The Annals of Statistics
2020, Vol. 48, No. 2, 863–885
https://doi.org/10.1214/19-AOS1826

FUNDAMENTAL LIMITS OF DETECTION IN THE SPIKED WIGNER MODEL

BY AHMED EL ALAOUI[1], FLORENT KRZAKALA[2] AND MICHAEL JORDAN[3]

[1]*Department of Electrical Engineering, Stanford University, elalaoui@stanford.edu*

[2]*Laboratoire de Physique Statistique, CNRS, PSL Universités, Ecole Normale Supérieure, Sorbonne Universités et Université Pierre & Marie Curie, florent.krzakala@ens.fr*

[3]*Department of Electrical Engineering & Computer Sciences, Department of Statistics, University of California, Berkeley, jordan@cs.berkeley.edu*

We study the fundamental limits of detecting the presence of an additive rank-one perturbation, or spike, to a Wigner matrix. When the spike comes from a prior that is i.i.d. across coordinates, we prove that the log-likelihood ratio of the spiked model against the nonspiked one is asymptotically normal below a certain *reconstruction threshold* which is not necessarily of a "spectral" nature, and that it is degenerate above. This establishes the maximal region of contiguity between the planted and null models. It is known that this threshold also marks a phase transition for estimating the spike: the latter task is possible above the threshold and impossible below. Therefore, both estimation and detection undergo the same transition in this random matrix model. Further information on the performance of the optimal test is also provided. Our proofs are based on Gaussian interpolation methods and a rigorous incarnation of the cavity method, as devised by Guerra and Talagrand in their study of the Sherrington–Kirkpatrick spin-glass model.

1. Introduction. Spiked models, which are distributions over matrices of the form "signal + noise," have been a mainstay in the statistical literature since their introduction by Johnstone (2001) as models for the study of high-dimensional principal component analysis. Spectral properties of these models have been extensively studied, in particular in random matrix theory, where they are known as *deformed ensembles* (Péché (2014)). Landmark investigations in this area (Baik, Ben Arous and Péché (2005), Baik and Silverstein (2006), Péché (2006), Féral and Péché (2007), Capitaine, Donati-Martin and Féral (2009), Bai and Yao (2012, 2008)) have established the existence of a *spectral threshold* above which the top eigenvalue detaches from the bulk of eigenvalues and becomes informative about the spike, and below which the top eigenvalue bears no information. Estimation using the top eigenvector undergoes the same transition, where it is known to "lose track" of the spike below the spectral threshold (Benaych-Georges and Nadakuditi (2011), Johnstone and Lu (2009), Nadler (2008), Paul (2007)). Although these spectral analyses have provided many insights, as have analyses based on more thoroughgoing usage of spectral data and/or more advanced optimization-based procedures (see Amini and Wainwright (2009), Berthet and Rigollet (2013), Dobriban (2017), Ledoit and Wolf (2002) and references therein), they do not characterize the fundamental limits of estimating the spike, or detecting its presence, from the observation of a sample matrix. Important progress on the detection problem was made by Onatski, Moreira and Hallin (2013, 2014) and Johnstone and Onatski (2015), who considered the spiked covariance model for a uniformly distributed

Received October 2017; revised February 2019.

MSC2010 subject classifications. Primary 62H25; secondary 62H15, 60G15, 60F05.

Key words and phrases. Hypothesis testing, random matrix models, contiguity, spin–glasses, Sherrington–Kirkpatrick model, replica–symmetry.

unit norm spike, and studied the asymptotics of the likelihood ratio (LR) of a spiked alternative against a spherical null. They showed asymptotic normality of the log-LR below the spectral threshold (also known as the BBP threshold, after Baik, Ben Arous and Péché (2005) in this setting), while it is degenerate, that is, exponentially small (large) under the null (alternative), above it. Their proof is intrinsically tied to the assumption of a spherical prior since it relies on the rotational symmetry of the model to express the LR exclusively in terms of the spectrum, the joint distribution of which is available in closed form.

We focus in this paper on the *spiked Wigner model*, which is the following symmetric random matrix model:

$$\boldsymbol{Y} = \sqrt{\frac{\lambda}{N}}\boldsymbol{x}^*\boldsymbol{x}^{*\top} + \boldsymbol{W}, \tag{1}$$

where $W_{ij} = W_{ji} \sim \mathcal{N}(0,1)$ and $W_{ii} \sim \mathcal{N}(0,\sigma^2)$, $\sigma > 0$, are independent for all $1 \le i \le j \le N$. The *spike* vector $\boldsymbol{x}^* \in \mathbb{R}^N$ represents the signal to be recovered, or its presence detected.

We assume that the entries x_i^* of the spike are i.i.d. from a prior distribution P_{x} on $\mathbb{R}$ having *bounded* support. The parameter $\lambda \ge 0$ plays the role of the signal-to-noise ratio, and the scaling by $\sqrt{N}$ is such that the signal and noise components of the observed data are of comparable magnitudes. Upon observing $\boldsymbol{Y}$, we want to test whether $\lambda > 0$ or $\lambda = 0$. We moreover want to understand the performance of the likelihood ratio test, which minimizes the sum of the Type-I and Type-II errors by the Neyman–Pearson lemma.

The testing problem becomes more subtle in our setting, where the spike comes from a product prior, since it is not clear that one does not lose power by discarding the eigenvectors of $\boldsymbol{Y}$. In fact, this situation presents a richer phenomenology: while the spherical case is characterized by the behavior of the spectrum, and the spectral threshold separates the regions of convergence and degeneracy of the LR, there are priors P_{x} in the i.i.d. case for which the spectral threshold loses its information-theoretic relevance. These priors exhibit a more subtle phase transition that happens *strictly before* the spike manifests its presence in the spectrum. A desire to understand this phenomenon is the main impetus for the present work.

This transition was discovered by Lesieur, Krzakala and Zdeborová (2015) while studying the estimation problem in the context of sparse PCA. Perry et al. (2018) and Banks et al. (2017) proved the possibility of both estimation and asymptotically certain—we will say "strong"—detection below the spectral threshold for certain sparse priors. However, their techniques—which are based on careful conditioning of the second moment of the LR—are not able to determine the phase transition threshold, the explicit form of which was conjectured by Lesieur et al.

Our contribution is to rigorously pin down this phase transition for the detection problem. We prove asymptotic normality of the log-LR below a certain *reconstruction threshold* λ_c and degeneracy above it. This allows us to show mutual contiguity of the null and the alternative below λ_c and to derive formulas for the Type-I and Type-II errors of the LR test, as well as the KL divergence and total variation distance, between the null and alternative. Our approach reposes on seminal work by Guerra and Talagrand in their study of the Sherrington–Kirkpatrick (SK) spin-glass model.

The paper is organized as follows: Section 2 sets up the problem, Section 3 contains our main results on LR fluctuations and the limits of detection, Section 4 provides background on essential concepts from spin-glass theory that are necessary for the proof, and Sections 5, 6 and 7 are devoted to the detailed proofs.

2. The LR, the RS formula and the reconstruction threshold.

2.1. *The LR.* We denote by $\mathbb{P}_\lambda$ the joint probability law of the observations, $\boldsymbol{Y} = \{Y_{ij} : 1 \le i \le j \le N\}$, as per (1) and we define the likelihood ratio or Radon–Nikodym derivative of $\mathbb{P}_\lambda$ with respect to $\mathbb{P}_0$ as

$$L(\cdot;\lambda) \equiv \frac{\mathrm{d}\mathbb{P}_\lambda}{\mathrm{d}\mathbb{P}_0}. \tag{2}$$

Conditioning on $\boldsymbol{x}^*$ and using the Gaussianity of $\boldsymbol{W}$ yields the formula

$$\begin{aligned} L(\boldsymbol{Y};\lambda) = \int \exp\Bigg(&\sum_{i<j}\sqrt{\frac{\lambda}{N}}Y_{ij}x_ix_j - \frac{\lambda}{2N}x_i^2x_j^2 \\ &+ \frac{1}{\sigma^2}\sum_{i=1}^N\sqrt{\frac{\lambda}{N}}Y_{ii}x_i^2 - \frac{\lambda}{2N}x_i^4\Bigg)\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}), \end{aligned} \tag{3}$$

for any fixed $\boldsymbol{Y}$. Define the *free energy* of the planted model $\mathbb{P}_\lambda$ as

$$f_N := \frac{1}{N}\mathbb{E}_{\mathbb{P}_\lambda}\log L(\boldsymbol{Y};\lambda) = \frac{1}{N}D_{\mathsf{KL}}(\mathbb{P}_\lambda,\mathbb{P}_0), \tag{4}$$

where D_{KL} is the Kullback–Leibler divergence between probability measures. The reconstruction threshold λ_c is defined as the largest positive number below which the limit of f_N vanishes. This latter limit, referred to as the *replica-symmetric* (RS) formula, provides a full characterization of the limits of estimating the spike with nontrivial accuracy (Barbier et al. (2016), Lelarge and Miolane (2019)).

2.2. *The* RS *formula.* For $r \ge 0$, consider the function

$$\psi(r) := \mathbb{E}_{x^*,z}\log\int\exp\left(\sqrt{r}zx + rxx^* - \frac{r}{2}x^2\right)\mathrm{d}P_{\mathrm{x}}(x), \tag{5}$$

where $z \sim \mathcal{N}(0,1)$, and $x^* \sim P_{\mathrm{x}}$. This is the KL divergence between the distributions of the random variables $y = \sqrt{r}x^* + z$ and z. We define the replica-symmetric potential

$$F(\lambda,q) := \psi(\lambda q) - \frac{\lambda q^2}{4}, \tag{6}$$

and the replica-symmetric formula

$$\phi_{\mathsf{RS}}(\lambda) := \sup_{q\ge 0} F(\lambda,q). \tag{7}$$

A central result in this context, which was conjectured by Lesieur, Krzakala and Zdeborová (2015), and then proved in a sequence of papers (Barbier et al. (2016), Deshpande, Abbé and Montanari (2016), El Alaoui and Krzakala (2018), Krzakala, Xu and Zdeborová (2016), Lelarge and Miolane (2019)), is that free energy f_N converges to the RS formula for all $\lambda \ge 0$:

$$f_N \quad\longrightarrow\quad \phi_{\mathsf{RS}}(\lambda). \tag{8}$$

In particular, the limit is independent of σ, that is, it is insensitive to $(Y_{ii})_{i=1}^N$.

The values of q that maximize the RS potential and their properties play an important role. Lelarge and Miolane (2019) proved that the map $q \mapsto F(\lambda,q)$ has a unique maximizer

$q^* = q^*(\lambda)$ for all $\lambda \in \mathcal{D}$ where $\mathcal{D} = \mathbb{R}_+ \setminus$ countable set. Moreover, they showed that the map $\lambda \in \mathcal{D} \mapsto q^*(\lambda)$ is nondecreasing, and

$$\lim_{\substack{\lambda \to 0 \\ \lambda \in \mathcal{D}}} q^*(\lambda) = \mathbb{E}_{P_{\mathrm{x}}}[X]^2 \quad \text{and} \quad \lim_{\substack{\lambda \to \infty \\ \lambda \in \mathcal{D}}} q^*(\lambda) = \mathbb{E}_{P_{\mathrm{x}}}[X^2], \tag{9}$$

where $X \sim P_{\mathrm{x}}$. One can interpret the value $q^*(\lambda)$ as the best overlap an estimator $\widehat{\theta}(\boldsymbol{Y})$ based on observing $\boldsymbol{Y}$ can have with the spike $\boldsymbol{x}^*$. Indeed, Lelarge and Miolane also showed that the squared overlap $(\frac{1}{N}\boldsymbol{x}^\top \boldsymbol{x}^*)^2$ between the spike $\boldsymbol{x}^*$ and a random draw $\boldsymbol{x}$ from the posterior $\mathbb{P}_\lambda(\cdot|\boldsymbol{Y})$ concentrates about $q^*(\lambda)^2$.

2.3. *The reconstruction threshold.* The first limit in (9) shows that when the prior P_{x} is not centered, it is always possible to have a nonzero overlap with $\boldsymbol{x}^*$ (just by guessing at random from the prior). An interesting situation then is when the prior has zero mean. Since q^* is a nondecreasing function of λ, it is useful to define the critical value of λ below which a nonzero overlap with $\boldsymbol{x}^*$ is impossible:

$$\begin{aligned} \lambda_c &:= \sup\{\lambda > 0 : q^*(\lambda) = 0\} \\ &= \sup\{\lambda > 0 : \phi_{\mathsf{RS}}(\lambda) = 0\}. \end{aligned} \tag{10}$$

The second equality follows by the a.e. uniqueness of the maximizer q^*. We refer to λ_c as the *reconstruction threshold*. The next lemma establishes a natural bound on λ_c.

LEMMA 1. *We have* $\lambda_c \cdot (\mathbb{E}_{P_{\mathrm{x}}}[X^2])^2 \le 1$.

PROOF. Indeed, assume that P_{x} is centered, and let $\lambda > (\mathbb{E}[X^2])^{-2}$. Since $\psi'(0) = \frac{1}{2}\mathbb{E}_{P_{\mathrm{x}}}[X]^2 = 0$ and $\psi''(0) = \frac{1}{2}(\mathbb{E}_{P_{\mathrm{x}}}[X^2])^2$, we see that $\partial_q F(\lambda, 0) = 0$ and $\partial_q^2 F(\lambda, 0) = \frac{\lambda}{2}(\lambda\, \mathbb{E}_{P_{\mathrm{x}}}[X^2]^2 - 1) > 0$. So $q = 0$ cannot be a maximizer of $F(\lambda, \cdot)$. Therefore, $q^*(\lambda) > 0$ and $\lambda \ge \lambda_c$. $\square$

The importance of Lemma 1 stems from the fact that the value $(\mathbb{E}_{P_{\mathrm{x}}}[X^2])^{-2}$ is the spectral threshold previously discussed. Above this value, the first eigenvalue of the matrix $\boldsymbol{Y}$ detaches from the bulk (Capitaine, Donati-Martin and Féral (2009), Féral and Péché (2007), Péché (2006)). This value also marks the limit below which the first eigenvector of $\boldsymbol{Y}$ captures no information about the spike $\boldsymbol{x}^*$ (Benaych-Georges and Nadakuditi (2011)). The inequality in Lemma 1 can be strict or turn into equality depending on the prior P_{x}. For instance, there is equality if the prior is Gaussian or Rademacher—so that the first eigenvector overlaps with the spike as soon as estimation becomes possible at all—and strict inequality in the case of the (sufficiently) sparse Rademacher prior $P_{\mathrm{x}} = \frac{\rho}{2}\delta_{-1/\sqrt{\rho}} + (1-\rho)\delta_0 + \frac{\rho}{2}\delta_{+1/\sqrt{\rho}}$. More precisely, there exists a value

$$\rho^* = \inf\{\rho \in (0,1) : \psi'''(0) < 0\} \approx 0.092,$$

such that $\lambda_c = 1$ for $\rho \ge \rho^*$, and $\lambda_c < 1$ for $\rho < \rho^*$. In the latter case, the spectral approach to estimating $\boldsymbol{x}^*$ fails for $\lambda \in (\lambda_c, 1)$, and it is believed that no polynomial time algorithm succeeds in this region (Banks et al. (2017), Krzakala, Xu and Zdeborová (2016), Lesieur, Krzakala and Zdeborová (2015)).

3. Fluctuations below the reconstruction threshold. In this section, we study the behavior of $\log L$. It can be seen by a standard concentration-of-measure argument that for all $\lambda > 0$, $\log L(\boldsymbol{Y}; \lambda)$ concentrates about its expectation with fluctuations of order $\sqrt{N}$. While this bound is likely to be of the right order above λ_c, it is very pessimistic below λ_c. Indeed,

we will show that the fluctuations are of constant order with a Gaussian limiting law in this regime. This behavior of unusually small fluctuations is often referred to as "superconcentration." We refer to Chatterjee (2014) for more on this topic. Throughout the rest of the paper, except in Section 8, we discard the diagonal terms Y_{ii} from the observations: we formally take $\sigma = +\infty$ in (3). (See the Remark below.)

THEOREM 2. *Assume that the prior P_{x} is centered, has unit variance and bounded support. Also, let $\sigma = +\infty$. For all $\lambda < \lambda_c$,*

$$\log L(\boldsymbol{Y};\lambda) \rightsquigarrow \mathcal{N}\Big(\pm\frac{1}{4}\big(-\log(1-\lambda)-\lambda\big), \frac{1}{2}\big(-\log(1-\lambda)-\lambda\big)\Big),$$

where the plus sign holds under the alternative $\boldsymbol{Y} \sim \mathbb{P}_\lambda$ and the minus sign under the null $\boldsymbol{Y} \sim \mathbb{P}_0$. The symbol "$\rightsquigarrow$" denotes convergence in distribution as $N \to \infty$.

REMARK. The assumption $\sigma = +\infty$ is only for convenience; its removal does not pose any additional technical difficulties. When the diagonal is kept, the limiting Gaussian is still of the form $\mathcal{N}(\pm\mu, 2\mu)$, but now $\mu = \frac{1}{4}(-\log(1-\lambda)-\lambda)(1+\frac{\kappa}{\sigma^2}) + \frac{\lambda}{2\sigma^2}$, $\kappa = \mathbb{E}_{P_{\mathrm{x}}}[X^3]^2$. We refer to Section 8 for a discussion of how this adjusted formula would appear in the proof.

We point out that a result of this form was originally proved in the case of the Sherrington–Kirkpatrick (SK) model: Aizenman, Lebowitz and Ruelle (1987) showed that the log-partition function of this model has Gaussian fluctuations in the "high temperature" regime (which corresponds to λ small enough.) In fact, Theorem 2, if specialized to the Rademacher prior $P_{\mathrm{x}} = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$, reduces to their result (with $\lambda_c = 1$) since the LR L is equal to the partition function of the SK model in that case.

Our result has a parallel in the work of Johnstone and Onatski (2015), Onatski, Moreira and Hallin (2013, 2014), who focused on spherical priors and studied the likelihood ratio of the joint eigenvalue densities under the spiked covariance model, showing its asymptotic normality below the spectral threshold. We also note that similar fluctuation results were recently proved by Baik and Lee (2016, 2017) for a spherical model where one integrates over the uniform measure on the sphere in the definition of L. Their model, due to its integrable nature, is amenable to analysis using tools from random matrix theory. The authors are thus able to also analyze a "low temperature" regime (absent from our problem) where the fluctuations are no longer Gaussian but given by the Tracy–Widom law. However, their techniques seem to be restricted to the spherical case. Closer to our setting is the recent work of Banerjee and Ma (2018) (see also Banerjee (2018)) who use a precisely conditioned second-moment argument to show asymptotic normality of similar log-likelihood ratios. However, this technique, in its current state, is not able to achieve the optimal threshold λ_c.

3.1. *Limits of strong and weak detection.* Consider the problem of deciding whether an array of observations $\boldsymbol{Y} = \{Y_{ij} : 1 \le i < j \le N\}$ is likely to have been generated from $\mathbb{P}_\lambda$ for a fixed $\lambda > 0$ or from $\mathbb{P}_0$. Let us denote by $\boldsymbol{H}_0 : \boldsymbol{Y} \sim \mathbb{P}_0$ the null hypothesis and $\boldsymbol{H}_\lambda : \boldsymbol{Y} \sim \mathbb{P}_\lambda$ the alternative hypothesis. We consider two formulations of this problem: one would like to construct a sequence of measurable tests $T : \mathbb{R}^{N(N-1)/2} \mapsto \{0, 1\}$ that returns 0 for $\boldsymbol{H}_0$ and 1 for $\boldsymbol{H}_\lambda$, for which either

$$\lim_{N\to\infty} \mathbb{P}_\lambda\big(T(\boldsymbol{Y}) = 0\big) \vee \mathbb{P}_0\big(T(\boldsymbol{Y}) = 1\big) = 0, \tag{11}$$

or less stringently, the total misclassification error or risk,

$$\mathsf{err}(T) := \mathbb{P}_\lambda\big(T(\boldsymbol{Y}) = 0\big) + \mathbb{P}_0\big(T(\boldsymbol{Y}) = 1\big), \tag{12}$$

is minimized among all possible tests T.

Strong detection. Using a second-moment argument (based on the computation of a truncated version of $\mathbb{E}\, L(\boldsymbol{Y};\lambda)^2$), Banks et al. (2017) and Perry et al. (2018) showed that $\mathbb{P}_\lambda$ and $\mathbb{P}_0$ are mutually contiguous when $\lambda < \lambda_0$, where the latter quantity equals λ_c for some priors P_{x} while it is suboptimal for others (e.g., the sparse Rademacher case; see further discussion below). It is easy to see that contiguity implies impossibility of strong detection since, for instance, if $\mathbb{P}_0(T(\boldsymbol{Y}) = 1) \to 0$ then $\mathbb{P}_\lambda(T(\boldsymbol{Y}) = 0) \to 1$. Here, we show that Theorem 2 provides a more powerful approach to contiguity.

COROLLARY 3. *Assume the prior P_{x} is centered, has unit variance and bounded support. Then for all $\lambda < \lambda_c$, $\mathbb{P}_\lambda$ and $\mathbb{P}_0$ are mutually contiguous.*

PROOF. A consequence of Theorem 2 is that if $\frac{\mathrm{d}\,\mathbb{P}_\lambda}{\mathrm{d}\,\mathbb{P}_0} \rightsquigarrow U$ under $\mathbb{P}_0$ along some subsequence and for some random variable U, then by the continuous mapping theorem we necessarily have $U = \exp \mathcal{N}(-\mu, 2\mu)$, where $\mu = \frac{1}{4}(-\log(1-\lambda) - \lambda)$. We have $\Pr(U > 0) = 1$, and $\mathbb{E}\, U = 1$. We now conclude using Le Cam's first lemma in both directions (Lemma 6.4 or Example 6.5, van der Vaart (1998)). $\square$

This approach allows one to circumvent second-moment computations which are not guaranteed to be tight in general, and necessitate careful and prior-specific conditioning that truncates away problematic atypical events. On the other hand, we prove (at the end of Section 7) that strong detection is possible above λ_c.

PROPOSITION 4. *Let $\lambda > \lambda_c$. If $\boldsymbol{Y} \sim \mathbb{P}_\lambda$, then $\frac{1}{N}\log L(\boldsymbol{Y};\lambda) > 0$ with probability approaching one as $N \to +\infty$. On the other hand, if $\boldsymbol{Y} \sim \mathbb{P}_0$ then $\frac{1}{N}\log L(\boldsymbol{Y};\lambda) \le 0$ with probability approaching one as $N \to +\infty$. Therefore, $\mathbb{P}_\lambda$ and $\mathbb{P}_0$ are mutually orthogonal above λ_c.*

REMARK. It is tempting to believe that $\overline{\lim} \frac{1}{N}\, \mathbb{E}_{\mathbb{P}_0} \log L(\boldsymbol{Y};\lambda) < 0$ above λ_c (the high-probability statement is a consequence of concentration), but we do not know of a simple proof of this. One can show, following Guerra (2003), that there is a nonincreasing sequence of thresholds $(\lambda_k)_{k\ge 1}$—each one corresponding to the point where the so-called "k-RSB" interpolation bound dips below zero—such that the above limit is strictly negative above $\lambda_\infty = \lim \lambda_k$. By our contiguity argument, it is necessarily true that $\lambda_\infty \ge \lambda_c$. Equality would follow if one can show overlap convergence (the analogue of Theorem 10 with $R_{1,2}$ replacing $R_{1,*}$) for all $\lambda < \lambda_\infty$ under the null model $\mathbb{P}_0$, but this goes beyond the scope of this paper.

We note that in the case of the sparse Rademacher prior, $P_{\mathrm{x}} = \frac{\rho}{2}\delta_{-1/\sqrt{\rho}} + (1-\rho)\delta_0 + \frac{\rho}{2}\delta_{+1/\sqrt{\rho}}$, we have $\lambda_c = 1$ if $\rho \ge \rho^* \approx 0.092$ and $\lambda_c < 1$ otherwise. Corollary 3 and Proposition 4 exactly pin down the regime of contiguity, thus closing the gaps in the results of Banks et al. (2017) and Perry et al. (2018). The behavior at the exact critical value $\lambda = \lambda_c$ is left open.

Weak detection. We have seen that strong detection is possible if and only if $\lambda > \lambda_c$. It is then natural to ask whether weak detection is possible below λ_c; that is, is it possible to test with accuracy *better than that of a random guess* below the reconstruction threshold? The answer is *yes*, and this is another consequence of Theorem 2. More precisely, the optimal test minimizing the risk (12) is the likelihood ratio test which rejects the null hypothesis $\boldsymbol{H}_0$ (i.e., returns "1") if $L(\boldsymbol{Y};\lambda) > 1$, and its error is

$$\mathsf{err}^*(\lambda) = \mathbb{P}_\lambda\big(L(\boldsymbol{Y};\lambda) \le 1\big) + \mathbb{P}_0\big(L(\boldsymbol{Y};\lambda) > 1\big) = 1 - D_{\mathsf{TV}}(\mathbb{P}_\lambda, \mathbb{P}_0). \tag{13}$$

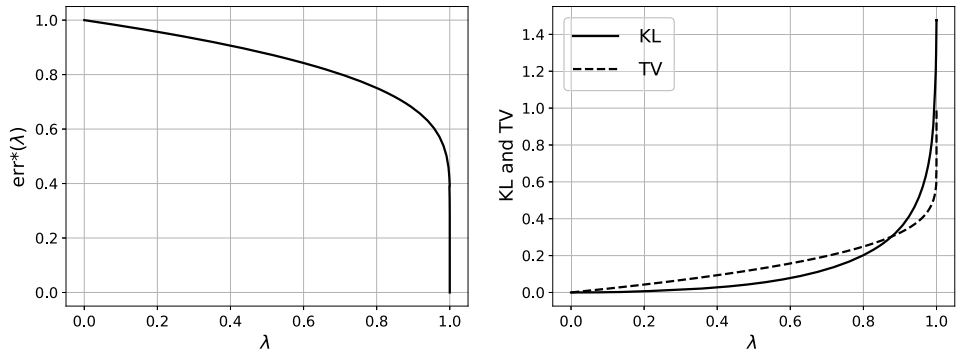

FIG. 1. *Plots of formulas* (14) *and* (15).

One can readily deduce from Theorem 2 the Type-I and Type-II errors of the likelihood ratio test. By symmetry of the means of the limiting Gaussians, the errors $\mathbb{P}_0(\log L(\boldsymbol{Y};\lambda) > 0)$ and $\mathbb{P}_\lambda(\log L(\boldsymbol{Y};\lambda) \le 0)$ converge to a common limit $\frac{1}{2}\mathsf{erfc}(\frac{\sqrt{\mu}}{2})$ for all $\lambda < \lambda_c$, where $\mu = \frac{1}{4}(-\log(1-\lambda) - \lambda)$ and $\mathsf{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,\mathrm{d}t$ is the complementary error function. Therefore, one obtains the following formula for $\mathsf{err}^*(\lambda)$ and the total variation distance between $\mathbb{P}_\lambda$ and $\mathbb{P}_0$ (plotted in Figure 1).

COROLLARY 5. *For all $\lambda < \lambda_c$ (and $\sigma = +\infty$), we have*

$$\lim_{N\to\infty} \mathsf{err}^*(\lambda) = 1 - \lim_{N\to\infty} D_{\mathsf{TV}}(\mathbb{P}_\lambda, \mathbb{P}_0) = \mathsf{erfc}\left(\frac{1}{4}\sqrt{-\log(1-\lambda) - \lambda}\right). \tag{14}$$

Moreover, the proof of Theorem 2 allows us to obtain a formula for the KL divergence between $\mathbb{P}_\lambda$ and $\mathbb{P}_0$ below the reconstruction threshold λ_c (see Figure 1).

COROLLARY 6 (of the proof). *Assume the prior P_{x} is centered, is of unit variance and has bounded support (and $\sigma = +\infty$.) Then for all $\lambda < \lambda_c$,*

$$\lim_{N\to\infty} D_{\mathsf{KL}}(\mathbb{P}_\lambda, \mathbb{P}_0) = \frac{1}{4}\big(-\log(1-\lambda) - \lambda\big). \tag{15}$$

Note that the above formulas are only valid up to λ_c. When $\lambda_c < 1$, TV and KL both witness an abrupt discontinuity at λ_c to 1 and ∞, respectively. When $\lambda_c = 1$, then the behavior is more smooth with an asymptote at 1.

4. Replicas, overlaps, Gibbs measures and Nishimori.

4.1. *Important notions.* A crucial component of the proof of our main results is the understanding of the convergence of the overlap $\boldsymbol{x}^\top \boldsymbol{x}^*/N$, where $\boldsymbol{x}$ is drawn from $\mathbb{P}_\lambda(\cdot|\boldsymbol{Y})$, to its limit $q^*(\lambda)$. By Bayes' rule, we see that

$$\mathrm{d}\mathbb{P}_\lambda(\boldsymbol{x}|\boldsymbol{Y}) = \frac{e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})}{\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})}, \tag{16}$$

where H is the Hamiltonian (recall that $\sigma = +\infty$)

$$-H(\boldsymbol{x}) := \sum_{i<j} \sqrt{\frac{\lambda}{N}} Y_{ij} x_i x_j - \frac{\lambda}{2N} x_i^2 x_j^2. \tag{17}$$

From the equations (3) and (4), it is straightforward to see that

$$f_N = \frac{1}{N}\,\mathbb{E}_{\mathbb{P}_\lambda} \log \int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}).$$

This provides another way of interpreting f_N as the expected log-partition function (or normalizing constant) of the posterior $\mathbb{P}_\lambda(\cdot|\boldsymbol{Y})$. For an integer $n \geq 1$ and $f : (\mathbb{R}^N)^{n+1} \mapsto \mathbb{R}$, we define the Gibbs average of f w.r.t. H as

$$\begin{aligned}&\langle f(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^*)\rangle \\ &\quad := \frac{\int f(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^*) \prod_{l=1}^n e^{-H(\boldsymbol{x}^{(l)})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}^{(l)})}{(\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}))^n}.\end{aligned} \tag{18}$$

This is simply the average of f with respect to $\mathbb{P}_\lambda(\cdot|\boldsymbol{Y})^n$. The variables $\boldsymbol{x}^{(l)}, l = 1, \dots, n$, are called *replicas*, and are interpreted as random variables drawn independently from the posterior. When $n = 1$, we simply write $f(\boldsymbol{x}, \boldsymbol{x}^*)$ instead of $f(\boldsymbol{x}^{(1)}, \boldsymbol{x}^*)$. Throughout this paper, we use the following notation: for $l, l' = 1, \dots, n, *$, we let

$$R_{l,l'} := \boldsymbol{x}^{(l)} \cdot \boldsymbol{x}^{(l')} = \frac{1}{N}\sum_{i=1}^N x_i^{(l)} x_i^{(l')}.$$

4.2. *The Nishimori property under* $\mathbb{P}_\lambda$. The fact that the Gibbs measure $\langle\cdot\rangle$ is a posterior distribution (16) has far-reaching consequences. A crucial implication is that the $n+1$-tuples $(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n+1)})$ and $(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^*)$ have the same law under $\mathbb{E}_{\mathbb{P}_\lambda}\langle\cdot\rangle$. To see this, let us perform the following experiment:

1. Construct $\boldsymbol{x}^* \in \mathbb{R}^N$ by independently drawing its coordinates from P_{x}.
2. Construct $\boldsymbol{Y}$ as $Y_{ij} = \sqrt{\frac{\lambda}{N}} x_i^* x_j^* + W_{ij}$, where $W_{ij} \sim \mathcal{N}(0, 1)$ are all independent for $i < j$. (Therefore, $\boldsymbol{Y}$ is distributed according to $\mathbb{P}_\lambda$.)
3. Draw $n+1$ independent random vectors $(\boldsymbol{x}^{(l)})_{l=1}^{n+1}$ from $\mathbb{P}_\lambda(\boldsymbol{x} \in \cdot|\boldsymbol{Y})$.

We have by the tower property of expectations $\mathbb{E}\,\psi(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^*) = \mathbb{E}[\mathbb{E}[\psi(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^*)|\boldsymbol{Y}]] = \mathbb{E}\,\psi(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n+1)})$ for any measurable real-valued function ψ. Therefore, the following equality of joint laws holds:

$$(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n+1)}) \overset{\mathrm{d}}{=} (\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^*). \tag{19}$$

This implies in particular that under the alternative $\mathbb{P}_\lambda$, the overlaps $R_{1,*}$ between a replica and the spike have the same distribution as the overlap $R_{1,2}$ between two replicas. The latter is a very important property of the planted model $\mathbb{P}_\lambda$, which is usually named after Nishimori (2001) in spin-glass theory. Property (19) substantially simplifies important technical arguments that are otherwise very difficult to conduct under the null. A recurring example in our context is the following: to prove the convergence of the overlap between two replicas, $\mathbb{E}\langle R_{1,2}^2\rangle \to 0$, it suffices to prove $\mathbb{E}\langle R_{1,*}^2\rangle \to 0$ since the two quantities are equal.

5. Proof of LR fluctuations. In this section, we prove Theorem 2. It suffices to prove the fluctuations under one of the hypotheses. Fluctuations under the remaining one come for free as a consequence of Le Cam's third lemma (van der Vaart (1998), Theorem 6.6). We choose to treat the planted case $\boldsymbol{Y} \sim \mathbb{P}_\lambda$. The reason is that it is easier to deal with the planted model, due to the Nishimori property (19).

5.1. *Fluctuations under* $\mathbb{P}_\lambda$. In this section, we prove Gaussian fluctuations of $\log L$ through the convergence of its characteristic function. Let $\mathrm{i}^2 = -1$ and $s \in \mathbb{R}$ be fixed. For λ and $\boldsymbol{Y} \sim \mathbb{P}_\lambda$, let

$$\phi_N(\lambda) = \mathbb{E}_{\mathbb{P}_\lambda}\big[e^{\mathrm{i}s \log L(\boldsymbol{Y};\lambda)}\big].$$

THEOREM 7. *For all $\lambda < \lambda_c$ and $s \in \mathbb{R}$, there exists a constant $K = K(\lambda, s) < \infty$ such that*

$$\big|\phi_N(\lambda) - e^{(\mathrm{i}s - s^2)\mu}\big| \le \frac{K}{\sqrt{N}},$$

where $\mu = \frac{1}{4}(-\log(1-\lambda) - \lambda)$.

The map $s \mapsto e^{(\mathrm{i}s - s^2)\mu}$ is the characteristic function of $\mathcal{N}(\mu, 2\mu)$.

LEMMA 8. *For all $\lambda \ge 0$,*

$$\phi_N'(\lambda) = \frac{\mathrm{i}s - s^2}{4}\,\mathbb{E}\big[\big(N\langle R_{1,*}^2\rangle - \langle x_N^2 x_N^{*2}\rangle\big)e^{\mathrm{i}s\log L}\big]. \tag{20}$$

PROOF. By differentiation with respect to λ, we obtain

$$\phi_N'(\lambda) = \mathrm{i}s\,\mathbb{E}\Big[\Big(\frac{\mathrm{d}}{\mathrm{d}\lambda}\log L\Big)e^{\mathrm{i}s\log L}\Big] = \mathrm{i}s\,\mathbb{E}\Big[\Big\langle -\frac{\mathrm{d}}{\mathrm{d}\lambda}H(\boldsymbol{x})\Big\rangle e^{\mathrm{i}s\log L}\Big],$$

where the Hamiltonian H is given in (17). Since $\boldsymbol{Y} \sim \mathbb{P}_\lambda$, we can write more explicitly $-H(\boldsymbol{x}) = \sum_{i<j}\sqrt{\frac{\lambda}{N}}W_{ij}x_ix_j + \frac{\lambda}{N}x_ix_jx_i^*x_j^* - \frac{\lambda}{2N}x_i^2x_j^2$. Therefore,

$$\begin{aligned}\phi_N'(\lambda) &= \mathrm{i}s\sum_{i<j}\frac{1}{2\sqrt{\lambda N}}\,\mathbb{E}\big[\langle W_{ij}x_ix_j\rangle e^{\mathrm{i}s\log L}\big] - \frac{1}{2N}\,\mathbb{E}\big[\langle x_i^2x_j^2\rangle e^{\mathrm{i}s\log L}\big]\\ &\quad + \mathrm{i}s\sum_{i<j}\frac{1}{N}\,\mathbb{E}\big[\langle x_ix_jx_i^*x_j^*\rangle e^{\mathrm{i}s\log L}\big].\end{aligned} \tag{21}$$

Now we perform Gaussian integration by parts with respect to each variable W_{ij} and obtain

$$\begin{aligned}\frac{1}{2\sqrt{\lambda N}}\,\mathbb{E}\big[\langle W_{ij}x_ix_j\rangle e^{\mathrm{i}s\log L}\big] &= \frac{1}{2N}\,\mathbb{E}\big[\langle x_i^2x_j^2\rangle e^{\mathrm{i}s\log L}\big] - \frac{1}{2N}\,\mathbb{E}\big[\langle x_ix_j\rangle^2 e^{\mathrm{i}s\log L}\big]\\ &\quad + \frac{\mathrm{i}s}{2N}\,\mathbb{E}\big[\langle x_ix_j\rangle^2 e^{\mathrm{i}s\log L}\big].\end{aligned}$$

Plugging this into (21) and rearranging, we obtain

$$\begin{aligned}\phi_N'(\lambda) &= -\frac{\mathrm{i}s + s^2}{4}\,\mathbb{E}\big[\big(N\langle R_{1,2}^2\rangle - \langle x_N^2\rangle^2\big)e^{\mathrm{i}s\log L}\big]\\ &\quad + \frac{\mathrm{i}s}{2}\,\mathbb{E}\big[\big(N\langle R_{1,*}^2\rangle - \langle x_N^2x_N^{*2}\rangle\big)e^{\mathrm{i}s\log L}\big].\end{aligned} \tag{22}$$

Since we are under the planted model $\mathbb{P}_\lambda$ and $e^{\mathrm{i}s\log L}$ depends only on $\boldsymbol{Y}$, we can use the Nishimori property (19) to replace $R_{1,2}$ and $x_N^{(1)}x_N^{(2)}$ by $R_{1,*}$ and $x_Nx_N^*$, respectively, in the first term in (22). $\square$

The derivative involves the average $\mathbb{E}[(N\langle R_{1,*}^2\rangle - \langle x_N^2x_N^{*2}\rangle)e^{\mathrm{i}s\log L}]$. A crucial step in the argument is to show that $e^{\mathrm{i}s\log L}$ and its prefactor in the above expression are asymptotically

independent, so that one can split the expectation of the product into the product of the expectations. More precisely, one should expect the quantities $N\langle R_{1,*}^2\rangle$ and $\langle x_N^2 x_N^{*2}\rangle$ to tightly concentrate about some deterministic values when $\lambda < \lambda_c$, such that the right-hand side in (20) is a multiple of $\mathbb{E}[e^{\mathrm{i}s\log L}] = \phi_N(\lambda)$. We will then be left with a simple differential equation whose solution is $s \mapsto e^{(\mathrm{i}s - s^2)\mu}$.

PROPOSITION 9. *For all $\lambda < \lambda_c$ and $s \in \mathbb{R}$, there exists $K = K(\lambda, s) < \infty$ such that*

$$\mathbb{E}\big[\big(N\langle R_{1,*}^2\rangle - \langle x_N^2 x_N^{*2}\rangle\big)e^{\mathrm{i}s\log L}\big] = \frac{\lambda}{1-\lambda}\,\mathbb{E}\big[e^{\mathrm{i}s\log L}\big] + \delta,$$

where $|\delta| \le K(s,\lambda)/\sqrt{N}$.

From here, we can prove the convergence of ϕ_N by integrating the differential equation given in Lemma 8.

PROOF OF THEOREM 7. Plugging the result of Proposition 9 into Lemma 8 yields

$$\phi_N'(\lambda) = \frac{\mathrm{i}s - s^2}{4}\,\frac{\lambda}{1-\lambda}\phi_N(\lambda) + \delta,$$

where $|\delta| \le K(s,\lambda)/\sqrt{N}$. Since $\phi_N(0) = 1$ and the primitive of $\lambda \mapsto \frac{\lambda}{1-\lambda}$ is $\lambda \mapsto -\lambda - \log(1-\lambda)$, integrating w.r.t. λ yields the result. $\square$

PROOF OF COROLLARY 6. We prove the convergence of $D_{\mathsf{KL}}(\mathbb{P}_\lambda, \mathbb{P}_0)$. By differentiation and use of the Nishimori property (19), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\mathbb{E}_{\mathbb{P}_\lambda}\log L(\boldsymbol{Y};\lambda) &= -\frac{1}{4}\,\mathbb{E}\big[\big(N\langle R_{1,2}^2\rangle - \langle x_N^2\rangle^2\big)\big] + \frac{1}{2}\,\mathbb{E}\big[\big(N\langle R_{1,*}^2\rangle - \langle x_N^2 x_N^{*2}\rangle\big)\big]\\ &= \frac{1}{4}\,\mathbb{E}\big[\big(N\langle R_{1,*}^2\rangle - \langle x_N^2 x_N^{*2}\rangle\big)\big].\end{aligned}$$

Now we use Proposition 9 with $s = 0$, and integrate w.r.t. λ to conclude. $\square$

It remains to prove Proposition 9. This will require the deployment of techniques from the theory of mean-field spin glasses.

5.2. *Sketch of proof of Proposition 9.* The idea is to show self-consistency relations among the quantities of interest. Namely, we will prove that for all $\lambda < 1$,

$$N\,\mathbb{E}\big[\langle R_{1,*}^2\rangle e^{\mathrm{i}s\log L}\big] = \frac{1}{1-\lambda}\,\mathbb{E}\big[\langle x_N^2 x_N^{*2}\rangle e^{\mathrm{i}s\log L}\big] + \delta, \tag{23}$$

and

$$\mathbb{E}\big[\langle x_N^2 x_N^{*2}\rangle e^{\mathrm{i}s\log L}\big] = \mathbb{E}\big[e^{\mathrm{i}s\log L}\big] + \delta, \tag{24}$$

where in both cases

$$|\delta| \le K(\lambda) N\,\mathbb{E}\langle |R_{1,*}|^3\rangle.$$

Next, we need to prove the convergence of the third moment of the overlap $R_{1,*}$ under $\mathbb{E}\langle\cdot\rangle$ at an optimal rate of $\mathcal{O}(1/N^{3/2})$:

THEOREM 10. *For all $\lambda < \lambda_c$, there exists a constant $K = K(\lambda) < \infty$ such that*

$$\mathbb{E}\langle R_{1,*}^4\rangle \le \frac{K}{N^2}.$$

This will allow us to conclude that $|\delta| \le K(\lambda)/\sqrt{N}$. It is interesting to note that while the self-consistent (or cavity) equations (23) and (24) hold for all $\lambda < 1$, the convergence of the overlap toward zero is only true up to λ_c.

6. Proof of asymptotic decoupling. We proceed to the proof of Proposition 9. As explained earlier, the argument is in two stages. We first prove (23) then (24).

6.1. *Preliminary bounds.* We make repeated use of interpolation arguments in our proofs. We state here a few elementary lemmas that we will invoke several times. We denote the overlaps between replicas where the last variable x_N is deleted by a superscript "$-$":

$$R^-_{l,l'} = \frac{1}{N}\sum_{i=1}^{N-1} x_i^{(l)} x_i^{(l')}.$$

Let $\{H_t : t \in [0,1]\}$ be a family of interpolating Hamiltonians. We let $\langle\cdot\rangle_t$ denote the corresponding Gibbs average, similar to (18). Following Talagrand's notation, we write

$$\nu_t(f) := \mathbb{E}\langle f\rangle_t,$$

for a generic function f of n replicas $\boldsymbol{x}^{(l)}$, $l = 1, \dots, n$. We abbreviate ν_1 by ν. The main tool we use is the following interpolation that isolates the last variable x_N from the rest of the system:

$$\begin{aligned} -H_t(\boldsymbol{x}) := &\sum_{1\le i<j\le N-1} \sqrt{\frac{\lambda}{N}} W_{ij} x_i x_j + \frac{\lambda}{N} x_i x_i^* x_j x_j^* - \frac{\lambda}{2N} x_i^2 x_j^2 \\ &+ \sum_{i=1}^{N-1} \sqrt{\frac{\lambda t}{N}} W_{iN} x_i x_N + \frac{\lambda t}{N} x_i x_i^* x_N x_N^* - \frac{\lambda t}{2N} x_i^2 x_N^2. \end{aligned} \tag{25}$$

At $t = 1$, we have $H_t = H$, and at $t = 0$ the variable x_N decouples from the rest of the variables. Moreover, the Nishimori property (19) is still valid under $\langle\cdot\rangle_t$: the last column of $\boldsymbol{Y}$ simply becomes $(\sqrt{\frac{\lambda t}{N}} x_i^* x_N^* + W_{iN})_{i=1}^{N-1}$.

LEMMA 11. *Let f be a function of n replicas $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}^*$. Then*

$$\begin{aligned} \nu_t'(f) = &\frac{\lambda}{2} \sum_{1\le l\ne l'\le n} \nu_t\big(R^-_{l,l'} y^{(l)} y^{(l')} f\big) - \lambda n \sum_{l=1}^{n} \nu_t\big(R^-_{l,n+1} y^{(l)} y^{(n+1)} f\big) \\ &+ \lambda n \sum_{l=1}^{n} \nu_t\big(R^-_{l,*} y^{(l)} y^* f\big) - \lambda n \nu_s\big(R^-_{n+1,*} y^{(n+1)} y^* f\big) \\ &+ \lambda \frac{n(n+1)}{2} \nu_t\big(R^-_{n+1,n+2} y^{(n+1)} y^{(n+2)} f\big), \end{aligned}$$

where we have written $y = x_N$.

PROOF. The computation relies on Gaussian integration by parts; see Talagrand ((2011a), Lemma 1.6.3), for the details of a similar computation. □

LEMMA 12. *If f is a bounded nonnegative function, then for all $t \in [0,1]$,*

$$\nu_t(f) \le K(\lambda, n)\nu(f).$$

PROOF. Since the variables and the overlaps are all bounded, using Lemma 11 we have for all $t \in [0, 1]$,

$$\left|\nu_t'(f)\right| \le K(\lambda, n)\nu_t(f).$$

Then we conclude using Grönwall's lemma. □

6.2. *The cavity method*. In its essence, the cavity method amounts to removing one variable from the system—in a manner akin to leave-one-out methods in statistics—and analyzing the influence of the remaining variables on the variable that has been removed. It was initially introduced to solve certain models of spin glasses (Mézard, Parisi and Virasoro (1987)), and was developed into a rigorous probabilistic theory by Talagrand (2011a, 2011b). To make use of the cavity method, we isolate the Nth variable from the rest (without loss of generality, by symmetry among the variables x_i) and compute

$$\mathbb{E}\big[\big(N\langle R_{1,*}^2\rangle - \langle x_N^2 x_N^{*2}\rangle\big)e^{\mathrm{i}s\log L}\big] = N\,\mathbb{E}\big[\langle x_N x_N^* R_{1,*}^-\rangle e^{\mathrm{i}s\log L}\big].$$

Let

$$X(t) := \exp\bigg(\mathrm{i}s\log\int e^{-H_t(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})\bigg),$$

where H_t is defined in (25). Note that we have $X(1) = e^{\mathrm{i}s\log L}$. We now consider the interpolative function

$$\varphi(t) := N\,\mathbb{E}\big[\langle x_N x_N^* R_{1,*}^-\rangle_t X(t)\big].$$

Our strategy is approximate $\varphi(1)$ by $\varphi(0) + \varphi'(0)$ via a Taylor expansion, which requires is to control the second derivative φ''. Notice that since the last variable decouples from the rest of the system at $t = 0$, we have

$$\begin{aligned}\varphi(0) &= N\,\mathbb{E}\big[\langle x_N x_N^*\rangle_0\big]\cdot\mathbb{E}\big[\langle R_{1,*}^-\rangle_0 X(0)\big]\\ &= N\,\mathbb{E}_{P_{\mathrm{x}}}[X]^2\cdot\mathbb{E}\big[\langle R_{1,*}^-\rangle_0 X(0)\big] = 0.\end{aligned}$$

The latter equality holds because P_{x} is centered. Next, a bit of algebra (similar to Lemma 11) shows that the derivative $\varphi'(t)$ is a linear combination of terms of the form

$$\lambda N\,\mathbb{E}\big[\langle x_N x_N^* x_N^{(a)} x_N^{(b)} R_{1,*}^- R_{a,b}^-\rangle_t X(t)\big], \tag{26}$$

where $(a, b) \in \{(1, *), (2, *), (1, 2), (2, 3)\}$. We see that at $t = 0$, if the above expression involves a variable $x_N^{(a)}$ of degree 1 then this term vanishes. Therefore, the only remaining term is the one where $(a, b) = (1, *)$. Therefore,

$$\begin{aligned}\varphi'(0) &= \lambda N\,\mathbb{E}\big[\langle x_N^2 x_N^{*2}\rangle_0\big]\cdot\mathbb{E}\big[\langle (R_{1,*}^-)^2\rangle_0 X(0)\big]\\ &= \lambda N\,\mathbb{E}_{P_{\mathrm{x}}}\big[X^2\big]^2\cdot\mathbb{E}\big[\langle (R_{1,*}^-)^2\rangle_0 X(0)\big]\\ &= \lambda N\,\mathbb{E}\big[\langle (R_{1,*}^-)^2\rangle_0 X(0)\big].\end{aligned} \tag{27}$$

The last equality holds because P_{x} has unit variance. Now we turn to $\varphi''(t)$. Taking another derivative generates monomials of degree three in the overlaps and the last variable, so $\varphi''(t)$ is a linear combination of terms

$$\lambda^2 N\,\mathbb{E}\big[\langle x_N x_N^* x_N^{(a)} x_N^{(b)} x_N^{(c)} x_N^{(d)} R_{1,2}^- R_{a,b}^- R_{c,d}^-\rangle_t X(t)\big], \tag{28}$$

where (a, b, c, d) range over a finite set of combinations. Our goal is to bound the second derivative independently of t, so that we are able to use Taylor's approximation

$$\left|\varphi(1) - \varphi(0) - \varphi'(0)\right| \le \sup_{0\le t\le 1}\left|\varphi''(t)\right|. \tag{29}$$

Since P_{x} has bounded support and $|X(t)| = 1$, Hölder's inequality and the Nishimori property imply that (28) is bounded in modulus by

$$\lambda N K\, \mathbb{E}[\langle |R_{1,2}^- R_{a,b}^- R_{c,d}^-| \rangle_t] \le \lambda N K\, \mathbb{E}[\langle |R_{1,*}^-|^3 \rangle_t]^{1/3}.$$

Using Lemma 12 and the convergence of the fourth moment, Theorem 10, we have

$$\mathbb{E}\langle |R_{1,*}^-|^3 \rangle_t \le K(\lambda)\, \mathbb{E}\langle (R_{1,*}^-)^4 \rangle^{3/4} \le \frac{K(\lambda)}{N^{3/2}}.$$

Therefore, by the above estimates we have

$$\sup_{0 \le t \le 1} |\varphi''(t)| \le \frac{K(\lambda)}{\sqrt{N}}. \tag{30}$$

Now, our next goal is to prove

$$|\varphi'(0) - \lambda N\, \mathbb{E}[\langle R_{1,*}^2 \rangle e^{\mathrm{i}s \log L}]| \le \frac{K(\lambda)}{\sqrt{N}}. \tag{31}$$

We consider the function

$$\psi(t) := \lambda N\, \mathbb{E}[\langle (R_{1,*}^-)^2 \rangle_t X(t)].$$

Observe that (27) tells us that $\psi(0) = \varphi'(0)$. On the other hand,

$$|\psi(1) - \lambda N\, \mathbb{E}[\langle R_{1,*}^2 \rangle e^{\mathrm{i}s \log L}]| \le 2\lambda\, \mathbb{E}\langle |R_{1,*}^- x_N x_N^*| \rangle + \frac{\lambda}{N}\, \mathbb{E}\langle (x_N x_N^*)^2 \rangle.$$

By boundedness of the prior, the first term in the RHS is bounded by

$$K(\lambda)\, \mathbb{E}\langle |R_{1,*}^-| \rangle \le K(\lambda)/\sqrt{N},$$

and the second term is bounded by $K(\lambda)/N$. So it suffices to show that

$$\sup_{0 \le t \le 1} |\psi'(t)| \le \frac{K(\lambda)}{\sqrt{N}}.$$

Similar to φ, the derivative of ψ is a sum of terms of the form

$$\lambda^2 N\, \mathbb{E}[\langle x_N^{(a)} x_N^{(b)} (R_{1,*}^-)^2 R_{a,b}^- \rangle_t X(t)].$$

It is clear that the same method used to bound φ'' (the generic term of which is (28)) also works in this case, so we obtain the desired bound on ψ'. Finally, using (29), (30) and (31), we obtain

$$N\, \mathbb{E}[\langle R_{1,*}^2 \rangle e^{\mathrm{i}s \log L}] - \mathbb{E}[\langle x_N^2 x_N^{*2} \rangle e^{\mathrm{i}s \log L}] = \lambda N\, \mathbb{E}[\langle R_{1,*}^2 \rangle e^{\mathrm{i}s \log L}] + \delta,$$

where $|\delta| \le K(\lambda)/\sqrt{N}$. This is equivalent to (23) and closes the first stage of the argument. Now we need to show that

$$\mathbb{E}[\langle x_N^2 x_N^{*2} \rangle e^{\mathrm{i}s \log L}] = \mathbb{E}[e^{\mathrm{i}s \log L}] + \delta.$$

We similarly consider the function $\psi(t) = \mathbb{E}[\langle x_N^2 x_N^{*2} \rangle_t X(t)]$. We have

$$\psi(0) = \mathbb{E}[\langle x_N^2 x_N^{*2} \rangle_0] \cdot \mathbb{E}[X(0)] = \mathbb{E}_{P_{\mathrm{x}}}[X^2]^2 \cdot \mathbb{E}[X(0)] = \mathbb{E}[X(0)].$$

The derivative of ψ is a sum of term of the form

$$\lambda\, \mathbb{E}[\langle x_N^2 x_N^{*2} x_N^{(a)} x_N^{(b)} R_{a,b}^- \rangle_t X(t)].$$

By our earlier argument, $|\psi'(t)| \le K(\lambda)/\sqrt{N}$ for all t, so that

$$|\psi(1) - \mathbb{E}[X(0)]| \le \frac{K(\lambda)}{\sqrt{N}}.$$

It remains to show that $|\mathbb{E}[X(0)] - \mathbb{E}[X(1)]| \le K/\sqrt{N}$, and this is done in exactly the same way: by bounding the derivative of $t \mapsto \mathbb{E}[X(t)]$. This yields (24) and concludes the proof.

7. Overlap convergence. In this section, we prove Theorem 10 on the convergence of the overlaps to zero under $\mathbb{P}_\lambda$, and below λ_c. At a high level, we will first prove that the overlap $R_{1,*}$ converges in probability to zero under $\mathbb{E}\langle\cdot\rangle$: for all $\epsilon > 0$,

$$\mathbb{E}\langle \mathbb{1}\{|R_{1,*}| \geq \epsilon\}\rangle \leq K e^{-cN}. \tag{32}$$

This will be achieved via two interpolation bounds combined with concentration of measure. The way the argument works is roughly as follows: for a fixed q we have

$$\begin{aligned}\mathbb{E}\langle \mathbb{1}\{R_{1,*} \simeq q\}\rangle &= \mathbb{E}\,\frac{\int \mathbb{1}\{R_{1,*} \simeq q\} e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})}{\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})} \\ &= \mathbb{E}\,\frac{\exp\{N \times \frac{1}{N}\log \int \mathbb{1}\{R_{1,*} \simeq q\} e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})\}}{\exp\{N \times \frac{1}{N}\log \int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})\}}.\end{aligned}$$

We invoke concentration-of-measure arguments to show that the logarithmic terms in the numerator and the denominator are close to their expectations, hence we obtain

$$\mathbb{E}\langle \mathbb{1}\{R_{1,*} \simeq q\}\rangle \simeq \exp\{N(f_N(q) - f_N)\},$$

where $f_N(q) = \frac{1}{N}\mathbb{E}\log \int \mathbb{1}\{R_{1,*} \simeq q\} e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})$ and f_N is the unconstrained free energy (with no indicator). It is now apparent that $R_{1,*}$ is exponentially unlikely to take values q such that $f_N(q) < f_N$. It remains to lower bound f_N and upper bound $f_N(q)$ by quantities that preserve a strict inequality for all $q \neq 0$. These quantities will naturally be the replica-symmetric formula $\phi_{\mathsf{RS}}(\lambda)$ and the replica-symmetric potential $F(\lambda, q)$ respectively, and the proof relies on Guerra's interpolation method.

Next, this convergence in probability is boosted to a statement of convergence of the second moment: $\mathbb{E}\langle R_{1,*}^2\rangle \leq \frac{K}{N}$, which is in turn boosted to a statement of convergence of the fourth moment: $\mathbb{E}\langle R_{1,*}^4\rangle \leq \frac{K}{N^2}$. The apparent recursive nature of this argument is a feature of the cavity method: one can control higher-order quantities once one knows how to control low-order ones and control certain error terms. We now present the interpolation bounds and then prove (32). The cavity arguments which allow us to convert this to convergence of moments are presented in the Supplementary Material (El Alaoui, Krzakala and Jordan (2020)), since they are very similar to the arguments already presented in Section 6.

7.1. *Guerra's interpolation bound.* We present the interpolation method of Guerra (2001); a main tool in our arguments.

PROPOSITION 13. *Recall $f_N = \frac{1}{N}\mathbb{E}_{\mathbb{P}_\lambda}\log L(\boldsymbol{Y};\lambda)$. For all $\lambda \geq 0$, there exist $K > 0$ such that*

$$f_N \geq \sup_{q\geq 0} F(\lambda, q) - \frac{K}{N} = \phi_{\mathsf{RS}}(\lambda) - \frac{K}{N}.$$

PROOF. Consider the family of interpolating Hamiltonians,

$$\begin{aligned}-H_t(\boldsymbol{x}) := &\sum_{i<j}\sqrt{\frac{t\lambda}{N}}W_{ij}x_ix_j + \frac{t\lambda}{N}x_ix_i^*x_jx_j^* - \frac{t\lambda}{2N}x_i^2x_j^2 \\ &+ \sum_{i=1}^N \sqrt{(1-t)r}z_ix_i + (1-t)rx_ix_i^* - \frac{(1-t)r}{2}x_i^2,\end{aligned} \tag{33}$$

where the z_i's are i.i.d. standard Gaussian r.v.'s independent of everything else, and $r = \lambda q^*(\lambda)$. We similarly define the Gibbs average $\langle\cdot\rangle_t$ as in (18) where H is replaced by H_t.

Note that the Nishimori property (19) is preserved under $\langle\cdot\rangle_t$ for all $t \in [0, 1]$. Indeed, the interpolation is constructed in such a way that $\langle\cdot\rangle_t$ is the posterior distribution of the signal $\boldsymbol{x}^*$ given the augmented set of observations

$$\begin{cases} Y_{ij} = \sqrt{\dfrac{t\lambda}{N}} x_i^* x_j^* + W_{ij}, & 1 \le i < j \le N, \\ y_i = \sqrt{(1-t)r} x_i^* + z_i, & 1 \le i \le N, \end{cases} \tag{34}$$

which can be interpreted as having side information about $\boldsymbol{x}^*$ from a scalar Gaussian channel with $r = \lambda q^*(\lambda)$. Now we consider the interpolating free energy

$$\varphi(t) := \frac{1}{N} \mathbb{E} \log \int e^{-H_t(\boldsymbol{x})} \, \mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}). \tag{35}$$

We see that $\varphi(1) = f_N$ and $\varphi(0) = \psi(\lambda q)$. This function is differentiable in t, and by differentiation, we have

$$\begin{aligned} \varphi'(t) &= \frac{1}{N} \mathbb{E}\left\langle -\frac{\mathrm{d}H_t(\boldsymbol{x})}{\mathrm{d}t} \right\rangle_t \\ &= \frac{1}{N} \mathbb{E}\left\langle -\frac{\lambda}{2N} \sum_{i<j} x_i^2 x_j^2 + \frac{1}{2}\sqrt{\frac{\lambda}{tN}} \sum_{i<j} W_{ij} x_i x_j + \frac{\lambda}{N} \sum_{i<j} x_i x_i^* x_j x_j^* \right\rangle_t \\ &\quad + \frac{1}{N} \mathbb{E}\left\langle \frac{\lambda q}{2} \sum_{i=1}^N x_i^2 - \frac{1}{2}\sqrt{\frac{\lambda q}{1-t}} \sum_{i=1}^N z_i x_i - \lambda q \sum_{i=1}^N x_i x_i^* \right\rangle_t. \end{aligned}$$

Now we use Gaussian integration by parts to eliminate the variables W_{ij} and z_i. The details of this computation are explained extensively in many sources (see, e.g., Krzakala, Xu and Zdeborová (2016), Lelarge and Miolane (2019), Talagrand (2011a)). We get

$$\begin{aligned} \varphi'(t) &= -\frac{\lambda}{2N^2} \mathbb{E}\left\langle \sum_{i<j} x_i^{(1)} x_j^{(1)} x_i^{(2)} x_j^{(2)} \right\rangle_t + \frac{\lambda}{N^2} \mathbb{E}\left\langle \sum_{i<j} x_i x_i^* x_j x_j^* \right\rangle_t \\ &\quad + \frac{\lambda q}{2N} \mathbb{E}\left\langle \sum_{i=1}^N x_i^{(1)} x_i^{(2)} \right\rangle_t - \frac{\lambda q}{N} \mathbb{E}\left\langle \sum_{i=1}^N x_i x_i^* \right\rangle_t. \end{aligned}$$

Completing the squares yields

$$\begin{aligned} \varphi'(t) &= -\frac{\lambda}{4} \mathbb{E}\langle (\boldsymbol{x}^{(1)} \cdot \boldsymbol{x}^{(2)} - q)^2 \rangle_t + \frac{\lambda}{4} q^2 + \frac{\lambda}{4N^2} \sum_{i=1}^N \mathbb{E}\langle {x_i^{(1)}}^2 {x_i^{(2)}}^2 \rangle_t \\ &\quad + \frac{\lambda}{2} \mathbb{E}\langle (\boldsymbol{x} \cdot \boldsymbol{x}^* - q)^2 \rangle_t - \frac{\lambda}{2} q^2 - \frac{\lambda}{2N^2} \sum_{i=1}^N \mathbb{E}\langle {x_i}^2 {x_i^*}^2 \rangle_t. \end{aligned} \tag{36}$$

The first line in the above expression involves overlaps between two independent replicas, while the second one involves overlaps between one replica and the planted solution. Using the Nishimori property, the derivative of φ can be written as

$$\varphi'(t) = \frac{\lambda}{4} \mathbb{E}\langle (R_{1,*} - q)^2 \rangle_t - \frac{\lambda}{4} q^2 - \frac{\lambda}{4N} \mathbb{E}\langle {x_N}^2 {x_N^*}^2 \rangle_t. \tag{37}$$

The last term follows by symmetry between variables. We finish the argument by noting that $\mathbb{E}\langle (R_{1,*} - q)^2 \rangle_t \ge 0$, and the product ${x_N}^2 {x_N^*}^2$ is bounded. We then integrate with respect to time to obtain the result. $\square$

7.2. *Guerra's interpolation at fixed overlap.* Let us first introduce the so-called *Franz–Parisi (FP) potential* (Franz and Parisi (1995, 1998)). For $\boldsymbol{x}^* \in \mathbb{R}^N$ fixed, $m \in \mathbb{R} \setminus \{0\}$ and $\epsilon > 0$ define the set

$$A = \begin{cases} R_{1,*} \in [m, m+\epsilon) & \text{if } m > 0, \\ R_{1,*} \in (m-\epsilon, m] & \text{if } m < 0. \end{cases}$$

Now define the FP potential as

$$\Phi_\epsilon(m, \boldsymbol{x}^*) := \frac{1}{N} \mathbb{E}_{\boldsymbol{W}} \log \int \mathbb{1}\{\boldsymbol{x} \in A\} e^{-H(\boldsymbol{x})} \, \mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}), \tag{38}$$

where the expectation is only over the Gaussian disorder $\boldsymbol{W}$. This is the free energy of a subsystem of configurations having an overlap close to a fixed value m with the planted signal $\boldsymbol{x}^*$.

For $r \geq 0$ and $s \in \mathbb{R}$, we let

$$\widehat{\psi}(r, s) := \mathbb{E}_z \log \int \exp\left(\sqrt{r} z x + s x - \frac{r}{2} x^2\right) \mathrm{d}P_{\mathrm{x}}(x) \tag{39}$$

and

$$\begin{aligned} \overline{\psi}(r, s) &:= \mathbb{E}_{x^*} \widehat{\psi}(r, s x^*) \\ &= \mathbb{E}_{x^*, z} \log \int \exp\left(\sqrt{r} z x + s x x^* - \frac{r}{2} x^2\right) \mathrm{d}P_{\mathrm{x}}(x). \end{aligned} \tag{40}$$

We see that $\overline{\psi}(r, r) = \psi(r)$, but unlike ψ, the function $\overline{\psi}$ does not have an interpretation as the KL between two distributions. The next lemma states a key property of this function that will be useful later on (see the Supplementary Material (El Alaoui, Krzakala and Jordan (2020)) for the proof).

LEMMA 14. *For all $r \geq 0$, $\overline{\psi}(r, -r) \leq \overline{\psi}(r, r)$.*

Additionally, for $\boldsymbol{x}^* \in \mathbb{R}^N$ fixed, we define the function

$$\widehat{F}(\lambda, m, q) := \frac{1}{N} \sum_{i=1}^N \widehat{\psi}(\lambda q, \lambda m x_i^*) - \frac{\lambda}{2} m^2 + \frac{\lambda}{4} q^2.$$

Recall that $\mathbb{E}_{x^*} \widehat{F}(\lambda, q, q)$ is the RS potential $F(\lambda, q)$ from (6).

PROPOSITION 15. *Fix $m \in \mathbb{R}$, $\epsilon > 0$ and $\lambda \geq 0$. There exist constants $K = K(\lambda) > 0$ such that*

$$\Phi_\epsilon(m; \boldsymbol{x}^*) \leq \widehat{F}(\lambda, |m|, m) + \frac{\lambda \epsilon^2}{2} + \frac{K}{N}.$$

PROOF. To obtain a bound on $\Phi_\epsilon(m; \boldsymbol{x}^*)$ we use the interpolation method with Hamiltonian

$$\begin{aligned} -H_t(\boldsymbol{x}) := &\sum_{i<j} \sqrt{\frac{t\lambda}{N}} W_{ij} x_i x_j + \frac{t\lambda}{N} x_i x_i^* x_j x_j^* - \frac{t\lambda}{2N} x_i^2 x_j^2 \\ &+ \sum_{i=1}^N \sqrt{(1-t)\lambda|m|} z_i x_i + (1-t)\lambda m x_i x_i^* - \frac{(1-t)\lambda|m|}{2} x_i^2. \end{aligned}$$

by varying $t \in [0, 1]$. The r.v.'s W, z are all i.i.d. standard Gaussians independent of everything else. We define

$$\varphi(t) := \frac{1}{N} \mathbb{E}_{W,z} \log \int \mathbb{1}\{\boldsymbol{x} \in A\} e^{-H_t(\boldsymbol{x})} \, \mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}).$$

We compute the derivative w.r.t. t, and use Gaussian integration by prts to obtain

$$\begin{aligned}\varphi'(t) = &-\frac{\lambda}{4} \mathbb{E}\langle (R_{1,2} - |m|)^2 \rangle_t + \frac{\lambda t}{4} |m|^2 + \frac{\lambda}{4N^2} \sum_{i=1}^N \mathbb{E}\langle {x_i^{(1)}}^2 {x_i^{(2)}}^2 \rangle_t \\ &+ \frac{\lambda}{2} \mathbb{E}\langle (R_{1,*} - m)^2 \rangle_t - \frac{\lambda}{2} m^2 - \frac{\lambda}{2N^2} \sum_{i=1}^N \mathbb{E}\langle {x_i}^2 {x_i^*}^2 \rangle_t,\end{aligned}$$

where $\langle \cdot \rangle_t$ is the Gibbs average w.r.t. the Hamiltonian $-H_t(\boldsymbol{x}) + \log \mathbb{1}\{\boldsymbol{x} \in A\}$. A few things now happen. Notice that the planted term (first term in the second line) is trivially smaller than $\lambda \epsilon^2/2$ due to the overlap restriction. Moreover, the last terms in both lines are of order $1/N$ since the variables x_i are bounded. The first term in the first line, which involves the overlap between two replicas, is more challenging. What makes this term difficult to control is that the Gibbs measure $\langle \cdot \rangle_t$ no longer satisfies the Nishimori property due to the overlap restriction, so the overlap between two replicas no longer has the same distribution as the overlap of one replica with the planted spike. Fortunately, this term is always nonpositive so we can ignore it altogether and obtain an upper bound:

$$\varphi'(t) \le -\frac{\lambda}{4} m^2 + \frac{\lambda \epsilon^2}{2} + \frac{\lambda K}{N}.$$

Integrating over t, we get

$$\Phi_\epsilon(m; \boldsymbol{x}^*) \le \varphi(0) - \frac{\lambda}{4} m^2 + \frac{\lambda \epsilon^2}{2} + \frac{\lambda K}{N}.$$

Finally, by dropping the indicator, we have

$$\begin{aligned}\varphi(0) &= \frac{1}{N} \mathbb{E}_z \log \int \mathbb{1}\{\boldsymbol{x} \in A\} e^{-H_0(\boldsymbol{x})} \, \mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) \\ &\le \frac{1}{N} \mathbb{E}_z \log \int e^{-H_0(\boldsymbol{x})} \, \mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_z \log \int \exp\left(\sqrt{\lambda |m|} z x_i + \lambda m x x_i^* - \frac{\lambda |m|}{2} x^2 \right) \mathrm{d}P_{\mathrm{x}}(x) \\ &= \frac{1}{N} \sum_{i=1}^N \widehat{\psi}(\lambda |m|, \lambda m x_i^*).\end{aligned}$$

□

7.3. *Convergence in probability of the overlaps.* As explained earlier, Propositions 13 and 15 imply convergence in probability of the overlaps.

PROPOSITION 16. *For all $\lambda < \lambda_c$ and $\epsilon > 0$, there exist constants $K = K(\lambda, \epsilon) \ge 0$, $c = c(\lambda, \epsilon, P_{\mathrm{x}}) \ge 0$ such that*

$$\mathbb{E}\langle \mathbb{1}\{|R_{1,*}| \ge \epsilon\} \rangle \le K e^{-cN}.$$

To prove the above proposition, we first show that the partition function of the model enjoys sub-Gaussian concentration on a logarithmic scale. This is an elementary consequence of two classical concentration-of-measure results: concentration of Lipschitz functions of Gaussian random variables, and concentration of *convex* Lipschitz functions of *bounded* random variables.

LEMMA 17. *Fix $\boldsymbol{x}^* \in \mathbb{R}^N$ and let A be a Borel subset of $\mathbb{R}^N$. Define the random variable*

$$Z := \int_A e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}),$$

where the randomness comes from the Gaussian disorder $\boldsymbol{W}$. There exists a constant $K > 0$ depending on λ and P_{x} such that for all $u \geq 0$,

$$\Pr\big(|\log Z - \mathbb{E}\log Z| \geq Nu\big) \leq 2e^{-Nu^2/K}.$$

PROOF. We notice that the map $\boldsymbol{W} \mapsto \frac{1}{N}\log Z$ is Lipschitz with constant $K\sqrt{\frac{\lambda}{N}}$ for every $\boldsymbol{x}^* \in \mathbb{R}^N$. Then we invoke the Borell–Tsirelson–Ibragimov–Sudakov inequality of concentration of Lipschitz functions of Gaussian r.v.'s. See Boucheron, Lugosi and Massart (2013). □

LEMMA 18. *Define the random variable*

$$\mathtt{f} := \frac{1}{N}\mathbb{E}_{\boldsymbol{W}}\log\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}),$$

where the randomness comes from the planted vector $\boldsymbol{x}^*$. There exist a constant $K > 0$ depending on λ and P_{x} such that for all $u \geq 0$,

$$\Pr\big(|\mathtt{f} - \mathbb{E}\,\mathtt{f}| \geq u\big) \leq 2e^{-Nu^2/K}.$$

PROOF. We notice that the map $\boldsymbol{x}^* \mapsto \mathtt{f}$ is Lipschitz with constant $K\frac{\lambda}{\sqrt{N}}$ and convex. Moreover, the coordinates x_i^* are bounded. We then invoke Talagrand's inequality on the concentration of convex Lipschitz functions of bounded r.v.'s. See Boucheron, Lugosi and Massart (2013). □

LEMMA 19. *There exists a constant $K > 0$ depending on λ, m and P_{x} such that for all $u \geq 0$,*

$$\Pr\left(\left|\sum_{i=1}^N \widehat{\psi}(\lambda|m|, \lambda m x_i^*) - \overline{\psi}(\lambda|m|, \lambda m)\right| \geq Nu\right) \leq 2e^{-Nu^2/K}.$$

PROOF. Since $|\partial_s\widehat{\psi}(r, sx^*)| \leq K^2$, $|\partial_r\widehat{\psi}(r, sx^*)| \leq K^2/2$ and $\widehat{\psi}(0,0) = 0$, where K is a bound on the radius of the support of $P_{\boldsymbol{x}}$, we have $|\widehat{\psi}(r, sx^*)| \leq K^2(r/2 + s)$. The claim now follows from Hoeffding's inequality. □

PROOF OF PROPOSITION 16. For $\epsilon, \epsilon' > 0$, we can write the decomposition

$$\begin{aligned}\mathbb{E}\langle\mathbb{1}\{|R_{1,*}| \geq \epsilon\}\rangle &= \sum_{l\geq 0}\mathbb{E}\langle\mathbb{1}\{R_{1,*} - \epsilon \in [l\epsilon', (l+1)\epsilon')\}\rangle \\ &\quad + \sum_{l\geq 0}\mathbb{E}\langle\mathbb{1}\{-R_{1,*} - \epsilon \in [l\epsilon', (l+1)\epsilon')\}\rangle,\end{aligned}$$

where the integer index l ranges over a finite set of size $\leq K/\epsilon'$ since the prior P_{x} has bounded support. We will only treat the first sum in the above expression since the argument extends trivially to the second sum.

Let $A = \{R_{1,*} - \epsilon \in [l\epsilon', (l+1)\epsilon')\}$ and write

$$\mathbb{E}\langle \mathbb{1}\{\boldsymbol{x} \in A\}\rangle = \mathbb{E}_{\boldsymbol{x}^*}\,\mathbb{E}_{\boldsymbol{W}}\left[\frac{\int_A e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})}{\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x})}\right]. \tag{41}$$

By virtue of Lemma 17, the two quantities in this fraction enjoy sub-Gaussian concentration on a logarithmic scale over the Gaussian disorder. For any given l and $u \geq 0$, we simultaneously have

$$\frac{1}{N}\log\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) \geq \frac{1}{N}\,\mathbb{E}_{\boldsymbol{W}}\log\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) - u,$$

and

$$\begin{aligned}\frac{1}{N}\log\int_A e^{-H_t(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) &\leq \frac{1}{N}\,\mathbb{E}_{\boldsymbol{W}}\log\int_A e^{-H_t(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) + u\\ &= \Phi_{\epsilon'}(\epsilon + l\epsilon'; \boldsymbol{x}^*) + u,\end{aligned}$$

with probability at least $1 - 2e^{-Nu^2/K}$. On the complement of this event, we simply bound the fraction in (41) by 1. Combining the above bounds, we obtain

$$\mathbb{E}\langle \mathbb{1}\{\boldsymbol{x} \in A\}\rangle \leq 2e^{-Nu^2/K} + \mathbb{E}_{\boldsymbol{x}^*}\big[e^{N(\Delta+2u)}\big],$$

where

$$\Delta = \Phi_{\epsilon'}(m; \boldsymbol{x}^*) - \frac{1}{N}\,\mathbb{E}_{\boldsymbol{W}}\log\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}),$$

with $m = \epsilon + l\epsilon'$. By Proposition 15, $\Phi_{\epsilon'}$ is upper bounded by a quantity that concentrates over the randomness of $\boldsymbol{x}^*$. We use Lemma 18 and Lemma 19 in the same way we used Lemma 17: for $u' \geq 0$, we simultaneously have

$$\Phi_{\epsilon'}(m; \boldsymbol{x}^*) \leq F(\lambda, |m|, m) + \frac{\lambda\epsilon^2}{2} + \frac{\lambda K}{N} + u',$$

and

$$\frac{1}{N}\,\mathbb{E}_{\boldsymbol{W}}\log\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) \geq f_N - u',$$

with probability at least $1 - 4e^{-Nu'^2/K}$, where

$$f_N = \mathbb{E}_{\boldsymbol{W},\boldsymbol{x}^*}\log\int e^{-H(\boldsymbol{x})}\,\mathrm{d}P_{\mathrm{x}}^N(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}_\lambda}\log L(\boldsymbol{Y}; \lambda).$$

Moreover, by Lemma 14, we have $F(\lambda, |m|, m) \leq F(\lambda, |m|, |m|) \equiv F(\lambda, m)$. Therefore,

$$\mathbb{E}_{\boldsymbol{x}^*}\big[e^{N\Delta}\big] \leq \exp\big(N\big(F(\lambda, |m|) - f_N + 2u'\big)\big) + 4e^{-Nu'^2/K}.$$

The second term is obtained by considering the low-probability complement event and noting that $\Delta \leq 0$. Now, by Proposition 13, $f_N \geq \sup_q F(\lambda, q) - K/N$. When $\lambda < \lambda_c$, $q = 0$ is the unique maximizer of the RS potential. Therefore, $F(\lambda, |m|) - f_N < -c(\epsilon) < 0$ for all $|m| > \epsilon$. We obtain

$$\mathbb{E}\langle \mathbb{1}\{\boldsymbol{x} \in A\}\rangle \leq 2e^{-Nu^2/K} + 4e^{-Nu'^2/K + 2Nu} + e^{N(-c(\epsilon)+2u+2u')}.$$

Finally, adjusting the parameters u, u' yields the desired result (e.g., $u' = c(\epsilon)/3$ and $u = c(\epsilon)^2/36 \wedge c(\epsilon)/9$). $\square$

PROOF OF PROPOSITION 4. Here, we prove possibility of strong detection above λ_c. From Proposition 13, we know that $\underline{\lim} \frac{1}{N} \mathbb{E}_{\mathbb{P}_\lambda} \log L \geq \phi_{\mathsf{RS}}(\lambda) > 0$ for $\lambda > \lambda_c$. On the other hand, $\mathbb{E}_{\mathbb{P}_0} \log L \leq 0$ by Jensen's inequality. Now it remains to argue that $\frac{1}{N} \log L$ concentrates about its expectation under both $\mathbb{P}_\lambda$ and $\mathbb{P}_0$. This is a consequence of Lemmas 17 and 18: we have, for all $u \geq 0$,

$$\mathbb{P}_\lambda(\log L - \mathbb{E}_{\mathbb{P}_\lambda} \log L \leq -Nu) \vee \mathbb{P}_0(\log L - \mathbb{E}_{\mathbb{P}_0} \log L \geq Nu) \leq 4e^{-Nu^2/K}.$$

This concludes the proof. (Note also that the tail decays fact enough to insure almost-sure convergence via the Borel–Cantelli lemma.) $\square$

8. When the diagonal is not discarded. When the variance of the diagonal noise entries W_{ii} is kept finite, one has to keep track of the contribution of the diagonal part $d(\boldsymbol{x}) = \frac{1}{\sigma^2} \sum_{i=1}^N \sqrt{\frac{\lambda}{N}} Y_{ii} x_i^2 - \frac{\lambda}{2N} x_i^4$ of the Hamiltonian. In this case, the derivative of the characteristic function $\phi_N(\lambda)$ of the log-LR w.r.t. λ displayed in Lemma 8 has an additional term:

$$\phi_N'(\lambda) = \frac{\mathrm{i}s - s^2}{4} \mathbb{E}\big[\big(N\langle R_{1,*}^2\rangle - \langle x_N^2 x_N^{*2}\rangle\big) e^{\mathrm{i}s \log L}\big] + \frac{\mathrm{i}s - s^2}{2\sigma^2} \mathbb{E}\big[\langle x_N^2 x_N^{*2}\rangle e^{\mathrm{i}s \log L}\big].$$

The cavity computations performed in Section 6 also need to be altered in a minor way: in the interpolation argument separating the last variable x_N from the rest of the variables, we also have to make $d(\boldsymbol{x})$ time-dependent by performing the change of variable $\lambda \to \lambda t$. As a result of the computation, equation (23) becomes

$$(1-\lambda) N \, \mathbb{E}\big[\langle R_{1,*}^2\rangle e^{\mathrm{i}s \log L}\big] = \mathbb{E}\big[\langle x_N^2 x_N^{*2}\rangle e^{\mathrm{i}s \log L}\big] + \frac{\lambda\kappa}{\sigma^2} \mathbb{E}\big[e^{\mathrm{i}s \log L}\big] + \delta,$$

with $|\delta| \leq K/\sqrt{N}$, $\kappa = \mathbb{E}_{P_{\mathrm{x}}}[X^3]^2$, while equation (24) remains intact. As a result of these changes, and the above formula for ϕ_N', we get

$$\phi_N'(\lambda) = \frac{\mathrm{i}s - s^2}{4} \left(\frac{1 + \lambda\kappa/\sigma^2}{1-\lambda} - 1 \right) \phi_N(\lambda) + \frac{\mathrm{i}s - s^2}{2\sigma^2} \phi_N(\lambda) + \delta,$$

and this leads to the formula claimed.

9. Conclusions. This paper investigates the fundamental limits of spike detection in the rank-one spiked Wigner model. We proved that the logarithm of the likelihood ratio has Gaussian fluctuations below the reconstruction threshold λ_c while it is exponentially large under the alternative above it. This establishes the maximal region of contiguity between the planted and null models: namely the open interval $(0, \lambda_c)$. This also pins down the performance of the optimal test, and provides formulae for the Kullback–Leibler and the total variation distances between the null and planted distributions. An important characteristic of this threshold is that it is not necessarily related to the spectrum of the observed matrix: there are cases where λ_c does not correspond to the point where the signal shows up in the spectrum.

Our proofs repose on the technology developed within spin-glass theory for the study of the SK model. It is of interest to extend these techniques to other spiked models, notably spiked covariance models where the perturbation is in the covariance matrix of the data. Partial progress establishing Gaussian fluctuations of the LR in a restricted regime was recently obtained by a subset of the authors (El Alaoui and Jordan (2018)). Reaching the optimal threshold—a conjectural formula of which is provided in this recent paper—remains an interesting problem.

Acknowledgments. We are grateful to Léo Miolane for insightful conversations and to Nike Sun for comments on an earlier version of this manuscript. We warmly thank the anonymous reviewers of their feedback. This research was initiated at the *Workshop on Statistical physics, Learning, Inference and Networks* at École de Physique des Houches, Winter 2017.

This work was supported by the Multidisciplinary University Research Initiative under Army Research Office Grant W911NF-17-1-0304.

This work was performed when the first author was affiliated with UC Berkeley.

The second author was supported by the EU (FP/2007-2013/ERC Grant 307087-SPARCS).

The third author was supported by the Mathematical Data Science program of the Office of Naval Research Grant N00014-15-1-2670.

SUPPLEMENTARY MATERIAL

Supplement to "Fundamental limits of detection in the spiked Wigner model" (DOI: 10.1214/19-AOS1826SUPP; .pdf). This supplement (El Alaoui, Krzakala and Jordan (2020)) contains the proof of convergence of the moments of the overlap $R_{1,*}$ thereby completing the proof of Theorem 10, and the proof of Lemma 14.

REFERENCES

AIZENMAN, M., LEBOWITZ, J. L. and RUELLE, D. (1987). Some rigorous results on the Sherrington–Kirkpatrick spin glass model. *Comm. Math. Phys.* **112** 3–20. MR0904135

AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450 https://doi.org/10.1214/08-AOS664

BAI, Z. and YAO, J. (2008). Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** 447–474. MR2451053 https://doi.org/10.1214/07-AIHP118

BAI, Z. and YAO, J. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.* **106** 167–177. MR2887686 https://doi.org/10.1016/j.jmva.2011.10.009

BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

BAIK, J. and LEE, J. O. (2016). Fluctuations of the free energy of the spherical Sherrington–Kirkpatrick model. *J. Stat. Phys.* **165** 185–224. MR3554380 https://doi.org/10.1007/s10955-016-1610-0

BAIK, J. and LEE, J. O. (2017). Fluctuations of the free energy of the spherical Sherrington–Kirkpatrick model with ferromagnetic interaction. *Ann. Henri Poincaré* **18** 1867–1917. MR3649446 https://doi.org/10.1007/s00023-017-0562-5

BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680 https://doi.org/10.1016/j.jmva.2005.08.003

BANERJEE, D. (2018). Contiguity and non-reconstruction results for planted partition models: The dense case. *Electron. J. Probab.* **23** Paper No. 18, 28. MR3771755 https://doi.org/10.1214/17-EJP128

BANERJEE, D. and MA, Z. (2018). Asymptotic normality and analysis of variance of log-likelihood ratios in spiked random matrix models. arXiv preprint arXiv:1804.00567.

BANKS, J., MOORE, C., VERSHYNIN, R., VERZELEN, N. and XU, J. (2017). Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. In *IEEE International Symposium on Information Theory* (*ISIT*) 1137–1141. IEEE.

BARBIER, J., DIA, M., MACRIS, N., KRZAKALA, F., LESIEUR, T. and ZDEBOROVÁ, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems* (*NIPS*) 424–432.

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201 https://doi.org/10.1016/j.aim.2011.02.007

BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849 https://doi.org/10.1214/13-AOS1127

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

CAPITAINE, M., DONATI-MARTIN, C. and FÉRAL, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann. Probab.* **37** 1–47. MR2489158 https://doi.org/10.1214/08-AOP394

CHATTERJEE, S. (2014). *Superconcentration and Related Topics*. *Springer Monographs in Mathematics*. Springer, Cham. MR3157205 https://doi.org/10.1007/978-3-319-03886-5

DESHPANDE, Y., ABBÉ, E. and MONTANARI, A. (2016). Asymptotic mutual information for the binary stochastic block model. In *IEEE International Symposium on Information Theory* (*ISIT*) 185–189.

DOBRIBAN, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Ann. Statist.* **45** 1810–1833. MR3670197 https://doi.org/10.1214/16-AOS1514

EL ALAOUI, A. and JORDAN, M. I. (2018). Detection limits in the high-dimensional spiked rectangular model. In *Proceedings of the* 31*st Conference on Learning Theory* (*COLT*) **75** 410–438.

EL ALAOUI, A. and KRZAKALA, F. (2018). Estimation in the spiked Wigner model: A short proof of the replica formula. In *IEEE International Symposium on Information Theory* (*ISIT*) 1874–1878.

EL ALAOUI, A., KRZAKALA, F. and JORDAN, M. (2020). Supplement to "Fundamental limits of detection in the spiked Wigner model." https://doi.org/10.1214/19-AOS1826SUPP.

FÉRAL, D. and PÉCHÉ, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.* **272** 185–228. MR2291807 https://doi.org/10.1007/s00220-007-0209-3

FRANZ, S. and PARISI, G. (1995). Recipes for metastable states in spin glasses. *J. Phys., I* **5** 1401–1415.

FRANZ, S. and PARISI, G. (1998). Effective potential in glassy systems: Theory and simulations. *Phys. A* **261** 317–339.

GUERRA, F. (2001). Sum rules for the free energy in the mean field spin glass model. In *Mathematical Physics in Mathematics and Physics* (*Siena*, 2000). *Fields Inst. Commun.* **30** 161–170. Amer. Math. Soc., Providence, RI. MR1867553

GUERRA, F. (2003). Broken replica symmetry bounds in the mean field spin glass model. *Comm. Math. Phys.* **233** 1–12. MR1957729 https://doi.org/10.1007/s00220-002-0773-5

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

JOHNSTONE, I. M. and ONATSKI, A. (2015). Testing in high-dimensional spiked models. arXiv preprint arXiv:1509.07269.

KRZAKALA, F., XU, J. and ZDEBOROVÁ, L. (2016). Mutual information in rank-one matrix estimation. In *Information Theory Workshop* (*ITW*) 71–75.

LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. MR1926169 https://doi.org/10.1214/aos/1031689018

LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2015). Phase transitions in sparse PCA. In *IEEE International Symposium on Information Theory* (*ISIT*) 1635–1639.

MÉZARD, M., PARISI, G. and VIRASORO, M. A. (1987). *Spin Glass Theory and Beyond*. *World Scientific Lecture Notes in Physics* **9**. World Scientific Co., Inc., Teaneck, NJ. MR1026102

NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013 https://doi.org/10.1214/08-AOS618

NISHIMORI, H. (2001). *Statistical Physics of Spin Glasses and Information Processing*: *An Introduction*. *International Series of Monographs on Physics* **111**. Oxford Univ. Press, New York. Translated from the 1999 Japanese original. MR2250384 https://doi.org/10.1093/acprof:oso/9780198509417.001.0001

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. MR3113808 https://doi.org/10.1214/13-AOS1100

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2014). Signal detection in high dimension: The multispiked case. *Ann. Statist.* **42** 225–254. MR3189485 https://doi.org/10.1214/13-AOS1181

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

PÉCHÉ, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields* **134** 127–173. MR2221787 https://doi.org/10.1007/s00440-005-0466-z

PÉCHÉ, S. (2014). Deformed ensembles of random matrices. In *Proceedings of the International Congress of Mathematicians—Seoul* 2014. *Vol. III* 1159–1174. Kyung Moon Sa, Seoul. MR3729069

PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.* **46** 2416–2451. MR3845022 https://doi.org/10.1214/17-AOS1625

TALAGRAND, M. (2011a). *Mean Field Models for Spin Glasses*. *Volume I*: *Basic Examples*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge*. *A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas*. 3*rd Series*. *A Series of Modern Surveys in Mathematics*] **54**. Springer, Berlin. MR2731561 https://doi.org/10.1007/978-3-642-15202-3

TALAGRAND, M. (2011b). *Mean Field Models for Spin Glasses*. *Volume II*: *Advanced Replica-Symmetry and Low Temperature*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge*. *A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas*. 3*rd Series*. *A Series of Modern Surveys in Mathematics*] **55**. Springer, Heidelberg. MR3024566

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