT.

B.2 Proof of Lemma 40

Proof of Lemma 40

(i) For any $\varepsilon > 0$, let $S^+(\varepsilon) = \{x : m(\mu, W, x) - m(\mu, W', x) > \varepsilon\}$. For $X \sim U([0,1])$, we have,

$$\begin{aligned}
\mathbb{P}(X \in S^+(\varepsilon)) &\le \frac{1}{\varepsilon}\mathbb{E}\Big[(m(\mu, W, X) - m(\mu, W', X))\mathbf{1}_{S^+(\varepsilon)}(X)\Big] \\
&= \frac{1}{\varepsilon}\mathbb{E}_{X,X'}\Big[(W(X,X') - W'(X,X'))\mathbf{1}_{S^+(\varepsilon)}(X)\mathbb{E}[U'|X']\Big] \\
&\le \frac{1}{\varepsilon}\|W - W'\|_\square.
\end{aligned}$$

Setting $S^-(\varepsilon) = \{x : m(\mu, W, x) - m(\mu, W', x) < -\varepsilon\}$, the same argument now yields that

$$\mathbb{P}[X \in S^-(\varepsilon)] \le \frac{1}{\varepsilon}\|W - W'\|_\square.$$

(ii) Let $(X_p, Z_p, U_p) \sim \nu_p$ and $(X, Z, U) \sim \nu$. Using Skorokhod Embedding Theorem, we assume that $(X_p, Z_p, U_p) \to (X, Z, U)$ a.s. as $p \to \infty$. Since $\int |W(x,y)|\mathrm{d}x\mathrm{d}y < \infty$, for any $\varepsilon > 0$ there exists W' continuous such that $\|W - W'\|_1 \le \varepsilon$. Then we have,

$$\mathbb{E}[|m(\mu, W, X) - m(\mu, W', X)|] \le \|W - W'\|_1 \le \varepsilon.$$

Also,

$$\begin{aligned}
&|m(\nu_p, W', x) - m(\nu, W', x)| \\
&= \Big|\mathbb{E}[W'(x, X_p)U_p] - \mathbb{E}[W'(x, X)U]\Big| \\
&\le \Big|\mathbb{E}[W'(x, X_p)U_p] - \mathbb{E}[W'(x, X_p)U]\Big| + \Big|\mathbb{E}[W'(x, X_p)U] - \mathbb{E}[W'(x, X)U]\Big| \\
&\le \|W'\|_\infty \mathbb{E}[|U_p - U|] + \sup_{x,y,z\in[0,1], |y-z| \le |X - X_p|} \Big|W'(x,y) - W'(x,z)\Big| = o(1)
\end{aligned}$$

using the uniform continuity of W'. The desired conclusion follows upon combining the two displays above.

■

Appendix C. Proofs of Examples

We establish Corollaries 9-11 and 26-29 in this section. Throughout this section $o_P(1)$ terms converge to zero in probability under the marginal distribution of the design matrix $\mathbf{X}$.

C.1 Accuracy of mean-field approximation

Proof of Corollary 9 This is immediate from Theorem 7. ■

Proof of Corollary 10

With A_p and D_p denoting the off-diagonal and diagonal parts of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, to invoke Theorem 7 we need to verify that A_p satisfies (5) and (6), and the empirical measure $\frac{1}{p}\sum_{i=1}^{p}\delta_{D_p(i,i)}$ is uniformly integrable. We verify these conditions below:

Since $\{\mathbf{x}_i\}_{1\le i\le n} \overset{i.i.d.}{\sim} N(0,\Gamma_p)$, and

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}},$$

invoking (Vershynin, 2012, Proposition 2.1) gives

$$\left\|\mathbf{X}^{\mathrm{T}}\mathbf{X} - \Gamma_p\right\|_2 = O_P\Big(\sqrt{\frac{p}{n}}\Big) = o_P(1), \tag{66}$$

where the last equality uses $p = o(n)$. Noting that

$$\|D_p - \Gamma_{p,\mathrm{diag}}\|_2 \le \|\mathbf{X}^{\mathrm{T}}\mathbf{X} - \Gamma_p\|_2$$

gives

$$\|A_p - \Gamma_{p,\mathrm{off}}\|_2 = o_P(1). \tag{67}$$

Consequently, A_p satisfies (5) with high probability, as $\mathrm{tr}(\Gamma_{p,\mathrm{off}}^2) = o(p)$. Further, since $\|\Gamma_p\|_2 = O(1)$, it follows from (66) gives that $\|A_p\|_2 = O_P(1)$. Finally, noting that

$$\max_{1\le i\le p}|D_p(i,i)| \le \|\mathbf{X}^{\mathrm{T}}\mathbf{X}\|_2 = \|\Gamma_p\|_2 + o_P(1)$$

we have $\frac{1}{p}\sum_{i=1}^{p}\delta_{D_p(i,i)}$ is uniformly integrable with high probability. This completes the proof of the corollary.

■

Proof of Corollary 11

It suffices to verify the same conditions on the matrices (A_p, D_p) as in Corollary 10. To this end, note that for any $i \neq j$ we have

$$A_p(i,j) = \frac{p}{n}\sum_{k=1}^{n} B(k,i)B(k,j), \tag{68}$$

and so

$$\begin{aligned}
\mathbb{E}\sum_{i\neq j} A_p(i,j)^2 =& \frac{p^2}{n^2}\sum_{i\neq j}\sum_{k,\ell=1}^{n}\mathbb{E}[B(k,i)B(k,j)B(\ell,i)B(\ell,j)] \\
=& \frac{p^2}{n^2}\sum_{i\neq j}\sum_{k\neq \ell}\mathbb{E}[B(k,i)B(k,j)B(\ell,i)B(\ell,j)] + \frac{p^2}{n^2}\sum_{i\neq j}\sum_{k=1}^{n}\mathbb{E}[B(k,i)B(k,j)] \\
\leq& \frac{p^2}{n^2}\times\frac{n^2p^2\lambda^4}{p^4} + \frac{p^2}{n^2}\times\frac{np^2\lambda^2}{p^2} = \lambda^4 + \frac{\lambda^2 p^2}{n},
\end{aligned}$$

which is $o(p)$ as $p = o(n)$. This verifies (5).

Also, (68) gives

$$\frac{1}{p}\mathbb{E}\sum_{i,j=1}^{p}|A_p(i,j)| = \frac{1}{n}\sum_{k=1}^{n}\sum_{i\neq j}\mathbb{E}[B(k,i)B(k,j)] \leq \lambda^2,$$

and so (6) holds with high probability. Finally we have

$$\mathbb{E}\frac{1}{p}\sum_{i=1}^{p}|D_p(i,i)|^2 = \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}\Big(\frac{p}{n}\sum_{k=1}^{n}B(k,i)\Big)^2 \leq \frac{p^2}{n^2}\Big[\frac{n^2\lambda^2}{p^2} + \frac{n\lambda}{p}\Big] = \lambda^2 + o(1),$$

and so $\frac{1}{p}\sum_{i=1}^{p} D_p(i,i)$ is uniformly integrable. ■

C.2 Limiting variational formula

Proof of Corollary 26

(a) The desired conclusion follows from Theorem 17, once we can verify

$$d_{L_1}(w_{\mathbf{D}}, 1) = o_P(1), \quad d_{\square}(W_{pA_p}, W) = o_P(1).$$

Here $\mathbf{D} = (D_p(1,1), \cdots, D_p(p,p))$ denotes the diagonal entries of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, and A_p denotes the off diagonal part of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Proceeding to verify the above display, invoking (66) we have $D_p(i,i) = 1 + \frac{1}{p}G(i/p)^2 + o_P(1)$, and so $w_{\mathbf{D}} \overset{L_1}{\to} 1$. Also, since $\Gamma_p(i,j) = \frac{1}{p}G(i/p)G(j/p)$ for all $i \neq j$, it follows that $d_{\square}(W_{p\Gamma_p}, W) \to 0$, where we use the almost sure continuity of $G(.,.)$. Since (67) gives $d_{\square}(W_{pA_p}, W_{p\Gamma_p}) = o_P(1)$, combining we get

$$d_{\square}(W_{pA_p}, W) \leq d_{\square}(W_{pA_p}, W_{p\Gamma_p}) + d_{\square}(W_{p\Gamma_p}, W) = o_P(1),$$

This completes the proof of part (a).

(b) We begin by verifying condition (11). To this effect, set

$$\overline{M}_p(\mathbf{u}) := -\frac{1}{2\sigma^2}\Big[\mathbf{u}^{\mathrm{T}}\Gamma_p\mathbf{u} - 2\mathbf{z}^{\mathrm{T}}\mathbf{u}\Big] - \sum_{i=1}^{p} G(u_i, d_i),$$

and use the fact that $\max_{i\in[p]}\sum_{j\neq i}\Gamma_p(i,j) \leq \lambda < \sigma^2$ along with part (a) of Lemma 25 to get the existence of $\mathbf{u}_p^*$ such that

$$\overline{M}_p(\mathbf{u}_p^*) - \overline{M}_p(\mathbf{u}) \geq \lambda\|\mathbf{u}_p^* - \mathbf{u}\|_2^2.$$

This in turn shows that for any $\varepsilon > 0$ we have

$$\limsup_{p\to\infty} \sup_{\mathbf{u}:\|\mathbf{u}-\mathbf{u}_p^*\|_2^2 > p\delta} \frac{1}{p}\Big\{\overline{M}_p(\mathbf{u}_p^*) - \overline{M}_p(\mathbf{u})\Big\} < 0. \tag{69}$$

Using (66) gives

$$\frac{1}{p}\sup_{\mathbf{u}\in[-1,1]^p} \frac{1}{p}\Big[\overline{M}_p(\mathbf{u}) - \widetilde{M}_p(\mathbf{u})\Big] = o_P(1). \tag{70}$$

Given (69), and (70), it follows that

$$\limsup_{p\to\infty} \sup_{\mathbf{u}:\|\mathbf{u}-\mathbf{u}_p^*\|_2^2 > p\delta} \frac{1}{p}\Big\{M_p(\mathbf{u}_p^*) - M_p(\mathbf{u})\Big\} < 0.$$

Thus we have verified (11). The desired conclusion then follows from Corollary 22. Part (b)(ii) follows from part (b)(ii) of Lemma 25 and Corollary 22.

■

Proof of Corollary 28

(a) The desired conclusion follows from Theorem 17, once we can verify

$$d_{L_1}(w_{\mathbf{D}}, \psi) = o_P(1), \quad d_{\square}(W_{pA_p}, W) = o_P(1).$$

Proceeding to verify the above display, note that

$$D_p(i,i) := (\mathbf{X}^{\mathrm{T}}\mathbf{X})_{ii} = \frac{p}{n\sigma^2}\sum_{k=1}^{n} B_{ki}.$$

Thus, setting $\widetilde{D}_i := \frac{1}{n}\sum_{k=1}^{n} G(k/n, i/p)$, Using Chernoff bounds it follows that

$$\mathbb{P}\left(\max_{1\leq i\leq p}|D_p(i,i) - \widetilde{D}_i| > \delta\right) \leq pe^{-c\delta^2\frac{n}{p}} \leq pe^{-c\delta^2\sqrt{p}} \to 0,$$

and so it follows that

$$d_{L_1}(w_{\mathbf{D}}, \psi) \leq d_{L_1}(w_{\mathbf{D}}, w_{\widetilde{D},p}) + d_{L_1}(w_{\widetilde{D},p}, \psi) = o_P(1)$$

where the last equality uses the almost sure continuity of $G(.,.)$.

It thus suffices to show that

$$d_{\square}(W_{pA_p}, W) \to 0.$$

To this end, note that for any $i \neq j$ we have

$$A_p(i,j) = \frac{p}{n}\sum_{k=1}^{n} B(k,i)B(k,j).$$

Using Lemma 42, setting $\widetilde{A}_p(i,j) := \mathbb{E}[A_p(i,j)] = \frac{1}{np}\sum_{k=1}^{n} G(\frac{k}{n},\frac{i}{p})G(\frac{k}{n},\frac{j}{p})$,

$$\mathbb{P}\Big(\max_{S,T\subseteq[p]}\Big|\frac{\sum_{i\in S,j\in T} pA_p(i,j)}{p^2} - \frac{\sum_{i\in S,j\in T} p\widetilde{A}_p(i,j)}{p^2}\Big| > \delta\Big) \leq 2^p e^{-C\sqrt{n}\delta}.$$

From this, using the condition $p = o(\sqrt{n})$ gives

$$d_{\square}(W_{pA_p}, W_{p\widetilde{A}_p}) = o_P(1),$$

which in turn gives

$$d_{\square}(W_{pA_p}, W) \leq d_{\square}(W_{pA_p}, W_{p\widetilde{A}_p}) + d_{\square}(W_{p\widetilde{A}_p}, W) = o_P(1),$$

where the last equality again uses the almost sure continuity of $G(\dot{;}\dot{)}$.

(b) (i) Once again, the desired conclusion of part (b) follows from Corollary 22, once we verify condition (11).

To this effect, fixing $\delta > 0$ and setting $\vartheta_i := \sum_{j\neq i}^{p} A_p(i,j)$, define a $p \times p$ matrix $B_{p,\delta}$ by setting

$$B_{p,\delta}(i,j) := A_p(i,j)1\{\vartheta_i \leq \sigma^2(1-\delta), \vartheta_j \leq \sigma^2(1-\delta)\},$$

and note that

$$\sup_{\mathbf{u}\in[-1,1]^p}\Big|\mathbf{u}'A_p\mathbf{u}_p - \mathbf{u}'B_{p,\delta}\mathbf{u}\Big| \leq \frac{2}{p}\sum_{i=1}^{p}\vartheta_i 1\{\vartheta_i > \sigma^2(1-\delta)\}.$$

We now claim that there exists $\delta > 0$ such that

$$\limsup_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\vartheta_i 1\{\vartheta_i > \sigma^2(1-\delta)\} = 0. \tag{71}$$

Given (71), it suffices to show that (11) holds for $\overline{M}_p : [-1,1]^p \mapsto \mathbb{R}$ defined by

$$\overline{M}_p(\mathbf{u}) := -\frac{1}{2\sigma^2}\Big[\mathbf{u}^{\mathrm{T}}B_{p,\delta}\mathbf{u} - 2\mathbf{z}^{\mathrm{T}}\mathbf{u}\Big] - \sum_{i=1}^{p} G(u_i, d_i).$$

But this is immediate from Lemma 25 part (a), on noting that

$$\max_{i\in[p]} \sum_{j\neq i} B_{p,\delta}(i,j) \leq \sigma^2(1-\delta).$$

It thus remains to verify (71). To this effect, recall from part (a) that W_{pA_p} converges to W in the cut metric, which in turn implies $\frac{1}{p}\sum_{i=1}^{p} \delta_{\vartheta_i}$ converges weakly in probability to the law of $S(X)$, where $X \sim U[0,1]$ (Borgs et al., 2015, Theorem 2.16). This gives

$$\limsup_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \vartheta_i 1\{\vartheta_i > \sigma^2(1-\delta)\} \leq \mathbb{E}S(X)1\{S(X) \geq \sigma^2(1-\delta)\}.$$

It thus suffices to show that there exists $\delta > 0$ such that the RHS above is 0. But this follows on noting that $\text{esssup}(S(X)) < \sigma^2$. This completes the proof of part (i).

Part (b)(ii) follows from part (b)(ii) of Lemma 25 and Corollary 22, as before.

■

Proof of Corollary 29

(a) A direct calculation gives $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \frac{1}{2}\mathbf{I} + A_p$, where

$$\begin{aligned} A_p(i,j) =& \frac{1}{p} \text{ if } i \leq \frac{p}{2}, j > \frac{p}{2} \text{ or } i > \frac{p}{2}, j \leq \frac{p}{2}, \\ =& 0 \text{ otherwise .} \end{aligned}$$

It then follows that $d_{\square}(W_{pA_p}, W) \to W$, and $d_{L_1}(w_{\mathbf{D}}, \psi) \to 0$. The desired conclusion then follows from Theorem 17 as before.

(b) Since $\limsup_{p\to\infty} \max_{i\in[p]} \sum_{j\neq i} |A_p(i,j)| = \frac{1}{2}$, the result is immediate from Corollary 22 and Lemma 25.

■

Appendix D. Relevant concentration inequalities

Lemma 42 *Let $B_{ik} \sim \text{Ber}(G(i/n, j/p)/p)$ are independent random variables with $G(x,y) \leq \lambda$. For any $\delta > 0$, there exists $C > 0$ (depending on δ) such that*

$$\mathbb{P}\Big(\max_{S,T\subset[p]} \Big| \sum_{k\in S, l\in T} (A_p(k,l) - \mathbb{E}[A_p(k,l)]) \Big| > p\delta \Big) \leq 2^p \exp\Big(- C\sqrt{n} \Big).$$

To prove Lemma 42, we first obsere that if X is sub-exponential, then X^2 is sub-Weibull Kuchibhotla and Chakrabortty (2018).

Lemma 43 *Let $Y_1, \cdots, Y_n$ be independent sub-exponential random variables with common sub-exponential parameter $c > 0$. For any $\delta > 0$, there exists $C := C(\delta) > 0$ such that*

$$\mathbb{P}\Big[\Big|\sum_{i=1}^{n}(Y_i^2 - \mathbb{E}[Y_i^2])\Big| > n\delta\Big] \leq \exp\Big(-C\sqrt{n}\Big).$$

Proof We first claim that Y_i^2 are sub-Weibull, i.e., there exists a constant $C' > 0$ (depending only on c) such that

$$\mathbb{P}[|Y_i^2 - \mathbb{E}[Y_i^2]| > t] \leq 2\exp(-C'\sqrt{t}). \tag{72}$$

We establish the upper tail deviation bound. A similar argument works for the lower tail, and is thus omitted. As the variables Y_i have a common sub-exponential constant, $\max_{i\leq n}\mathbb{E}[Y_i^2]$ is uniformly bounded in n. For any $t > 0$ with $\sqrt{t} \geq (\sqrt{2}-1)^2\max_{i\leq n}\mathbb{E}^2[Y_i]$,

$$\begin{aligned}\mathbb{P}[Y_i^2 > \mathbb{E}[Y_i^2] + t] &= \mathbb{P}[Y_i - \mathbb{E}[Y_i] > \sqrt{t + \mathbb{E}[Y_i^2]} - \mathbb{E}[Y_i]] \\ &\leq \mathbb{P}[Y_i - \mathbb{E}[Y_i] > \sqrt{t + \mathbb{E}^2[Y_i]} - \mathbb{E}[Y_i]].\end{aligned}$$

We now claim that

$$\sqrt{t + \mathbb{E}^2[Y_i]} - \mathbb{E}[Y_i] > \frac{\sqrt{t} - (\sqrt{2}-1)\mathbb{E}[Y_i]}{\sqrt{2}}. \tag{73}$$

Using (73), along with the deviation bound above, we have,

$$\mathbb{P}[Y_i^2 > \mathbb{E}[Y_i^2] + t] \leq \mathbb{P}\Big[Y_i - \mathbb{E}[Y_i] > \frac{\sqrt{t} - (\sqrt{2}-1)\mathbb{E}[Y_i]}{\sqrt{2}}\Big] \leq \exp(-C'\sqrt{t}),$$

for some constant $C' > 0$. The last inequality uses that Y_i is sub-exponential. (73) can be verified by direct computation.

Using (72) and (Kuchibhotla and Chakrabortty, 2018, Proposition A.3), $\max_{i\leq n}\|Y_i^2\|_{\psi_{1/2}}$ is uniformly bounded in n. Consequently, using (Kuchibhotla and Chakrabortty, 2018, Theorem 3.1), for any $t > 0$, we have,

$$\mathbb{P}\Big[\Big|\sum_{i=1}^{n}(Y_i^2 - \mathbb{E}[Y_i^2])\Big| > C_1\|b\|_2(\sqrt{t} + L_n t^2)\Big] \leq \exp(-t),$$

where C_1 is a constant free of n, and $L_n = C_2\|b\|_\infty/\|b\|_2$ for some constant $C_2 > 0$ independent of n. Here, $b = (\|Y_1^2\|_{\psi_{1/2}}, \cdots, \|Y_n^2\|_{\psi_{1/2}})$. We set $C_1\|b\|_2(\sqrt{t} + L_n t^2) = n\delta$ for some $\varepsilon > 0$. Direct calculation yields that $t(\delta) = \sqrt{\frac{n\delta}{C_1C_2\|b\|_\infty}}(1 + o(1))$, so that

$$\mathbb{P}\Big[\Big|\sum_{i=1}^{n}(Y_i^2 - \mathbb{E}[Y_i^2])\Big| > n\delta\Big] \leq \exp\Big(-C_3\sqrt{n}\Big),$$

where $C_3 > 0$ depends on δ. This concludes the proof. ■

Proof of Lemma 42 Recall that $B_{ik} \sim \mathrm{Ber}(G(i/n, j/p)/p)$ are independent random variables. For $S \subseteq [p]$, define $Y_i(S) = \sum_{k\in S} B_{ik}$. By Chernoff bound, the collection $\{Y_i(S) : 1 \le i \le n, S \subseteq p\}$ are sub-exponential with a common sub-exponential parameter (depending only on λ).

By Chernoff bounds, for any fixed $S \subseteq [p]$, $Y_i = \sum_{k\in S} B_{ik}$ is a sub-exponential random variable. Using Lemma 43,

$$\mathbb{P}\Big(\Big|\sum_{k,l\in S}(A_p(k,l) - \mathbb{E}[A_p(k,l)])\Big| > p\delta\Big) = \mathbb{P}\Big(\Big|\sum_{i=1}^{n}(Y_i(S)^2 - \mathbb{E}[Y_i(S)^2])\Big| > n\delta\Big) \le \exp\Big(-C\sqrt{n}\Big).$$

A union bound then concludes

$$\mathbb{P}\Big(\max_{S\subseteq[p]}\Big|\sum_{k,l\in S}(A_p(k,l) - \mathbb{E}[A_p(k,l)])\Big| > p\delta\Big) \le 2^p \exp\Big(-C\sqrt{n}\Big). \tag{74}$$

We set $\varepsilon_{kl} := A_p(k,l) - \mathbb{E}[A_p(k,l)]$ and claim that

$$\max_{S,T\subseteq[p]} |\sum_{k\in S,l\in T}\varepsilon_{kl}| \le \frac{5}{2}\max_{S\subseteq[p]}|\sum_{k,l\in S}\varepsilon_{kl}| \tag{75}$$

The required conclusion follows from (75) and the deviation bound (74). It thus remains to prove (75). To this effect, note that

$$\sum_{k\in S,l\in T}\varepsilon_{kl} = \sum_{k\in S\setminus T,l\in T\setminus S}\varepsilon_{kl} + \sum_{k\in S\setminus T,l\in S\cap T}\varepsilon_{kl} + \sum_{k\in S\cap T,l\in T\setminus S}\varepsilon_{kl} + \sum_{k\in S\cap T,l\in S\cap T}\varepsilon_{kl}.$$

$$\sum_{k,l\in S\cup T}\varepsilon_{kl} = \sum_{k,l\in S\setminus T}\varepsilon_{kl} + \sum_{k,l\in T\setminus S}\varepsilon_{kl} + \sum_{k,l\in S\cap T}\varepsilon_{kl} + 2\Big(\sum_{k\in S\setminus T,l\in T\setminus S}\varepsilon_{kl} + \sum_{k\in S\setminus T,l\in S\cap T}\varepsilon_{kl} + \sum_{k\in S\cap T,l\in T\setminus S}\varepsilon_{kl}\Big).$$

Thus

$$\sum_{k\in S,l\in T}\varepsilon_{kl} = \frac{1}{2}\Big[\sum_{k,l\in S\cup T}\varepsilon_{kl} - \Big(\sum_{k,l\in S\setminus T}\varepsilon_{kl} + \sum_{k,l\in T\setminus S}\varepsilon_{kl} + \sum_{k,l\in S\cap T}\varepsilon_{kl}\Big)\Big] + \sum_{k\in S\cap T,l\in S\cap T}\varepsilon_{kl}$$

which on using triangle inequality gives (75).

■

References

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497, 2020.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.

Fanny Augeri. A transportation approach to the mean-field approximation. *arXiv preprint arXiv:1903.08021*, 2019.

Tim Austin. The structure of low-complexity gibbs measures on product spaces. *The Annals of Probability*, 47(6):4002–4023, 2019.

Sayantan Banerjee, Ismaël Castillo, and Subhashis Ghosal. Bayesian inference in high-dimensional models. *arXiv preprint arXiv:2101.04491*, 2021.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Anirban Basak and Sumit Mukherjee. Universality of the mean-field for the potts model. *Probability Theory and Related Fields*, 168(3):557–600, 2017.

Anirban Bhattacharya, Antik Chakraborty, and Bani K Mallick. Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042, 2016.

Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs ii. multiway cuts and statistical physics. *Annals of Mathematics*, pages 151–219, 2012.

Christian Borgs, Jennifer T Chayes, Henry Cohn, and Shirshendu Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015.

Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An l^p theory of sparse graph convergence ii: Ld convergence, quotients and right convergence. *The Annals of Probability*, 46(1):337–396, 2018.

Christian Borgs, Jennifer Chayes, Henry Cohn, and Yufei Zhao. An l^p theory of sparse graph convergence i: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062, 2019.

Peter Carbonetto and Matthew Stephens. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1):73–108, 2012.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Sourav Chatterjee and Amir Dembo. Nonlinear large deviations. *Advances in Mathematics*, 299:396–450, 2016.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

Nicholas A Cook, Amir Dembo, and Huy Tuan Pham. Regularity method and large deviation principles for the erdos-renyi hypergraph. *arXiv preprint arXiv:2102.09100*, 2021.

Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. *Geometric and Functional Analysis*, 28(6):1548–1596, 2018.

Richard S Ellis, James L Monroe, and Charles M Newman. The ghs and other correlation inequalities for a class of even ferromagnets. *Communications in Mathematical Physics*, 46(2):167–182, 1976.

Zhou Fan, Song Mei, and Andrea Montanari. Tap free energy, spin glasses, and variational inference. *arXiv preprint arXiv:1808.07890*, 2018.

Sebastian Farquhar, Lewis Smith, and Yarin Gal. Liberty or depth: Deep bayesian neural nets do not need complex weight posterior approximations. *Advances in Neural Information Processing Systems*, 33:4346–4357, 2020.

Augusto Fasano, Daniele Durante, and Giacomo Zanella. Scalable and accurate variational bayes for high-dimensional binary regression models. *arXiv preprint arXiv:1911.06743*, 2019.

Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33:15897–15908, 2020.

Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

Robert Gallager. Low-density parity-check codes. *IRE Transactions on information theory*, 8(1):21–28, 1962.

Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. In *International conference on machine learning*, pages 2221–2231. PMLR, 2019.

Subhashis Ghosal. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331, 1999.

Peter Hall, Tung Pham, Matt P Wand, and Shen SJ Wang. Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics*, 39(5): 2502–2532, 2011.

Wei Han and Yun Yang. Statistical inference in mean-field variational bayes. *arXiv preprint arXiv:1911.01525*, 2019.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2): 949–986, 2022.

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of statistics*, 33(2):730–773, 2005.

Vishesh Jain, Frederic Koehler, and Elchanan Mossel. The mean-field approximation: Information inequalities, algorithms, and complexity. *arXiv preprint arXiv:1802.06126*, 2018.

Vishesh Jain, Frederic Koehler, and Andrej Risteski. Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1226–1236, 2019.

Norman L Johnson and Samuel Kotz. *Continuous univariate distributions: Distributions in statistics*. John Wiley and Sons, 1971.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

Se Yoon Lee, Debdeep Pati, and Bani K Mallick. Continuous shrinkage prior revisited: a collapsing behavior and remedy. *arXiv preprint arXiv:2007.02192*, 2020.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

László Lovász and Balázs Szegedy. Szemerédi's lemma for the analyst. *GAFA Geometric And Functional Analysis*, 17(1):252–270, 2007.

David JC MacKay. Good error-correcting codes based on very sparse matrices. *IEEE transactions on Information Theory*, 45(2):399–431, 1999.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

Sarah E Neville, John T Ormerod, and MP Wand. Mean field variational bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, 8(1): 1113–1151, 2014.

John T Ormerod, Chong You, and Samuel Müller. A variational bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549–3594, 2017.

Edward Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

Kolyan Ray and Botond Szabó. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, pages 1–12, 2021.

Kolyan Ray, Botond Szabo, and Gabriel Clara. Spike and slab variational bayes for high dimensional logistic regression. *arXiv preprint arXiv:2010.11665*, 2020.

Qifan Song and Faming Liang. Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*, 2017.

Huayang Tang, Xin Jin, Yang Li, Hui Jiang, Xianfa Tang, Xu Yang, Hui Cheng, Ying Qiu, Gang Chen, Junpu Mei, et al. A large-scale screen for coding variants predisposing to psoriasis. *Nature genetics*, 46(1):45–50, 2014.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Bo Wang and D Michael Titterington. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

Yixin Wang and David Blei. Variational bayes under model misspecification. In *Advances in Neural Information Processing Systems*, pages 13357–13367, 2019a.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019b.

T Westling and TH McCormick. Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics*, 28(4):778–789, 2019.

Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

Jun Yan. Nonlinear large deviations: Beyond the hypercube. *Annals of Applied Probability*, 30(2):812–846, 2020.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905, 2020.

Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *Annals of Statistics*, 48(5): 2575–2598, 2020.