TAP free energy, spin glasses, and variational inference

Zhou Fan[*] Song Mei[†] Andrea Montanari[‡]

August 23, 2018

Abstract

We consider the Sherrington-Kirkpatrick model of spin glasses with ferromagnetically biased couplings. For a specific choice of the couplings mean, the resulting Gibbs measure is equivalent to the Bayesian posterior for a high-dimensional estimation problem known as '$\mathbb{Z}_2$ synchronization'. Statistical physics suggests to compute the expectation with respect to this Gibbs measure (the posterior mean in the synchronization problem), by minimizing the so-called Thouless-Anderson-Palmer (TAP) free energy, instead of the mean field (MF) free energy. We prove that this identification is correct, provided the ferromagnetic bias is larger than a constant (i.e. the noise level is small enough in synchronization). Namely, we prove that the scaled ℓ_2 distance between any low energy local minimizers of the TAP free energy and the mean of the Gibbs measure vanishes in the large size limit. Our proof technique is based on upper bounding the expected number of critical points of the TAP free energy using the Kac-Rice formula.

1 Introduction and main results

Computing expectations of a high-dimensional probability distribution is a central problem in computer science, statistics, and statistical physics. While a number of mathematical and computational techniques have been developed for this task, fundamental questions remain unanswered even for seemingly simple models.

Consider the problem of estimating a vector $\boldsymbol{x} \in \{+1, -1\}^n$ from the observation $\boldsymbol{Y} \in \mathbb{R}^{n \times n}$ given by

$$\boldsymbol{Y} = \frac{\lambda}{n} \boldsymbol{x} \boldsymbol{x}^\mathsf{T} + \boldsymbol{W} \,. \tag{1}$$

Here $\boldsymbol{W}$ is an unknown noise matrix, which is distributed according to the Gaussian orthogonal ensemble (GOE). Namely, $\boldsymbol{W} = \boldsymbol{W}^\mathsf{T} \in \mathbb{R}^{n \times n}$ where $(W_{ij})_{1 \le i \le j \le n}$ are independent centered Gaussian random variables with off-diagonal entries having variance $1/n$ and diagonal entries having variance $2/n$. (In what follows, we will write $\boldsymbol{W} \sim \mathrm{GOE}(n)$.) The parameter $\lambda \in (0, \infty)$ is assumed to be known and corresponds to a signal-to-noise ratio.

This problem is known as $\mathbb{Z}_2$ synchronization [Sin11] and is closely related to correlation clustering [BBC04]. It has applications to determining the orientation of a manifold [SW11], and as a simplified model for the two-groups stochastic block model [HLL83] and for topic models [Ble12, GJM18]. $\mathbb{Z}_2$ synchronization is known to undergo a phase transition at $\lambda = 1$: for $\lambda < 1$, no estimator can achieve a correlation with the true signal $\boldsymbol{x}$ which is bounded away from zero. Viceversa, for $\lambda > 1$, there exist estimators that achieve a strictly positive correlation, see e.g. [MS16, DAM16].

In this paper, we consider the regime $\lambda > 1$ and study the optimal estimator. Notice that –even in absence of noise– the vector $\boldsymbol{x}$ can be determined from the observation $\boldsymbol{Y}$ only up to a sign. In order to resolve this ambiguity, we set as our goal to estimate the $n \times n$ matrix of relative signs $\boldsymbol{X} = \boldsymbol{x} \boldsymbol{x}^\mathsf{T}$. An estimator will be a map $\widehat{\boldsymbol{X}} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$, $\boldsymbol{Y} \mapsto \widehat{\boldsymbol{X}}(\boldsymbol{Y})$. We will see that the optimal $\widehat{\boldsymbol{X}}$ is approximately of rank one, i.e. $\widehat{\boldsymbol{X}}(\boldsymbol{Y}) \approx \boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T}$, and therefore $\boldsymbol{m}_\star \in \mathbb{R}^n$ can be viewed as an estimator of $\pm \boldsymbol{x}$.

[*]Department of Statistics and Data Science, Yale University
[†]Institute for Computational and Mathematical Engineering, Stanford University
[‡]Department of Electrical Engineering and Department of Statistics, Stanford University

By the symmetry of the problem, it is reasonable to consider Bayes estimation with respect to the uniform prior[1]. Namely, we assume $\boldsymbol{x} \sim \text{Unif}(\{+1, -1\}^n)$, and try to minimize the matrix mean square error, defined by

$$\text{MMSE}_n(\lambda) = \inf_{\widehat{\boldsymbol{X}}} \frac{1}{n^2} \mathbb{E}\Big[\big\|\widehat{\boldsymbol{X}}(\boldsymbol{Y}) - \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\big\|_F^2\Big]. \tag{2}$$

Decision theory immediately implies that the minimum of $\text{MMSE}_n(\lambda)$ is achieved by the posterior mean,

$$\widehat{\boldsymbol{X}}_{\text{Bayes}} = \widehat{\boldsymbol{X}}_{\text{Bayes}}(\boldsymbol{Y}) = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T} \mid \boldsymbol{Y}]. \tag{3}$$

The asymptotics of the Bayes risk $\lim_{n\to\infty} \text{MMSE}_n(\lambda)$ was calculated in [DAM16].

Computing the posterior expectation requires summing functions over $\boldsymbol{x} \in \{+1, -1\}^n$, an example of the high-dimensional integration problem mentioned above. While exact integration is expected to be intractable, a number methods have been developed that attempt to compute the posterior expectation approximately. Two main strategies are Markov Chain Monte Carlo (see, e.g., the surveys in [Dia09, Sin12, ADFDJ03]) and variational inference (e.g., [WJ08, MM09, BKM17]). Markov Chain Monte Carlo usually requires a large number of steps to get an accurate approximation. Further, despite remarkable mathematical progresses, Markov chain mixing is well understood only in simple cases.

Variational inference attempts to compute the marginals of a high-dimensional distribution by minimizing a suitable 'free energy' function. This approach is usually faster than MCMC, leading to a broad range of applications, from topic modeling [Ble12] to computer vision, and inference in graphical models [KF09]. Of course, its accuracy relies in a crucial way on the accuracy of the free energy construction. For instance, several applications make use of the so-called 'naive mean field' free energy [Ble12]. However, it was shown in [GJM18] that, for topic modeling (and even the simpler $\mathbb{Z}_2$ synchronization problem), naive mean field can return wrong information about the posterior distribution. Theorem 1.3 of this paper confirms this conclusion, by showing that naive mean field does not compute the correct posterior mean for $\mathbb{Z}_2$ synchronization when λ is a large enough constant.

Methods from spin glass theory [MPV87] can overcome these limitations. Consider the Gibbs measure of the Sherrington-Kirkpatrick (SK) model of spin glasses [KS78], with ferromagnetic bias aligned with $\boldsymbol{x}$,

$$G_{\beta,\lambda}(\boldsymbol{\sigma}) = \frac{1}{Z_n(\beta,\lambda)} \exp\{\beta\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\} = \frac{1}{Z_n(\beta,\lambda)} \exp\{\beta\lambda\langle\boldsymbol{\sigma}, \boldsymbol{x}\rangle^2/2 + \beta\langle\boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma}\rangle/2\}. \tag{4}$$

It is easy to check that the posterior distribution of $\boldsymbol{x}$ given observation $\boldsymbol{Y}$ is given by the biased SK measure with[2] $\beta = \lambda$

$$p(\boldsymbol{\sigma}|\boldsymbol{Y}) = G_{\lambda,\lambda}(\boldsymbol{\sigma}) \propto \exp\big\{\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\big\}. \tag{5}$$

Following the statistical physics terminology, we shall refer to the variables $\sigma_1, \ldots, \sigma_n$ as 'spins'.

For the SK model, naive mean field is known to be a poor approximation of the actual free energy, which explains the failures mentioned above. In 1977, Thouless, Anderson and Palmer [TAP77] proposed a variational formula for the SK free energy, whose decision variables $\boldsymbol{m} = (m_1, \ldots, m_n) \in [-1, 1]^n$ encode the expectation of the spins $\boldsymbol{\sigma}$. This variational formula is known as the TAP free energy, and its first-order stationarity conditions are known as the TAP equations. The relationship between the TAP equations and the Gibbs measure has been studied in the physics and mathematics literature for the last 40 years, and still presents a number of outstanding challenges. A brief overview is presented in Section 2.

Explicitly, the TAP free energy for the SK model (4) is the function $\mathcal{F}_{\beta,\lambda} : [-1, 1]^n \to \mathbb{R}$, defined by

$$\mathcal{F}_{\beta,\lambda}(\boldsymbol{m}) = -\frac{1}{n}\sum_{i=1}^n h(m_i) - \frac{\beta}{2n}\langle\boldsymbol{Y}, \boldsymbol{m}\boldsymbol{m}^\mathsf{T}\rangle - \frac{\beta^2}{4}[1 - Q(\boldsymbol{m})]^2, \tag{6}$$

where $Q(\boldsymbol{m}) = \|\boldsymbol{m}\|_2^2/n$, and $h : [-1, 1] \to \mathbb{R}$ is the binary entropy function

$$h(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right).$$

---

[1]In particular, the Bayes estimator under uniform prior is minimax optimal over $\boldsymbol{x} \in \{+1, -1\}^n$.
[2]In physics language, the condition $\beta = \lambda$ is referred to as the 'Nishimori line' [Nis01].

The corresponding TAP equations (first order stationary condition for $\mathcal{F}_{\beta,\lambda}$) are given by

$$\boldsymbol{m} = \tanh\left(\beta\boldsymbol{Y}\boldsymbol{m} - \beta^2[1 - Q(\boldsymbol{m})]\boldsymbol{m}\right). \tag{7}$$

The first two terms in Eq. (6) correspond to the naive mean field free energy, while the term $-\beta^2(1 - Q(\boldsymbol{m}))^2/4$ is known as 'Onsager's correction,' and is the main innovation introduced in [TAP77]. Indeed, heuristically, TAP free energy can be understood using Plefka's expansion [Ple82], which provides a series expansion of the log partition function. The first three terms of the expansion correspond to $\mathcal{F}_{\beta,\lambda}(\boldsymbol{m})$, and the others are expected to be negligible as $n \to \infty$.

Statistical physics suggests an ambitious general conjecture on the role of the TAP free energy, namely, the Gibbs measure (4) is a convex combination of a number of pure states (each pure state $G_{\beta,\lambda}^\alpha$ is a probability measure on $\{+1, -1\}^n$ of the form $G_{\beta,\lambda}^\alpha(\boldsymbol{\sigma}) \propto G_{\beta,\lambda}(\boldsymbol{\sigma})\mathbf{1}_{\boldsymbol{\sigma}\in\Omega_\alpha}$), which are in correspondence with the local minima of $\mathcal{F}_{\beta,\lambda}$. Namely, each pure state α corresponds to a local minimum $\boldsymbol{m}^\alpha$. We refer to Section 2 for pointers to this line of work.

Understanding the geometry of the critical points of the TAP free energy $\mathcal{F}_{\beta,\lambda}$ would help us elucidate the structure of the Gibbs measure $G_{\beta,\lambda}$. A remarkable sequence of papers within the statistical mechanics literature [BM80, PP95, CGPM03, CLPR03, CLPR04, CLR05] computed the expected number of critical points of the TAP free energy for the SK model ($\lambda = 0$) using non-rigorous but sophisticated tools from physics. More precisely, these authors obtain the exponential growth rate of this expectation, as a function of the free energy (the value of $\mathcal{F}_{\beta,\lambda}(\boldsymbol{m})$), a quantity that is known as the 'complexity' of the spin model. In our first main result, we prove rigorously that the formula given by [CLR05] is indeed an upper bound for the expected number of critical points, in the general setting $\lambda \geq 0$. To the best of our knowledge, a similar rigorous calculation of the complexity was only obtained so far for the p-spin spherical model, which is significantly simpler because of the spherical symmetry [AAČ13].

In the context of the Bayesian model (5), the above picture simplifies. The model is known to be replica symmetric [DAM16], and hence we expect the global minimum of $\mathcal{F}_{\lambda,\lambda}$ to correspond to the posterior expectation. Our second main result proves this conjecture for all λ larger than a big enough constant. This implies that minimizing the TAP free energy is a viable approach to optimal estimation in the present model.

While the relevance of TAP free energy for statistical inference was pointed out several times in the past (see Section 2), this is the first rigorous result confirming this connection.

1.1 The complexity of the TAP free energy

For $U \subseteq \mathbb{R}^4$ and $V \subseteq (-1, 1)$, define the number of critical points of $\mathcal{F}_{\beta,\lambda}$ with $\boldsymbol{m}$ in the region defined by U and V by

$$\mathrm{Crit}_n(U, V) = \sum_{\boldsymbol{m}:\nabla\mathcal{F}_{\beta,\lambda}(\boldsymbol{m})=\mathbf{0}} \mathbf{1}\{(Q(\boldsymbol{m}), M(\boldsymbol{m}), A(\boldsymbol{m}), E(\boldsymbol{m})) \in U, \ \boldsymbol{m} \in V^n\}, \tag{8}$$

where Q, M, A, E are the functions

$$\begin{aligned}
Q(\boldsymbol{m}) =& \frac{1}{n}\|\boldsymbol{m}\|_2^2, \\
M(\boldsymbol{m}) =& \frac{1}{n}\langle\boldsymbol{x}, \boldsymbol{m}\rangle, \\
A(\boldsymbol{m}) =& \frac{1}{n}\sum_{i=1}^{n} m_i \operatorname{arctanh}(m_i), \\
E(\boldsymbol{m}) =& -\frac{1}{n}\sum_{i=1}^{n}\left[h(m_i) + \frac{1}{2}m_i\operatorname{arctanh}(m_i)\right] - \frac{\beta^2}{4}(1 - Q(\boldsymbol{m})^2).
\end{aligned} \tag{9}$$

Note that at any point $\boldsymbol{m}$ where $\nabla\mathcal{F}_{\beta,\lambda}(\boldsymbol{m}) = \mathbf{0}$, we have from Eq. (7) that $\mathcal{F}_{\beta,\lambda}(\boldsymbol{m}) = E(\boldsymbol{m})$.

Define the function $S_\star : (0, 1] \times [-1, 1] \times \mathbb{R}^2 \to \mathbb{R}$ by

$$S_\star(q, \varphi, a, e) = \inf_{(\mu,\nu,\tau,\gamma)\in\mathbb{R}^4} S(q, \varphi, a, e; \mu, \nu, \tau, \gamma), \tag{10}$$

where

$$S(q, \varphi, a, e; \mu, \nu, \tau, \gamma) = \frac{1}{4\beta^2}\left[\frac{a}{q} - \frac{\beta\lambda\varphi^2}{q} - \beta^2(1-q)\right]^2 - q\mu - \varphi\nu - a\tau - [u(q,a) - e]\gamma + \log I, \qquad (11)$$

$$u(q, a) = -\frac{\beta^2}{4}(1-q^2) + \frac{a}{2}, \qquad (12)$$

and

$$I = \int_{-\infty}^{\infty} \frac{1}{(2\pi\beta^2 q)^{1/2}} \exp\left\{-\frac{(x - \beta\lambda\varphi)^2}{2\beta^2 q} + \mu\tanh^2(x) + \nu\tanh(x) + \tau x\tanh(x) + \gamma\log[2\cosh(x)]\right\}\mathrm{d}x. \quad (13)$$

Theorem 1.1. *Fix any $\beta > 0$, $\lambda \geq 0$, $\eta, b \in (0, 1)$, and closed set $U \subseteq [\eta, 1] \times \mathbb{R}^3$. Let $V_n = [-1 + e^{-n^b}, 1 - e^{-n^b}]$. Then*

$$\limsup_{n\to\infty} n^{-1}\log\mathbb{E}[\mathrm{Crit}_n(U, V_n)] \leq \sup_{(q,\varphi,a,e)\in U} S_\star(q, \varphi, a, e).$$

The main proof strategy of Theorem 1.1 is to calculate $\mathbb{E}[\mathrm{Crit}_n(U, V_n)]$ using the Kac-Rice formula [AT09, Theorem 11.2.1]. This is the same strategy as pioneered by Auffinger, Ben Arous and Černý in [AAČ13] (building on early insights by Fyodorov [Fyo04]) in the context of the spherical p-spin model. However, several new challenges arise. First of all, we cannot adopt [AT09, Theorem 11.2.1] directly because of a degeneracy of the conditional Hessian. We present a proof for the Kac-Rice upper bound with degeneracy in Appendix A.

We then evaluate the Kac-Rice formula to leading exponential order. As usual, the most difficult step requires to evaluate the expected determinant of the Hessian. In the spherical model, this is distributed as the determinant of a Wigner matrix, shifted by a term proportional to the identity, and an exact formula was given in [AAČ13]. A slightly more complicated calculation arises for the p-spin spherical model with biased coupling, see [AMMN18], where the Hessian is a low-rank deformation of a Wigner matrix, shifted by a term proportional to the identity. In the present case, the Hessian is distributed as a low-rank deformation of a Wigner matrix *plus a diagonal matrix*, which depends on the point $\boldsymbol{m} \in [-1, +1]^n$. Unlike in earlier work, the bulk of the spectral distribution of the Hessian depends now on the point $\boldsymbol{m}$ (instead of $\|\boldsymbol{m}\|_2$). We give an upper bound of the expected determinant of the Hessian using free probability theory in Appendix B. Finally, we use this expression to upper bound the exponential growth rate of $\mathbb{E}[\mathrm{Crit}_n(U, V_n)]$ using Sanov's Theorem and Varadhan's Lemma in Appendix C. We refer to Section 3 for details of the proof.

Remark 1.1. *We impose the technical conditions that $U \subseteq [\eta, 1] \times \mathbb{R}^3$ and $V_n = [-1 + e^{-n^b}, 1 - e^{-n^b}]$ because of the singular behaviors of the TAP free energy at zero and at the boundary, which are subtle to deal with using the Kac-Rice formula. Instead, the behaviors of $\mathcal{F}_{\beta,\lambda}$ near zero and near the boundary can be easily analyzed by direct arguments.*

Indeed, $\boldsymbol{m} = \boldsymbol{0}$ is always a critical point of $\mathcal{F}_{\beta,\lambda}$. In one interesting case $\beta > 1$ and $\lambda = 0$, it can be easily shown that, $\mathcal{F}_{\beta,\lambda}$ is strongly convex over $D_\eta \equiv \{\boldsymbol{m} \in [-1, 1]^n : Q(\boldsymbol{m}) \leq \eta\}$ with high probability, so that there is no other critical points of $\mathcal{F}_{\beta,\lambda}$ inside D_η. In another interesting case, namely when $\beta = \lambda$ are large enough, the critical points inside D_η have high function value with high probability (see Lemma E.1), so that they are not relevant.

Near the boundary, $\nabla\mathcal{F}_{\beta,\lambda}$ diverges due to the entropy term. It is easy to show that there is no critical points of $\mathcal{F}_{\beta,\lambda}$ within region $\{\boldsymbol{m} : |\|\boldsymbol{m}\|_\infty - 1| \leq \exp(-n^{2/3})\}$ with high probability for sufficiently large n (see Lemma 4.4 for the proof of the case $\beta = \lambda$).

Remark 1.2. *It is interesting to specialize Theorem 1.1 to $\lambda = 0$ (the SK model with zero-mean couplings), and analytically maximize over φ (the value of the magnetization), which is equivalent to setting $\nu = \varphi = 0$. Finally we replace the variable a by Δ via $a = \beta^2 q(1-q) + 2q\Delta$. This results in the reduced expression*

$$S_{0,\star,\beta}(q, \Delta, e) = \inf_{(\mu,\tau,\gamma)\in\mathbb{R}^3} S_0(q, \Delta, e; \mu, \tau, \gamma), \qquad (14)$$

$$S_0(q, \Delta, e; \mu, \tau, \gamma) = \frac{\Delta^2}{\beta^2} - q\mu + e\gamma - \left(\tau + \frac{\gamma}{2}\right)\left(\beta^2 q(1-q) + 2q\Delta\right) + \frac{\beta^2}{4}(1-q^2)\gamma + \log I_0, \qquad (15)$$

$$I_0 = \int_{-\infty}^{\infty} \frac{1}{(2\pi\beta^2 q)^{1/2}} \exp\left\{-\frac{x^2}{2\beta^2 q} + \mu\tanh^2(x) + \tau x\tanh(x) + \gamma\log[2\cosh(x)]\right\}\mathrm{d}x. \qquad (16)$$

4

This coincides with the expression in the statistical physics literature, cf. [BM80, PP95, CGPM03, CLPR03, CLPR04, CLR05]. We refer in particular to [CLPR03] which compares different theoretical physics approaches. Equations (29)-(31) [CLPR03] can be recovered from the above expression by setting $\mu = \lambda_{\text{CLPR}} - (\Delta^2/(2\beta^2 q))$, $\gamma = -u_{\text{CLPR}}$, $\tau = (u_{\text{CLPR}}/2) + (\Delta/(\beta^2 q))$. Minimization over (μ, τ, γ) is then replaced by minimization over $(\lambda_{\text{CLPR}}, u_{\text{CLPR}})$.

Notice that the expression of [CLPR03] involves one extra parameter (denoted by B). However, the authors set it to 0 on the basis of physical considerations motivated by the 'Plefka's criterion' [Ple82].

Remark 1.3. *Still considering (for the sake of simplicity) the case $\lambda = 0$, denote by $H_{\text{SK}}(\boldsymbol{\sigma}) = -\langle \boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma} \rangle / 2$ (where $\boldsymbol{\sigma} \in \{+1, -1\}^n$) the Hamiltonian of the SK model. Substituting $\boldsymbol{m} = \boldsymbol{\sigma}$ in Eq. (6), we obtain the following lower bound on the ground state energy:*

$$\frac{1}{n} \min_{\boldsymbol{\sigma} \in \{-1,+1\}^n} H_{\text{SK}}(\boldsymbol{\sigma}) \geq \frac{1}{\beta} \min_{\boldsymbol{m} \in [-1,+1]^n} \mathcal{F}_{\beta,0}(\boldsymbol{m}). \tag{17}$$

Note that $(1/n) \min_{\boldsymbol{\sigma} \in \{-1,+1\}^n} H_{\text{SK}}(\boldsymbol{\sigma})$ concentrates exponentially around its expectation by Gaussian concentration [BLM13]. Using Markov's inequality and the fact that local minima of $\mathcal{F}_{\beta,0}(\boldsymbol{m})$ occur in the interior of $[-1,+1]^n$ (see Lemma 4.4), we get the following lower bound on the expected ground state energy

$$\lim_{n \to \infty} \inf \frac{1}{n} \mathbb{E} \min_{\boldsymbol{\sigma} \in \{-1,+1\}^n} H_{\text{SK}}(\boldsymbol{\sigma}) \geq F_{\text{1RSB}}(\beta) \equiv \frac{1}{\beta} \inf \left\{ e : \sup_{q, \Delta} S_{0,\star,\beta}(q, \Delta, e) \geq 0 \right\}. \tag{18}$$

(Here $S_{0,\star,\beta}(q, \Delta, e)$ is given by Eq. (14).) In [CGPM03], Cavagna, Giardina, Parisi, and Mézard identify $F_{\text{1RSB}}(\beta)$ with the 'one step replica symmetry breaking' (1RSB) formula for the free energy of the SK model. Hence, our result provides an alternative route to prove the celebrated 1RSB lower bound –first established by Guerra in [Gue03]– in the zero temperature ($\beta = \infty$) case.

1.2 Bayes estimation in $\mathbb{Z}_2$ synchronization

We now return to the $\mathbb{Z}_2$ synchronization model of Eq. (1). As mentioned above, the posterior distribution of $\boldsymbol{x}$ given observation $\boldsymbol{Y}$ is given (under the uniform prior) by the biased SK Gibbs measure (4) with $\beta = \lambda$, cf. Eq. (5). Accordingly, we fix $\beta = \lambda$ throughout this section.

Given the relation between TAP free energy and the Bayes posterior, it is natural to consider the estimator

$$\widehat{\boldsymbol{x}}^{\text{TAP}}(\boldsymbol{Y}) \equiv \arg \min_{\boldsymbol{m} \in [-1,+1]^n} \mathcal{F}_{\lambda,\lambda}(\boldsymbol{m}). \tag{19}$$

Our next theorem provides a characterization of the landscape of the TAP free energy $\mathcal{F}_{\lambda,\lambda}$ for λ a sufficiently large constant, implying in particular that $\widehat{\boldsymbol{x}}^{\text{TAP}}(\boldsymbol{Y})$ is close to the Bayes optimal estimator. More precisely, we show that any critical point $\boldsymbol{m}$ for which $\mathcal{F}_{\lambda,\lambda}(\boldsymbol{m})$ is below a certain constant energy value is such that $\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} \approx \widehat{\boldsymbol{X}}_{\text{Bayes}}$. This implies that any such $\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}$ also asymptotically achieves the Bayes risk MMSE(λ). This result suggests that a good minimizer of $\mathcal{F}_{\lambda,\lambda}$ can be computed by applying a gradient based optimization algorithm, provided the initialization has a value of $\mathcal{F}_{\lambda,\lambda}$ below a constant.

Theorem 1.2. *Denote $\mathcal{C}_{\lambda,n} = \{\boldsymbol{m} \in [-1,1]^n : \nabla \mathcal{F}_{\lambda,\lambda}(\boldsymbol{m}) = \boldsymbol{0}, \mathcal{F}_{\lambda,\lambda}(\boldsymbol{m}) \leq -\lambda^2/3\}$. There exists a constant $\lambda_0 > 0$, such that for any constant $\lambda > \lambda_0$, we have $\mathcal{C}_{\lambda,n} \neq \emptyset$ with high probability, and*

$$\lim_{n \to \infty} \mathbb{E}\left[\left(\sup_{\boldsymbol{m} \in \mathcal{C}_{\lambda,n}} \frac{1}{n^2} \|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\text{Bayes}}\|_F^2\right) \wedge 1\right] = 0. \tag{20}$$

We refer to Section 4 for the proof of this theorem.

Remark 1.4. *Since the TAP free energy is non-convex we do not know whether it can be optimized efficiently. However, the papers [DAM16, MV17] develop an approximate message passing (AMP) algorithm that attempts to construct the solution of the TAP equation (7). The algorithm is iterative, and proceeds as follows, with $\boldsymbol{m}^{-1} = \boldsymbol{0}$:*

$$\begin{aligned} \boldsymbol{m}^0 &= \tanh\left(c_0(\lambda)\sqrt{n}\, \boldsymbol{v}_1(\boldsymbol{Y})\right), \\ \boldsymbol{m}^{k+1} &= \tanh(\lambda \boldsymbol{Y} \boldsymbol{m}^k - \lambda^2[1 - Q(\boldsymbol{m}^k)]\boldsymbol{m}^{k-1}). \end{aligned} \tag{21}$$

Here $\boldsymbol{v}_1(\boldsymbol{Y})$ is the principal eigenvector of $\boldsymbol{Y}$ with unit norm, and $c_0(\lambda)$ is a suitable normalization (see [MV17]). Building on earlier work by Bolthausen [Bol14], as well as on [DAM16], it is shown in [MV17] that the iterates $\boldsymbol{m}^k$ asymptotically achieve the Bayes risk, for any $\lambda > 1$,

$$\lim_{k \to \infty} \lim_{n \to \infty} \frac{1}{n^2} \|\boldsymbol{m}^k (\boldsymbol{m}^k)^\mathsf{T} - \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\|_F^2 = \lim_{n \to \infty} \mathrm{MMSE}_n(\lambda). \tag{22}$$

Together with Theorem 1.2, this result implies that the algorithm is constructing an approximation of $\widehat{\boldsymbol{x}}^{\mathrm{TAP}}(\boldsymbol{Y})$:

$$\lim_{k \to \infty} \lim_{n \to \infty} \mathbb{E}\Big[\frac{1}{n^2} \|\boldsymbol{m}^k (\boldsymbol{m}^k)^\mathsf{T} - \widehat{\boldsymbol{x}}^{\mathrm{TAP}} (\widehat{\boldsymbol{x}}^{\mathrm{TAP}})^\mathsf{T}\|_F^2\Big] = 0 \,. \tag{23}$$

This leaves open two interesting questions: (i) Can we construct an algorithm that converges to $\widehat{\boldsymbol{x}}^{\mathrm{TAP}}(\boldsymbol{Y})$, for fixed n (i.e. can we invert the order of limits in the last equation)? (ii) Do generic optimization algorithms (e.g., gradient descent) converge to $\widehat{\boldsymbol{x}}^{\mathrm{TAP}}(\boldsymbol{Y})$, when applied to $\mathcal{F}_{\lambda,\lambda}$?

Remark 1.5. *While –as discussed in the last remark– the AMP algorithm of [MV17] achieves the Bayes risk, the optimization-based formulation of Eq. (19) has important practical advantages. In particular, we do not know whether how robust is AMP with respect to deviations from the model (1). On the other hand, we expect $\widehat{\boldsymbol{x}}^{\mathrm{TAP}}(\boldsymbol{Y})$ to be robust and meaningful even if the model for the data $\boldsymbol{Y}$ is incorrect.*

The TAP free energy (6) may be contrasted with the "mean field free energy"

$$\mathcal{F}_{\mathrm{MF}}(\boldsymbol{m}) = -\frac{1}{n} \sum_{i=1}^n h(m_i) - \frac{\lambda}{2n} \langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m} \rangle,$$

whose stationary points satisfy the mean field equations

$$\boldsymbol{m} = \tanh(\lambda \boldsymbol{Y} \boldsymbol{m}) \,. \tag{24}$$

Notice that the mean field equations omit the Onsager correction term $-\lambda^2 [1 - Q(\boldsymbol{m})]\boldsymbol{m}$. To clarify the importance of this correction term, we establish the following negative result for critical points $\boldsymbol{m}$ of $\mathcal{F}_{\mathrm{MF}}$.

Theorem 1.3. *Denote $\mathcal{S}_{\lambda,n} = \{\boldsymbol{m} \in (-1,1)^n : \nabla \mathcal{F}_{\mathrm{MF}}(\boldsymbol{m}) = \boldsymbol{0}\}$. There exists a constant $\lambda_0 > 0$ and a constant $\varepsilon(\lambda) > 0$ for every $\lambda > \lambda_0$, such that*

$$\lim_{n \to \infty} \mathbb{P}\Big(\inf_{\boldsymbol{m} \in \mathcal{S}_{\lambda,n}} \frac{1}{n^2} \|\boldsymbol{m}\boldsymbol{m}^\mathsf{T} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2 \geq \varepsilon(\lambda) \Big) = 1. \tag{25}$$

The proof of this theorem is presented in Section 5.

1.3 Notations

We use boldface for vectors and matrices, e.g. $\boldsymbol{X}$ for a matrix and $\boldsymbol{x}$ for a vector. For a univariate function $f : \mathbb{R} \to \mathbb{R}$ and a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n$, we reload the function $f : \mathbb{R}^n \to \mathbb{R}^n$ with $f(\boldsymbol{x}) = (f(x_1), f(x_2), \ldots, f(x_n))^\mathsf{T}$. For example, we write $\boldsymbol{x}^2 = (x_1^2, x_2^2, \ldots, x_n^2)^\mathsf{T}$. We denote $\mathsf{B}^n(\boldsymbol{x}, r)$ to be the Euclidean ball in $\mathbb{R}^n$ centered at $\boldsymbol{x} \in \mathbb{R}^n$ with radius r.

2 Related literature

The Sherrington-Kirkpatrick (SK) model [KS78] is a mean field model of spin glasses. The asymptotics of its log-partition function $\lim_{n \to \infty} n^{-1} \log Z_n(\beta, \lambda)$ (cf. Eq. (4)) was first computed using the replica method [Par79, Par80, Par83], and is expressed using the so-called Parisi's formula. Thouless, Anderson, and Palmer [TAP77] proposed a variational principle in terms of the TAP free energy of Eq. (6).

The relationship between the Gibbs measure (4) (and the partition function) and the TAP free energy have been studied in physics literature [BMY84, CGPM03, DDY83]. The local minimizers of the TAP free

energy are interpreted as the local magnetizations of pure states. As a partial confirmation of this prediction, Talagrand [Tal10] and Chatterjee [Cha10] proved that the TAP equation holds for local magnetizations at high temperature (up to a small error term). Bolthausen [Bol14] derived an iterative algorithm for solving the TAP equations at high temperature. This is an instance of a broader class of iterative algorithms that are now known as approximate message passing (AMP), see e.g. [BM11]. At low temperature regime, Auffinger and Jagannath [AJ16] proved that the TAP equations are satisfied –approximately– by local magnetizations of pure states. Recently, Chen and Panchenko [CP17] proved that the minimum of the TAP free energy (under a suitable constraint on $\boldsymbol{m}$) coincides indeed with the normalized log partition function $n^{-1} \log Z_n(\beta, \lambda)$ up to an error vanishing as $n \to \infty$.

The exponential growth rate of the expected number of critical points of the TAP free energy is also known as 'annealed complexity.' It was first computed by Bray and Moore in [BM80] using the replica method, while an alternative 'supersymmetric' approach was developed in [CGPM03]. The two methods give equivalent formal expressions, which correspond to the function S_0 of Eq. (15). These two papers however report different result for the overall complexity as a function of the free energy, which result from two different ways to set the Lagrange parameters in S_0. This discrepancy was further investigated in [CLPR03] and [CLR05] which suggest that the specific choice should depend on the value of the free energy. This is consistent with the prescription of our Theorem 1.1.

So far, the annealed complexity function has been computed rigorously only for the spherical p-spin model [AAČ13]. For the same model, Subag [Sub17] used the second moment method to show that the typical number of critical points coincides with its expected value for energies in a certain interval. As mentioned above, the main technical challenge in extending these results to the SK model is to compute the conditional expectation of the absolute value of the determinant of the Hessian. (Let us mention that –apart from using non-rigorous methods– all of the physics derivations drop this absolute value, without justification.) Our calculation of this quantity relies on tools from free probability theory [Voi91, Bia97, CDMFF11].

The $\mathbb{Z}_2$ synchronization problem has proven a useful testing ground for comparing different approaches. The Bayes risk was computed in [DAM16]. In particular, a non-trivial correlation with the signal can only be achieved for $\lambda > 1$. The ideal threshold $\lambda = 1$ is achieved both by spectral methods [BAP05] and semidefinite programming [MS16, JMRT16]. However, neither of these approaches achieves the Bayes optimal error.

Variational methods based on free-energy have been the object of recent interest within the statistics literature. Zhang and Zhou [ZZ17] study mean field variational inference applied to the stochastic block model. They prove optimal error rates in a regime that is equivalent to the $\lambda \to \infty$ regime for $\mathbb{Z}_2$ synchronization. In the opposite direction, [GJM18] shows that for $\lambda \in (1/2, 1)$, the minimizer of the mean field free energy does not approximate the Bayes posterior mean. Replacing the mean field with the TAP free energy solves this problem.

3 Proof of Theorem 1.1

In this section we prove Theorem 1.1. There are several key technical results that are needed in the proof. We defer the proof of those results to various appendices.

By symmetry with respect to the sign change for any coordinate

$$x_i \mapsto -x_i, \qquad m_i \mapsto -m_i,$$

we may assume without loss of generality throughout the proof of Theorem 1.1 that $\boldsymbol{x} = \boldsymbol{1}$. Let $f_n(\boldsymbol{m}) = n\mathcal{F}_{\beta,\lambda}(\boldsymbol{m})$ be the scaled TAP free energy. Then f_n and its gradient and Hessian are given by

$$f_n(\boldsymbol{m}) = -\sum_{i=1}^n h(m_i) - \frac{\beta\lambda}{2n}\langle \boldsymbol{1}, \boldsymbol{m} \rangle^2 - \frac{\beta}{2}\langle \boldsymbol{W}, \boldsymbol{m}\boldsymbol{m}^\mathsf{T} \rangle - \frac{n\beta^2}{4}[1 - Q(\boldsymbol{m})]^2, \tag{26}$$

$$\boldsymbol{g}_n(\boldsymbol{m}) = \operatorname{arctanh}(\boldsymbol{m}) - \frac{\beta\lambda}{n}\langle \boldsymbol{1}, \boldsymbol{m} \rangle\boldsymbol{1} - \beta\boldsymbol{W}\boldsymbol{m} + \beta^2[1 - Q(\boldsymbol{m})]\boldsymbol{m}, \tag{27}$$

$$\boldsymbol{H}_n(\boldsymbol{m}) = \operatorname{diag}(1/(1 - \boldsymbol{m}^2)) - \frac{\beta\lambda}{n}\boldsymbol{1}\boldsymbol{1}^\mathsf{T} - \beta\boldsymbol{W} + \beta^2[1 - Q(\boldsymbol{m})]\mathbf{I}_n - \frac{2\beta^2}{n}\boldsymbol{m}\boldsymbol{m}^\mathsf{T}. \tag{28}$$

3.1 Kac-Rice formula

We prove the following statement in Section A, using a version of the Kac-Rice formula.

Proposition 3.1. *Fix any $\beta > 0$ and $\lambda \geq 0$. Denote by $p_{\boldsymbol{m}}(\boldsymbol{0})$ the Lebesgue density of $\boldsymbol{g}_n(\boldsymbol{m})$ at $\boldsymbol{0}$. Then for any $\delta \equiv \delta_n > 0$ and any $U \subseteq [\delta, 1] \times \mathbb{R}^3$ and $V \subseteq [-1 + \delta, 1 - \delta]$, we have*

$$\mathbb{E}[\mathrm{Crit}_n(U, V)] \leq \int_{V^n} \mathbf{1}_{\{(Q(\boldsymbol{m}), M(\boldsymbol{m}), A(\boldsymbol{m}), E(\boldsymbol{m})) \in U\}} \mathbb{E}\big[|\det(\boldsymbol{H}_n(\boldsymbol{m}))|\big|\boldsymbol{g}_n(\boldsymbol{m}) = \boldsymbol{0}\big] p_{\boldsymbol{m}}(\boldsymbol{0}) \mathrm{d}\boldsymbol{m}. \qquad (29)$$

3.2 Determinant of Hessian and density of gradient

We next analyze the expectation on the right side of (29) using techniques from random matrix theory. The proof of the following result is deferred to Section B.

Proposition 3.2. *Fix $\beta > 0$, $\lambda \geq 0$, and $\eta, b \in (0, 1)$. Then there is a constant $n_0 \equiv n_0(\beta, \lambda, \eta, b) > 0$ and $C_0 \equiv C_0(\beta, \lambda, \eta, b)$, such that the following holds for all $n \geq n_0$. Let $\boldsymbol{m}$ satisfy $\|\boldsymbol{m}\|_\infty \leq 1 - e^{-n^b}$ and $Q(\boldsymbol{m}) \in [\eta, 1]$. Define*

$$L(\boldsymbol{m}) = \frac{\beta^2[1 - Q(\boldsymbol{m})]^2}{2} + \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{1 - m_i^2}\right). \qquad (30)$$

Then

$$\mathbb{E}\big[|\det(\boldsymbol{H}_n(\boldsymbol{m}))|\big|\boldsymbol{g}_n(\boldsymbol{m}) = \boldsymbol{0}\big] \leq \exp(n \cdot L(\boldsymbol{m}) + C_0 \cdot n^{\max(0.9, b)}).$$

Define

$$\boldsymbol{u}(\boldsymbol{m}) = [u_1(\boldsymbol{m}), \ldots, u_n(\boldsymbol{m})]^\mathsf{T} = \mathrm{arctanh}(\boldsymbol{m}) - \beta\lambda/n \cdot \langle \mathbf{1}, \boldsymbol{m}\rangle \cdot \mathbf{1} + \beta^2[1 - Q(\boldsymbol{m})]\boldsymbol{m}.$$

The quantity $p_{\boldsymbol{m}}(\boldsymbol{0})$ in (29) may also be evaluated as follows.

Lemma 3.3.

$$p_{\boldsymbol{m}}(\boldsymbol{0}) = \int_{\mathbb{R}} \frac{1}{\{2\pi\beta^2 Q(\boldsymbol{m})\}^{n/2}} \exp\left\{-\sum_{i=1}^{n} \frac{(u_i(\boldsymbol{m}) - ym_i)^2}{2\beta^2 Q(\boldsymbol{m})}\right\} \cdot \sqrt{\frac{n}{2\pi\beta^2}} \exp\left\{-\frac{ny^2}{2\beta^2}\right\}\mathrm{d}y.$$

Proof. First, we observe that

$$\mathbb{E}[(\boldsymbol{W}\boldsymbol{m})_i(\boldsymbol{W}\boldsymbol{m})_j] = Q(\boldsymbol{m}) \cdot \mathbf{1}\{i = j\} + m_i m_j/n.$$

Therefore, we have

$$\boldsymbol{W}\boldsymbol{m} \stackrel{d}{=} Q(\boldsymbol{m})^{1/2}\boldsymbol{v} + Y \cdot \boldsymbol{m},$$

where $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $Y \sim \mathcal{N}(0, 1/n)$ are independent. As a consequence, by (27),

$$\boldsymbol{g}_n(\boldsymbol{m}) \stackrel{d}{=} \boldsymbol{u}(\boldsymbol{m}) - \beta(Q(\boldsymbol{m})^{1/2}\boldsymbol{v} + Y \cdot \boldsymbol{m}),$$

and the form of $p_{\boldsymbol{m}}(\boldsymbol{0})$ follows by first writing the density of $\boldsymbol{g}_n(\boldsymbol{m})$ conditional on βY, and then integrating over the law of βY. $\qquad \square$

Combining Proposition 3.1, Proposition 3.2, and Lemma 3.3 above, we have the following immediate corollary.

Corollary 3.4. *Fix $\beta > 0$, $\lambda \geq 0$, and $\eta, b \in (0, 1)$. There exist $n_0 \equiv n_0(\beta, \lambda, \eta, b)$ and $C_0 \equiv C_0(\beta, \lambda, \eta, b)$ such that for all $n \geq n_0$, any $U \subseteq [\eta, 1] \times \mathbb{R}^3$, and $V_n = [-1 + e^{-n^b}, 1 - e^{-n^b}]$, we have*

$$\mathbb{E}[\mathrm{Crit}_n(U, V_n)] \leq T(U, V_n)$$

where

$$T(U, V_n) = \sqrt{\frac{n}{2\pi\beta^2}} \int_{(-1,1)^n \times \mathbb{R}} \exp\left(nJ(\boldsymbol{m}, y) - \frac{ny^2}{2\beta^2} + C_0 \cdot n^{\max(0.9, b)}\right) \cdot \mathbf{1}_{(Q(\boldsymbol{m}), M(\boldsymbol{m}), A(\boldsymbol{m}), E(\boldsymbol{m})) \in U} \, \mathrm{d}\boldsymbol{m} \, \mathrm{d}y,$$

8

$$J(\boldsymbol{m}, y) = \frac{\beta^2 (1 - Q(\boldsymbol{m}))^2}{2} - \frac{1}{2} \log(2\pi\beta^2 Q(\boldsymbol{m})) + \frac{1}{n} \sum_{i=1}^{n} \log g(m_i; M(\boldsymbol{m}), Q(\boldsymbol{m}), y),$$

and

$$g(x; \varphi, q, y) = \frac{1}{1 - x^2} \cdot \exp\left(-\frac{(\operatorname{arctanh}(x) - \beta\lambda\varphi + \beta^2(1-q)x - yx)^2}{2\beta^2 q} \right). \tag{31}$$

3.3 Variational upper bound

We next analyze the integral $T(U, V_n)$ in Corollary 3.4 using large deviations techniques. Let $\mathcal{P}$ be the space of Borel probability measures on $(-1, 1)$. For $\rho \in \mathcal{P}$, define

$$A(\rho) = \int_{-1}^{1} x \operatorname{arctanh}(x)\, \rho(\mathrm{d}x), \qquad M(\rho) = \int_{-1}^{1} x\, \rho(\mathrm{d}x), \qquad Q(\rho) = \int_{-1}^{1} x^2\, \rho(\mathrm{d}x),$$

$$E(\rho) = -\int_{-1}^{1} \left[h(x) + \frac{1}{2} x \operatorname{arctanh}(x) \right] \rho(\mathrm{d}x) - \frac{\beta^2}{4}(1 - Q(\rho)^2).$$

Here, $A(\rho)$ may equal ∞ and $E(\rho)$ may equal $-\infty$. For $\rho \in \mathcal{P}$ such that $Q(\rho) > 0$ and for $y \in \mathbb{R}$, define the extended real-valued functional

$$J(\rho, y) = \frac{\beta^2 (1 - Q(\rho))^2}{2} - \frac{1}{2} \log(2\pi\beta^2 Q(\rho)) + \int_{-1}^{1} \log g(x; M(\rho), Q(\rho), y)\, \rho(\mathrm{d}x).$$

Note that $\log[1/(1 - x^2)]$ grows at a slower rate than $\operatorname{arctanh}(x)^2$ as $x \to \pm 1$, so $J(\rho, y) < \infty$ always, and $J(\rho, y) > -\infty$ if and only if $\int_{-1}^{1} \operatorname{arctanh}(x)^2\, \rho(\mathrm{d}x) < \infty$. These functionals extend the previous definitions upon identifying $\boldsymbol{m} \in (-1, 1)^n$ with the empirical distribution of its coordinates. We establish the following upper bound in Section C.

Proposition 3.5. *Fix $\beta > 0$, $\lambda \geq 0$, $\eta, b \in (0, 1)$, and any closed set $U \subseteq [\eta, 1] \times \mathbb{R}^3$. Let V_n and $T(U, V_n)$ be as in Corollary 3.4, let π_0 be the uniform distribution on $(-1, 1)$, and let $H(\cdot | \pi_0)$ be the relative entropy to π_0. Then*

$$\limsup_n \frac{1}{n} \log T(U, V_n) \leq \sup_{\rho \in \mathcal{P}:\, (Q(\rho), M(\rho), A(\rho), E(\rho)) \in U} \; \sup_{y \in \mathbb{R}} \left\{ J(\rho, y) - H(\rho | \pi_0) - \frac{y^2}{2\beta^2} + \log 2 \right\}.$$

3.4 Proof of Theorem 1.1

Finally, we conclude the proof of Theorem 1.1 by analyzing the optimization problem in Proposition 3.5 over y and ρ.

Proof. Fix $\rho \in \mathcal{P}$ such that $\sup_{y \in \mathbb{R}}[J(\rho, y) - H(\rho | \pi_0)] > -\infty$. This implies $J(\rho, y) > -\infty$ for some y, so $\int_{-1}^{1} \operatorname{arctanh}(x)^2\, \rho(\mathrm{d}x) < \infty$. Furthermore $H(\rho | \pi_0) < \infty$, so in particular $\rho(\mathrm{d}x) = f(x)\mathrm{d}x$ for some density function $f(x)$ on $(-1, 1)$. Let $(q, \varphi, a, e) = (Q(\rho), M(\rho), A(\rho), E(\rho))$. Then

$$\int \log g(x; \varphi, q, y)\, \rho(\mathrm{d}x) = \chi_1 y^2 + \chi_2 y + \chi_3,$$

where

$$\chi_1 = -\int \frac{x^2}{2\beta^2 q}\rho(\mathrm{d}x) = -\frac{1}{2\beta^2},$$

$$\chi_2 = \int \frac{x(\operatorname{arctanh}(x) - \beta\lambda\varphi + \beta^2(1-q)x)}{\beta^2 q}\rho(\mathrm{d}x) = \frac{a}{\beta^2 q} - \frac{\lambda}{\beta}\frac{\varphi^2}{q} + 1 - q,$$

$$\chi_3 = \int \left(\log \frac{1}{1 - x^2} - \frac{(\operatorname{arctanh}(x) - \beta\lambda\varphi + \beta^2(1-q)x)^2}{2\beta^2 q} \right) \rho(\mathrm{d}x)$$

$$= \int \left(\log \frac{1}{1 - x^2} - \frac{\operatorname{arctanh}(x)^2}{2\beta^2 q} + \frac{\lambda\varphi\operatorname{arctanh}(x)}{\beta q} \right) \rho(\mathrm{d}x) + \frac{(2\beta\lambda(1-q) - \lambda^2)\varphi^2 - 2a(1-q)}{2q} - \frac{\beta^2(1-q)^2}{2}.$$

9

Let $\rho_{\varphi,q}$ be the law of $\tanh(z)$ where $z \sim \mathcal{N}(\beta\lambda\varphi, \beta^2 q)$. Note that

$$\int_{-1}^{1} \frac{1}{1-x^2} \exp\left(-\frac{\operatorname{arctanh}(x)^2}{2\beta^2 q} + \frac{\lambda\varphi \operatorname{arctanh}(x)}{\beta q}\right) \mathrm{d}x = \int_{\mathbb{R}} e^{-y^2/(2\beta^2 q)+\lambda\varphi y/(\beta q)} dy = e^{\lambda^2\varphi^2/(2q)}(2\pi\beta^2 q)^{1/2},$$

and

$$h_{\varphi,q}(x) = e^{-\lambda^2\varphi^2/(2q)}(2\pi\beta^2 q)^{-1/2}\frac{1}{1-x^2}\exp\left(-\frac{\operatorname{arctanh}(x)^2}{2\beta^2 q} + \frac{\lambda\varphi \operatorname{arctanh}(x)}{\beta q}\right)$$

is the density function of the law $\rho_{\varphi,q}$. Then

$$\chi_3 = \int \log h_{\varphi,q}(x)\rho(\mathrm{d}x) + \frac{1}{2}\log(2\pi\beta^2 q) + \chi_4, \quad \chi_4 = \frac{(\beta\lambda\varphi^2 - a)(1-q)}{q} - \frac{\beta^2(1-q)^2}{2}.$$

Note further that

$$\sup_{y\in\mathbb{R}}\left[\left(\chi_1 - \frac{1}{2\beta^2}\right)y^2 + \chi_2 y\right] = \sup_{y\in\mathbb{R}}\left[-\frac{1}{\beta^2}y^2 + \chi_2 y\right] = \frac{\beta^2\chi_2^2}{4},$$

and $H(\rho|\pi_0) = \log 2 + \int_{-1}^{1} \log f(x)\,\rho(\mathrm{d}x)$, where $f(x)$ is the density of ρ. We hence obtain

$$\sup_{y\in\mathbb{R}}\left[J(\rho, y) - H(\rho|\pi_0) - y^2/(2\beta^2) + \log 2\right] = \frac{\beta^2(1-q)^2}{2} + \frac{\beta^2\chi_2^2}{4} + \chi_4 - \int_{-1}^{1}\log\frac{f(x)}{h_{\varphi,q}(x)}\rho(\mathrm{d}x)$$

$$= \frac{1}{4\beta^2}\left(\beta^2(1-q) + \frac{\beta\lambda\varphi^2}{q} - \frac{a}{q}\right)^2 - H(\rho|\rho_{\varphi,q}),$$

where the second line applies some algebra to simplify $\beta^2(1-q)^2/2 + \beta^2\chi_2^2/4 + \chi_4$. Then, combining this with Corollary 3.4 and Proposition 3.5,

$$\limsup_{n} n^{-1}\log \mathbb{E}[\operatorname{Crit}_n(U, V_n)] \leq \sup_{(q,\varphi,a,e)\in U}\left[\frac{1}{4\beta^2}\left(\beta^2(1-q) + \frac{\beta\lambda\varphi^2}{q} - \frac{a}{q}\right)^2 - \inf_{\rho\in\mathcal{P}(q,\varphi,a,e)} H(\rho|\rho_{\varphi,q})\right]$$

where $\mathcal{P}(q,\varphi,a,e) = \{\rho \in \mathcal{P} : (Q(\rho), M(\rho), A(\rho), E(\rho)) = (q,\varphi,a,e)\}$.

Recalling $u(q,a)$ from (12) and introducing Lagrange multipliers for these constraints, we have

$$\inf_{\rho\in\mathcal{P}(q,\varphi,a,e)} H(\rho|\rho_{\varphi,q})$$

$$= \inf_{\rho\in\mathcal{P}}\sup_{\mu,\nu,\tau,\gamma} H(\rho|\rho_{\varphi,q}) - \mu(Q(\rho) - q) - \nu(M(\rho) - \varphi) - \tau(A(\rho) - a) - \gamma[u(Q(\rho), A(\rho)) - E(\rho) - u(q,a) + e]$$

$$\geq \sup_{\mu,\nu,\tau,\gamma} \mu q + \nu\varphi + \tau a + \gamma[u(q,a) - e] + \inf_{\rho\in\mathcal{P}} F(\rho; \mu, \nu, \tau, \gamma),$$

where

$$F(\rho; \mu, \nu, \tau, \gamma) = H(\rho|\rho_{\varphi,q}) - \mu Q(\rho) - \nu M(\rho) - \tau A(\rho) - \gamma[u(Q(\rho), A(\rho)) - E(\rho)]$$

$$= \int_{-1}^{1}\left(\log\frac{f(x)}{h_{\varphi,q}(x)} - \mu x^2 - \nu x - \tau x\operatorname{arctanh}(x) - \gamma\left(-\frac{1}{2}\log\frac{1+x}{2} - \frac{1}{2}\log\frac{1-x}{2}\right)\right)\rho(\mathrm{d}x)$$

$$= H(\rho|\rho_{\mu,\nu,\tau,\gamma}) - \log I.$$

Here,

$$\rho_{\mu,\nu,\tau,\gamma}(\mathrm{d}x) = I^{-1}\exp\left(\mu x^2 + \nu x + \tau x\operatorname{arctanh}(x) + \gamma\left(-\frac{1}{2}\log\frac{1+x}{2} - \frac{1}{2}\log\frac{1-x}{2}\right)\right)h_{\varphi,q}(x)\mathrm{d}x$$

is a suitably defined exponential family density with base measure $h_{\varphi,q}(x)\mathrm{d}x$, and I as defined in (13) is the normalizing constant for this density. Thus

$$\inf_{\rho\in\mathcal{P}} F(\rho; \mu, \nu, \tau, \gamma) = -\log I$$

with the infimum achieved at $\rho = \rho_{\mu,\nu,\tau,\gamma}$, and Theorem 1.1 follows. $\qquad\square$

4 Proof of Theorem 1.2

Throughout the proof, we write $\mathcal{F}$ as shorthand for $\mathcal{F}_{\lambda,\lambda}$. Let $q_\star$ be the largest real solution of the equation

$$q_\star = \mathbb{E}_G[\tanh(\lambda^2 q_\star + \sqrt{\lambda^2 q_\star}G)^2], \tag{32}$$

for $G \sim \mathcal{N}(0,1)$. Note that when $\lambda > 1$, we have $q_\star > 0$ [DAM16, Lemma 4.2]. Denote

$$\varphi_\star = q_\star, \qquad a_\star = \lambda^2 q_\star, \qquad e_\star = -\frac{\lambda^2}{4}(1 - 2q_\star - q_\star^2) - \mathbb{E}_G\{\log[2\cosh(\lambda^2 q_\star + \sqrt{\lambda^2 q_\star}G)]\}, \tag{33}$$

and recall

$$Q(\boldsymbol{m}) = \frac{1}{n}\|\boldsymbol{m}\|_2^2, \qquad M(\boldsymbol{m}) = \frac{1}{n}\langle \boldsymbol{x}, \boldsymbol{m}\rangle.$$

Recall $\mathcal{C}_{\lambda,n}$ the set of critical points of $\mathcal{F}$ with energy below $-\lambda^2/3$, and define

$$\mathcal{C}_{\lambda,n}^+ \equiv \{\boldsymbol{m} \in [-1,1]^n : \nabla\mathcal{F}(\boldsymbol{m}) = \boldsymbol{0}, \ \mathcal{F}(\boldsymbol{m}) \leq -\lambda^2/3, \ M(\boldsymbol{m}) \geq 0\}$$

so that $\mathcal{C}_{\lambda,n} = \mathcal{C}_{\lambda,n}^+ \cup -\mathcal{C}_{\lambda,n}^+$. Denote $\Gamma_\varepsilon \subseteq [-1,1]^n$ to be the pre-image of $(Q, M, \mathcal{F})$ at a neighborhood of $(q_\star, \varphi_\star, e_\star)$:

$$\Gamma_\varepsilon \equiv \{\boldsymbol{m} \in [-1,1]^n : |Q(\boldsymbol{m}) - q_\star| \leq \varepsilon, \ |M(\boldsymbol{m}) - \varphi_\star| \leq \varepsilon, \ |\mathcal{F}(\boldsymbol{m}) - e_\star| < \varepsilon\}. \tag{34}$$

The following shows that $\emptyset \neq \mathcal{C}_{\lambda,n}^+ \subseteq \Gamma_\varepsilon$ with high probability:

Proposition 4.1. *There exists $\lambda_0 < \infty$ such that for any $\lambda \geq \lambda_0$ and $\varepsilon > 0$, for some constants $n_0 \equiv n_0(\lambda, \varepsilon)$ and $c_0 \equiv c_0(\lambda, \varepsilon)$, with probability at least $1 - e^{-c_0 n}$ for all $n \geq n_0$, we have: $\mathcal{C}_{\lambda,n}^+ \subseteq \Gamma_\varepsilon$, and some global minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}$ satisfies $\boldsymbol{m}_\star \in \mathcal{C}_{\lambda,n}^+ \subseteq \Gamma_\varepsilon$.*

4.1 Proof of Proposition 4.1

Recall that in Theorem 1.1, for $\lambda > 0$, choosing $\beta = \lambda$, taking $\eta, b \in (0,1)$, and taking closed sets $U \subseteq [\eta, 1] \times \mathbb{R}^3$, $V_n = [-1 + e^{-n^b}, 1 - e^{-n^b}]$, we have

$$\limsup_{n\to\infty} n^{-1}\log\mathbb{E}[\mathrm{Crit}_n(U, V_n)] \leq \sup_{(q,\varphi,a,e)\in U} S_\star(q, \varphi, a, e).$$

The following proposition bounds the function $S_\star(q, \varphi, a, e)$ for (q, φ) in a small neighborhood of $(q_\star, \varphi_\star)$. Its proof is contained in Section D.

Proposition 4.2. *For any $\lambda > 0$, we have*

$$S_\star(q_\star, \varphi_\star, a_\star, e_\star) = 0.$$

Furthermore, there exist constants $c_0, \lambda_0 > 0$ and a function $c(\lambda) > 0$ such that for any $\lambda \geq \lambda_0$ and any (q, φ) satisfying $|q - q_\star|, |\varphi - \varphi_\star| \leq c_0/\lambda^2$, we have

$$S_\star(q, \varphi, a, e) \leq -c(\lambda)[(a - a_\star)^2 + (q - q_\star)^2 + (\varphi - \varphi_\star)^2 + (e - e_\star)^2]. \tag{35}$$

Note that for c_0, λ_0 as above and $\lambda \geq \lambda_0$, this implies

$$\sup\{S_\star(q, \varphi, a, e) : |q - q_\star|, |\varphi - \varphi_\star| \leq c_0/\lambda^2, (a, e) \in \mathbb{R}^2\} \leq 0,$$

with the supremum uniquely attained at $(q_\star, \varphi_\star, a_\star, e_\star)$.

The following result localizes any critical point of $\mathcal{F}$ satisfying Theorem 4.1 to the above neighborhood of $(q_\star, \varphi_\star)$. Its proof is in Section E.

Proposition 4.3. *Fix any positive integer k. Then there exist $\lambda_0, C_0 > 0$ and functions $c_0(\lambda), n_0(\lambda) > 0$ such that for all $\lambda \geq \lambda_0$ and $n \geq n_0(\lambda)$, with probability at least $1 - e^{-c_0(\lambda)n}$, any $\boldsymbol{m} \in \mathcal{C}_{\lambda,n}^+$ also satisfies $Q(\boldsymbol{m}), M(\boldsymbol{m}) \geq 1 - C_0/\lambda^k$.*

Finally, the following elementary lemma shows that with high probability, there are no critical points with any coordinate outside the set V_n in Theorem 1.1.

Lemma 4.4. *There exist constants n_0, and $c_0 > 0$ so that for $n \geq n_0$, with probability at least $1 - e^{-c_0 n}$, all points $\boldsymbol{m} \in (-1, 1)^n$ satisfying $\boldsymbol{g}_n(\boldsymbol{m}) = \boldsymbol{0}$ also satisfy*

$$\|\boldsymbol{m}\|_\infty \leq 1 - e^{-4\lambda^2 - 6\lambda\sqrt{n}}. \tag{36}$$

Proof. Note that when $\boldsymbol{g}_n(\boldsymbol{m}) = \boldsymbol{0}$, we have

$$\|\operatorname{arctanh}(\boldsymbol{m})\|_\infty = \left\| -\lambda^2/n \cdot \langle \boldsymbol{1}, \boldsymbol{m} \rangle \boldsymbol{1} - \lambda \cdot \boldsymbol{W}\boldsymbol{m} + \lambda^2[1 - Q(\boldsymbol{m})]\boldsymbol{m} \right\|_\infty \leq 2\lambda^2 + \lambda\|\boldsymbol{W}\|_{\mathrm{op}}\sqrt{n}.$$

With probability at least $1 - e^{-c_0 n}$ for $n \geq n_0$ and some $n_0 > 0$ and $c_0 > 0$, we have $\|\boldsymbol{W}\|_{\mathrm{op}} \leq 3$, and hence $\|\operatorname{arctanh}(\boldsymbol{m})\|_\infty \leq 2\lambda^2 + 3\lambda\sqrt{n}$. Applying $\tanh(x) \leq 1 - e^{-2x}$ for $x \geq 0$, with probability at least $1 - e^{-c_0 n}$, we have

$$\|\boldsymbol{m}\|_\infty \leq \tanh(\|\operatorname{arctanh}(\boldsymbol{m})\|_\infty) \leq 1 - e^{-4\lambda^2 - 6\lambda\sqrt{n}}. \tag{37}$$

$\square$

Proof of Proposition 4.1. According to Lemma 4.4 and Proposition 4.3, there exist constants λ_0, C_0 and functions $c_0(\lambda), n_0(\lambda)$ such that for $\lambda \geq \lambda_0$ and $n \geq n_0(\lambda)$, the following happens with probability at least $1 - e^{-c_0 n}$: For any $\boldsymbol{m} \in \mathcal{C}_{\lambda,n}^+$, we have $\|\boldsymbol{m}\|_\infty \leq 1 - e^{-4\lambda^2 - 6\lambda\sqrt{n}} \leq 1 - e^{-\lambda^2\sqrt{n}}$ and $Q(\boldsymbol{m}), M(\boldsymbol{m}) \geq 1 - C_0/\lambda^3$.

Define

$$U_\varepsilon = \{(q, \varphi, a, e) : (q, \varphi) \in [1 - C_0/\lambda^3, 1]^2, (a, e) \in \mathbb{R}^2\} \cap \{(q, \varphi, a, e) : \max\{|q - q_\star|, |\varphi - \varphi_\star|, |e - e_\star|\} \geq \varepsilon\},$$

and $V_n = [-1 + e^{-\lambda^2\sqrt{n}}, 1 - e^{-\lambda^2\sqrt{n}}]$. According to Theorem 1.1, we have

$$\limsup_{n \to \infty} n^{-1} \log \mathbb{E}[\operatorname{Crit}_n(U_\varepsilon, V_n)] \leq \sup_{(q,\varphi,a,e) \in U_\varepsilon} S_\star(q, \varphi, a, e).$$

According to Proposition 4.2, for $\lambda \geq \lambda_0$ with λ_0 large enough, we have

$$0 > \sup_{(q,\varphi,a,e) \in U_\varepsilon} S_\star(q, \varphi, a, e) \equiv -s_0(\lambda, \varepsilon).$$

Therefore, there exists $n_0(\lambda, \varepsilon)$ large enough so that for $n \geq n_0(\lambda, \varepsilon)$,

$$n^{-1} \log \mathbb{E}[\operatorname{Crit}_n(U_\varepsilon, V_n)] \leq -s_0(\lambda, \varepsilon)/2.$$

Accordingly, by Markov's inequality, we have

$$\mathbb{P}(\operatorname{Crit}_n(U_\varepsilon, V_n) \geq 1) \leq \mathbb{E}[\operatorname{Crit}_n(U_\varepsilon, V_n)] \leq \exp\{-s_0 n/2\}.$$

Combining the above statements concludes that $\mathcal{C}_{\lambda,n}^+ \subseteq \Gamma_\varepsilon$.

Next we show that a global minimizer $\boldsymbol{m}_\star \in \mathcal{C}_{\lambda,n}^+$. The gradient of $\mathcal{F}$ diverges at the boundary of $[-1, 1]^n$ and points outside $[-1, 1]^n$ because of the entropy term. Hence, any global minimizer $\boldsymbol{m}_\star$ belongs to $(-1, 1)^n$ and satisfies $\nabla\mathcal{F}(\boldsymbol{m}_\star) = \boldsymbol{0}$. Furthermore, note $\mathcal{F}(\boldsymbol{x}) = -\lambda^2/2 - \lambda\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x} \rangle/(2n) \leq -\lambda^2/2 + \lambda\|\boldsymbol{W}\|_{\mathrm{op}}/2$, and $\mathbb{P}(\|\boldsymbol{W}\|_{\mathrm{op}} \leq 3) \geq 1 - e^{-c_0 n}$ for $n \geq n_0$ with some constant c_0 and n_0. Accordingly, we have $\mathcal{F}(\boldsymbol{m}_\star) \leq \mathcal{F}(\boldsymbol{x}) \leq -\lambda^2/2 + 3\lambda/2 \leq -\lambda^2/3$ for $\lambda \geq 9$ and $n \geq n_0$ with probability at least $1 - e^{-c_0 n}$. Taking $-\boldsymbol{m}_\star$ in place of $\boldsymbol{m}_\star$ if necessary, we ensure $M(\boldsymbol{m}_\star) \geq 0$. $\square$

4.2 Proof of Theorem 1.2

In this regime, [DAM16] showed that $\operatorname{MMSE}_n(\lambda)$ converges to a limiting value which admits the following characterization.

Proposition 4.5 ([DAM16]). *Let $\lambda > 0$ be a fixed constant. Let $q_\star$ be the largest nonnegative solution to the equation*

$$q_\star = \mathbb{E}_{G \sim \mathcal{N}(0,1)}\Big[\tanh(\lambda^2 q_\star + \sqrt{\lambda^2 q_\star}\, G)^2\Big]. \tag{38}$$

Then

$$\lim_{n \to \infty} \mathrm{MMSE}_n(\lambda) = 1 - q_\star^2. \tag{39}$$

The value $q_\star$ is monotonically increasing in λ, with $q_\star = 0$ when $\lambda \leq 1$ and $q_\star \in (0,1)$ when $\lambda > 1$.

Proof of Proposition 4.5. See [DAM16, Eqs. (167) and (143)] for (39). (The notational identification with [DAM16] is $\gamma_*/\lambda \leftrightarrow q_\star$ and $\lambda \leftrightarrow \lambda^2$. [DAM16] establishes this result with additional direct observations under a binary erasure channel with erasure probability $1 - \varepsilon$, [DAM16, Eq. (25)], and the statement for $\varepsilon = 0$ follows from continuity in ε.) For the properties of $q_\star$, see [DAM16, Lemma 4.2]. $\qquad\square$

Combining Proposition 4.1 and Proposition 4.5, we immediately derive that any local minimizer of the TAP free energy asymptotically achieves the Bayes risk, and hence is close to the Bayes posterior mean.

Proof of Theorem 1.2. Note that $\mathcal{C}_{\lambda,n}$ is compact. Hence, when $\mathcal{C}_{\lambda,n} \neq \emptyset$, the maximum value of $\|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2$ over $\mathcal{C}_{\lambda,n}$ can be attained by some $\boldsymbol{m} \in [-1,1]^n$. We denote $\boldsymbol{m}_\star \in \mathbb{R}^n$ to be the following

$$\boldsymbol{m}_\star = \begin{cases} \text{a random element within the set } \arg\max_{\boldsymbol{m} \in \mathcal{C}(\lambda,n)} \|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2, \text{ for } \mathcal{C}_{\lambda,n} \neq \emptyset; \\ \boldsymbol{0}, \text{ for } \mathcal{C}_{\lambda,n} = \emptyset. \end{cases}$$

For this definition of $\boldsymbol{m}_\star$, we have $\boldsymbol{m}_\star$ is independent of $\boldsymbol{x}$ conditional on $\boldsymbol{Y}$.

By Proposition 4.1, for any constants $\varepsilon, \delta > 0$, for sufficiently large n, with probability at least $1 - \delta$, we have $\mathcal{C}_{\lambda,n} \neq \emptyset$, and

$$n^{-1}\|\boldsymbol{m}_\star\|^2 \in [q_\star - \varepsilon, q_\star + \varepsilon], \qquad n^{-1}|\langle \boldsymbol{m}_\star, \boldsymbol{x}\rangle| \in [q_\star - \varepsilon, q_\star + \varepsilon]. \tag{40}$$

Denoting the above good event to be $\mathcal{E}$. Then we have

$$n^{-2}\mathbb{E}\Big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2 \Big| \mathcal{E}\Big] = n^{-2}\mathbb{E}\Big[\|\boldsymbol{x}\|^4 + \|\boldsymbol{m}_\star\|^4 - 2|\langle \boldsymbol{m}_\star, \boldsymbol{x}\rangle|^2 \Big| \mathcal{E}\Big]$$
$$\leq \Big(1 + (q_\star + \varepsilon)^2 - 2(q_\star - \varepsilon)^2\Big) \leq 1 - q_\star^2 + 6\varepsilon,$$

so that

$$n^{-2}\mathbb{E}\Big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2\Big] \leq n^{-2}\mathbb{E}\Big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2 \Big| \mathcal{E}\Big]\mathbb{P}(\mathcal{E}) + 2 \cdot \mathbb{P}(\mathcal{E}^c) \leq 1 - q_\star^2 + 6\varepsilon + 2\delta.$$

As $\boldsymbol{m}_\star$ is independent of $\boldsymbol{x}$ conditional on $\boldsymbol{Y}$, we have

$$n^{-2}\mathbb{E}\big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2\big] = n^{-2}\mathbb{E}\big[\mathbb{E}\big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2 | \boldsymbol{Y}\big]\big]$$
$$= n^{-2}\mathbb{E}\Big[\mathbb{E}\Big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\boldsymbol{Y}]\|_F^2 + \|\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\boldsymbol{Y}] - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2 \Big| \boldsymbol{Y}\Big]\Big]$$
$$= n^{-2}\mathbb{E}\Big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2\Big] + \mathrm{MMSE}_n(\lambda).$$

Then Proposition 4.5 implies, for all $n \geq n_0(\lambda, \varepsilon, \delta)$,

$$n^{-2}\mathbb{E}\Big[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2\Big] < 7\varepsilon + 2\delta.$$

Then by the definition of $\boldsymbol{m}_\star$, we have

$$\Big(n^{-2} \sup_{\boldsymbol{m} \in \mathcal{C}_{\lambda,n}} \|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2\Big) \wedge 1 \leq n^{-2}\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2,$$

so that (20) holds. $\qquad\square$

5 Proof of Theorem 1.3

Proof of Theorem 1.3. Let $\boldsymbol{m}_\star$ be any critical point of the TAP free energy $\mathcal{F} = \mathcal{F}_{\lambda,\lambda}$, and let $\boldsymbol{m}$ be any critical point of $\mathcal{F}_{\mathrm{MF}}$. By the conditions $\boldsymbol{0} = \nabla\mathcal{F}(\boldsymbol{m}_\star)$ and $\boldsymbol{0} = \nabla\mathcal{F}_{\mathrm{MF}}(\boldsymbol{m})$, we have

$$\boldsymbol{m}_\star = \tanh(\lambda \cdot \boldsymbol{Y}\boldsymbol{m}_\star - \lambda^2[1 - Q(\boldsymbol{m}_\star)]\boldsymbol{m}_\star), \tag{41}$$

$$\boldsymbol{m} = \tanh(\lambda \cdot \boldsymbol{Y}\boldsymbol{m}). \tag{42}$$

For constants $C_0, t > 0$ to be chosen later, consider the event $\mathcal{E}$ where

$$\mathcal{E} = \Big\{ \sup_{\boldsymbol{u} \in \mathsf{B}^n(\boldsymbol{0},1)} \Big[\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|(\boldsymbol{W}\boldsymbol{u})_i| \geq t/\sqrt{n}\} \Big] < C_0/t^2 \Big\},$$

and $\mathsf{B}^n(\boldsymbol{0},1)$ denotes the unit ball around $\boldsymbol{0}$. Note that $\boldsymbol{W} = (\boldsymbol{G} + \boldsymbol{G}^\mathsf{T})/\sqrt{2n}$ with $G_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for $1 \leq i, j \leq n$, so that

$$\sup_{\boldsymbol{u} \in \mathsf{B}^n(\boldsymbol{0},1)} \Big[\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|(\boldsymbol{W}\boldsymbol{u})_i| \geq \frac{t}{\sqrt{n}}\} \Big]$$

$$\leq \sup_{\boldsymbol{u} \in \mathsf{B}^n(\boldsymbol{0},1)} \Big[\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|(\boldsymbol{G}\boldsymbol{u})_i| \geq \frac{t}{\sqrt{2}}\} \Big] + \sup_{\boldsymbol{u} \in \mathsf{B}^n(\boldsymbol{0},1)} \Big[\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|(\boldsymbol{G}^\mathsf{T}\boldsymbol{u})_i| \geq \frac{t}{\sqrt{2}}\} \Big].$$

Then by Lemma E.2, for some constant $c_0, C_0 > 0$ and for any t sufficiently large, $\mathcal{E}$ holds with probability at least $1 - e^{-c_0 n}$. Define the (random) index set

$$\mathcal{I} = \Big\{ i \in \{1, \dots, n\} : |(\boldsymbol{W}\boldsymbol{m})_i| < t \text{ and } |(\boldsymbol{W}\boldsymbol{m}_\star)_i| < t \Big\}.$$

As $\boldsymbol{m}/\sqrt{n} \in \mathsf{B}^n(\boldsymbol{0},1)$ and similarly for $\boldsymbol{m}_\star$, we have on $\mathcal{E}$ that $|\mathcal{I}|/n \geq 1 - 2C_0/t^2$. Applying $|(\boldsymbol{Y}\boldsymbol{m})_i| \leq |\lambda| + |(\boldsymbol{W}\boldsymbol{m})_i|$ by (1), for $i \in \mathcal{I}$ we have

$$|\lambda \cdot (\boldsymbol{Y}\boldsymbol{m}_\star)_i - \lambda^2[1 - Q(\boldsymbol{m}_\star)]m_{\star,i}| \leq 2\lambda^2 + \lambda t,$$

$$|\lambda \cdot (\boldsymbol{Y}\boldsymbol{m})_i| \leq \lambda^2 + \lambda t.$$

Then taking the difference of (41) and (42) and applying the lower bound $\tanh'(x) \geq c(\lambda, t)$ for all $|x| \leq 2\lambda^2 + \lambda t$ and a constant $c(\lambda, t) > 0$, we have

$$|m_{\star,i} - m_i| \geq c(\lambda, t)\Big| \lambda \cdot (\boldsymbol{Y}\boldsymbol{m}_\star - \boldsymbol{Y}\boldsymbol{m})_i - \lambda^2[1 - Q(\boldsymbol{m}_\star)]m_{\star,i} \Big|.$$

Then

$$\lambda^4[1 - Q(\boldsymbol{m}_\star)]^2 m_{\star,i}^2 \leq 2c(\lambda, t)^{-2}|m_{\star,i} - m_i|^2 + 2\lambda^2 \cdot (\boldsymbol{Y}\boldsymbol{m}_\star - \boldsymbol{Y}\boldsymbol{m})_i^2.$$

Summing over $i \in \{1, \dots, n\}$ and using the trivial bound $m_{\star,i}^2 \leq 1$ for $i \notin \mathcal{I}$, we obtain

$$\lambda^4[1 - Q(\boldsymbol{m}_\star)]^2 \|\boldsymbol{m}_\star\|_2^2 \leq \lambda^4[1 - Q(\boldsymbol{m}_\star)]^2 (n - |\mathcal{I}|) + 2c(\lambda, t)^{-2}\|\boldsymbol{m}_\star - \boldsymbol{m}\|_2^2 + 2\lambda^2\|\boldsymbol{Y}\|_{\mathrm{op}}^2\|\boldsymbol{m}_\star - \boldsymbol{m}\|_2^2.$$

Suppose λ is sufficiently large such that the conclusion $|Q(\boldsymbol{m}_\star) - q_\star| < q_\star/4$ from 4.1 holds with probability $1 - \delta/4$. Choose t large enough such that $2C_0/t^2 < q_\star/2$, so $|\mathcal{I}|/n > 1 - q_\star/2$ on the event $\mathcal{E}$. Note also that with probability $1 - e^{-c_0 n}$ for some constant $c_0 > 0$, we have $\|\boldsymbol{Y}\|_{\mathrm{op}} \leq \lambda + 3$. Then on a combined event of probability $1 - \delta/2$, for any critical points $\boldsymbol{m}_\star$ and $\boldsymbol{m}$ as above, we obtain

$$\frac{1}{n}\|\boldsymbol{m}_\star - \boldsymbol{m}\|_2^2 \geq c(\lambda)$$

for some constant $c(\lambda) > 0$.

Consider now any global minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}$. Note that $-\boldsymbol{m}_\star$ is also a global minimizer of $\mathcal{F}$, so on this combined event, we also have $n^{-1}\|\boldsymbol{m}_\star + \boldsymbol{m}\|_2^2 \geq c(\lambda)$. Then

$$
\begin{aligned}
\frac{1}{n^2}\|\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} - \boldsymbol{m}\boldsymbol{m}^\mathsf{T}\|_F^2 &= \frac{1}{n^2}\left(\|\boldsymbol{m}_\star\|_2^4 + \|\boldsymbol{m}\|_2^4 - 2|\langle \boldsymbol{m}, \boldsymbol{m}_\star\rangle|^2\right) \\
&\geq \frac{1}{2n^2}\|\boldsymbol{m}_\star - \boldsymbol{m}\|_2^2\|\boldsymbol{m}_\star + \boldsymbol{m}\|_2^2 \geq c(\lambda)^2/2.
\end{aligned}
\tag{43}
$$

On the other hand, note that

$$
\mathcal{F}(\boldsymbol{m}_\star) \leq \mathcal{F}(\boldsymbol{x}) = -\frac{\lambda^2}{2} - \frac{\lambda}{2n}\langle \boldsymbol{W}, \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\rangle \leq -\frac{\lambda^2}{2} + \lambda,
$$

where the last inequality holds with probability at least $1 - e^{-n}$ by the fact that $\langle \boldsymbol{W}, \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\rangle \sim \mathcal{N}(0, 2n)$. For λ sufficiently large, we have $-\lambda^2/2 + \lambda \leq -\lambda^2/3$. Then Theorem 1.2 implies, with probability at least $1 - \delta/2$,

$$
\frac{1}{n^2}\|\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2 < c(\lambda)^2/8.
$$

Combining this with (43) and setting $\varepsilon(\lambda) = c(\lambda)^2/8$, with probability at least $1 - \delta$, for any critical point $\boldsymbol{m}$ of $\mathcal{F}_{\mathrm{MF}}$ we obtain

$$
\frac{1}{n^2}\|\boldsymbol{m}\boldsymbol{m}^\mathsf{T} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2 \geq \frac{1}{2n^2}\|\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} - \boldsymbol{m}\boldsymbol{m}^\mathsf{T}\|_F^2 - \frac{1}{n^2}\|\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2 > \varepsilon(\lambda).
$$

$\square$

Acknowledgments

A Proof of Proposition 3.1

Fix n and write as shorthand $\boldsymbol{g} \equiv \boldsymbol{g}_n$ and $\boldsymbol{H} \equiv \boldsymbol{H}_n$. For any measurable $T \subseteq (-1,1)^n$, define

$$\mathrm{Crit}(T) = \sum_{\boldsymbol{m} : \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}} \mathbf{1}\{\boldsymbol{m} \in T\}.$$

We wish to apply the Kac-Rice formula [AT09, Theorem 11.2.1] for $\mathbb{E}[\mathrm{Crit}(T)]$. The statement of [AT09, Theorem 11.2.1] does not directly apply in our setting, as $\boldsymbol{g}(\boldsymbol{m})$ and $\boldsymbol{H}(\boldsymbol{m}) = \nabla \boldsymbol{g}(\boldsymbol{m})$ do not admit a joint density on $\mathbb{R}^n \times \mathbb{R}^{n(n+1)/2}$. (Conditional on $\boldsymbol{H}(\boldsymbol{m})$, the gradient $\boldsymbol{g}(\boldsymbol{m})$ is deterministic.) We will instead adapt the proof presented in [AT09] to handle this technicality. For simplicity, we prove only the upper bound, which is all that we require for our application.

Let $\mathsf{B}(\delta)$ be the open ball of radius δ around $\boldsymbol{0}$. Proposition 3.1 is an immediate consequence of the following lemma.

Lemma A.1. *Let $p_{\boldsymbol{m}}$ be the Lebesgue density of $\boldsymbol{g}(\boldsymbol{m})$. Let $\delta \equiv \delta_n > 0$. Then for any Borel measurable set $T \subseteq V_\delta \equiv [-1+\delta, 1-\delta]^n \setminus \mathsf{B}(\delta)$,*

$$\mathbb{E}[\mathrm{Crit}(T)] \le \int \mathbf{1}\{\boldsymbol{m} \in T\} \, \mathbb{E}\Big[\big|\det \boldsymbol{H}(\boldsymbol{m})\big| \, \big| \, \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}\Big] p_{\boldsymbol{m}}(\boldsymbol{0}) \mathrm{d}\boldsymbol{m}.$$

To prove this lemma, we apply the following result from [AT09]: Define the smoothed delta function

$$\delta_\varepsilon(\boldsymbol{m}) = \begin{cases} \mathrm{Vol}(\mathsf{B}(\varepsilon))^{-1} & \boldsymbol{m} \in \mathsf{B}(\varepsilon) \\ 0 & \boldsymbol{m} \notin \mathsf{B}(\varepsilon) \end{cases}$$

which integrates to 1 over $\mathbb{R}^n$.

Lemma A.2. *Suppose $\boldsymbol{g} : (-1,1)^n \to \mathbb{R}^n$ is deterministic and continuously differentiable, and $T \subset (-1,1)^n$ is compact. Suppose furthermore that there are no points $\boldsymbol{m} \in T$ satisfying both $\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}$ and $\det \nabla \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}$, and also no points $\boldsymbol{m} \in \partial T$ satisfying $\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}$. Then*

$$\mathrm{Crit}(T) = \lim_{\varepsilon \to 0} \int \mathbf{1}\{\boldsymbol{m} \in T\} \delta_\varepsilon(\boldsymbol{g}(\boldsymbol{m})) \, |\det \nabla \boldsymbol{g}(\boldsymbol{m})| \mathrm{d}\boldsymbol{m}.$$

Proof. See [AT09, Theorem 11.2.3]. $\square$

The below verifies that the conditions required for Lemma A.2 hold almost surely.

Lemma A.3. *Let $\boldsymbol{g} \equiv \boldsymbol{g}_n$ be defined by (27), and let $T \subset (-1,1)^n$ be compact with $\boldsymbol{0} \notin T$ and ∂T having Lebesgue measure 0. Then the conditions of Lemma A.2 hold with probability 1.*

Proof. Let us first verify that with probability 1, no point $\boldsymbol{m} \in \partial T$ satisfies $\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}$. Fix $C_0 > 0$ and consider the event $\mathcal{E} = \{\|\boldsymbol{W}\|_{\mathrm{op}} < C_0\}$. As T is compact and does not contain $\boldsymbol{0}$, it belongs to $K_{2\delta} \equiv [-1+2\delta, 1-2\delta]^n \setminus (-2\delta, 2\delta)^n$ for some (n,T)-dependent quantity $\delta > 0$. Since ∂T has Lebesgue outer measure 0, for any $\varepsilon > 0$ there exists a countable collection of balls $\{\mathsf{B}(\boldsymbol{m}_i, r_i) : i \in \mathcal{I}\}$ such that

$$\sum_{i \in \mathcal{I}} r_i^n < \varepsilon, \qquad \partial T \subset \bigcup_{i \in \mathcal{I}} \mathsf{B}(\boldsymbol{m}_i, r_i),$$

where each $\mathsf{B}(\boldsymbol{m}_i, r_i)$ is the open ball of radius r_i around $\boldsymbol{m}_i$. Taking $\varepsilon < \delta^n$ so that $r_i < \delta$ for each i, we may assume without loss of generality that each center $\boldsymbol{m}_i$ belongs to K_δ (because otherwise the ball has empty intersection with $K_{2\delta}$ and therefore with T). On the event $\mathcal{E}$, $\boldsymbol{g}(\boldsymbol{m})$ is L-Lipschitz over K_δ for some (n, δ, C_0)-dependent quantity $L > 0$. Hence

$$\mathbb{P}\left[\,\mathcal{E} \cap \{\text{there exists } \boldsymbol{m} \in \partial T : \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}\}\right] \le \sum_{i \in \mathcal{I}} \mathbb{P}\left[\|\boldsymbol{g}(\boldsymbol{m}_i)\|_2 < L r_i\right].$$

16

Observe that for each fixed $\boldsymbol{m}_i$, the vector $\boldsymbol{g}(\boldsymbol{m}_i)$ has a multivariate normal distribution with covariance

$$\beta^2 \left(n^{-1} \|\boldsymbol{m}_i\|_2^2 \mathbf{I} + n^{-1} \boldsymbol{m}_i \boldsymbol{m}_i^{\mathsf{T}} \right) \succeq \beta^2 \delta^2 \mathbf{I}.$$

Then the density of $\boldsymbol{g}(\boldsymbol{m}_i)$ is bounded as $\varphi_i(\boldsymbol{x}) \leq (2\pi\beta^2\delta^2)^{-n/2}$, and for some $C = C(\beta) > 0$ not depending on ε, δ, we have

$$\mathbb{P}\left[\|\boldsymbol{g}(\boldsymbol{m}_i)\|_2 < Lr_i \right] \leq \left(\frac{CLr_i}{\delta} \right)^n.$$

Hence

$$\mathbb{P}\left[\mathcal{E} \cap \{\text{there exists } \boldsymbol{m} \in \partial T : \boldsymbol{g}(\boldsymbol{m}) = 0\} \right] \leq \sum_{i \in \mathcal{I}} \left(\frac{CLr_i}{\delta} \right)^n \leq \left(\frac{CL}{\delta} \right)^n \varepsilon.$$

As $\varepsilon > 0$ is arbitrary, the above probability must be 0. Then

$$\mathbb{P}\left[\text{there exists } \boldsymbol{m} \in \partial T : \boldsymbol{g}(\boldsymbol{m}) = 0\right] \leq 1 - \mathbb{P}[\mathcal{E}].$$

Now taking $C_0 \to \infty$,
$$\mathbb{P}\left[\text{there exists } \boldsymbol{m} \in \partial T : \boldsymbol{g}(\boldsymbol{m}) = 0\right] = 0.$$

Next, let us verify that with probability 1, no point $\boldsymbol{m} \in T$ satisfies both $\boldsymbol{g}(\boldsymbol{m}) = 0$ and $\det \boldsymbol{H}(\boldsymbol{m}) = 0$. Define the set

$$S = \left\{ \boldsymbol{m} \in (-1,1)^n : \frac{m_i}{1 - m_i^2} - \frac{2\beta^2}{n} \|\boldsymbol{m}\|_2^2 \cdot m_i - \operatorname{arctanh}(m_i) = 0 \quad \forall i \in [n] \right\},$$

and suppose first that $T \cap S = \emptyset$. Note that for any $r > 0$, we may construct a maximal packing $\{\mathsf{B}(\boldsymbol{m}_i, r/2) : i \in \mathcal{I}\}$ where $\boldsymbol{m}_i \in T$ for each i. Namely these balls do not intersect, and no additional ball $\mathsf{B}(\boldsymbol{m}, r/2)$ with $\boldsymbol{m} \in T$ may be added to the packing. Then $\{\mathsf{B}(\boldsymbol{m}_i, r) : i \in \mathcal{I}\}$ is a finite cover of T, and a volume argument shows $|\mathcal{I}| \leq C/r^n$ for an n-dependent constant $C > 0$. As T is compact and S is closed, T is included in the set
$$U_\delta = \{\boldsymbol{m} \in [-1 + \delta, 1 - \delta]^n : \operatorname{dist}(\boldsymbol{m}, S) \geq \delta, \ \|\boldsymbol{m}\|_2 \geq \delta\}$$

for some (n, T)-dependent $\delta > 0$. Fixing $C_0 > 0$ and defining the event $\mathcal{E} = \{\|\boldsymbol{W}\|_{\mathrm{op}} < C_0\}$, we have on $\mathcal{E}$ that $\boldsymbol{g}(\cdot)$ and $\det(\boldsymbol{H}(\cdot))$ are both L-Lipschitz over $\boldsymbol{m} \in U_\delta$ for some $L = L(n, \delta, C_0) > 0$. Hence

$$\mathbb{P}\left[\mathcal{E} \cap \{\text{there exists } \boldsymbol{m} \in T : \boldsymbol{g}(\boldsymbol{m}) = 0, \ \det \boldsymbol{H}(\boldsymbol{m}) = 0\} \right] \leq \sum_{i \in \mathcal{I}} \mathbb{P}\left[\mathcal{E}, \ \|\boldsymbol{g}(\boldsymbol{m}_i)\|_2 < Lr, \ |\det \boldsymbol{H}(\boldsymbol{m}_i)| < Lr \right].$$

Consider a fixed index i, and let $\boldsymbol{v}$ and $\boldsymbol{A}$ be such that

$$\boldsymbol{g}(\boldsymbol{m}_i) = -\beta(\boldsymbol{W}\boldsymbol{m}_i + \boldsymbol{v}), \qquad \boldsymbol{H}(\boldsymbol{m}_i) = -\beta(\boldsymbol{W} + \boldsymbol{A})$$

for $\boldsymbol{g}, \boldsymbol{H}$ as defined in (27) and (28). Since $\boldsymbol{m}_i \in U_\delta$, the conditions of Lemma A.4 below are satisfied for some quantities $c_0 \equiv c_0(n, \delta, \lambda, \beta) > 0$ and $C_0 \equiv C_0(n, \delta, \lambda, \beta) > 0$. Then we obtain, for all r sufficiently small and some $C > 0$ independent of r,

$$\mathbb{P}\left[\mathcal{E}, \ \|\boldsymbol{g}(\boldsymbol{m}_i)\|_2 < Lr, \ |\det \boldsymbol{H}(\boldsymbol{m}_i)| < Lr \right] < Cr^{n + 1/(3n)}.$$

Applying this and $|\mathcal{I}| \leq C/r^n$ to the above, and taking $r \to 0$ followed by $C_0 \to \infty$, we obtain

$$\mathbb{P}\left[\text{there exists } \boldsymbol{m} \in T : \boldsymbol{g}(\boldsymbol{m}) = 0, \ \det \boldsymbol{H}(\boldsymbol{m}) = 0\right] = 0.$$

If $T \cap S \neq \emptyset$, this argument holds for the compact set

$$T \setminus \{\boldsymbol{m} : \operatorname{dist}(\boldsymbol{m}, S) < \delta\}$$

17

and any $\delta > 0$. Taking a union bound over a countable sequence $\delta \to 0$, and noting that $T \cap S$ is closed, we obtain

$$\mathbb{P}\left[\text{there exists } \boldsymbol{m} \in T \setminus S : \boldsymbol{g}(\boldsymbol{m}) = 0, \ \det \boldsymbol{H}(\boldsymbol{m}) = 0\right] = 0.$$

Finally, note that S is the zero set of a non-trivial real analytic function, and thus S has Lebesgue measure 0 [Mit15]. The same argument as for ∂T shows

$$\mathbb{P}\left[\text{there exists } \boldsymbol{m} \in T \cap S : \boldsymbol{g}(\boldsymbol{m}) = 0\right] = 0,$$

and combining the above,

$$\mathbb{P}\left[\text{there exists } \boldsymbol{m} \in T : \boldsymbol{g}(\boldsymbol{m}) = 0, \ \det \boldsymbol{H}(\boldsymbol{m}) = 0\right] = 0.$$

$\square$

Lemma A.4. *For any (n-dependent) quantities $c_0, C_0 > 0$, there exist $C \equiv C(n, c_0, C_0) > 0$ and $\varepsilon_0 \equiv \varepsilon_0(n, c_0, C_0) > 0$ such that the following holds: Let $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{m} \in (-1,1)^n$, and $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be any deterministic vectors/matrices such that $\boldsymbol{A}$ is symmetric, $\|\boldsymbol{A}\|_{\mathrm{op}} < C_0$, $\|\boldsymbol{A}\boldsymbol{m} - \boldsymbol{v}\|_2 > c_0$, and $\|\boldsymbol{m}\|_2 > c_0$. Let $\boldsymbol{W} \sim \mathrm{GOE}(n)$. Then for all $\varepsilon \in (0, \varepsilon_0)$,*

$$\mathbb{P}\left[\|\boldsymbol{W}\boldsymbol{m} + \boldsymbol{v}\|_2 < \varepsilon \ \text{and} \ |\det(\boldsymbol{W} + \boldsymbol{A})| < \varepsilon \ \text{and} \ \|\boldsymbol{W}\|_{\mathrm{op}} < C_0\right] < C\varepsilon^{n + 1/(3n)}.$$

Proof. Throughout the proof, C and $c > 0$ denote arbitrary (n, c_0, C_0)-dependent constants that may change from line to line.

Note that $\boldsymbol{W}\boldsymbol{m} + \boldsymbol{v}$ has a multivariate normal distribution with covariance

$$n^{-1}\|\boldsymbol{m}\|_2^2 \mathbf{I} + n^{-1}\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} \succeq n^{-1}c_0^2 \mathbf{I}.$$

Then the density of $\boldsymbol{W}\boldsymbol{m} + \boldsymbol{v}$ is bounded by an n-dependent constant, so $\mathbb{P}[\|\boldsymbol{W}\boldsymbol{m} + \boldsymbol{v}\|_2 < \varepsilon] < C\varepsilon^n$. Hence it suffices to show

$$\mathbb{P}\left[|\det(\boldsymbol{W} + \boldsymbol{A})| < \varepsilon, \ \|\boldsymbol{W}\|_{\mathrm{op}} < C_0 \,\Big|\, \boldsymbol{W}\boldsymbol{m} = \boldsymbol{w}\right] < C\varepsilon^{1/(3n)} \tag{44}$$

for any deterministic $\boldsymbol{w}$ satisfying $\|\boldsymbol{w} + \boldsymbol{v}\|_2 < \varepsilon < \varepsilon_0$.

For this, define a deterministic orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{n \times n}$ such that its first column is $\boldsymbol{m}/\|\boldsymbol{m}\|_2$, and the span of its first two columns contains $\boldsymbol{m}$ and $\boldsymbol{A}\boldsymbol{m} + \boldsymbol{w}$. Set

$$\check{\boldsymbol{W}} = \boldsymbol{O}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{O}, \qquad \check{\boldsymbol{w}} = \boldsymbol{O}^{\mathsf{T}}\boldsymbol{w}, \qquad \check{\boldsymbol{A}} = \boldsymbol{O}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{O}.$$

Then, rotating coordinates and denoting by $\boldsymbol{e}_i$ the ith standard basis vector,

$$\mathbb{P}\left[|\det(\boldsymbol{W} + \boldsymbol{A})| < \varepsilon, \ \|\boldsymbol{W}\|_{\mathrm{op}} < C_0 \,\Big|\, \boldsymbol{W}\boldsymbol{m} = \boldsymbol{w}\right] = \mathbb{P}\left[|\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})| < \varepsilon, \ \|\check{\boldsymbol{W}}\|_{\mathrm{op}} < C_0 \,\Big|\, \|\boldsymbol{m}\|_2 \check{\boldsymbol{W}}\boldsymbol{e}_1 = \check{\boldsymbol{w}}\right].$$

Conditional on $\|\boldsymbol{m}\|_2 \check{\boldsymbol{W}}\boldsymbol{e}_1 = \check{\boldsymbol{w}}$, the first column (and also first row) of $\check{\boldsymbol{W}} + \check{\boldsymbol{A}}$ is deterministic and given by

$$(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})\boldsymbol{e}_1 = \frac{\check{\boldsymbol{w}}}{\|\boldsymbol{m}\|_2} + \check{\boldsymbol{A}}\boldsymbol{e}_1 = \frac{1}{\|\boldsymbol{m}\|_2}\boldsymbol{O}^{\mathsf{T}}(\boldsymbol{w} + \boldsymbol{A}\boldsymbol{m}) = \alpha_1 \boldsymbol{e}_1 + \alpha_2 \boldsymbol{e}_2,$$

where the last equality holds by construction of $\boldsymbol{O}$ for some scalars α_1, α_2 which satisfy $\alpha_1^2 + \alpha_2^2 = \|\boldsymbol{A}\boldsymbol{m} + \boldsymbol{w}\|_2^2/\|\boldsymbol{m}\|_2^2$. Then, denoting by $\boldsymbol{H}^{(1)} = \check{\boldsymbol{W}}^{(1)} + \check{\boldsymbol{A}}^{(1)}$ and $\boldsymbol{H}^{(12)} = \check{\boldsymbol{W}}^{(12)} + \check{\boldsymbol{A}}^{(12)}$ the lower-right $(n-1) \times (n-1)$ and $(n-2) \times (n-2)$ submatrices of $\boldsymbol{W} + \boldsymbol{A}$, and expanding the determinant along the first column,

$$\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}}) = \alpha_1 \det \boldsymbol{H}^{(1)} - \alpha_2^2 \det \boldsymbol{H}^{(12)}.$$

For sufficiently small ε_0, the given conditions and $\|\boldsymbol{w} + \boldsymbol{v}\|_2 < \varepsilon_0$ imply $\alpha_1^2 + \alpha_2^2 > c$. We consider two cases:

18

- Case 1: $|\alpha_1| < \varepsilon^{1/3}$. Then $\alpha_2^2 > c$ for a constant $c > 0$. For $\|\boldsymbol{W}\|_{\mathrm{op}} \leq C_0$ and $\|\boldsymbol{A}\|_{\mathrm{op}} \leq C_0$, we have $\|\boldsymbol{H}^{(1)}\|_{\mathrm{op}} \leq 2C_0$ and hence $|\det(\boldsymbol{H}^{(1)})| \leq (2C_0)^{n-1}$. Combining with $|\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})| < \varepsilon$, we have $|\det \boldsymbol{H}^{(12)}| < C\varepsilon^{1/3}$ for some constant $C > 0$. Also, $\|\boldsymbol{W}\|_{\mathrm{op}} < C_0$ and $\|\boldsymbol{A}\|_{\mathrm{op}} \leq C_0$ imply $\|\boldsymbol{H}^{(12)}\| < 2C_0$. Hence

$$\mathbb{P}\Big[|\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})| < \varepsilon, \ \|\check{\boldsymbol{W}}\|_{\mathrm{op}} < C_0 \Big| \|\boldsymbol{m}\|_2 \check{\boldsymbol{W}} \boldsymbol{e}_1 = \check{\boldsymbol{w}}\Big] \leq \mathbb{P}\Big[|\det \boldsymbol{H}^{(12)}| < C\varepsilon^{1/3}, \ \|\boldsymbol{H}^{(12)}\|_{\mathrm{op}} < 2C_0\Big]$$

(45)

for a constant $C > 0$ and sufficiently small ε. Writing the spectral decomposition $\boldsymbol{H}^{(12)} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{T}}$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n-2})$, and applying the change of variables $\mathrm{d}\boldsymbol{H}^{(12)} = (1/Z')\prod_{i<j}|\lambda_i - \lambda_j|\,\mathrm{d}\boldsymbol{\Lambda}\,\mathrm{d}\boldsymbol{U}$ [AGZ10], the joint density of ordered eigenvalues $\lambda_1, \ldots, \lambda_{n-2}$ of $\boldsymbol{H}^{(12)}$ is given by

$$(1/Z)\mathbf{1}\{\lambda_1 \leq \ldots \leq \lambda_{n-2}\}\prod_{i<j}|\lambda_j - \lambda_i| \cdot \int \exp\left(-\frac{n}{4}\,\mathrm{Tr}\left[(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{T}} - \check{\boldsymbol{A}}^{(12)})^2\right]\right)\mathrm{d}\boldsymbol{U},$$

where the integral is over the orthogonal group of dimension $n-2$. This density is bounded by an n-dependent constant over the set

$$\left\{\lambda_1 \leq \ldots \leq \lambda_{n-2} : \prod_i |\lambda_i| < C\varepsilon^{1/3}, \ \max_i |\lambda_i| < 2C_0\right\},$$

and the Lebesgue volume of this set is bounded by $C\varepsilon^{1/(3n)}$ since at least one coordinate $|\lambda_i|$ is less than $C\varepsilon^{1/(3n)}$. Then the right side of (45) is at most $C\varepsilon^{1/(3n)}$ for a constant $C > 0$.

- Case 2: $|\alpha_1| \geq \varepsilon^{1/3}$. Considering separately the events $|\det \boldsymbol{H}^{(12)}| < \varepsilon^{1/3}$ and $|\det \boldsymbol{H}^{(12)}| \geq \varepsilon^{1/3}$, and handling the first event by the argument of Case 1 above, we obtain

$$\mathbb{P}\Big[|\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})| < \varepsilon, \ \|\check{\boldsymbol{W}}\|_{\mathrm{op}} < C_0 \Big| \|\boldsymbol{m}\|_2 \check{\boldsymbol{W}} \boldsymbol{e}_1 = \check{\boldsymbol{w}}\Big]$$
$$\leq C\varepsilon^{1/(3n)} + \mathbb{P}\Big[|\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})| < \varepsilon, \ |\det \boldsymbol{H}^{(12)}| \geq \varepsilon^{1/3} \Big| \|\boldsymbol{m}\|_2 \check{\boldsymbol{W}} \boldsymbol{e}_1 = \check{\boldsymbol{w}}\Big].$$

For the probability on the right side, let us further condition on all entries of $\check{\boldsymbol{W}}$ except $\check{W}_{22}$. We have

$$\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}}) = \alpha_1 \check{W}_{22} \det \boldsymbol{H}^{(12)} + \mathrm{const}$$

where const denotes a quantity that does not depend on $\check{W}_{22}$. Then, when $|\alpha_1| \geq \varepsilon^{1/3}$ and $|\det \boldsymbol{H}^{(12)}| \geq \varepsilon^{1/3}$, the quantity $\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})/\varepsilon$ is conditionally normally distributed with variance at least $2\varepsilon^{-2/3}/n$. This normal distribution has density upper bounded by $C\varepsilon^{1/3}$, so

$$\mathbb{P}\Big[|\det(\check{\boldsymbol{W}} + \check{\boldsymbol{A}})| < \varepsilon, \ |\det \boldsymbol{H}^{(12)}| \geq \varepsilon^{1/3} \Big| \|\boldsymbol{m}\|_2 \check{\boldsymbol{W}} \boldsymbol{e}_1 = \check{\boldsymbol{w}}\Big] \leq C\varepsilon^{1/3}.$$

Combining these cases yields (44) as desired. $\qquad\square$

Proof of Lemma A.1. First consider $T \subseteq V_\delta \equiv [-1 + \delta, 1 - \delta]^n \setminus \mathsf{B}(\delta)$ to be a closed hyperrectangle. By Lemma A.2, Lemma A.3, Fatou's Lemma, and Fubini's Theorem,

$$\mathbb{E}[\mathrm{Crit}(T)] \leq \liminf_{\varepsilon \to 0} \int \mathbf{1}\{\boldsymbol{m} \in T\}\mathbb{E}\Big[\delta_\varepsilon(\boldsymbol{g}(\boldsymbol{m}))\,|\det \boldsymbol{H}(\boldsymbol{m})|\Big]\mathrm{d}\boldsymbol{m}.$$

Denoting by $p_{\boldsymbol{m}}$ the density of $\boldsymbol{g}(\boldsymbol{m})$, we have

$$\mathbb{E}\Big[\delta_\varepsilon(\boldsymbol{g}(\boldsymbol{m}))\,|\det \boldsymbol{H}(\boldsymbol{m})|\Big] = \mathrm{Vol}(\mathsf{B}(\varepsilon))^{-1}\int_{\boldsymbol{u} \in \mathsf{B}(\varepsilon)} \mathbb{E}\Big[|\det \boldsymbol{H}(\boldsymbol{m})|\Big|\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{u}\Big]p_{\boldsymbol{m}}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}.$$

Define

$$D(\boldsymbol{u}, \boldsymbol{m}) \equiv \mathbb{E}\Big[|\det \boldsymbol{H}(\boldsymbol{m})|\Big|\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{u}\Big]p_{\boldsymbol{m}}(\boldsymbol{u}).$$

19

For any fixed $\boldsymbol{m} \in V_\delta$, the vector $\boldsymbol{g}(\boldsymbol{m})$ is a Gaussian random vector, and $\mathbb{E}[\boldsymbol{g}(\boldsymbol{m})]$ and $\mathbb{E}[\boldsymbol{g}(\boldsymbol{m})\boldsymbol{g}(\boldsymbol{m})^\mathsf{T}]$ are continuous functions in $\boldsymbol{m}$. Hence, the density $p_{\boldsymbol{m}}(\boldsymbol{u})$ is a continuous function of $(\boldsymbol{u}, \boldsymbol{m}) \in \overline{\mathsf{B}(\varepsilon)} \times T$. Moreover, by Lemma B.1, we have $[\boldsymbol{H}(\boldsymbol{m})|\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{u}] \overset{d}{=} -\beta\mathsf{P}_{\boldsymbol{m}}^\perp \boldsymbol{W}\mathsf{P}_{\boldsymbol{m}}^\perp + A(\boldsymbol{u}, \boldsymbol{m})$, where $\boldsymbol{W} \sim \mathrm{GOE}(n)$, $\mathsf{P}_{\boldsymbol{m}}^\perp$ is the projection orthogonal to $\boldsymbol{m}$, and $A(\boldsymbol{u}, \boldsymbol{m})$ is continuous in $(\boldsymbol{u}, \boldsymbol{m}) \in \overline{\mathsf{B}(\varepsilon)} \times T$. Therefore, $D(\boldsymbol{u}, \boldsymbol{m})$ is continuous in $(\boldsymbol{u}, \boldsymbol{m}) \in \overline{\mathsf{B}(\varepsilon)} \times T$, and hence it is bounded on $\overline{\mathsf{B}(\varepsilon)} \times T$. The bounded convergence theorem and the continuity of $D(\boldsymbol{u}, \boldsymbol{m})$ yield

$$\mathbb{E}[\mathrm{Crit}(T)] \leq \int \mathbf{1}\{\boldsymbol{m} \in T\} \liminf_{\varepsilon \to 0} \mathbb{E}\Big[\delta_\varepsilon(\boldsymbol{g}(\boldsymbol{m})) \,|\det \boldsymbol{H}(\boldsymbol{m})|\Big] \mathrm{d}\boldsymbol{m} = \int \mathbf{1}\{\boldsymbol{m} \in T\} D(\boldsymbol{0}, \boldsymbol{m}) \mathrm{d}\boldsymbol{m}.$$

Finally, we generalize the above inequality to a general Lebesgue measurable set $E \subseteq V_\delta$. This follows by standard measure theory machinery, and we sketch the proof in the following.

For any Lebesgue measurable set $E \subseteq V_\delta$, define $\nu_0(E) = \mathbb{E}[\mathrm{Crit}(E)]$ and $\nu(E) = \int \mathbf{1}\{\boldsymbol{m} \in E\} D(\boldsymbol{0}, \boldsymbol{m}) \mathrm{d}\boldsymbol{m}$. Define $\mathcal{T} \equiv \{T \subseteq V_\delta : T \text{ is closed hyperrectangle}\}$. Let $\nu_\star$ be the outer measure generated by the set function $\nu : \mathcal{T} \to \mathbb{R}$, i.e. for any $E \subset V_\delta$,

$$\nu_\star(E) \equiv \inf \left\{ \sum_{i=1}^\infty \nu(T_i) : \ E \subseteq \bigcup_i T_i, \ T_i \in \mathcal{T} \right\}. \tag{46}$$

Note that we have shown $\nu_0(T) \leq \nu(T)$ for any closed hyperrectangle $T \in \mathcal{T}$. Then for any Lebesgue measurable set $E \subseteq V_\delta$,

$$\nu_0(E) \overset{(i)}{\leq} \inf \left\{ \sum_{i=1}^\infty \nu_0(T_i) : \ E \subseteq \bigcup_i T_i, \ T_i \in \mathcal{T} \right\} \overset{(ii)}{\leq} \inf \left\{ \sum_{i=1}^\infty \nu(T_i) : \ E \subseteq \bigcup_i T_i, \ T_i \in \mathcal{T} \right\} \overset{(iii)}{=} \nu_\star(E).$$

The inequality (i) is given by the definition $\nu_0(E) = \mathbb{E}[\mathrm{Crit}(E)]$, the nonnegativity and additivity of $\mathrm{Crit}(\cdot)$, and the linearity of expectation. The inequality (ii) is given by $\nu_0(T_i) \leq \nu(T_i)$ for each $T_i \in \mathcal{T}$. The equality (iii) is given by the definition $\nu_\star$.

Since $D(\boldsymbol{0}, \boldsymbol{m})$ is bounded over $\boldsymbol{m} \in V_\delta$, we have that ν is absolutely continuous with respect to Lebesgue measure, and $\nu_\star(E) = \nu(E)$ for any Lebesgue measurable set $E \subseteq V_\delta$ by a standard argument. This concludes the proof. $\qquad \square$

B   Proof of Proposition 3.2

The following lemma is standard, and can be found for example in [BM11, Lemma 11].

Lemma B.1. *Let $\boldsymbol{W} \sim \mathrm{GOE}(n)$, and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Denote by $\mathsf{P}_{\boldsymbol{x}}^\perp = \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^\mathsf{T}/\|\boldsymbol{x}\|_2^2$ the projection orthogonal to $\boldsymbol{x}$. Then we have*

$$\boldsymbol{W}|_{\boldsymbol{W}\boldsymbol{x}=\boldsymbol{y}} \overset{d}{=} \mathsf{P}_{\boldsymbol{x}}^\perp \boldsymbol{W}\mathsf{P}_{\boldsymbol{x}}^\perp + (\boldsymbol{x}\boldsymbol{y}^\mathsf{T} + \boldsymbol{y}\boldsymbol{x}^\mathsf{T})/\|\boldsymbol{x}\|_2^2 - \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\langle \boldsymbol{x}, \boldsymbol{y}\rangle/\|\boldsymbol{x}\|_2^4.$$

In other words, for any continuous integrable function, we have

$$\mathbb{E}[f(\boldsymbol{W})|\boldsymbol{W}\boldsymbol{x} = \boldsymbol{y}] = \mathbb{E}[f(\mathsf{P}_{\boldsymbol{x}}^\perp \boldsymbol{W}\mathsf{P}_{\boldsymbol{x}}^\perp + (\boldsymbol{x}\boldsymbol{y}^\mathsf{T} + \boldsymbol{y}\boldsymbol{x}^\mathsf{T})/\|\boldsymbol{x}\|_2^2 - \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\langle \boldsymbol{x}, \boldsymbol{y}\rangle/\|\boldsymbol{x}\|_2^4)]. \tag{47}$$

Denoting $\boldsymbol{u} \equiv \boldsymbol{u}(\boldsymbol{m}) = \mathrm{arctanh}(\boldsymbol{m}) - \beta\lambda/n \cdot \langle \mathbf{1}, \boldsymbol{m}\rangle\mathbf{1} + \beta^2[1 - Q(\boldsymbol{m})]\boldsymbol{m}$, the condition $\boldsymbol{g}_n(\boldsymbol{m}) = \boldsymbol{0}$ is equivalent to $\boldsymbol{W}\boldsymbol{m} = \beta^{-1}\boldsymbol{u}$. Then by Eqs. (27–28), conditional on $\boldsymbol{g}_n(\boldsymbol{m}) = \boldsymbol{0}$, the Hessian $\boldsymbol{H}_n(\boldsymbol{m})$ is equal in law to

$$\boldsymbol{Z}(\boldsymbol{m}) = \boldsymbol{D}(\boldsymbol{m}) - \beta\boldsymbol{W} + \Delta(\boldsymbol{m})$$

where

$$\boldsymbol{D}(\boldsymbol{m}) = \mathrm{diag}(d_1, \ldots, d_n), \qquad d_i = \frac{1}{1 - m_i^2} + \beta^2[1 - Q(\boldsymbol{m})], \tag{48}$$

and

$$\Delta(\boldsymbol{m}) = \beta \boldsymbol{W}\boldsymbol{m}\boldsymbol{m}^\mathsf{T}/\|\boldsymbol{m}\|_2^2 + \beta \boldsymbol{m}\boldsymbol{m}^\mathsf{T}\boldsymbol{W}/\|\boldsymbol{m}\|_2^2 - \beta \boldsymbol{m}\boldsymbol{m}^\mathsf{T}\langle \boldsymbol{m}, \boldsymbol{W}\boldsymbol{m}\rangle/\|\boldsymbol{m}\|_2^4$$
$$- \frac{\beta\lambda}{n}\cdot \mathbf{1}\mathbf{1}^\mathsf{T} - 2\beta^2/n\cdot \boldsymbol{m}\boldsymbol{m}^\mathsf{T} - (\boldsymbol{m}\boldsymbol{u}^\mathsf{T} + \boldsymbol{u}\boldsymbol{m}^\mathsf{T})/\|\boldsymbol{m}\|_2^2 + \boldsymbol{m}\boldsymbol{m}^\mathsf{T}\langle \boldsymbol{m}, \boldsymbol{u}\rangle/\|\boldsymbol{m}\|_2^4. \tag{49}$$

Let σ_β be the semicircle law with support $[-2\beta, 2\beta]$, let $\mu_{\boldsymbol{D}(\boldsymbol{m})}$ be the empirical spectral distribution of $\boldsymbol{D}(\boldsymbol{m})$, and define their additive free convolution [Voi91]

$$\nu_{\boldsymbol{m}} = \mu_{\boldsymbol{D}(\boldsymbol{m})} \boxplus \sigma_\beta.$$

We expect $\nu_{\boldsymbol{m}}$ to approximate the bulk spectral distribution of $\boldsymbol{Z}(\boldsymbol{m})$ for large n [Pas72]. Denote $\mathbf{i} = \sqrt{-1}$ and let log denote the complex logarithm with branch cut on the negative real axis. For $\varepsilon > 0$, let

$$l_\varepsilon(x) = \log|x - \mathbf{i}\varepsilon| = \mathrm{Re}\log(x - \mathbf{i}\varepsilon).$$

Denote

$$\mathrm{Tr}\, l_\varepsilon(\boldsymbol{Z}) = \sum_{i=1}^n l_\varepsilon(\lambda_i(\boldsymbol{Z}))$$

where $\lambda_1(\boldsymbol{Z}), \ldots, \lambda_n(\boldsymbol{Z})$ are the eigenvalues of $\boldsymbol{Z}$.

Recall $L(\boldsymbol{m})$ as defined in (30). We prove Proposition 3.2 via the following three lemmas. We defer the proofs of the first two to Sections B.1 and B.2.

Lemma B.2. *Fix $\beta > 0$, $\lambda \geq 0$, $\eta, b \in (0,1)$, and $a \in (0, 1/9)$. Let $\boldsymbol{m}$ satisfy $\|\boldsymbol{m}\|_\infty \leq 1 - e^{-n^b}$ and $Q(\boldsymbol{m}) \in [\eta, 1]$, and set $\varepsilon = n^{-a}$. Then there exist constants $C \equiv C(\beta, \lambda, \eta, b, a) > 0$ and $n_0 \equiv n_0(\beta, \lambda, \eta, b, a) > 0$ such that for all $n \geq n_0$,*

$$\mathbb{E}[\mathrm{Tr}\, l_\varepsilon(\boldsymbol{Z}(\boldsymbol{m}))] \leq n \int l_\varepsilon(x)\nu_{\boldsymbol{m}}(\mathrm{d}x) + C(\varepsilon^{-5} + n^b + \log n).$$

Lemma B.3. *For any $\beta > 0$ and $\boldsymbol{m} \in (-1, 1)^n$, there exists a constant $C \equiv C(\beta) > 0$ depending only on β such that for any $\varepsilon > 0$,*

$$\int l_\varepsilon(x)\nu_{\boldsymbol{m}}(\mathrm{d}x) \leq L(\boldsymbol{m}) + C\varepsilon.$$

Lemma B.4. *For any $\varepsilon > 0$, any $t > 0$, any $\boldsymbol{m} \in (-1, 1)^n$, and all $n \geq 1$,*

$$\mathbb{P}[|\mathrm{Tr}\, l_\varepsilon(\boldsymbol{Z}(\boldsymbol{m})) - \mathbb{E}\,\mathrm{Tr}\, l_\varepsilon(\boldsymbol{Z}(\boldsymbol{m}))| \geq nt] \leq 2\exp(-n^2\varepsilon^2 t^2/(8\beta^2)).$$

Proof of Lemma B.4. By [AGZ10, Lemma 2.3.1], if $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, then the map $\boldsymbol{Z} \mapsto \mathrm{Tr}\, f(\boldsymbol{Z})$ is $L\sqrt{2n}$-Lipschitz with respect to the Frobenius norm of $\boldsymbol{Z}$. Then for each fixed $\boldsymbol{m}$, the quantity $\mathrm{Tr}\, f(\boldsymbol{Z}(\boldsymbol{m}))$ is a $2L\beta$-Lipschitz function of the $n(n-1)/2$ standard Gaussian variables which parametrize $\mathsf{P}_{\boldsymbol{m}}^\perp \boldsymbol{W} \mathsf{P}_{\boldsymbol{m}}^\perp$. By Gaussian concentration of measure, for any $t > 0$, we have

$$\mathbb{P}[|\mathrm{Tr}\, f(\boldsymbol{Z}(\boldsymbol{m})) - \mathbb{E}\,\mathrm{Tr}\, f(\boldsymbol{Z}(\boldsymbol{m}))| \geq nt] \leq 2\exp\left(-\frac{n^2 t^2}{8L^2\beta^2}\right).$$

The result follows from observing that l_ε is differentiable, with derivative at each $x \in \mathbb{R}$ satisfying

$$\left|\frac{\mathrm{d}}{\mathrm{d}x}l_\varepsilon(x)\right| = \left|\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{Re}\log(x - \mathbf{i}\varepsilon)\right| = \left|\mathrm{Re}\frac{1}{x - \mathbf{i}\varepsilon}\right| \leq \frac{1}{\varepsilon}.$$

$\square$

Proof of Proposition 3.2. Taking $\varepsilon = n^{-0.11}$, we obtain from Lemmas B.2, B.3, and B.4, for some constants $C, c, n_0 > 0$, all $n \geq n_0$, and all $t > 0$,

$$\mathbb{P}\left[\mathrm{Tr}\, l_\varepsilon(\boldsymbol{Z}(\boldsymbol{m})) \geq nL(\boldsymbol{m}) + Cn^{\max(0.89, b)} + nt\right] \leq 2\exp(-cn^{1.78}t^2).$$

21

Then, setting $c_n = nL(\boldsymbol{m}) + Cn^{\max(0.89,b)}$,

$$\mathbb{E}[|\det \boldsymbol{Z}(\boldsymbol{m})|] \leq \mathbb{E}[\exp(\operatorname{Tr} l_\varepsilon(\boldsymbol{Z}(\boldsymbol{m})))]$$

$$\leq \exp(c_n + n^{0.89}) + \int_{\exp(c_n + n^{0.89})}^\infty \mathbb{P}[\exp(\operatorname{Tr} l_\varepsilon(\boldsymbol{Z}(\boldsymbol{m}))) \geq t]\,\mathrm{d}t$$

$$= \exp(c_n + n^{0.89}) + e^{c_n} \int_{n^{0.89}}^\infty e^s\,\mathbb{P}[\operatorname{Tr} l_\varepsilon(\boldsymbol{Z}(\boldsymbol{m})) \geq c_n + s]\,\mathrm{d}s$$

$$\leq \exp(c_n + n^{0.89}) + e^{c_n} \int_{n^{0.89}}^\infty 2\exp(s - cs^2/n^{0.22})\mathrm{d}s$$

$$< \exp(nL(\boldsymbol{m}) + C_0 \cdot n^{\max(0.9,b)}),$$

the last line holding for sufficiently large n_0 and C_0. $\qquad\qquad\square$

In the remainder of this section, we prove Lemmas B.2 and B.3.

B.1 Proof of Lemma B.2

We first establish an analogue of Lemma B.2 for a matrix of the form $\boldsymbol{D} - \beta\boldsymbol{W}$.

Lemma B.5. *Fix $\beta > 0$ and $a \in (0, 1/9)$, and let $\varepsilon = n^{-a}$. Then there exist constants $C \equiv C(\beta, a) > 0$ and $n_0 \equiv n_0(\beta, a) > 0$ such that for all $n \geq n_0$, the following holds: Let $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ be any real symmetric matrix, let $\boldsymbol{H} = \boldsymbol{D} - \beta\boldsymbol{W}$ where $\boldsymbol{W} \sim \operatorname{GOE}(n)$, and let $\nu = \mu_{\boldsymbol{D}} \boxplus \sigma_\beta$. Then*

$$\mathbb{E}[\operatorname{Tr} l_\varepsilon(\boldsymbol{H})] \leq n \int l_\varepsilon(x)\nu(\mathrm{d}x) + C(\varepsilon^{-5} + \log(\|\boldsymbol{D}\|_{\mathrm{op}} + 1) + \log n).$$

By the results of [Pas72], the Stieltjes transform

$$g(z) = \int \frac{1}{x - z}\nu(\mathrm{d}x) \tag{50}$$

of the above measure ν satisfies, for each $z \in \mathbb{C}^+$, the fixed-point equation

$$g(z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{d_i - z - \beta^2 g(z)}, \tag{51}$$

with $(d_i)_{i \leq n}$ denoting the eigenvalues of $\boldsymbol{D}$. We first provide a quantitative estimate of the approximation of the spectral law of $\boldsymbol{H}$ by ν, in terms of their Stieltjes transforms at points $z \in \mathbb{C}^+$ where $\operatorname{Im} z \gtrsim n^{-1/9}$.

Lemma B.6. *Fix $\beta > 0$, let $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ be any real symmetric matrix, and let $\boldsymbol{H}$ and ν be as in Lemma B.5. For $z \in \mathbb{C}^+$, denote the Stieltjes transforms of $\boldsymbol{H}$ and ν as*

$$\hat{g}(z) = \frac{1}{n}\operatorname{Tr}(\boldsymbol{H} - z\mathbf{I})^{-1}, \qquad g(z) = \int \frac{1}{x - z}\nu(\mathrm{d}x).$$

If $n \geq 40\max(\beta/\operatorname{Im} z, 1)^9$, then

$$|\mathbb{E}\hat{g}(z) - g(z)| < \frac{12}{n\operatorname{Im} z}\max(\beta/\operatorname{Im} z, 1)^5.$$

Proof. We follow the approach of [Pas05, Theorem 3.1], applying integration by parts and the Poincaré inequality. Fix $z \in \mathbb{C}^+$ and denote $\eta = \operatorname{Im} z$. Denote the resolvents of $\boldsymbol{H}$ and $\boldsymbol{D}$ as

$$G_{\boldsymbol{H}}(z) = (\boldsymbol{H} - z\mathbf{I})^{-1}, \qquad G_{\boldsymbol{D}}(z) = (\boldsymbol{D} - z\mathbf{I})^{-1}.$$

Applying $\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{B}^{-1}$, we obtain

$$G_{\boldsymbol{H}}(z) - G_{\boldsymbol{D}}(z) = \beta G_{\boldsymbol{H}}(z)\boldsymbol{W}G_{\boldsymbol{D}}(z).$$

As $G_{\boldsymbol{D}}(z)$ is deterministic, this yields

$$\mathbb{E}\,G_{\boldsymbol{H}}(z) = G_{\boldsymbol{D}}(z) + \beta\,\mathbb{E}[G_{\boldsymbol{H}}(z)\boldsymbol{W}]G_{\boldsymbol{D}}(z). \tag{52}$$

We use integration by parts on the term $\mathbb{E}[G_{\boldsymbol{H}}(z)\boldsymbol{W}]$. For any differentiable bounded function $f : \mathbb{R} \to \mathbb{R}$, when $\xi \sim \mathcal{N}(0, \sigma^2)$,

$$\mathbb{E}[\xi f(\xi)] = \sigma^2 \mathbb{E}[f'(\xi)]. \tag{53}$$

Consider indices $i, j, k \in [n]$. If $j < k$, then setting $\boldsymbol{X}$ as the matrix with (j, k) and (k, j) entries equal to 1 and remaining entries 0, we have

$$\begin{aligned}
\frac{\partial G_{\boldsymbol{H}}(z)_{ij}}{\partial W_{jk}} &= \lim_{\delta \to 0} \delta^{-1}\Big(\big(\boldsymbol{H} - z\mathbf{I} - \beta\delta\boldsymbol{X}\big)^{-1} - \big(\boldsymbol{H} - z\mathbf{I}\big)^{-1}\Big)_{ij} \\
&= \beta(G_{\boldsymbol{H}}(z)\boldsymbol{X}G_{\boldsymbol{H}}(z))_{ij} = \beta(G_{\boldsymbol{H}}(z)_{ij}G_{\boldsymbol{H}}(z)_{kj} + G_{\boldsymbol{H}}(z)_{ik}G_{\boldsymbol{H}}(z)_{jj}).
\end{aligned} \tag{54}$$

For $z \in \mathbb{C}^+$ with $\operatorname{Im} z = \eta$, we have $|G_{\boldsymbol{H}}(z)_{ij}| \le 1/\eta$. Then (53) yields

$$\mathbb{E}[G_{\boldsymbol{H}}(z)_{ij}W_{jk}] = \frac{1}{n}\mathbb{E}\left[\frac{\partial G_{\boldsymbol{H}}(z)_{ij}}{\partial W_{jk}}\right] = \frac{\beta}{n}\mathbb{E}[G_{\boldsymbol{H}}(z)_{ij}G_{\boldsymbol{H}}(z)_{kj} + G_{\boldsymbol{H}}(z)_{ik}G_{\boldsymbol{H}}(z)_{jj}].$$

Similar arguments yield the same identity when $j = k$. Then applying $G_{\boldsymbol{H}}(z)_{kj} = G_{\boldsymbol{H}}(z)_{jk}$ and summing over j, we obtain

$$\mathbb{E}[G_{\boldsymbol{H}}(z)\boldsymbol{W}] = \frac{\beta}{n}\mathbb{E}[G_{\boldsymbol{H}}(z)^2] + \beta\,\mathbb{E}[\hat{g}(z)G_{\boldsymbol{H}}(z)].$$

Denoting $\delta(z) = \hat{g}(z) - \mathbb{E}\hat{g}(z)$ and substituting the above into (52),

$$(\mathbb{E}\,G_{\boldsymbol{H}}(z))[\mathbf{I} - \beta^2(\mathbb{E}\hat{g}(z))G_{\boldsymbol{D}}(z)] = \left(\mathbf{I} + \frac{\beta^2}{n}\mathbb{E}[G_{\boldsymbol{H}}(z)^2] + \beta^2\,\mathbb{E}[\delta(z)G_{\boldsymbol{H}}(z)]\right)G_{\boldsymbol{D}}(z).$$

Multiplying on the right by $G_{\boldsymbol{D}}(z)^{-1} = \boldsymbol{D} - z\mathbf{I}$,

$$(\mathbb{E}\,G_{\boldsymbol{H}}(z))[\boldsymbol{D} - z\mathbf{I} - \beta^2(\mathbb{E}\hat{g}(z))\mathbf{I}] = \mathbf{I} + \frac{\beta^2}{n}\mathbb{E}[G_{\boldsymbol{H}}(z)^2] + \beta^2\,\mathbb{E}[\delta(z)G_{\boldsymbol{H}}(z)].$$

Now multiplying on the right by $[\boldsymbol{D} - z\mathbf{I} - (\beta^2\mathbb{E}\hat{g}(z))\mathbf{I}]^{-1} = G_{\boldsymbol{D}}(z + \beta^2\mathbb{E}\hat{g}(z))$, taking $n^{-1}\operatorname{Tr}$, and rearranging,

$$\mathbb{E}\hat{g}(z) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i - z - \beta^2\mathbb{E}\hat{g}(z)} = r(z) \tag{55}$$

where

$$r(z) = \frac{\beta^2}{n^2}\mathbb{E}\operatorname{Tr}\left(G_{\boldsymbol{H}}(z)^2 G_{\boldsymbol{D}}(z + \beta^2\mathbb{E}\hat{g}(z))\right) + \frac{\beta^2}{n}\mathbb{E}\left[\delta(z)\operatorname{Tr}\left(G_{\boldsymbol{H}}(z)G_{\boldsymbol{D}}(z + \beta^2\mathbb{E}\hat{g}(z))\right)\right] =: \mathrm{I} + \mathrm{II}.$$

Let us bound this remainder $r(z)$. Noting $\|(\boldsymbol{X} - z\mathbf{I})^{-1}\|_{\mathrm{op}} \le 1/\operatorname{Im} z$ for any real-symmetric $\boldsymbol{X}$, observing $\operatorname{Im} z + \beta^2\mathbb{E}\hat{g}(z) \ge \operatorname{Im} z$, and applying $|\operatorname{Tr}\boldsymbol{X}| \le n\|\boldsymbol{X}\|_{\mathrm{op}}$, we obtain

$$|\mathrm{I}| \le \frac{\beta^2}{n\eta^3}.$$

For II, applying these bounds and Cauchy-Schwarz,

$$|\mathrm{II}| \le \frac{\beta^2}{n}\mathbb{E}\big[|\delta(z)|^2\big]^{1/2}\frac{n}{\eta^2} = \frac{\beta^2}{\eta^2}\operatorname{Var}(\hat{g}(z))^{1/2},$$

where $\operatorname{Var}(\hat{g}) = \mathbb{E}[|\hat{g} - \mathbb{E}\hat{g}|^2] = \mathbb{E}[|\hat{g}|^2] - |\mathbb{E}\hat{g}|^2$. We apply the Poincaré inequality to bound $\operatorname{Var}(\hat{g}(z))$. For i.i.d. $\mathcal{N}(0, 1)$ variables $\xi_1, \ldots, \xi_k$ and $f : \mathbb{R}^k \to \mathbb{C}$ with bounded partial derivatives,

$$\operatorname{Var}(f(\xi_1, \ldots, \xi_k)) \le \mathbb{E}[\|\nabla f(\xi_1, \ldots, \xi_k)\|_2^2]. \tag{56}$$

For $j < k$, a computation similar to (54) yields

$$\frac{1}{\sqrt{n}}\frac{\partial \hat{g}(z)}{\partial W_{jk}} = \frac{1}{n^{3/2}}\sum_{i=1}^{n}\frac{\partial G_{\boldsymbol{H}}(z)_{ii}}{\partial W_{jk}} = \frac{2\beta}{n^{3/2}}[G_{\boldsymbol{H}}(z)^2]_{jk}, \qquad \frac{1}{\sqrt{n/2}}\frac{\partial \hat{g}(z)}{\partial W_{jj}} = \frac{\sqrt{2}\beta}{n^{3/2}}[G_{\boldsymbol{H}}(z)^2]_{jj}.$$

Noting $(\sqrt{n/2}\,W_{jj}: j \in [n]) \cup (\sqrt{n}\,W_{jk}: 1 \leq j < k \leq n)$ are i.i.d. $\mathcal{N}(0,1)$ variables, (56) yields

$$\mathrm{Var}(\hat{g}(z)) \leq \mathbb{E}\left[\sum_{j=1}^{n}\frac{2\beta^2}{n^3}|[G_{\boldsymbol{H}}(z)^2]_{jj}|^2 + \sum_{j<k}\frac{4\beta^2}{n^3}|[G_{\boldsymbol{H}}(z)^2]_{jk}|^2\right] = \frac{2\beta^2}{n^3}\,\mathbb{E}\,\mathrm{Tr}\left(G_{\boldsymbol{H}}(z)^2\overline{(G_{\boldsymbol{H}}(z)^2)}^\mathsf{T}\right) \leq \frac{2\beta^2}{n^2\eta^4}.$$

Combining the above,

$$|r(z)| \leq \frac{\beta^2}{n\eta^4}(\eta + \sqrt{2}\beta). \tag{57}$$

Finally, we compare (55) with the fixed-point equation (51) satisfied by $g(z)$ to obtain a bound on $\mathbb{E}\hat{g}(z) - g(z)$: Denoting $(x)_+ = \max(x,0)$, we have

$$|r(z)| = \left|\mathbb{E}\hat{g}(z) - g(z) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{d_i - z - \beta^2\mathbb{E}\hat{g}(z)} - \frac{1}{d_i - z - \beta^2 g(z)}\right)\right|$$

$$= \left|\mathbb{E}\hat{g}(z) - g(z)\right|\left|1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\beta^2}{(d_i - z - \beta^2\mathbb{E}\hat{g}(z))(d_i - z - \beta^2 g(z))}\right|$$

$$\geq \left|\mathbb{E}\hat{g}(z) - g(z)\right|\left(1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\beta^2}{|d_i - z - \beta^2\mathbb{E}\hat{g}(z)|^2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\beta^2}{|d_i - z - \beta^2 g(z)|^2}}\right)_+, \tag{58}$$

the last step applying $|1 - x| \geq (1 - |x|)_+$ and Cauchy-Schwarz. Taking the imaginary part of (51) and rearranging,

$$\frac{\beta^2 \mathrm{Im}\, g(z)}{\eta + \beta^2 \mathrm{Im}\, g(z)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\beta^2}{|d_i - z - \beta^2 g(z)|^2}.$$

We have

$$\mathrm{Im}\, g(z) = \mathrm{Im}\int\frac{1}{x-z}\nu(\mathrm{d}x) = \int\frac{\eta}{|x-z|^2}\nu(\mathrm{d}x) \leq 1/\eta,$$

and hence

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\beta^2}{|d_i - z - \beta^2 g(z)|^2} \leq \frac{\beta^2}{\eta^2 + \beta^2}.$$

Similarly, taking imaginary parts of (55) and rearranging,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\beta^2}{|d_i - z - \beta^2\mathbb{E}\hat{g}(z)|^2} = \frac{\beta^2 \mathrm{Im}\,\mathbb{E}\hat{g}(z) - \beta^2 \mathrm{Im}\, r(z)}{\eta + \beta^2 \mathrm{Im}\,\mathbb{E}\hat{g}(z)} \leq \frac{\beta^2}{\eta^2 + \beta^2} + \frac{\beta^2|r(z)|}{\eta}.$$

Applying $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and the above bounds to (58), we obtain

$$|r(z)| \geq \left|\mathbb{E}\hat{g}(z) - g(z)\right|\left(1 - \frac{\beta^2}{\eta^2 + \beta^2} - \sqrt{\frac{\beta^2}{\eta^2 + \beta^2}\cdot\frac{\beta^2|r(z)|}{\eta}}\right)_+ \geq \left|\mathbb{E}\hat{g}(z) - g(z)\right|\left(\frac{\eta^2}{\eta^2 + \beta^2} - \sqrt{\frac{\beta^2|r(z)|}{\eta}}\right)_+.$$

When

$$n \geq 40\max\left(\frac{\beta}{\eta}, 1\right)^9 \geq 20\left(\frac{\beta^4}{\eta^4} + \frac{\beta^9}{\eta^9}\right) > \frac{4\beta^4(\eta + \sqrt{2}\beta)(\eta^2 + \beta^2)^2}{\eta^9},$$

we verify from (57) that

$$\sqrt{\frac{\beta^2|r(z)|}{\eta}} \leq \frac{\eta^2}{2(\eta^2 + \beta^2)}.$$

24

Then the above implies

$$|\mathbb{E}\hat{g}(z) - g(z)| \le \frac{2|r(z)|(\eta^2 + \beta^2)}{\eta^2} \le \frac{2\beta^2(\eta + \sqrt{2}\beta)(\eta^2 + \beta^2)}{n\eta^6} < \frac{6}{n\eta}\left(\frac{\beta^2}{\eta^2} + \frac{\beta^5}{\eta^5}\right) \le \frac{12}{n\eta}\max\left(\frac{\beta}{\eta}, 1\right)^5.$$

$\square$

Proof of Lemma B.5. Denote by $\hat{\nu}$ the empirical spectral measure of $\boldsymbol{H}$, and by $\hat{g}(z)$ its Stieltjes transform. Let $t_n = n(\|\boldsymbol{D}\|_{\mathrm{op}} + \beta + 1)$. Then we have

$$\mathrm{Tr}\, l_\varepsilon(\boldsymbol{H}) = n\int \mathrm{Re}\log(x - \mathbf{i}\varepsilon)\,\hat{\nu}(\mathrm{d}x)$$

$$= n\int \mathrm{Re}\left(\log(x - \mathbf{i}t_n) - \int_\varepsilon^{t_n} \frac{-\mathbf{i}}{x - \mathbf{i}t}\mathrm{d}t\right)\hat{\nu}(\mathrm{d}x)$$

$$= n\int \log|x - \mathbf{i}t_n|\,\hat{\nu}(\mathrm{d}x) + n\,\mathrm{Re}\left(\mathbf{i}\int_\varepsilon^{t_n}\hat{g}(\mathbf{i}t)\mathrm{d}t\right).$$

For $\varepsilon = n^{-a}$ where $a \in (0, 1/9)$, we may apply Lemma B.6 to get

$$\left|\int_\varepsilon^{t_n}\mathbb{E}[\hat{g}(\mathbf{i}t)]\mathrm{d}t - \int_\varepsilon^{t_n}g(\mathbf{i}t)\mathrm{d}t\right| \le \frac{12}{n}\left(\int_\varepsilon^\beta \frac{\beta^5}{t^6}\mathrm{d}t + \int_\beta^{t_n}\frac{1}{t}\mathrm{d}t\right) \le C\left(\frac{\varepsilon^{-5} + \log(\|\boldsymbol{D}\|_{\mathrm{op}} + 1) + \log n}{n}\right)$$

for some (β, a)-dependent constants $C, n_0 > 0$ and all $n \ge n_0$. Reversing the above steps, we may write

$$\mathrm{Re}\left(\mathbf{i}\int_\varepsilon^{t_n}g(\mathbf{i}t)\mathrm{d}t\right) = \int\left(-\log|x - \mathbf{i}t_n| + \log|x - \mathbf{i}\varepsilon|\right)\nu(\mathrm{d}x),$$

so combining the above yields

$$\mathbb{E}\,\mathrm{Tr}\, l_\varepsilon(\boldsymbol{H}) \le n\int l_\varepsilon(x)\nu(\mathrm{d}x) + r_1 + r_2, \tag{59}$$

where

$$r_1 = \mathbb{E}\left[n\int \log|x - \mathbf{i}t_n|\,\hat{\nu}(\mathrm{d}x)\right] - n\int \log|x - \mathbf{i}t_n|\,\nu(\mathrm{d}x), \qquad r_2 = C(\varepsilon^{-5} + \log(\|\boldsymbol{D}\|_{\mathrm{op}} + 1) + \log n).$$

To bound r_1, note that since $\nu = \mu_{\boldsymbol{D}} \boxplus \sigma_\beta$, we have $\sup\{|x| : x \in \mathrm{supp}(\nu)\} \le \|\boldsymbol{D}\|_{\mathrm{op}} + 2\beta$. Applying

$$\log|x - \mathbf{i}t_n| - \log t_n = \mathrm{Re}\log(x - \mathbf{i}t_n) - \mathrm{Re}\log(-\mathbf{i}t_n) = \mathrm{Re}\,\frac{x}{\tilde{x} - \mathbf{i}t_n}$$

for some $\tilde{x} \in \mathbb{R}$ with $|\tilde{x}| \le |x|$, we obtain

$$\left|\int \log|x - \mathbf{i}t_n|\,\nu(\mathrm{d}x) - \log t_n\right| \le \frac{\|\boldsymbol{D}\|_{\mathrm{op}} + 2\beta}{t_n} \le \frac{2}{n}. \tag{60}$$

A similar argument holds for $\hat{\nu}$. Indeed, for all $t > 0$, we have

$$\mathbb{P}[\|\boldsymbol{H}\|_{\mathrm{op}} > \|\boldsymbol{D}\|_{\mathrm{op}} + 2\beta + t] \le \mathbb{P}[\beta\|\boldsymbol{W}\|_{\mathrm{op}} > 2\beta + t] \le 2\exp\left(-\frac{nt^2}{4\beta^2}\right), \tag{61}$$

see e.g. [DS01, Theorem II.11]. Let $\mathcal{E}$ be the event where $\|\boldsymbol{H}\|_{\mathrm{op}} \le \|\boldsymbol{D}\|_{\mathrm{op}} + 2\beta + 1$. Then we have

$$S_1 := \left|\mathbb{E}\left[\int \log|x - \mathbf{i}t_n|\,\hat{\nu}(\mathrm{d}x)\Big|\mathcal{E}\right] - \log t_n\right| \le \frac{\|\boldsymbol{D}\|_{\mathrm{op}} + 2\beta + 1}{t_n} \le \frac{2}{n},$$

and

$$S_2 := \mathbb{E}\left[\int \log|x - \mathbf{i}t_n|\,\hat{\nu}(\mathrm{d}x)\Big|\mathcal{E}^c\right] \le \mathbb{E}[\log(\|\boldsymbol{H}\|_{\mathrm{op}})|\mathcal{E}^c] + \log(t_n).$$

25

Note that, by (61),

$$\mathbb{E}[\log(\|\boldsymbol{H}\|_{\mathrm{op}}) \mid \mathcal{E}^c]\mathbb{P}[\mathcal{E}^c] = \mathbb{E}[\log(\|\boldsymbol{H}\|_{\mathrm{op}})\mathbf{1}\{\mathcal{E}^c\}]$$

$$= \int_0^\infty \mathbb{P}\Big[\|\boldsymbol{H}\|_{\mathrm{op}} \geq \max(e^t, \|\boldsymbol{D}\|_{\mathrm{op}} + 2\beta + 1)\Big]\mathrm{d}t$$

$$\leq \log(\|\boldsymbol{D}\|_{\mathrm{op}} + 2\beta + 1) \cdot 2e^{-n/(4\beta^2)} + \int_{\log(\|\boldsymbol{D}\|_{\mathrm{op}}+2\beta+1)}^\infty \mathbb{P}[\|\boldsymbol{H}\|_{\mathrm{op}} > e^t]\,\mathrm{d}t$$

$$< [\log(\|\boldsymbol{D}\|_{\mathrm{op}} + 1) + \log n] \cdot e^{-cn}$$

for a constant $c \equiv c(\beta) > 0$ and $n \geq n_0(\beta) > 0$. Then

$$\left|\mathbb{E}\int \log|x - \mathbf{i}t_n|\,\hat{\nu}(\mathrm{d}x) - \log t_n\right| \leq S_1 \cdot \mathbb{P}[\mathcal{E}] + [S_2 + \log(t_n)] \cdot \mathbb{P}[\mathcal{E}^c] \leq \frac{2}{n} + [\log(\|\boldsymbol{D}\|_{\mathrm{op}} + 1) + \log n] \cdot e^{-cn}.$$

Combining this bound with Eq. (60), we obtain $|r_1| \leq \log(\|\boldsymbol{D}\|_{\mathrm{op}} + 1) + \log n$ for sufficiently large n_0 and all $n \geq n_0$. Then the lemma follows from (59). $\qquad\square$

We now apply this result to establish Lemma B.2.

Proof of Lemma B.2. Fix $\boldsymbol{m}$ and denote $\boldsymbol{H} = \boldsymbol{D} - \beta\boldsymbol{W}$, so that $\boldsymbol{Z} = \boldsymbol{H} + \Delta$. Lemma B.5 gives

$$\mathbb{E}[\mathrm{Tr}\,l_\varepsilon(\boldsymbol{H})] \leq n\int l_\varepsilon(x)\nu(\mathrm{d}x) + C(\varepsilon^{-5} + \log(\|\boldsymbol{D}\|_{\mathrm{op}} + 1) + \log n). \tag{62}$$

From (49), the operator norm of Δ can be bounded by

$$\|\Delta\|_{\mathrm{op}} = \|\beta\boldsymbol{W}\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}/\|\boldsymbol{m}\|_2^2\|_{\mathrm{op}} + \|\beta\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}\boldsymbol{W}/\|\boldsymbol{m}\|_2^2\|_{\mathrm{op}} + \|\beta\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}\langle\boldsymbol{m}, \boldsymbol{W}\boldsymbol{m}\rangle/\|\boldsymbol{m}\|_2^4\|_{\mathrm{op}}$$

$$+ \|\beta\lambda/n \cdot \mathbf{1}\mathbf{1}^{\mathsf{T}}\|_{\mathrm{op}} + \|2\beta^2/n \cdot \boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}\|_{\mathrm{op}} + \|(\boldsymbol{m}\boldsymbol{u}^{\mathsf{T}} + \boldsymbol{u}\boldsymbol{m}^{\mathsf{T}})/\|\boldsymbol{m}\|_2^2\|_F + \|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}}\langle\boldsymbol{m}, \boldsymbol{u}\rangle/\|\boldsymbol{m}\|_2^4\|_F$$

$$\leq 3\beta\|\boldsymbol{W}\|_{\mathrm{op}} + 2\beta^2 + \beta\lambda + 3\|\boldsymbol{u}\|_2/\|\boldsymbol{m}\|_2.$$

For $Q(\boldsymbol{m}) \in [\eta, 1]$ and $\boldsymbol{m} \in [-1 + e^{-n^b}, 1 - e^{-n^b}]$, we have

$$\|\boldsymbol{u}\|_2 \leq \|\operatorname{arctanh}(\boldsymbol{m})\|_2 + \|\beta\lambda/n \cdot \langle\mathbf{1}, \boldsymbol{m}\rangle\mathbf{1}\|_2 + \|\beta^2(1 - Q(\boldsymbol{m}))\boldsymbol{m}\|_2$$

$$\leq \sqrt{n}\|\operatorname{arctanh}(\boldsymbol{m})\|_\infty + (\beta^2 + \beta\lambda)\sqrt{n} \leq \sqrt{n}(n^b + (\beta^2 + \beta\lambda)),$$

and $\|\boldsymbol{m}\|_2 \geq \sqrt{n\eta}$. Accordingly, we have

$$\|\Delta\|_{\mathrm{op}} \leq 3\lambda\|\boldsymbol{W}\|_{\mathrm{op}} + (2\beta^2 + \beta\lambda) + 3(n^b + (\beta^2 + \beta\lambda))/\sqrt{\eta}. \tag{63}$$

Note that Δ is given in Eq. (49), which has rank at most 8. Suppose it has r_+ positive eigenvalues and r_- negative eigenvalues, where $r_+ + r_- \leq 8$ (it is possible that r_+ or r_- is zero). Denote the eigenvalues of $\boldsymbol{H}$ as $\lambda_1(\boldsymbol{H}) \geq \lambda_2(\boldsymbol{H}) \geq \cdots \geq \lambda_k(\boldsymbol{H}) \geq 0 > \lambda_{k+1}(\boldsymbol{H}) \geq \cdots \geq \lambda_n(\boldsymbol{H})$, and those of $\boldsymbol{Z}$ as $\lambda_1(\boldsymbol{Z}) \geq \lambda_2(\boldsymbol{Z}) \geq \cdots \geq \lambda_n(\boldsymbol{Z})$. We apply the following bounds (we use the convention that, if $k = 0$, the set $\{1, 2, \ldots, k\} = \emptyset$):

- For $i \in S \equiv \{1, \ldots, r_+\} \cup \{n + 1 - r_-, \ldots, n\} \cup \{k - r_- + 1, \ldots, k + r_+\}$, we have

$$|\lambda_i(\boldsymbol{Z})| \leq \|\boldsymbol{H}\|_{\mathrm{op}} + \|\Delta\|_{\mathrm{op}}.$$

- The rest of the eigenvalues of $\boldsymbol{Z}$ satisfy, by Weyl's eigenvalue interlacing,

$$\lambda_{i-r_+}(\boldsymbol{H}) \geq \lambda_i(\boldsymbol{Z}) \geq 0, \quad i \in \{r_+ + 1, r_+ + 2, \ldots, k - r_-\},$$
$$\lambda_{i+r_-}(\boldsymbol{H}) \leq \lambda_i(\boldsymbol{Z}) \leq 0, \quad i \in \{k + r_+ + 1, k + r_+ + 2, \ldots, n - r_-\}.$$

Then

$$\mathrm{Tr}\,l_\varepsilon(\boldsymbol{Z}) \leq \mathrm{Tr}\,l_\varepsilon(\boldsymbol{H}) - \sum_{i \in S} l_\varepsilon(\lambda_i(\boldsymbol{H})) + 16\log(\|\boldsymbol{H}\|_{\mathrm{op}} + \|\Delta\|_{\mathrm{op}})$$

$$\leq \mathrm{Tr}\,l_\varepsilon(\boldsymbol{H}) - 16\log\varepsilon + 16\log(\|\boldsymbol{D}\|_{\mathrm{op}} + \lambda\|\boldsymbol{W}\|_{\mathrm{op}} + \|\Delta\|_{\mathrm{op}}).$$

Note that for $\boldsymbol{m}$ with $\|\boldsymbol{m}\|_\infty \leq 1 - e^{-n^b}$, we have $\|\boldsymbol{D}\|_{\mathrm{op}} \leq e^{2n^b} + \beta^2$. Then, taking expectations of the above and applying Jensen's inequality, the result follows from Eq. (62) and (63), and $\mathbb{E}[\|\boldsymbol{W}\|_{\mathrm{op}}] \leq 3$. $\qquad\square$

B.2 Proof of Lemma B.3

Fix $\lambda > 0$ and $\boldsymbol{m} \in (-1,1)^n$, and write as shorthand $\nu \equiv \nu_{\boldsymbol{m}}$. Lemma B.3 follows from several properties of ν and its support:

Lemma B.7. *The Stieltjes transform $g(z)$ of ν extends continuously from $\mathbb{C}^+$ to $\mathbb{C}^+ \cup \mathbb{R}$. Denoting this extension also by $g(z)$, the measure ν admits a continuous density given by*

$$f(z) = \frac{1}{\pi} \operatorname{Im} g(z), \qquad z \in \mathbb{R}.$$

At each $z \in \mathbb{R}$, letting $z_0 \in \mathbb{R}$ denote the closest point where $f(z_0) = 0$, we have

$$f(z) \leq \left(\frac{3}{4\pi^3 \beta} |z - z_0| \right)^{1/3}. \tag{64}$$

Proof. See [Bia97, Corollaries 1, 2, 5]. □

Lemma B.8. *The support of ν is contained in $[0, d_{\max} + 2\beta]$, where $d_{\max} = \max_{i \leq n} d_i$.*

Proof. We note that (51) holds also for $z \in \mathbb{R}$ by continuity, where $g(z)$ for $z \in \mathbb{R}$ is the continuous extension defined in Lemma B.7. Fix any $z \leq 0$, and consider the function

$$F(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{d_i - z - \beta^2 g}$$

defining the right side of (51). Let $p_1 < \ldots < p_k$ be the distinct values among

$$\{(d_i - z)/\beta^2 : i = 1, \ldots, n\}.$$

Then $F(g)$ is a rational function with poles $p_1, \ldots, p_k$.

The equation $g = F(g)$ is rearranged as a polynomial equation of degree $k+1$, and hence it as most $k+1$ complex roots counting multiplicity. From the graph of $F(g)$, there is at least one real root between p_j and p_{j+1} for each $j = 1, \ldots, k-1$. Furthermore, when $z \leq 0$, we have $p_1 > 1 - Q(\boldsymbol{m})$ and

$$F(1 - Q(\boldsymbol{m})) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(1 - m_i^2)^{-1} - z} \leq \frac{1}{n} \sum_{i=1}^{n} (1 - m_i^2) = 1 - Q(\boldsymbol{m}). \tag{65}$$

Then from the graph of $F(g)$ on $g \in (0, p_1)$, the equation $g = F(g)$ has at least two real roots in this interval, counting multiplicity. So all $k+1$ roots of $g = F(g)$ are real, there is exactly one root in each interval (p_j, p_{j+1}), and exactly two roots in $(0, p_1)$. In particular, $g(z)$ is real. Then by Lemma B.7, the density of ν is 0 for all $z \leq 0$, so ν is supported on $[0, \infty)$. The upper bound $d_{\max} + 2\beta$ follows from $\sup\{|x| : x \in \operatorname{supp}(\nu)\} \leq \|\boldsymbol{D}\|_{\operatorname{op}} + 2\beta$, as $\nu = \mu_{\boldsymbol{D}} \boxplus \sigma_\beta$. □

Lemma B.9. *At $z = 0$, the Stieltjes transform $g(0)$ is the smallest real root of (51) and satisfies $0 < g(0) \leq 1 - Q(\boldsymbol{m})$. Furthermore, for $L(\boldsymbol{m})$ defined in (30), we have*

$$\int \log(x) \nu(\mathrm{d}x) \leq L(\boldsymbol{m}).$$

Proof. From the proof of Lemma B.8, the equation $g = F(g)$ has exactly one root $g \in (p_j, p_{j+1})$ for each $j = 1, \ldots, k-1$ and exactly two roots $g \in (0, p_1)$ counting multiplicity. From the graph of F, this implies $F'(g) > 1$ for each root $g \in (p_j, p_{j+1})$, and also the two roots $g_1 \leq g_2$ in $(0, p_1)$ satisfy $F'(g_1) \leq 1$ and $F'(g_2) \geq 1$, with $F'(g_2) = 1$ if and only if $g_1 = g_2$. At any $z < 0$, since $z \notin \operatorname{supp}(\nu)$, $g(z)$ is analytic in a neighborhood of z and is given by (50). Differentiating (50) in z, we verify $g'(z) > 0$. On the other hand, differentiating (51) in z, we obtain

$$g'(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1 + \beta^2 g'(z)}{(d_i - z - \beta^2 g(z))^2},$$

so the condition $g'(z) > 0$ implies

$$1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\beta^2}{(d_i - z - \beta^2 g(z))^2} > 0.$$

Thus $F'(g(z)) < 1$. This holds for all $z < 0$, so we must have $F'(g(0)) \leq 1$ at $z = 0$ by continuity. Then $g(0)$ is the smallest real root to $g = F(g)$ at $z = 0$. Since $1 - Q(\boldsymbol{m})$ is one such root by (65), we have $0 < g(0) \leq 1 - Q(\boldsymbol{m})$.

To prove the bound on $\int \log(x)\nu(\mathrm{d}x)$, we write for any $T > 0$

$$\int \log(x)\nu(\mathrm{d}x) = \int \left(\log(x - \mathbf{i}T) - \int_0^T \frac{-\mathbf{i}}{x - \mathbf{i}t}\mathrm{d}t\right)\nu(\mathrm{d}x)$$

$$= \int \log(x - \mathbf{i}T)\nu(\mathrm{d}x) + \mathbf{i}\int_0^T g(\mathbf{i}t)\mathrm{d}t.$$

For each $t \in (0, T)$, applying (51), we have

$$g(\mathbf{i}t) = g(\mathbf{i}t)(1 + \beta^2 g'(\mathbf{i}t)) - \beta^2 g(\mathbf{i}t)g'(\mathbf{i}t)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1 + \beta^2 g'(\mathbf{i}t)}{d_i - \mathbf{i}t - \beta^2 g(\mathbf{i}t)} - \beta^2 g(\mathbf{i}t)g'(\mathbf{i}t).$$

Then $-\mathbf{i}g(\mathbf{i}t) = B'(t)$ for

$$B(t) := \frac{1}{n}\sum_{i=1}^{n}\log(d_i - \mathbf{i}t - \beta^2 g(\mathbf{i}t)) + \frac{\beta^2}{2}g(\mathbf{i}t)^2,$$

so

$$\mathbf{i}\int_0^T g(\mathbf{i}t)\mathrm{d}t = B(0) - B(T).$$

We next take $T \to \infty$, using $|g(\mathbf{i}T)| \leq \sup_x(1/|x - \mathbf{i}T|) \leq 1/T$ from Eq. (50). Applying a Taylor expansion of $\log(x - \mathbf{i}T)$, we have

$$\int \log(x - \mathbf{i}T)\nu(\mathrm{d}x) = \log(-\mathbf{i}T) + O(1/T), \qquad B(T) = \log(-\mathbf{i}T) + O(1/T).$$

(In the first equation, we use the fact that ν has bounded support, cf. Lemma B.8, and in the second we use the definition of B, together with $|g(\mathbf{i}T)| \leq 1/T$.) Combining the above, we obtain $\int \log(x)\nu(\mathrm{d}x) = B(0)$.

Finally, note that for F as defined in the proof of Lemma B.8, the equation $g = F(g)$ is equivalent to $0 = R'(g)$ where

$$R(g) = \frac{\beta^2 g^2}{2} + \frac{1}{n}\sum_{i=1}^{n}\log(d_i - \beta^2 g).$$

Recall that either $g(0) = 1 - Q(\boldsymbol{m})$, or $g(0) < 1 - Q(\boldsymbol{m})$ and these are the two roots of this equation $0 = R'(g)$ in $(-\infty, p_1)$. Since $R(g) \to \infty$ when $g \to -\infty$, we obtain in both cases $B(0) = R(g(0)) \leq R(1 - Q(\boldsymbol{m})) = L(\boldsymbol{m})$. $\qquad\square$

Proof of Lemma B.3. For any $x > 0$,

$$l_\varepsilon(x) - \log x = \mathrm{Re}\log(x - \mathbf{i}\varepsilon) - \mathrm{Re}\log(x) = \log\sqrt{x^2 + \varepsilon^2} - \log x = \frac{\sqrt{x^2 + \varepsilon^2} - x}{x + \tilde{\varepsilon}}$$

for some $\tilde{\varepsilon} \in [0, \sqrt{x^2 + \varepsilon^2} - x]$. Then $|l_\varepsilon(x) - \log x| \leq \varepsilon/x$, so

$$\int l_\varepsilon(x)\nu(\mathrm{d}x) \leq \int \log(x)\nu(\mathrm{d}x) + \varepsilon\int(1/x)\nu(\mathrm{d}x).$$

The result follows from Lemma B.9, Lemma B.8, and the density decay condition (64) at the smallest edge of ν in the case where $0 \in \mathrm{supp}(\nu)$. $\qquad\square$

C   Proof of Proposition 3.5

We equip $\mathcal{P}$ with the topology of weak convergence and the corresponding Borel σ-algebra. The proof applies Sanov's Theorem and Varadhan's Lemma, with modifications to handle the technicality that $J(\rho, y)$, $A(\rho)$, and $E(\rho)$ are not (weakly) continuous over $\rho \in \mathcal{P}$.

Lemma C.1.

   (a) *For any constants $\eta > 0$ and $\beta > 0$, there exists a constant $C \equiv C(\beta, \eta) < \infty$ such that $J(\rho, y) \leq C + y^2/(4\beta^2)$ for all $\rho \in \mathcal{P}$ with $Q(\rho) \geq \eta$.*

   (b) *$J : \{\rho \in \mathcal{P} : Q(\rho) > 0\} \times \mathbb{R} \to [-\infty, \infty)$ is upper semi-continuous.*

Proof. For (a), we apply the inequality $(c+1)a^2 + (c^{-1}+1)b^2 \geq (a+b)^2$ with $c = 2$, $a = \operatorname{arctanh}(x) - \beta\lambda\varphi + \beta^2(1-q)x - yx$, and $b = yx$. Recalling $g(x; \varphi, q, y)$ from Eq. (31), this yields

$$\log g(x; \varphi, q, y) \leq \log \frac{1}{1-x^2} - \frac{(\operatorname{arctanh}(x) - \beta\lambda\varphi + \beta^2(1-q)x)^2}{6\beta^2 q} + \frac{(yx)^2}{4\beta^2 q}.$$

By the boundedness of $M(\rho)$ and $Q(\rho)$ and the comparison of $\log 1/(1-x^2)$ and $\operatorname{arctanh}(x)^2$ as $x \to \pm 1$, we see that for $\varphi = M(\rho)$ and $q = Q(\rho)$, the first two terms above are together upper-bounded by an η-dependent constant. Then

$$J(\rho, y) \leq C + \int \frac{(yx)^2}{4\beta^2 Q(\rho)} \rho(\mathrm{d}x) = C + \frac{y^2}{4\beta^2}$$

for a constant $C \equiv C(\beta, \eta) < \infty$.

For (b), fix $y, y_1, y_2, \ldots \in \mathbb{R}$ such that $y_i \to y$ and $\rho, \rho_1, \rho_2, \ldots \in \mathcal{P}$ such that $\rho_i \to \rho$ weakly, $Q(\rho) > 0$, and $Q(\rho_i) > 0$ for all i. Then $Q(\rho_i) \to Q(\rho)$, so there is a lower bound $\eta > 0$ on all $Q(\rho_i)$, as well as a finite upper bound on all y_i^2. Fix a constant $\alpha \in \mathbb{R}$ and define $f_{i,\alpha}(x) = \max(\alpha, \log g(x; M(\rho_i), Q(\rho_i), y_i))$ and $f_\alpha(x) = \max(\alpha, \log g(x; M(\rho), Q(\rho), y))$. Then $f_{i,\alpha}$ and f_α are uniformly bounded above and below over all i. Furthermore, there is a value $\delta > 0$ such that $f_{i,\alpha}(x) = f_\alpha(x) = \alpha$ for all $x < -1 + \delta$ and $x > 1 - \delta$ and all i. As $f_{i,\alpha}(x) \to f_\alpha(x)$ uniformly over $x \in [-1+\delta, 1-\delta]$, this implies $f_{i,\alpha}(x) \to f_\alpha(x)$ uniformly over $x \in (-1, 1)$. Then

$$\left| \int f_{i,\alpha}(x)\rho_i(\mathrm{d}x) - \int f_\alpha(x)\rho(\mathrm{d}x) \right| \leq \int |f_{i,\alpha}(x) - f_\alpha(x)|\rho_i(\mathrm{d}x) + \left| \int f_\alpha(x)(\rho_i - \rho)(\mathrm{d}x) \right| \to 0.$$

Hence

$$\limsup_i J(\rho_i, y_i) \leq \lim_i \left(\frac{\beta^2(1 - Q(\rho_i))^2}{2} - \frac{1}{2}\log(2\pi\beta^2 Q(\rho_i)) + \int f_{i,\alpha}(x)\rho_i(\mathrm{d}x) \right)$$

$$= \frac{\beta^2(1 - Q(\rho))^2}{2} - \frac{1}{2}\log(2\pi\beta^2 Q(\rho)) + \int f_\alpha(x)\rho(\mathrm{d}x).$$

The left side is independent of α. As $x \mapsto g(x; M(\rho), Q(\rho), y)$ is bounded above, the monotone convergence theorem yields for the right side $\lim_{\alpha \to -\infty} \int f_\alpha(x)\rho(\mathrm{d}x) = \int \log g(x; M(\rho), Q(\rho), y)\, \rho(\mathrm{d}x)$. Then $\limsup_i J(\rho_i, y_i) \leq J(\rho, y)$, so J is upper semi-continuous. $\square$

Lemma C.2. *Fix any $\alpha \in (0, \infty)$, and suppose $\rho, \rho_1, \rho_2, \ldots \in \mathcal{P}$ are such that $\rho_i \to \rho$ weakly and*

$$\int_{-1}^1 \operatorname{arctanh}(x)^2\, \rho(\mathrm{d}x) \leq \alpha, \qquad \int_{-1}^1 \operatorname{arctanh}(x)^2\, \rho_i(\mathrm{d}x) \leq \alpha$$

for all i. Then $A(\rho_i) \to A(\rho)$ and $E(\rho_i) \to E(\rho)$.

Proof. As x^2 and $h(x)$ are continuous and bounded over $(-1, 1)$, it suffices to show

$$\int x \operatorname{arctanh}(x)\, \rho_i(\mathrm{d}x) \to \int x \operatorname{arctanh}(x)\, \rho(\mathrm{d}x).$$

For any $\delta \in (0, 1)$ such that $-1 + \delta$ and $1 - \delta$ are continuity points of ρ, we have by weak convergence

$$\int_{[-1+\delta, 1-\delta]} x \operatorname{arctanh}(x)\, \rho_i(\mathrm{d}x) \to \int_{[-1+\delta, 1-\delta]} x \operatorname{arctanh}(x)\, \rho(\mathrm{d}x).$$

Denoting $I_\delta = (-1, -1 + \delta) \cup (1 - \delta, 1)$, by Cauchy-Schwarz

$$\left(\int_{I_\delta} x \operatorname{arctanh}(x)\, \rho(\mathrm{d}x) \right)^2 \leq \int_{I_\delta} x^2 \rho(\mathrm{d}x) \cdot \int_{I_\delta} \operatorname{arctanh}(x)^2\, \rho(\mathrm{d}x) \leq \alpha \int_{I_\delta} x^2 \rho(\mathrm{d}x),$$

and similarly for ρ_i. As $\int_{I_\delta} x^2 \rho_i(\mathrm{d}x) \to \int_{I_\delta} x^2 \rho(\mathrm{d}x)$ also by weak convergence, this yields

$$\limsup_i \left| \int x \operatorname{arctanh}(x) \rho_i(\mathrm{d}x) - \int x \operatorname{arctanh}(x) \rho(\mathrm{d}x) \right| \leq 2\alpha \int_{I_\delta} x^2 \rho(\mathrm{d}x).$$

Taking $\delta \to 0$ yields the claim. $\qquad\square$

Proof of Proposition 3.5. Let ρ_n denote the empirical measure of $m_1, \ldots, m_n \overset{iid}{\sim} \pi_0$, where π_0 is the uniform distribution on $[-1, 1]$, and let $Y_n \sim \mathcal{N}(0, \beta^2/n)$ be independent of ρ_n. Then

$$T(U, V_n) = 2^n \cdot \mathbb{E}\left[\exp(nJ(\rho_n, Y_n) + C_0 n^{\max(0.9, b)}) \mathbf{1}\{(Q(\rho_n), M(\rho_n), A(\rho_n), E(\rho_n)) \in U\} \right]. \tag{66}$$

By Sanov's Theorem, ρ_n satisfies a large deviation principle (LDP) with rate n and rate function $\rho \mapsto H(\rho|\pi_0)$. From the form of the normal density, Y_n satisfies an LDP also with rate n and rate function $y \mapsto y^2/(2\beta^2)$. Then (ρ_n, Y_n) satisfies the LDP with rate function $R(\rho, y) = H(\rho|\pi_0) + y^2/(2\beta^2)$, and in particular, for any closed $S_{\mathrm{prod}} \subseteq \mathcal{P} \times \mathbb{R}$ we have

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}[(\rho_n, Y_n) \in S_{\mathrm{prod}}] \leq - \inf_{(\rho, y) \in S_{\mathrm{prod}}} R(\rho, y). \tag{67}$$

We now follow an argument similar to Varadhan's Lemma, see e.g. [DZ10, Lemma 4.3.6]. Note that $R : \mathcal{P} \times \mathbb{R} \to [0, \infty]$ is lower semi-continuous. Fix $\alpha \in (0, \infty)$ and set

$$\Xi(\alpha) = \{(\rho, y) \in \mathcal{P} \times \mathbb{R} : (Q(\rho), M(\rho), A(\rho), E(\rho)) \in U, \ J(\rho, y) \geq -\alpha\},$$
$$\Psi(\alpha) = \{(\rho, y) \in \mathcal{P} \times \mathbb{R} : R(\rho, y) \leq \alpha\}.$$

Note that $\Psi(\alpha)$ is compact, by compactness of sub-level sets of $H(\cdot|\pi_0)$. We claim that $\Xi(\alpha)$ is closed: Suppose $(\rho, y), (\rho_1, y_1), (\rho_2, y_2), \ldots \in \mathcal{P} \times \mathbb{R}$ are such that $(\rho_i, y_i) \to (\rho, y)$ and $(\rho_i, y_i) \in \Xi(\alpha)$ for all i. Then $M(\rho_i) \to M(\rho)$ and $Q(\rho_i) \to Q(\rho)$ by weak convergence. In particular $Q(\rho) \geq \eta$, so J is well-defined at ρ and $J(\rho, y) \geq \limsup_i J(\rho_i, y_i) \geq -\alpha$ by the upper semi-continuity established in Lemma C.1(b). Applying $(a - b)^2 \geq (a^2/2) - b^2$ and $\log 1/(1 - x^2) \leq \operatorname{arctanh}(x)^2/(8\beta^2) + c$ for a constant $c \equiv c(\beta) < \infty$, we have

$$\log g(x; \varphi, q, y) \leq c + \operatorname{arctanh}(x)^2 \left(\frac{1}{8\beta^2} - \frac{1}{4\beta^2 q} \right) + \frac{(\beta\lambda\varphi - \beta^2(1-q)x + yx)^2}{2\beta^2 q}.$$

Since $y_i \to y < \infty$, the above bound and the conditions $J(\rho_i, y_i) \geq -\alpha$, $J(\rho, y) \geq -\alpha$, $Q(\rho_i) \in [\eta, 1]$, and $Q(\rho) \in [\eta, 1]$ imply $\int_{-1}^1 \operatorname{arctanh}(x)^2\, \rho_i(\mathrm{d}x) \leq \kappa$ and $\int_{-1}^1 \operatorname{arctanh}(x)^2\, \rho(\mathrm{d}x) \leq \kappa$ for all i and some $\kappa < \infty$. Then Lemma C.2 implies $A(\rho_i) \to A(\rho)$ and $E(\rho_i) \to E(\rho)$. As U is closed, this implies $(Q(\rho), M(\rho), A(\rho), E(\rho)) \in U$, so $\Xi(\alpha)$ is closed as desired.

Now fix $\delta > 0$. By semi-continuity of R and J and finiteness of these functionals on $\Psi(\alpha) \cap \Xi(\alpha)$, for each $(\rho, y) \in \Psi(\alpha) \cap \Xi(\alpha)$ there exists an open neighborhood $O_{\rho, y} \subset \mathcal{P} \times \mathbb{R}$ such that Q is strictly positive on $\overline{O_{\rho, y}}$ and

$$\inf_{(\rho', y') \in \overline{O_{\rho, y}}} R(\rho', y') \geq R(\rho, y) - \delta, \qquad \sup_{(\rho', y') \in \overline{O_{\rho, y}}} J(\rho', y') \leq J(\rho, y) + \delta. \tag{68}$$

By compactness of $\Psi(\alpha) \cap \Xi(\alpha)$, we may take a finite collection $(\rho_1, y_1), \ldots, (\rho_K, y_K)$ such that $O = \bigcup_{i=1}^{K} O_{\rho_i, y_i}$ covers $\Psi(\alpha) \cap \Xi(\alpha)$. Then

$$\limsup_n n^{-1} \log \mathbb{E}\left[e^{nJ(\rho_n, Y_n)} \mathbf{1}\{(Q(\rho_n), M(\rho_n), A(\rho_n), E(\rho_n)) \in U\} \right]$$

$$\leq \limsup_n n^{-1} \log \mathbb{E}\left[\sum_{i=1}^{K} e^{nJ(\rho_n, Y_n)} \mathbf{1}\{(\rho_n, Y_n) \in O_{\rho_i, y_i}\} \right.$$

$$\left. + e^{nJ(\rho_n, Y_n)} \mathbf{1}\{J(\rho_n, Y_n) < -\alpha\} + e^{nJ(\rho_n, Y_n)} \mathbf{1}\{(\rho_n, Y_n) \in \Xi(\alpha) \setminus O\} \right]$$

$$\leq \max(E_1, \ldots, E_K, R_1, R_2)$$

where

$$E_i = \limsup_n n^{-1} \log \mathbb{E}\left[e^{nJ(\rho_n, Y_n)} \mathbf{1}\{(\rho_n, Y_n) \in O_{\rho_i, y_i}\} \right],$$

$$R_1 = \limsup_n n^{-1} \log \mathbb{E}\left[e^{nJ(\rho_n, Y_n)} \mathbf{1}\{J(\rho_n, Y_n) < -\alpha\} \right],$$

$$R_2 = \limsup_n n^{-1} \log \mathbb{E}\left[e^{nJ(\rho_n, Y_n)} \mathbf{1}\{(\rho_n, Y_n) \in \Xi(\alpha) \setminus O\} \right].$$

Clearly $R_1 \leq -\alpha$. Applying (67) and (68),

$$E_i \leq J(\rho_i, y_i) + \delta + \limsup_n n^{-1} \log \mathbb{P}\left[(\rho_n, Y_n) \in \overline{O_{\rho_i, y_i}} \right] \leq J(\rho_i, y_i) - R(\rho_i, y_i) + 2\delta.$$

By Holder's inequality, for any $\kappa > 1$,

$$R_2 \leq \limsup_n n^{-1} \log \left(\mathbb{E}\left[e^{\kappa n J(\rho_n, Y_n)} \mathbf{1}\{Q(\rho_n) \geq \eta\} \right]^{1/\kappa} \mathbb{P}\left[(\rho_n, Y_n) \in \Xi(\alpha) \setminus O \right]^{1-1/\kappa} \right).$$

Note that O is open, so $\Xi(\alpha) \setminus O$ is closed, and $R(\rho, y) > \alpha$ outside O. Then again applying (67) and Lemma C.1(a), for $\kappa \in (1, 2)$ and some $C \equiv C(\beta, \eta) < \infty$,

$$R_2 \leq \frac{1}{\kappa} \limsup_n n^{-1} \log \mathbb{E}\left[e^{\kappa n (C + Y_n^2/(4\beta^2))} \right] - (1 - 1/\kappa)\alpha = C - (1 - 1/\kappa)\alpha.$$

Choosing $\kappa = 3/2$ and combining with (66), we obtain

$$\limsup_n n^{-1} \log T(U, V_n)$$

$$\leq \max \left(\sup_{(\rho, y) \in \mathcal{P} \times \mathbb{R} : (Q(\rho), M(\rho), A(\rho), E(\rho)) \in U} J(\rho, y) - R(\rho, y) + 2\delta, \; -\alpha, \; C - \alpha/3 \right) + \log 2.$$

Taking $\delta \to 0$ and $\alpha \to \infty$ concludes the proof. $\qquad\square$

D   Proof of Proposition 4.2

Denote $\xi \sim \mathcal{N}(\lambda^2 \varphi_\star, \lambda^2 q_\star)$. Note that $\varphi_\star = q_\star$ implies $\mathbb{E}\{f(-\xi)\} = \mathbb{E}\{e^{-2\xi} f(\xi)\}$ whence $\mathbb{E}\{\tanh(\xi)^2\} = \mathbb{E}\{\tanh(\xi)\}$. Applying Gaussian integration by parts, we may verify from (38) and (33) that

$$\varphi_\star = \mathbb{E}[\tanh(\xi)], \qquad q_\star = \mathbb{E}[\tanh^2(\xi)], \qquad a_\star = \mathbb{E}[\xi \tanh(\xi)], \qquad u(q_\star, a_\star) - e_\star = \mathbb{E}[\log 2 \cosh(\xi)]. \quad (69)$$

Note that

$$S(q_\star, \varphi_\star, a_\star, e_\star; \mu, \nu, \tau, \gamma) = -q_\star \mu - \varphi_\star \nu - a_\star \tau - [u(q_\star, a_\star) - e_\star]\gamma + \log I,$$

where

$$I = \mathbb{E}\left[\exp\left(\mu \tanh^2(\xi) + \nu \tanh(\xi) + \tau \xi \tanh(\xi) + \gamma \log 2 \cosh(\xi) \right) \right].$$

31

Then $S(q_\star, \varphi_\star, a_\star, e_\star; \mu, \nu, \tau, \gamma)$ is a convex function of (μ, ν, τ, γ), with derivative with respect to (μ, ν, τ, γ) equal to 0 at $(\mu, \nu, \tau, \gamma) = 0$. Thus $S_\star(q_\star, \varphi_\star, a_\star, e_\star) = 0$ follows.

In the remainder of this section, we establish Eq. (35). Note that $q_\star \to 1$ as $\lambda \to \infty$ (see, e.g. [DAM16, Lemma 3.2]). Throughout the proof, we assume λ_0 is large enough such that the conditions of the proposition guarantee $(q, \varphi) \in [1/(1+\varepsilon), 1]^2$ for $\varepsilon \equiv 0.01$. Denote

$$S_+(q, \varphi, a) = \frac{1}{4\lambda^2 q^2}[a - \lambda^2\varphi^2 - \lambda^2(1-q)q]^2,$$

and

$$S_-(q, \varphi, a, e; \mu, \nu, \tau, \gamma) = -q\mu - \varphi\nu - a\tau - [u(q,a) - e]\gamma + \log I,$$

so that $S = S_+ + S_-$ and

$$S_\star(q, \varphi, a, e) = S_+(q, \varphi, a) + \inf_{\mu, \nu, \tau, \gamma} S_-(q, \varphi, a, e; \mu, \nu, \tau, \gamma).$$

Denote $Q(x) = \tanh^2 x$, $M(x) = \tanh x$, $A(x) = x\tanh x$, and $U(x) = \log[2\cosh(x)]$.

Step 1. Upper bound S_+. Noting $a_\star = \lambda^2\varphi_\star^2 + \lambda^2 q_\star(1 - q_\star)$, we have

$$S_+(q, \varphi, a) = \frac{1}{4\lambda^2 q^2}\{(a - a_\star) - (\lambda^2\varphi^2 - \lambda^2\varphi_\star^2) - [\lambda^2 q(1-q) - \lambda^2 q_\star(1-q_\star)]\}^2.$$

By the inequality
$$(x_1 + x_2 + x_3)^2 \le (1 + \varepsilon)^2 x_1^2 + (2 + 2/\varepsilon)^2 x_2^2 + (2 + 2/\varepsilon)^2 x_3^2,$$

and the conditions $q, \varphi \in [1/(1+\varepsilon), 1]$ and $\varepsilon = 0.01$, we have

$$S_+(q, \varphi, a) \le (1+\varepsilon)^2(a - a_\star)^2/(4\lambda^2 q^2) + (2 + 2/\varepsilon)^2\lambda^4\{(\varphi^2 - \varphi_\star^2)^2 + [q(1-q) - q_\star(1-q_\star)]^2\}/(4\lambda^2 q^2)$$
$$\le (1+\varepsilon)^4(a - a_\star)^2/(4\lambda^2) + (10/\varepsilon^2)\lambda^2(\varphi - \varphi_\star)^2 + (10/\varepsilon^2)\lambda^2(q - q_\star)^2. \tag{70}$$

Step 2. Bound the moment generating function of Q, M, A, and U.

Let $x \sim \mathcal{N}(\lambda^2\varphi, \lambda^2 q)$. Let $\mathbb{E}$ and Var be taken with respect to the randomness of x. We bound $\mathbb{E}[e^{\mu(Q - \mathbb{E}Q)}]$, $\mathbb{E}[e^{\nu(M - \mathbb{E}M)}]$, $\mathbb{E}[e^{\tau(A - \mathbb{E}A)}]$, and $\mathbb{E}[e^{\gamma(U - \mathbb{E}U)}]$.

Since $|Q| \le 1$, we have for $|\mu| < 1$

$$\mathbb{E}[\exp\{\mu(Q - \mathbb{E}Q)\}] \le 1 + \frac{1}{2}\mu^2\text{Var}(Q) + \sum_{k=3}^{\infty}\frac{1}{k!}|\mu|^k\mathbb{E}[|Q - \mathbb{E}Q|^k]$$
$$\le 1 + \frac{1}{2}\mu^2\text{Var}(Q)\sum_{k=2}^{\infty}|\mu|^{k-2} \le \exp\{\mu^2\text{Var}(Q)/[2(1 - |\mu|)]\}. \tag{71}$$

Similarly, we have for $|\nu| < 1$

$$\mathbb{E}[\exp\{\nu(M - \mathbb{E}M)\}] \le \exp\{\nu^2\text{Var}(M)/[2(1 - |\nu|)]\}. \tag{72}$$

To bound $\mathbb{E}[e^{\tau(A - \mathbb{E}A)}]$, note that $A(x) = x\tanh(x)$ is L_a-Lipschitz. Indeed, simple calculus shows that $\sup_t A'(t) = A'(A^{-1}(1)) = A^{-1}(1)$, whence $L_a = A^{-1}(1) \le 1.2$. Applying Gaussian concentration of measure, see e.g. [BLM13, Theorem 5.5], we have

$$\mathbb{E}[\exp\{\tau(A - \mathbb{E}A)\}] \le \exp\{\lambda^2 q\tau^2 L_a^2/2\} \le \exp\{\lambda^2\tau^2 L_a^2/2\}. \tag{73}$$

Similarly, to bound $\mathbb{E}[e^{\gamma(U - \mathbb{E}U)}]$, note $U(x) = \log[2\cosh(x)]$ is 1-Lipschitz. Hence we have

$$\mathbb{E}[\exp\{\gamma(U - \mathbb{E}U)\}] \le \exp\{\lambda^2 q\gamma^2/2\} \le \exp\{\lambda^2\gamma^2/2\}. \tag{74}$$

Step 3. Upper bound S_-.

Set $\alpha = 1 + \varepsilon$ and $\kappa = (3 + 3\varepsilon)/\varepsilon$, where $\varepsilon = 0.01$ as above. By Holder's inequality

$$\mathbb{E}[e^{\mu(Q - \mathbb{E}Q) + \nu(M - \mathbb{E}M) + \tau(A - \mathbb{E}A) + \gamma(U - \mathbb{E}U)}]$$
$$\leq \mathbb{E}[e^{\alpha\tau(A - \mathbb{E}A)}]^{1/\alpha} \mathbb{E}[e^{\kappa\mu(Q - \mathbb{E}Q)}]^{1/\kappa} \mathbb{E}[e^{\kappa\nu(M - \mathbb{E}M)}]^{1/\kappa} \mathbb{E}[e^{\kappa\gamma(U - \mathbb{E}U)}]^{1/\kappa}.$$

Given estimates (71), (72), (73), and (74), we have for $|\mu|, |\nu| < 1/\kappa$

$$\mathbb{E}[e^{\mu(Q - \mathbb{E}Q) + \nu(M - \mathbb{E}M) + \tau(A - \mathbb{E}A) + \gamma(U - \mathbb{E}U)}] \leq \exp\left\{\frac{\alpha\tau^2\lambda^2 L_a^2}{2} + \frac{\kappa\mu^2\mathrm{Var}(Q)}{2(1 - \kappa|\mu|)} + \frac{\kappa\nu^2\mathrm{Var}(M)}{2(1 - \kappa|\nu|)} + \frac{\kappa\gamma^2\lambda^2}{2}\right\},$$

and hence

$$
\begin{aligned}
S_-(q, \varphi, a, e; \mu, \nu, \tau, \gamma) \leq & -\tau(a - \mathbb{E}[A]) + \frac{\alpha\tau^2\lambda^2 L_a^2}{2} - \mu(q - \mathbb{E}[Q]) + \frac{\kappa\mu^2\mathrm{Var}(Q)}{2(1 - \kappa|\mu|)} \\
& -\nu(\varphi - \mathbb{E}[M]) + \frac{\kappa\nu^2\mathrm{Var}(M)}{2(1 - \kappa|\nu|)} - \gamma(u(q, a) - e - \mathbb{E}[U]) + \frac{\kappa\gamma^2\lambda^2}{2}.
\end{aligned}
\tag{75}
$$

To bound $\mathrm{Var}(Q)$ and $\mathrm{Var}(M)$, note that $\tanh(x) \geq 1 - e^{-x}$. Then for $(q, \varphi) \in [1/(1 + \varepsilon), 1]^2$ with $\varepsilon = 0.01$, there exists a universal constant λ_0 such that when $\lambda \geq \lambda_0$, we have (denoting $\phi(x) = \mathbb{P}(|G| \geq x)$ for $G \sim \mathcal{N}(0, 1)$)

$$
\begin{aligned}
\mathbb{E}[1 - \tanh x] = & \mathbb{E}[(1 - \tanh x)\mathbf{1}\{x \geq \lambda^2/2\}] + \mathbb{E}[(1 - \tanh x)\mathbf{1}\{x < \lambda^2/2\}] \\
\leq & e^{-\lambda^2/2} + 2\mathbb{P}(x < \lambda^2/2) \leq e^{-\lambda^2/2} + \phi(\lambda(\varphi - 1/2)) \leq e^{-\lambda^2/5},
\end{aligned}
\tag{76}
$$

and hence
$$\mathrm{Var}(M) = \mathbb{E}[\tanh^2 x] - \mathbb{E}[\tanh x]^2 \leq (1 + \mathbb{E}[\tanh x])(1 - \mathbb{E}[\tanh x]) \leq e^{-\lambda^2/10}.$$
Similarly, $\mathrm{Var}(Q) \leq e^{-\lambda^2/10}$.

Now we take $L_q^2 \geq e^{-\lambda^2/10}$ to be determined, and we take τ, μ, ν, γ to be

$$\tau = \frac{a - \mathbb{E}[A]}{\alpha\lambda^2 L_a^2}, \qquad \mu = \frac{q - \mathbb{E}[Q]}{2\kappa L_q^2}, \qquad \nu = \frac{\varphi - \mathbb{E}[M]}{2\kappa L_q^2}, \qquad \gamma = \frac{u(q, a) - e - \mathbb{E}[U]}{\kappa\lambda^2}$$

Then, as long as

$$\max\{|q - \mathbb{E}[Q]|, |\varphi - \mathbb{E}[M]|\} \leq L_q^2, \tag{77}$$

we have $|\mu|, |\nu| \leq 1/(2\kappa)$, and according to Eq. (75), we have

$$\inf_{\mu, \nu, \tau, \gamma} S_-(q, \varphi, a, e; \mu, \nu, \tau, \gamma) \leq -\frac{(a - \mathbb{E}[A])^2}{2\alpha\lambda^2 L_a^2} - \frac{(q - \mathbb{E}[Q])^2}{4\kappa L_q^2} - \frac{(\varphi - \mathbb{E}[M])^2}{4\kappa L_q^2} - \frac{(u(q, a) - e - \mathbb{E}[U])^2}{2\kappa\lambda^2}. \tag{78}$$

Step 4. Bound $\mathbb{E}A - a_\star$, $\mathbb{E}Q - q_\star$, $\mathbb{E}M - \varphi_\star$, and $\mathbb{E}U - (u(q_\star, a_\star) - e_\star)$.

Recall that the expectations in (78) are with respect to $x \sim \mathcal{N}(\lambda^2\varphi, \lambda^2 q)$ while the quantities of (69) are defined with $\xi \sim \mathcal{N}(\lambda^2\varphi_\star, \lambda^2 q_\star)$. To bound this difference, define $D_{F, \#} = \sup_{q, \varphi \in [0.9, 1]^2} |(\mathrm{d}/\mathrm{d}\#)\mathbb{E}_G[F(\lambda^2\varphi + \sqrt{\lambda^2 q}G)])|$ for $F = A, Q, M, U$ and $\# = q, \varphi$, where $G \sim \mathcal{N}(0, 1)$. Now we bound $D_{F, \#}$. We denote $x = \lambda^2\varphi + \sqrt{\lambda^2 q}G$, where (to simplify notation) in each line below $\varphi, q \geq 0.9$ are chosen to maximize the

corresponding expression. Applying (76) and Gaussian integration by parts, we obtain:

$$D_{A,\varphi} \le |(\mathrm{d}/\mathrm{d}\varphi)\mathbb{E}[x\tanh x]| = \lambda^2|\mathbb{E}[\tanh x + x(1-\tanh^2 x)]| \le 2\lambda^2,$$

$$D_{A,q} \le |(\mathrm{d}/\mathrm{d}q)\mathbb{E}[x\tanh x]| \le |\mathbb{E}\{[\tanh x + x(1-\tanh^2 x)]\lambda/(2\sqrt{q})G\}|$$
$$\le \lambda^2|\mathbb{E}[(1-x\tanh x)(1-\tanh^2 x)]| \le \lambda^2,$$

$$D_{M,\varphi} \le |(\mathrm{d}/\mathrm{d}\varphi)\mathbb{E}[\tanh x]| \le \lambda^2\mathbb{E}[1-\tanh^2 x] \le 2\lambda^2 e^{-\lambda^2/5},$$

$$D_{M,q} \le |(\mathrm{d}/\mathrm{d}q)\mathbb{E}[\tanh x]| \le \lambda/(2\sqrt{q})|\mathbb{E}[(1-\tanh^2 x)G]| \le \lambda^2|\mathbb{E}[\tanh x(1-\tanh^2 x)]| \le 2\lambda^2 e^{-\lambda^2/5},$$

$$D_{Q,\varphi} \le |(\mathrm{d}/\mathrm{d}\varphi)\mathbb{E}[\tanh^2 x]| \le \lambda^2|\mathbb{E}[2\tanh x(1-\tanh^2 x)] \le 4\lambda^2 e^{-\lambda^2/5},$$

$$D_{Q,q} \le |(\mathrm{d}/\mathrm{d}q)\mathbb{E}[\tanh^2 x]| \le \lambda/(2\sqrt{q})|\mathbb{E}[2\tanh x(1-\tanh^2 x)G]|$$
$$\le \lambda^2|\mathbb{E}[(1-3\tanh^2 x)(1-\tanh^2 x)]| \le 4\lambda^2 e^{-\lambda^2/5},$$

$$D_{U,\varphi} \le |(\mathrm{d}/\mathrm{d}\varphi)\mathbb{E}[\log 2\cosh x]| \le \lambda^2|\mathbb{E}[\tanh x]| \le \lambda^2,$$

$$D_{U,q} \le |(\mathrm{d}/\mathrm{d}q)\mathbb{E}[\log 2\cosh x]| \le \lambda/(2\sqrt{q})|\mathbb{E}[G\tanh x]| \le (\lambda^2/2)\mathbb{E}[1-\tanh^2 x] \le \lambda^2 e^{-\lambda^2/5}.$$

Moreover, denoting $D_{u,q}$ as the Lipschitz constant of $u(q,a)$ with respect to q, we have

$$D_{u,q} \le \sup_{q\in[0.9,1]}|(\mathrm{d}/\mathrm{d}q)u(q,a)| \le \sup_{q\in[0.9,1]}\lambda^2 q/2 \le \lambda^2/2.$$

For $(f,F) = (q,Q)$ or (φ,M), we have

$$|f - \mathbb{E}[F(x)]| \le |f - f_\star| + D_{F,q}|q - q_\star| + D_{F,\varphi}|\varphi - \varphi_\star|.$$

For $\lambda \ge 10$, we have $D_{Q,q} \vee D_{Q,\varphi} \vee D_{M,q} \vee D_{M,\varphi} \le 1/2$, and hence

$$\max\{|q - q_\star|, |\varphi - \varphi_\star|\} \le L_q^2/2 \tag{79}$$

implies (77).

Moreover, for $(f,F) = (q,Q)$, (φ,M), or (a,A), we have

$$(f - \mathbb{E}[F(x)])^2 \ge (|f - f_\star| - D_{F,q}|q - q_\star| - D_{F,\varphi}|\varphi - \varphi_\star|)_+^2$$
$$\ge (f - f_\star)^2 - 2|f - f_\star| \cdot (D_{F,q}|q - q_\star| + D_{F,\varphi}|\varphi - \varphi_\star|)$$
$$\ge (1-\varepsilon)(f - f_\star)^2 - (1/\varepsilon)(D_{F,q}|q - q_\star| + D_{F,\varphi}|\varphi - \varphi_\star|)^2$$
$$\ge (1-\varepsilon)(f - f_\star)^2 - (2/\varepsilon)D_{F,q}^2(q - q_\star)^2 - (2/\varepsilon)D_{F,\varphi}^2(\varphi - \varphi_\star)^2,$$

and

$$(u(q,a) - e - \mathbb{E}[U])^2 \ge (|(e - e_\star) - (a - a_\star)/2| - (D_{U,q} + D_{u,q})|q - q_\star| - D_{U,\varphi}|\varphi - \varphi_\star|)_+^2$$
$$\ge (1-\varepsilon)[(e - e_\star) - (a - a_\star)/2]^2 - (2/\varepsilon)(D_{U,q} + D_{u,q})^2(q - q_\star)^2 - (2/\varepsilon)D_{U,\varphi}^2(\varphi - \varphi_\star)^2.$$

Accordingly, for λ_0 sufficiently large, using Eq. (78), we obtain

$$\inf_{\mu,\nu,\tau,\gamma} S_-(q,\varphi,a,e;\mu,\nu,\tau,\gamma)$$
$$\le -\frac{1-\varepsilon}{2\alpha L_a^2\lambda^2}(a - a_\star)^2$$
$$-\left[\frac{1-\varepsilon}{4\kappa L_q^2} - \frac{2D_{A,q}^2}{2\varepsilon\alpha L_a^2\lambda^2} - \frac{2D_{Q,q}^2}{4\varepsilon\kappa L_q^2} - \frac{2D_{M,q}^2}{4\varepsilon\kappa L_q^2} - \frac{2(D_{U,q} + D_{u,q})^2}{2\varepsilon\kappa\lambda^2}\right](q - q_\star)^2 \tag{80}$$
$$-\left[\frac{1-\varepsilon}{4\kappa L_q^2} - \frac{2D_{A,\varphi}^2}{2\varepsilon\alpha L_a^2\lambda^2} - \frac{2D_{Q,\varphi}^2}{4\varepsilon\kappa L_q^2} - \frac{2D_{M,\varphi}^2}{4\varepsilon\kappa L_q^2} - \frac{2D_{U,\varphi}^2}{2\varepsilon\kappa\lambda^2}\right](\varphi - \varphi_\star)^2$$
$$-\frac{1-\varepsilon}{2\kappa\lambda^2}[(e - e_\star) - (a - a_\star)/2]^2.$$

as long as (q,φ) satisfies (79).

Step 5. Finish the proof. Let us apply

$$S_\star(q, \varphi, a, e) = S_+(q, \varphi, a) + \inf_{\mu, \nu, \tau, \gamma} S_-(q, \varphi, a, e; \mu, \nu, \tau, \gamma)$$

and add the upper bounds from (70) and (80). As $\varepsilon = 0.01$, $\alpha = 1 + \varepsilon = 1.01$, $\kappa = (3 + 3\varepsilon)/\varepsilon = 303$, and $L_a \leq 1.2$, the coefficient for the $(a - a_\star)^2$ term

$$-\frac{1 - \varepsilon}{2\alpha L_a^2 \lambda^2} + \frac{(1 + \varepsilon)^4}{4\lambda^2}$$

is a negative function of λ. Now take $L_q^2 = 2c_0/\lambda^2$ for some small constant $c_0 > 0$. For λ_0 large enough, this implies $L_q^2 \geq \exp\{-\lambda^2/10\}$. Note $D_{Q,q}, D_{Q,\varphi}, D_{M,q}, D_{M,\varphi}$ are exponentially small in λ, and $D_{A,q}, D_{A,\varphi}, D_{U,q}, D_{u,q}, D_{U,\varphi} \leq 2\lambda^2$. Hence for c_0 small enough and λ_0 large enough, the coefficients for the $(q - q_\star)^2$ and $(\varphi - \varphi_\star)^2$ terms are also negative functions of λ; moreover, $\max\{|q - q_\star|, |\varphi - \varphi_\star|\} \leq c_0/\lambda^2$ implies (79). Finally, observe that for any $c_1(\lambda), c_2(\lambda) > 0$, there exists $c(\lambda) > 0$ such that

$$-c_1(\lambda)(a - a_\star)^2 - c_2(\lambda)[(e - e_\star) - (a - a_\star)/2]^2 \leq -c(\lambda)(a - a_\star)^2 - c(\lambda)(e - e_\star)^2.$$

This concludes the proof of (35).

E   Proof of Proposition 4.3

Lemma E.1. *Fix any $\alpha \in (1/2, 1)$. Then there exist constants λ_0, n_0, and $c_0 > 0$ so that for all $\lambda \geq \lambda_0$ and $n \geq n_0$, with probability at least $1 - e^{-c_0 n}$, every $\boldsymbol{m} \in (-1, 1)^n$ with $\mathcal{F}(\boldsymbol{m}) \leq -\alpha\lambda^2/2$ satisfies*

$$Q(\boldsymbol{m}) \geq (2\alpha - 1 - 6/\lambda - 4/\lambda^2)^{1/2},$$
$$|M(\boldsymbol{m})| \geq (2\alpha - 1 - 6/\lambda - 4/\lambda^2)^{1/4}.$$

Proof. Note that we have

$$\mathcal{F}(\boldsymbol{m}) \geq -1 - \lambda/(2n) \cdot \langle \boldsymbol{m}, \boldsymbol{W}\boldsymbol{m} \rangle - (\lambda^2/2)M(\boldsymbol{m})^2 - (\lambda^2/4)(1 - Q(\boldsymbol{m}))^2$$
$$\geq -1 - \lambda\|\boldsymbol{W}\|_{\mathrm{op}}/2 - (\lambda^2/2)M(\boldsymbol{m})^2 - (\lambda^2/4)(1 - Q(\boldsymbol{m}))^2.$$

With probability at least $1 - e^{-c_0 n}$ for $n \geq n_0$ and some $n_0, c_0 > 0$, we have $\|\boldsymbol{W}\|_{\mathrm{op}} < 3$. On this event, for any $\boldsymbol{m}$ such that $\mathcal{F}(\boldsymbol{m}) \leq -\alpha\lambda^2/2$, we have

$$M(\boldsymbol{m})^2 + 1/2 \cdot (1 - Q(\boldsymbol{m}))^2 \geq \alpha - 2/\lambda^2 - \|\boldsymbol{W}\|_{\mathrm{op}}/\lambda > \alpha - \varepsilon,$$

where $\varepsilon = 2/\lambda^2 + 3/\lambda$. By Cauchy-Schwarz, $Q(\boldsymbol{m}) \geq M(\boldsymbol{m})^2$. Then

$$Q(\boldsymbol{m}) + 1/2 \cdot (1 - Q(\boldsymbol{m}))^2 \geq \alpha - \varepsilon.$$

As $2\alpha - 1 - 2\varepsilon > 0$ for any fixed $\alpha > 1/2$ and for sufficiently large λ_0, we obtain

$$Q(\boldsymbol{m}) \geq (2\alpha - 1 - 2\varepsilon)^{1/2}.$$

Then also $(1 - Q(\boldsymbol{m}))^2 \leq (1 - (2\alpha - 1 - 2\varepsilon)^{1/2})^2$, so

$$M(\boldsymbol{m})^2 \geq \alpha - \varepsilon - 1/2 \cdot (1 - Q(\boldsymbol{m}))^2 \geq (2\alpha - 1 - 2\varepsilon)^{1/2}.$$

$\square$

Lemma E.2. *There exist constants $\lambda_0, n_0, C_0, c_0 > 0$ such that for any $\lambda \geq \lambda_0$ and $n \geq n_0$, we have*

$$\mathbb{P}\left(\sup_{\boldsymbol{u} \in \mathsf{B}^n(\mathbf{0}, 1)} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{|\langle \boldsymbol{g}_i, \boldsymbol{u} \rangle| \geq \lambda\}\right] \geq C_0/\lambda^2\right) \leq \exp\{-c_0 n\}, \tag{81}$$

where $\{\boldsymbol{g}_i\}_{i \in [n]} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_n)$ and $\mathsf{B}^n(\mathbf{0}, 1)$ is the n-dimensional unit ball centered at $\mathbf{0}$.

Proof. Let $N(\varepsilon) = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|N(\varepsilon)|}\}$ be an ε-net of $\mathsf{B}^n(\boldsymbol{0}, 1)$, with $|N(\varepsilon)| \leq (3/\varepsilon)^n$. That is, for any $\boldsymbol{v} \in \mathsf{B}^n(\boldsymbol{0}, 1)$, there exists $\boldsymbol{v}_\star \in N(\varepsilon)$, such that $\|\boldsymbol{v} - \boldsymbol{v}_\star\|_2 \leq \varepsilon$. Then, for any $\boldsymbol{u} \in \mathsf{B}^n(\boldsymbol{0}, 1)$, there exists a sequence $\{\boldsymbol{u}_j\}_{j\geq 0} \subseteq N(\varepsilon)$, such that $\boldsymbol{u} = \sum_{j=0}^\infty \varepsilon^j \boldsymbol{u}_j$. As a consequence, for any vector $\boldsymbol{g}$, we have

$$\mathbf{1}\{|\langle \boldsymbol{g}, \boldsymbol{u}\rangle| \geq \lambda\} \leq \sum_{j=0}^\infty \mathbf{1}\{\varepsilon^j |\langle \boldsymbol{g}, \boldsymbol{u}_j\rangle| \geq \lambda/2^{j+1}\}, \tag{82}$$

and hence

$$\sup_{\boldsymbol{u}\in\mathsf{B}^n(\boldsymbol{0},1)} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\langle \boldsymbol{g}_i, \boldsymbol{u}\rangle| \geq \lambda\} \leq \sum_{j=0}^\infty \sup_{\boldsymbol{u}\in N(\varepsilon)} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\langle \boldsymbol{g}_i, \boldsymbol{u}\rangle| \geq \lambda/[2(2\varepsilon)^j]\}. \tag{83}$$

We fix $\varepsilon = 1/3$ throughout the proof.

First, we show that for any $\chi \geq 4$ and $\delta \leq 1$, we have (for $\varepsilon = 1/3$)

$$\mathbb{P}\Big(\sup_{\boldsymbol{u}\in N(\varepsilon)} \Big[\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\langle \boldsymbol{g}_i, \boldsymbol{u}\rangle| \geq \chi\}\Big] \geq 16/(\delta\chi^2)\Big) \leq \exp\{-n/\delta\}. \tag{84}$$

For a fixed $\boldsymbol{u}$ with $\|\boldsymbol{u}\|_2 = 1$, applying the Chernoff-Hoeffding inequality and denoting $\phi(\chi) = \mathbb{P}(|G| \geq \chi)$ for $G \sim \mathcal{N}(0,1)$, we have for all $t > \phi(\chi)$

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\langle \boldsymbol{g}_i, \boldsymbol{u}\rangle| \geq \chi\} \geq t\Big) \leq \exp\{-n\mathsf{D}_{\mathrm{kl}}(t\|\phi(\chi))\},$$

where for $a, b \in (0, 1)$, we define $\mathsf{D}_{\mathrm{kl}}(a\|b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$ to be the relative entropy of Bernoulli distribution with parameters a and b. Taking union bound over $\boldsymbol{u} \in N(\varepsilon)$ and applying $\phi(\chi) \geq \phi(\chi/\|\boldsymbol{u}\|_2)$,

$$\mathbb{P}\Big(\sup_{\boldsymbol{u}\in N(\varepsilon)} \Big[\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\langle \boldsymbol{g}_i, \boldsymbol{u}\rangle| \geq \chi\}\Big] \geq t\Big)$$
$$\leq \sum_{\boldsymbol{u}\in N(\varepsilon)} \exp\{-n\mathsf{D}_{\mathrm{kl}}(t\|\phi(\chi/\|\boldsymbol{u}\|_2))\} \leq \exp\{n[-\mathsf{D}_{\mathrm{kl}}(t\|\phi(\chi)) + \log(3/\varepsilon)]\}.$$

Applying $\varepsilon = 1/3$ and $\phi(\chi) \leq 2\exp\{-\chi^2/2\}$, we have

$$-\mathsf{D}_{\mathrm{kl}}(t\|\phi(\chi)) + \log(3/\varepsilon) = t\log(\phi(\chi)) - t\log t + (1-t)\log[1 + (t - \phi(\chi))/(1-t)] + \log 9$$
$$\leq t\log(\phi(\chi)) - t\log t + (t - \phi(\chi)) + \log 9$$
$$\leq t(2 - \chi^2/2) + 3.$$

Now take $t = 16/(\delta\chi^2) \geq \phi(\chi)$ for $\chi \geq 4$ and $\delta \leq 1$. Then $-\mathsf{D}_{\mathrm{kl}}(t\|\phi(\chi)) + \log(3/\varepsilon) \leq -1/\delta$. This proves Eq. (84). Applying Eq. (84) with $\chi_j = \lambda/[2(2\varepsilon)^j]$ (requiring $\lambda \geq 8$ so that $\chi_j \geq 4$) and $\delta_j = (2\varepsilon)^j \leq 1$ for $j \geq 0$, we have

$$\mathbb{P}\Big(\sup_{\boldsymbol{u}\in N(\varepsilon)} \Big[\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\langle \boldsymbol{g}_i, \boldsymbol{u}\rangle| \geq \lambda/[2(2\varepsilon)^j]\}\Big] \geq 64(2\varepsilon)^j/\lambda^2\Big) \leq \exp\{-n/(2\varepsilon)^j\}.$$

Finally, taking a union bound over $j \geq 0$ and applying

$$\sum_{j=0}^\infty \exp\{-n/(2\varepsilon)^j\} \leq \exp\{-c_0 n\}, \qquad \sum_{j=0}^\infty 64(2\varepsilon)^j/\lambda^2 \leq C_0/\lambda^2$$

for $n \geq n_0$ with some $C_0, c_0, n_0 > 0$ concludes the proof. $\qquad\square$

Lemma E.3. *Fix any positive integer k. Then there exist $\lambda_0, C_0 > 0$ and functions $c_0(\lambda), n_0(\lambda) > 0$ such that for all $\lambda \geq \lambda_0$ and $n \geq n_0(\lambda)$, with probability at least $1 - e^{-c_0(\lambda)n}$, all points $\boldsymbol{m} \in (-1,1)^n$ which satisfy $M(\boldsymbol{m}) + Q(\boldsymbol{m}) \geq 1.01$ and $\nabla\mathcal{F}(\boldsymbol{m}) = \boldsymbol{0}$ also satisfy $M(\boldsymbol{m}), Q(\boldsymbol{m}) \geq 1 - C_0/\lambda^k$.*

Proof. Since $\boldsymbol{W} \sim \mathrm{GOE}(n)$, we can write $\boldsymbol{W} = (\boldsymbol{G}+\boldsymbol{G}^\mathsf{T})/\sqrt{2}$, where $\boldsymbol{G} = \{G_{ij}\}_{i,j\in[n]}$ with $G_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1/n)$. Note for any $\boldsymbol{m} \in (-1,1)^n$, we have $\|\boldsymbol{m}-\boldsymbol{1}\|_2/\sqrt{n} \le 2$. According to Lemma E.2, there exist constants C_0, c_0, λ_0, and n_0 such that for any $n \ge n_0$ and $\lambda \ge \lambda_0$, we have

$$\mathbb{P}\Big(\sup_{\boldsymbol{m}\in(-1,1)^n}\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{|[\boldsymbol{G}(\boldsymbol{m}-\boldsymbol{1})]_i| \ge 0.001\lambda\}\Big] \ge C_0/\lambda^2\Big) \le \exp\{-c_0 n\},$$

$$\mathbb{P}\Big(\sup_{\boldsymbol{m}\in(-1,1)^n}\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{|[\boldsymbol{G}^\mathsf{T}(\boldsymbol{m}-\boldsymbol{1})]_i| \ge 0.001\lambda\}\Big] \ge C_0/\lambda^2\Big) \le \exp\{-c_0 n\}. \tag{85}$$

Moreover, by a simple Chernoff bound, we have (denoting $\phi(\lambda) = \mathbb{P}(|G| \ge \lambda) \le 2\exp\{-\lambda^2/2\}$ for $G \sim \mathcal{N}(0,1)$)

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{|[\boldsymbol{G}\boldsymbol{1}]_i| \ge 0.001\lambda\} \ge 3\phi(0.001\lambda)\Big) \le \exp\{-n\phi(0.001\lambda)\},$$

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{|[\boldsymbol{G}^\mathsf{T}\boldsymbol{1}]_i| \ge 0.001\lambda\} \ge 3\phi(0.001\lambda)\Big) \le \exp\{-n\phi(0.001\lambda)\}. \tag{86}$$

For any $\boldsymbol{m} \in (-1,1)^n$, note that $\nabla\mathcal{F}(\boldsymbol{m}) = \boldsymbol{0}$ implies

$$\boldsymbol{m} = \tanh(\lambda^2 M(\boldsymbol{m})\boldsymbol{1} + \lambda\cdot\boldsymbol{W}\boldsymbol{1} + \lambda\cdot\boldsymbol{W}(\boldsymbol{m}-\boldsymbol{1}) - \lambda^2[1-Q(\boldsymbol{m})]\boldsymbol{m}. \tag{87}$$

When the bad events in (86) and (85) do not happen, and $M(\boldsymbol{m})+Q(\boldsymbol{m}) \ge 1.01$, then at least $1 - 2C_0/\lambda^2 - 6\phi(0.001\lambda)$ fraction of coordinates of $\boldsymbol{m}$ satisfy

$$\begin{aligned} m_j &= \tanh(\lambda^2 M(\boldsymbol{m}) + \lambda[\boldsymbol{W}\boldsymbol{1}]_j + \lambda[\boldsymbol{W}(\boldsymbol{m}-\boldsymbol{1})]_j - \lambda^2(1-Q(\boldsymbol{m}))m_j) \\ &\ge \tanh((M(\boldsymbol{m})+Q(\boldsymbol{m})-1)\lambda^2 - 0.004\lambda^2) \ge \tanh(0.006\lambda^2) \\ &\ge 1 - e^{-0.006\lambda^2}, \end{aligned} \tag{88}$$

and the remaining coordinates of $\boldsymbol{m}$ satisfy $m_j \ge -1$. Therefore, for sufficiently large λ_0, we have

$$\begin{aligned} M(\boldsymbol{m}) = \langle\boldsymbol{m},\boldsymbol{1}\rangle/n &\ge (1 - 2C_0/\lambda^2 - 6\phi(0.001\lambda))(1 - e^{-0.006\lambda^2}) - 2C_0/\lambda^2 - 6\phi(0.001\lambda) \\ &\ge 1 - 6C_0/\lambda^2. \end{aligned}$$

Hence we also have

$$\|\boldsymbol{m}-\boldsymbol{1}\|_2/\sqrt{n} \le (2 - 2\langle\boldsymbol{m},\boldsymbol{1}\rangle/n)^{1/2} \le (12C_0)^{1/2}/\lambda, \tag{89}$$

with probability at least $1 - e^{-c_1(\lambda)n}$ for all $n \ge n_1(\lambda)$.

In the following, we apply the above argument recursively to prove the lemma. Suppose we already know that $\|\boldsymbol{m}-\boldsymbol{1}\|_2/\sqrt{n} \le K/\lambda^k$ for some constants $K > 0$ and $k \ge 1$, with probability $1 - e^{-c_k(\lambda)n}$ for $n \ge n_k(\lambda)$. Applying again Lemma E.2, for λ_0 sufficiently large such that $\varepsilon(\lambda) \equiv K/\lambda^k \le 2$ for all $\lambda \ge \lambda_0$, we have

$$\mathbb{P}\Big(\sup_{\|\boldsymbol{m}-\boldsymbol{1}\|_2/\sqrt{n}\le\varepsilon(\lambda)}\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{|[\boldsymbol{G}(\boldsymbol{m}-\boldsymbol{1})]_i| \ge 0.001\lambda\}\Big] \ge C_0\varepsilon(\lambda)^2/\lambda^2\Big) \le \exp\{-c_0 n\},$$

$$\mathbb{P}\Big(\sup_{\|\boldsymbol{m}-\boldsymbol{1}\|_2/\sqrt{n}\le\varepsilon(\lambda)}\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{|[\boldsymbol{G}^\mathsf{T}(\boldsymbol{m}-\boldsymbol{1})]_i| \ge 0.001\lambda\}\Big] \ge C_0\varepsilon(\lambda)^2/\lambda^2\Big) \le \exp\{-c_0 n\}. \tag{90}$$

When the bad events in (86) and (90) do not happen, for any $\boldsymbol{m}$ such that $\|\boldsymbol{m}-\boldsymbol{1}\|_2/\sqrt{n} \le \varepsilon(\lambda)$, by (87) at least $1 - 2C_0\varepsilon(\lambda)^2/\lambda^2 - 6\phi(0.001\lambda)$ fraction of coordinates of $\boldsymbol{m}$ satisfy (88). Therefore, for sufficiently large λ_0 (depending on K and k), we have

$$\begin{aligned} M(\boldsymbol{m}) = \langle\boldsymbol{m},\boldsymbol{1}\rangle/n &\ge (1 - 2C_0\varepsilon(\lambda)^2/\lambda^2 - 6\phi(0.001\lambda))(1 - e^{-0.006\lambda^2}) - 2C_0\varepsilon(\lambda)^2/\lambda^2 - 6\phi(0.001\lambda) \\ &\ge 1 - 6C_0\varepsilon(\lambda)^2/\lambda^2. \end{aligned}$$

Hence we have

$$\|\boldsymbol{m} - \mathbf{1}\|_2/\sqrt{n} \le (2 - 2\langle \boldsymbol{m}, \mathbf{1}\rangle/n)^{1/2} \le (12C_0)^{1/2}\varepsilon(\lambda)/\lambda = (12C_0)^{1/2}K/\lambda^{k+1}. \tag{91}$$

This holds with probability at least $1 - e^{-c_{k+1}(\lambda)n}$ for $n \ge n_{k+1}(\lambda)$ for some $0 < c_{k+1}(\lambda) \le c_k(\lambda)$ and $n_{k+1}(\lambda) \ge n_k(\lambda)$.

Thus for any fixed k, there are k-dependent constants $\lambda_0, C_0 > 0$ and k-dependent functions $c_0(\lambda), n_0(\lambda)$ such that $M(\boldsymbol{m}) \ge 1 - C_0/\lambda^k$ with probability at least $1 - e^{-c_0(\lambda)n}$ for $\lambda \ge \lambda_0$ and $n \ge n_0(\lambda)$. Applying $Q(\boldsymbol{m}) \ge M(\boldsymbol{m})^2$ by Cauchy-Schwarz, we obtain also the statement for $Q(\boldsymbol{m})$. □

Proof of Proposition 4.3. Applying Lemma E.1 with $\alpha = 2/3$, for any $\lambda \ge \lambda_0$ and $n \ge n_0$, with probability at least $1 - e^{-c_0 n}$, all $\boldsymbol{m} \in (-1,1)^n$ with $\mathcal{F}(\boldsymbol{m}) \le -\lambda^2/3$ satisfy $Q(\boldsymbol{m}) \ge M(\boldsymbol{m})^2 \ge (1/3 - 6/\lambda - 4/\lambda^2)^{1/2}$. For λ_0 large enough and $M(\boldsymbol{m}) \ge 0$, this implies $Q(\boldsymbol{m}) + M(\boldsymbol{m}) \ge 1.01$. Hence the proposition holds by Lemma E.3.

□

References

[AAČ13] Antonio Auffinger, Gérard Ben Arous, and Jiří Černỳ. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

[ADFDJ03] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, Cambridge New York, 2010.

[AJ16] Antonio Auffinger and Aukosh Jagannath. Thouless-anderson-palmer equations for conditional gibbs measures in the generic p-spin glass model. *arXiv preprint arXiv:1612.06359*, 2016.

[AMMN18] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics*, 2018. to appear, arXiv:1711.05424.

[AT09] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

[BAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[BBC04] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[Bia97] Philippe Biane. On the free convolution with a semi-circular distribution. *Indiana University Mathematics Journal*, pages 705–718, 1997.

[BKM17] 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.

[Ble12] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[BLM13] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BM80] AJ Bray and Michael A Moore. Metastable states in spin glasses. *Journal of Physics C: Solid State Physics*, 13(19):L469, 1980.

[BM11]   Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[BMY84]   AJ Bray, Michael A. Moore, and A. Peter Young. Weighted averages of tap solutions and parisi's q (x). *Journal of Physics C: Solid State Physics*, 17(5):L155, 1984.

[Bol14]   Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[CDMFF11] Mireille Capitaine, Catherine Donati-Martin, Delphine Féral, and Maxime Février. Free convolution with a semicircular distribution and eigenvalues of spiked deformations of wigner matrices. *Electronic Journal of Probability*, 16:1750–1792, 2011.

[CGPM03]   Andrea Cavagna, Irene Giardina, Giorgio Parisi, and Marc Mézard. On the formal equivalence of the TAP and thermodynamic methods in the SK model. *Journal of Physics A: Mathematical and General*, 36(5):1175, 2003.

[Cha10]   Sourav Chatterjee. Spin glasses and steins method. *Probability theory and related fields*, 148(3-4):567–600, 2010.

[CLPR03]   Andrea Crisanti, Luca Leuzzi, Giorgio Parisi, and Tommaso Rizzo. Complexity in the sherrington-kirkpatrick model in the annealed approximation. *Physical Review B*, 68(17):174401, 2003.

[CLPR04]   Andrea Crisanti, Luca Leuzzi, Giorgio Parisi, and Tommaso Rizzo. Spin-glass complexity. *Physical review letters*, 92(12):127203, 2004.

[CLR05]   Andrea Crisanti, Luca Leuzzi, and Tommaso Rizzo. Complexity in mean-field spin-glass models: Ising p-spin. *Physical Review B*, 71(9):094202, 2005.

[CP17]   Wei-Kuo Chen and Dmitry Panchenko. On the TAP free energy in the mixed p-spin models. arXiv:1709.03468, 2017.

[DAM16]   Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2016.

[DDY83]   Cyrano De Dominicis and A. Peter Young. Weighted averages and order parameters for the infinite range ising spin glass. *Journal of Physics A: Mathematical and General*, 16(9):2063, 1983.

[Dia09]   Persi Diaconis. The Markov Chain Monte Carlo Revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[DS01]   Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

[DZ10]   Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Springer-Verlag, Berlin, 2010.

[Fyo04]   Yan V Fyodorov. Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices. *Physical review letters*, 92(24):240601, 2004.

[GJM18]   Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. arXiv:1802.00568, 2018.

[Gue03]   Francesco Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12, 2003.

[HLL83]   Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[JMRT16]   Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, 2016.

[KF09]   Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[KS78]   Scott Kirkpatrick and David Sherrington. Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384, 1978.

[Mit15]   Boris Mityagin. The zero set of a real analytic function. arXiv:1512.07276, 2015.

[MM09]   Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[MPV87]   Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[MS16]   Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827. ACM, 2016.

[MV17]   Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. arXiv:1711.01682, 2017.

[Nis01]   Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[Par79]   Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979.

[Par80]   Giorgio Parisi. A sequence of approximated solutions to the sk model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980.

[Par83]   Giorgio Parisi. Order parameter for spin-glasses. *Physical Review Letters*, 50(24):1946, 1983.

[Pas72]   Leonid A Pastur. On the spectrum of random matrices. *Theoretical and Mathematical Physics*, 10(1):67–74, 1972.

[Pas05]   Leonid A. Pastur. A simple approach to the global regime of gaussian ensembles of random matrices. *Ukrainian Mathematical Journal*, 57(6):936–966, 2005.

[Ple82]   T Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, 15(6):1971, 1982.

[PP95]   Giorgio Parisi and Marc Potters. On the number of metastable states in spin glasses. *EPL (Europhysics Letters)*, 32(1):13, 1995.

[Sin11]   Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20, 2011.

[Sin12]   Alistair Sinclair. *Algorithms for random generation and counting: a Markov chain approach*. Springer Science & Business Media, 2012.

[Sub17]   Eliran Subag. The complexity of spherical p-spin modelsa second moment approach. *The Annals of Probability*, 45(5):3385–3450, 2017.

[SW11]   Amit Singer and Hau-tieng Wu. Orientability and diffusion maps. *Applied and computational harmonic analysis*, 31(1):44–58, 2011.

[Tal10]   Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

[TAP77]   David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[Voi91]   Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104(1):201–220, 1991.

[WJ08]   Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[ZZ17]   Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. arXiv:1710.11268, 2017.