

# A multiscale cavity method for sublinear-rank symmetric matrix factorization

Jean Barbier, Justin Ko, and Anas A. Rahman

**Abstract.** We consider a statistical model for symmetric matrix factorization with additive Gaussian noise in the high-dimensional regime, where the rank of the signal matrix to infer  $M$  scales with its size  $N$  as  $M = o(\sqrt{\ln N})$ . Allowing for an  $N$ -dependent rank offers new challenges and requires new methods. Working in the Bayes-optimal setting, we show that whenever the signal has i.i.d. entries, the limiting mutual information between signal and data is given by a variational formula involving a rank-*one* replica symmetric potential. In other words, from the information-theoretic perspective, the case of a (slowly) growing rank is the same as when  $M = 1$  (namely, the standard spiked Wigner model). The proof is primarily based on a novel *multiscale cavity method* allowing for growing rank along with some information-theoretic identities on worst noise for the vector Gaussian channel. We believe that the cavity method developed here will play a role in the analysis of a broader class of inference and spin models where the degrees of freedom are large arrays instead of vectors.

## 1. Introduction

Reconstructing a low-rank signal matrix observed through a noisy channel with highest possible accuracy is of fundamental importance across a range of disciplines including statistical inference, machine learning, signal processing, and information theory. This task encompasses a variety of models that have attracted considerable attention over the last decade, such as sparse principal component analysis and  $\mathbb{Z}_2$  synchronization [23, 45], the stochastic block model for community detection [1, 22], and submatrix localization [20, 35], just to cite a few. A common theme unifying these models is that they are either examples of, or are otherwise closely related to, the general notion of spiked matrix models [5, 38, 59] of the form “signal plus noise” that were introduced at the turn of this century.

---

*Mathematics Subject Classification 2020:* 62F15 (primary); 62H12, 82D30 (secondary).

*Keywords:* spiked Wigner model, sublinear-rank, multiscale cavity method, rank-one reduction.

In this work, we consider the archetypal *spiked Wigner model*, where one observes a symmetric deformation of a standard Wigner matrix. The data is generated as

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}} \mathbf{X}_0 \mathbf{X}_0^\top + \mathbf{Z}, \quad (1.1)$$

where  $\mathbf{X}_0 \in \mathbb{R}^{N \times M}$  is the signal matrix distributed according to some prior distribution  $\mathbb{P}_{\mathbf{X}, N, M}$ ,  $\mathbf{Z} \in \mathbb{R}^{N \times N}$  is a standard Wigner matrix with  $\mathbf{Z}_{ii} \sim \mathcal{N}(0, 2)$  ( $1 \leq i \leq N$ ) and  $\mathbf{Z}_{ij} = \mathbf{Z}_{ji} \sim \mathcal{N}(0, 1)$  ( $1 \leq i < j \leq N$ ) independently and identically distributed (i.i.d.),  $\lambda \geq 0$  is the signal-to-noise ratio (SNR), and the scaling by  $\sqrt{N}$  is such that the signal and noise are of comparable magnitudes. From another perspective, we may say that we observe the spike  $\mathbf{X}_0$  corrupted by additive Gaussian noise; note that results concerning this model extend to a wide range of noisy observation channels due to recently proven universality principles [31, 32, 42, 44].

We study the *free entropy* (equivalently, up to a simple additive term, the mutual information  $I(\mathbf{X}_0; \mathbf{Y})$ ; see, e.g., [43])

$$F_N(\lambda) := F_{N, M}(\lambda) = \frac{1}{NM} \mathbb{E}_{\mathbf{Z}, \mathbf{X}_0} \ln Z_{N, M} \quad (1.2)$$

of the model (1.1) in the thermodynamic limit  $N, M \rightarrow \infty$  with the rank  $M = M_N$  growing sufficiently slowly with  $N$ . Here,

$$Z_{N, M} := \int_{\mathbb{R}^{N \times M}} e^{H_N(\mathbf{X})} d\mathbb{P}_{\mathbf{X}, N, M}(\mathbf{X}) \quad (1.3)$$

denotes the *partition function* associated with the *Hamiltonian*

$$H_N(\mathbf{X}) := \frac{1}{2} \text{Tr} \left( \sqrt{\frac{\lambda}{N}} \mathbf{Z} \mathbf{X} \mathbf{X}^\top + \frac{\lambda}{N} \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X} \mathbf{X}^\top - \frac{\lambda}{2N} \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \right); \quad (1.4)$$

in the language of Bayesian inference rather than statistical physics, the partition function is simply the normalization of the posterior distribution  $\mathbb{P}_{\mathbf{X}|\mathbf{Y}, N, M}(\mathbf{X} | \mathbf{Y}(\mathbf{X}_0, \mathbf{Z}))$  and the Hamiltonian is the log-likelihood. In order to take advantage of replica symmetry (i.e., concentration of measure) [6, 10, 43], we moreover work in the *Bayes-optimal setting*, which means that in addition to the observed data  $\mathbf{Y}$ , the statistician charged with reconstructing  $\mathbf{X}_0$  from  $\mathbf{Y}$  also has knowledge of the form (1.1) of  $\mathbf{Y}$ , the prior distribution  $\mathbb{P}_{\mathbf{X}, N, M}$  of the signal  $\mathbf{X}_0$ , the SNR  $\lambda$ , and the rank  $M$  of  $\mathbf{X}_0 \mathbf{X}_0^\top$ .

Focusing on the Bayesian viewpoint, various applications of (the adaptive) interpolation methods, the cavity method, and the Approximate Message Passing (AMP) algorithm have been used to study (various generalizations of) the spiked Wigner model. Studies commenced in the rank  $M = 1$  case [8, 22, 41, 42, 46] before progressing to the low-rank (i.e., finite-rank) case [43, 47], with recent studies venturing into the regime of growing rank [13, 37, 48, 61]. In particular, an  $M$ -dimensional

variational formula for the limiting free entropy  $\lim_{N \rightarrow \infty} F_N(\lambda)$  can be surmised from [61] when  $M = o(N^{1/20})$ . Our present goal is to continue this line of research by proving a scalar equivalent of the aforementioned formula (thereby circumventing the intractability of the  $M$ -dependence seen therein) with regime of veracity  $M = o(\sqrt{\ln N})$ . This is a significant milestone towards understanding the recent conjecture based on the replica method of [60] that the rank- $M$  spiked Wigner model should behave as its rank-one counterpart, so long as the rank is sublinear, i.e., when  $M = o(N)$ .

It has long been known and utilized that there is a close interplay between statistical inference and the theory of spin glasses. In particular, the free entropy of the finite-rank spiked Wigner model is closely related to the free energies of vector spin glass models, with the main technical challenge in studies of the latter being the absence of replica symmetry. Variational formulae for free energies of various vector spin glass models have been proven using synchronization of overlap matrices [12, 26, 39, 40, 57, 58] and Hamilton–Jacobi equations [17–19, 54]; the relevant order parameters are increasing matrix paths, which are more challenging to deal with than the matrix-valued order parameters typical of Bayes-optimal inference problems. Said variational formulae have moreover been recently analyzed and simplified in [4, 11, 16] through exploitation of inherent symmetries within the studied models in order to reduce dimensionality of the order parameters. Our main technical contribution in this line of work is a generalization of the *Aizenman–Sims–Starr scheme* [3] to multiscale mean field models, which in this context are vector spin glass models with dimension-dependent vector spin coordinates. Using the cavity method instead of, say, the adaptive interpolation method [47, 61] means that we require only posterior (thermal) concentration of our order parameter, the overlap matrix, rather than full (quenched) concentration. This in turn enables us to employ a scalar perturbation parameter (as opposed to a matrix one) in the proof of concentration, leading to more tractable computations and better rates of convergence. Our secondary contribution is a proof of the fact that our variational formula for the limiting free entropy reduces to that of the rank-one spiked Wigner model, owing to a set of simple information-theoretic inequalities. Said identities are, to the best of our knowledge, not in the literature and we expect Corollary 2.4, say, to be a valuable addition to the literature on worst noises [24, 25, 63].

The reduction to rank one mentioned above follows from the fact that the entries of the signal  $X_0$  are i.i.d., which induces exchangeability in the entries of the overlap matrix. This is similar to recent simplifications of vector spin Parisi formulae obtained in [11, 16] through the utilization of underlying symmetries in the Potts spin glass. Combining these simplifications with our multiscale cavity method has the potential to extend the corresponding finite-rank vector spin models to sublinear-rank multiscale models, which is a natural step towards understanding the challenging

extensive-rank regime  $M = \Theta(N)$ ; see [9, 13–15, 48, 60, 65]. In this vein, we foresee that the relative tractability of our computations will also enable the theory of asymmetric matrix and tensor factorization [27, 28, 46, 47, 50, 52, 53] to be pushed into the growing rank regime. Indeed, we find some potential extensions of our methods to particular related settings are readily observable: In the case of asymmetric matrix factorization [27, 28, 50, 51, 53], we expect the finite-rank cavity method of [50] to extend to the growing rank regime via our multiscale formalism. Likewise for the cavity method employed in the study of symmetric tensor factorization [46]. The missing ingredient in both of these cases is our result reducing the rank- $M$  model to its rank-one counterpart. In the present setting, this is done by showing equivalence between the suprema of a rank- $M$  potential and its rank-one counterpart, introduced below. In [50], one is instead interested in the supremum of the sum of two such potentials coupled through a quadratic term, while in [46], the relevant potential involves Hadamard powers of the matrix argument  $\mathbf{Q}$ , so that the eigenvalue-based arguments of Section 3 cannot be used. Further investigation is needed to ascertain if the ideas of Section 3 can be adapted to overcome these challenges.

## 2. Setting and main results

### 2.1. The sublinear-rank spiked Wigner model and the replica symmetric potential

We study the  $N, M \rightarrow \infty$  limit of the free entropy  $F_N(\lambda)$  (see (1.2)) of the spiked Wigner model (1.1) with rank  $M = o(\sqrt{\ln N})$  and signal of i.i.d. entries, i.e., the prior distribution  $\mathbb{P}_{X,N,M}$  on  $\mathbf{X}_0 = (\mathbf{x}_{0,i} \in \mathbb{R}^M)_{i \leq N} = (\mathbf{x}_{0,ij} \in \mathbb{R})_{i \leq N, j \leq M}$  fully factorizes:

$$\mathbb{P}_{X,N,M}(\mathbf{X}_0) = \prod_{i=1}^N \mathbb{P}_{X,M}(\mathbf{x}_{0,i}) = \prod_{i=1}^N \prod_{j=1}^M \mathbb{P}_X(\mathbf{x}_{0,ij}). \quad (2.1)$$

In addition, we take  $\mathbb{P}_X$  to be a centered distribution with  $D$ -bounded support: There exists  $0 \leq D < \infty$  such that  $\text{supp } \mathbb{P}_X \subseteq [-D, D]$ .

An application of the (non-rigorous) replica method [49, 64] suggests that the large  $N$  limit of  $F_N(\lambda)$  is given by a variational formula written in terms of the rank- $M$  replica symmetric potential  $F_M^{\text{RS}}: \mathcal{S}_M \times \mathbb{R}_+ \mapsto \mathbb{R}$  defined by (with  $\mathcal{S}_M$  the set of symmetric positive semidefinite  $M \times M$  matrices)

$$\begin{aligned} F_M^{\text{RS}}(\mathbf{Q}, \lambda) &:= \frac{1}{M} \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \ln Z_M^{\text{RS}} - \frac{\lambda}{4M} \text{Tr } \mathbf{Q}^2, \\ Z_M^{\text{RS}} &:= \int_{\mathbb{R}^M} e^{\sqrt{\lambda} \mathbf{x}^\top \sqrt{\mathbf{Q}} \mathbf{z} + \lambda \mathbf{x}_0^\top \mathbf{Q} \mathbf{x} - \frac{\lambda}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x}} d\mathbb{P}_{X,M}(\mathbf{x}), \end{aligned} \quad (2.2)$$

where  $\mathbf{z} \in \mathbb{R}^M$  is a standard Gaussian vector and  $\mathbf{x}_0 \sim \mathbb{P}_{X,M}$  (see [46,47]). However, we find that it is instead given by a formula in terms of the simpler rank-one analog

$$\begin{aligned} F_1^{\text{RS}}(q, \lambda) &:= \mathbb{E}_{z, x_0} \ln Z_1^{\text{RS}} - \frac{\lambda}{4} q^2, \\ Z_1^{\text{RS}} &:= \int_{-\infty}^{\infty} e^{\sqrt{\lambda} q z x + \lambda q x_0 x - \frac{1}{2} \lambda q x^2} d\mathbb{P}_X(x), \end{aligned} \quad (2.3)$$

with  $z \sim \mathcal{N}(0, 1)$  and  $x_0 \sim \mathbb{P}_X$ . Thus, we have as our main result that, under our hypotheses above in addition to two further technical hypotheses concerning vector and scalar analogs of our channel (1.1), the  $N \rightarrow \infty$  limit of the free entropy  $F_N(\lambda)$  is given by the rank-one replica symmetric potential  $F_1^{\text{RS}}(q, \lambda)$  at its supremum over  $q$ .

**Theorem 2.1** (Rank-one replica formula for the growing-rank spiked Wigner model). *Assume the following hypotheses:*

- (H1) *The rank  $M = M_N \in \mathbb{N}$  is such that  $M = o(\sqrt{\ln N})$  and  $M_{N+1} - M_N \leq 1$  for all  $N \in \mathbb{N}$ .*
- (H2) *The prior distribution  $\mathbb{P}_{X,N,M}$  on  $\mathbf{X}_0$  fully factorizes as in equation (2.1).*
- (H3) *The distribution  $\mathbb{P}_X$  is centered with  $D$ -bounded support.*
- (H4) *The distribution  $\mathbb{P}_X$  is such that for  $\mathbb{P}_{X,M} = \mathbb{P}_X^{\otimes M}$  and all constant  $m \leq M$ ,*

$$\phi_m(\lambda) := \sup_{\mathcal{Q} \in \mathcal{S}_m} F_m^{\text{RS}}(\mathcal{Q}, \lambda) \quad (2.4)$$

*is real analytic in  $\lambda$  on  $[0, \infty)$  except for possibly one critical point (phase transition)  $0 < \lambda_c < \infty$ , independent of  $m$ .*

- (H5) *Letting  $x_0 \sim \mathbb{P}_X$  and  $z \sim \mathcal{N}(0, 1)$ , the minimum mean-square error (MMSE) of the scalar Gaussian channel  $y = \sqrt{\lambda} x_0 + z$ ,*

$$\text{mmse}(\lambda) := \mathbb{E}_{z, x_0} [(x_0 - \mathbb{E}[x_0 | y])^2], \quad (2.5)$$

*is such that  $\lambda^2 \text{mmse}(\lambda)$  is monotone on  $[\lambda_L'', \infty)$  for some  $\lambda_L'' > 0$ .*

*Setting  $\rho := \mathbb{E}_{\mathbb{P}_X} X^2$ , the limiting free entropy (1.2) of the spiked Wigner model (1.1) is then given in terms of the replica symmetric potential (2.3) by*

$$\lim_{N \rightarrow \infty} F_N(\lambda) = \lim_{N \rightarrow \infty} F_{N, M_N}(\lambda) = \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda). \quad (2.6)$$

*As a consequence, we have the following formula for the limiting minimum mean-square error of the spiked Wigner model:*

$$\begin{aligned} \lim_{N \rightarrow \infty} \text{MMSE}_{N, M_N}(\lambda) &= \rho^2 - q^*(\lambda)^2, \quad 0 \leq \lambda \neq \lambda_c, \\ \text{MMSE}_{N, M}(\lambda) &:= \frac{1}{N^2 M} \mathbb{E}_{\mathbf{Z}, \mathbf{X}_0} \left\| \mathbf{X}_0 \mathbf{X}_0^{\text{T}} - \mathbb{E}[\mathbf{X}_0 \mathbf{X}_0^{\text{T}} | \mathbf{Y}] \right\|_{\text{F}}^2, \end{aligned} \quad (2.7)$$

where  $q^*(\lambda) := \arg \max_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda)$  is the (unique [43, Proposition 17]) maximizer of  $F_1^{\text{RS}}(q, \lambda)$  and  $\|\cdot\|_F$  denotes the Frobenius norm.

The statement on the minimum mean-square error follows from the I-MMSE formula [33] and the arguments of [43].

We comment on our hypotheses:

(1) We require  $M = o(N^{1/4})$  to achieve thermal concentration of the overlap matrix

$$\mathbf{R}_{10} := \frac{1}{N} \mathbf{X}^\top \mathbf{X}_0 = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_{0,i}^\top \quad (2.8)$$

with respect to the Gibbs average corresponding to a particular perturbation of the Hamiltonian  $H_N(\mathbf{X})$  (1.4) and  $M = o(\sqrt{\ln N})$  for the error terms arising from the cavity computation to vanish in the large  $N$  limit. We believe that this growth rate for the rank is not fundamental and is rather a limitation of the proof methods. Indeed, we conjecture that Theorem 2.1 holds for  $M = o(N)$ , in line with [37]. The requirement that  $M_{N+1} - M_N \leq 1$  is in keeping with the natural notion that the sequence of matrices  $\mathbf{X}_0$  has its dimensions  $M, N$  growing by increments of 1, with  $N$  growing steadily.

(2) Symmetry properties of  $\mathbb{P}_{X,N,M}$  owing to its factorized nature encourage the thermal concentration of  $\mathbf{R}_{10}$  onto a scalar multiple of the identity, which in turn induces an equivalence

$$\sup_{q \in [0, \rho]} F_M^{\text{RS}}(q I_M, \lambda) \equiv \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda).$$

The factorization of the prior is thus key to reducing the dimensionality of the variational formula (2.6).

(3) Requiring  $\mathbb{P}_X$  to have  $D$ -bounded support is not very restrictive and is only a technical constraint; prior distributions  $\mathbb{P}_X$  with unbounded support can be treated by truncating them to have  $D$ -bounded support, working in the finite  $D$  regime as in this paper, and then taking  $D \rightarrow \infty$ ; see, e.g., [43, 47].

(4) It is believed that for sufficiently slowly growing rank,  $\lim_{N \rightarrow \infty} F_N(\lambda)$  exhibits a single rank-independent phase transition for many natural choices of prior  $\mathbb{P}_X$  (uniform, Rademacher, Gaussian, simple mixtures, etc.); see [51] and the references therein; for rank proportional to  $N$ , one still observes a unique phase transition, but it has a rank-dependent location [7]. We assume there is at most one rank-independent phase transition and translate this assumption on the model into one for  $\phi_m(\lambda)$  ( $m \leq M$ ). It is difficult to prove this hypothesis in general and it may not hold for certain (exotic) priors. Nonetheless, a convincing check for its validity is to numerically evaluate  $\phi_2(\lambda)$ , which is easy, and look for a unique first or second order phase

transition (a discontinuity in its first or second order derivative, which are the standard transitions in high-dimensional inference problems [69]). The same behavior is then expected for any fixed  $m$  – additional phase transitions can only appear in a large system limit. This is sufficient for our argument (in particular, Theorem 2.6). To improve clarity, we will often refer to hypothesis (H4) when it suffices to take a weaker hypothesis, such as when we require real analyticity of  $\phi_m(\lambda)$  only at low or high SNR, or at exactly  $m = M$ .

(5) As scalar channels are far more tractable than higher-dimensional ones, it is relatively easy to check monotonicity of  $\lambda^2 \text{mmse}(\lambda)$  in the high SNR regime. In particular, hypothesis (H5) holds when  $\text{mmse}(\lambda)$  is real analytic on  $[\lambda_L'', \infty)$  since  $\text{mmse}(\lambda) = o(1/\lambda)$  as  $\lambda \rightarrow \infty$  (see [67]). This includes exponential, finite-alphabet, and Gaussian distributed signals [34].

Theorem 2.1 manifests equivalence with its rank-one analog, rigorously analyzed in [8, 43, 47], due to the  $M$ -independence of the right-hand side of equation (2.6). Similar results concerning sublinear-rank models behaving as their finite-rank counterparts exist for mesoscopic spiked perturbations of random matrices [36] and for spherical integrals [37]. Our results complement these in the context of Bayesian inference. If hypothesis (H4) is weakened, then Theorem 2.1 holds for all SNR  $\lambda$  greater (smaller) than the largest (smallest) non-analytic value of  $\lambda$ . This follows from straightforward changes to the proof of Theorem 2.6 given at the end of Section 3.

The proof of Theorem 2.1 comprises a lower bound on  $F_N(\lambda)$  (which is valid for  $M, N \in \mathbb{N}$ ), hence

$$\liminf_{N \rightarrow \infty} F_N(\lambda),$$

obtained by a standard application of Guerra’s interpolation method [30], and a matching upper bound on

$$\limsup_{N \rightarrow \infty} F_N(\lambda),$$

obtained through a multiscale application of the cavity method prescribed by the Aizenman–Sims–Starr scheme [3]. The novelty of the argument lies in two key ingredients: Firstly, we show in Section 3, and summarize in Section 2.2, how the supremum of the rank- $M$  replica symmetric potential  $\phi_M(\lambda)$  (2.4) reduces to its rank-one analog  $\phi_1(\lambda)$  – this is used to obtain agreement between the aforementioned upper and lower bounds and moreover improves the tractability of the right-hand side of equation (2.6). Secondly, we develop in Section 5 a cavity method that is able to treat sequences  $F_{N,M}(\lambda)$  indexed by *two* growing dimensions instead of one, as is usually the case. In addition to the rank-one reduction of Section 3, said cavity computations also use the lower bound on  $F_N(\lambda)$  that we obtain via Guerra’s interpolation method and thermal concentration of the overlap matrix  $\mathbf{R}_{10}$ . We present these two prerequisite results, which are proven through relatively standard methods, in Section 4; the

contents of Sections 4 and 5 are summarized in Section 2.3. We emphasize that all results but one in this paper hold for  $M$  growing at certain polynomial rates in  $N$ .

For convenience of later analysis, it is helpful to re-express the replica symmetric potential in terms of the mutual information of the  $M$ -dimensional vector Gaussian channel (see Appendix A):

$$F_M^{\text{RS}}(\mathbf{Q}, \lambda) = -\frac{1}{M}I(\mathbf{x}_0; \sqrt{\lambda}\mathbf{Q}\mathbf{x}_0 + \mathbf{z}) - \frac{\lambda}{4M}\|\mathbf{Q} - \rho I_M\|_{\text{F}}^2 + \frac{\lambda\rho^2}{4}, \quad (2.9)$$

where  $\mathbf{z} \sim \mathcal{N}(0, I_M)$  is standard Gaussian noise and  $\mathbf{x}_0 \sim \mathbb{P}_{X, M}$ .

**Remark 2.2** (Properties of the replica symmetric potential). The squared Frobenius norm  $\|\mathbf{Q} - \rho I_M\|_{\text{F}}^2$  is a convex paraboloid with minimum at  $\mathbf{Q} = \rho I_M$ , while the mutual information  $I(\mathbf{x}_0; \sqrt{\lambda}\mathbf{Q}\mathbf{x}_0 + \mathbf{z})$  is a concave, non-decreasing monotone function of  $\mathbf{Q}$  (see [33, 62]): for  $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathcal{S}_M$ ,  $0 \leq t \leq 1$ , and  $\mathbf{z}, \mathbf{z}', \mathbf{z}'' \sim \mathcal{N}(0, I_M)$  independent, we have

$$\begin{aligned} & tI(\mathbf{x}_0; \sqrt{\lambda}\mathbf{Q}_1\mathbf{x}_0 + \mathbf{z}) + (1-t)I(\mathbf{x}_0; \sqrt{\lambda}\mathbf{Q}_2\mathbf{x}_0 + \mathbf{z}') \\ & \leq I(\mathbf{x}_0; \sqrt{\lambda(t\mathbf{Q}_1 + (1-t)\mathbf{Q}_2)}\mathbf{x}_0 + \mathbf{z}''). \end{aligned}$$

The problem of maximizing  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  over  $\mathbf{Q}$  therefore translates to one of comparing these two competing terms. For each  $m \in \mathbb{N}$  and  $\mathbf{Q} \in \mathcal{S}_m$ ,  $F_m^{\text{RS}}(\mathbf{Q}, \lambda) - \lambda\rho^2/4$  is a non-increasing, monotone, continuous, convex function of  $\lambda$ . Thus, the same holds true for  $\phi_m(\lambda) - \lambda\rho^2/4 = \sup_{\mathbf{Q} \in \mathcal{S}_m} F_m^{\text{RS}}(\mathbf{Q}, \lambda) - \lambda\rho^2/4$ .

## 2.2. Equivalence of the suprema of the rank- $M$ and rank-one replica symmetric potentials

We link the supremum of the rank- $M$  replica symmetric potential  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  to its rank-one equivalent through a series of intermediary results. The first two are of particular interest due to their being a statement on the worst possible additive Gaussian noise in a vector channel with i.i.d. inputs.

**Lemma 2.3** (Off-diagonal entries of the noise covariance bolster information). *Assume hypothesis (H2) and let  $\mathbf{x}_0 \sim \mathbb{P}_{X, M}$ ,  $x_0 \sim \mathbb{P}_X$ ,  $\mathbf{z} \sim \mathcal{N}(0, I_M)$ , and  $z \sim \mathcal{N}(0, 1)$ . Then, for  $\Sigma \in \mathcal{S}_M$ , the mutual information between  $\mathbf{x}_0$  and the output of the vector channel with additive Gaussian noise  $\Sigma^{1/2}\mathbf{z}$  of covariance  $\Sigma$  satisfies*

$$I(\mathbf{x}_0; \mathbf{x}_0 + \Sigma^{1/2}\mathbf{z}) \geq \sum_{i=1}^M I(x_0; x_0 + \sqrt{\Sigma_{ii}}z). \quad (2.10)$$

*This inequality is an equality if  $\Sigma$  is diagonal.*

We prove this lemma in Section 3.1 and use it in Section 3.3, with  $\Sigma = (\lambda \mathbf{Q})^{-1}$ , to give a partial proof of Theorem 2.6 below. Indeed, due to convexity properties of the squared Frobenius norm in equation (2.9), Lemma 2.3 only supplies a proof of Theorem 2.6 when the eigenvalues of  $\mathbf{Q}$  are restricted to be sufficiently large. However, we are able to utilize a convexity result of [66] on signals drawn from distributions with  $D$ -bounded supports in order to formulate a parallel proof for when  $\mathbf{Q}$  has small eigenvalues and obtain the following corollary of Lemma 2.3.

**Corollary 2.4** (Worst Gaussian noise with covariance of fixed trace). *Let  $\Sigma \in \mathcal{S}_M$  be such that  $\Sigma_{ii} \geq D^2$  for each  $1 \leq i \leq M$  and let  $\sigma := \text{Tr } \Sigma / M$  be its normalized trace. Then, in the setting of Lemma 2.3 with  $\mathbb{P}_X$  having  $D$ -bounded support, we have that*

$$I(\mathbf{x}_0; \mathbf{x}_0 + \Sigma^{1/2} \mathbf{z}) \geq MI(\mathbf{x}_0; \mathbf{x}_0 + \sqrt{\sigma} \mathbf{z}). \quad (2.11)$$

*This inequality is an equality if  $\Sigma = \sigma I_M$ .*

The partial proofs of Theorem 2.6 emerging from Lemma 2.3 and Corollary 2.4 are interestingly dual to each other in the sense that the first proof focuses on the diagonal entries of  $\Sigma$ , while the second focuses instead on its eigenvalues; note that these parameters coincide when  $\Sigma$  is diagonal. This shift in focus means that the roles of the mutual information and the squared Frobenius norm in the expression (2.9) are reversed in the two partial proofs. Importantly, said partial proofs correspond respectively to regimes of high and low SNR due to the following properties, proven in Section 3.2, of maximizers  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  linking the regime of  $\mathbf{Q}$  having large (small) eigenvalues to that of high (low) SNR.

**Proposition 2.5** (Properties of  $\mathbf{Q}^*(\lambda)$ ). *Assume hypotheses (H2) and (H4). Let  $\mathbb{P}_X$  be centered with bounded second moment  $\rho = \mathbb{E}_{\mathbb{P}_X} X^2$  and let  $\mathbf{Q}^*(\lambda) \in \mathcal{S}_M$  be such that*

$$F_M^{\text{RS}}(\mathbf{Q}^*(\lambda), \lambda) = \phi_M(\lambda).$$

*Then, we have that  $\mathbf{Q}^*(\lambda)$  and its eigenvalues  $q_1^*, \dots, q_M^*$  have the following properties:*

- (1)  $0 \leq q_1^*, \dots, q_M^* \leq \rho$  for all  $\lambda \geq 0$ .
- (2)  $\|\mathbf{Q}^*(\lambda)\|_{\mathbb{F}}$  is continuous for all  $\lambda \geq 0$  except possibly at  $\lambda = \lambda_c$ .
- (3) Given  $\rho'_S \in (0, \rho)$ , there exists  $0 < \lambda'_S < \lambda_c$  such that  $0 \leq q_1^*, \dots, q_M^* < \rho'_S$  for all  $0 \leq \lambda < \lambda'_S$ .
- (4) Given  $\rho'_L \in (0, \rho)$ , there exists  $\lambda'_L > \lambda_c$  such that  $\rho'_L < q_1^*, \dots, q_M^* \leq \rho$  for all  $\lambda > \lambda'_L$ .

Thus, in Section 3.3, we first prove the following theorem in the high and low SNR regimes before finally using analytic continuation in  $\lambda$  of  $\phi_m(\lambda)$  ( $m \leq M$ ), assumed

to be real analytic for all  $\lambda \neq \lambda_c$  by hypothesis, to extend the regime of veracity to the entirety of the positive real line.

**Theorem 2.6** (Rank-one reduction). *Assume hypotheses (H2)–(H5). We then have for all SNR  $\lambda \geq 0$  that*

$$\sup_{\mathbf{Q} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{Q}, \lambda) = \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda).$$

### 2.3. Bounding the limiting free entropy via the interpolation and multiscale cavity methods

The proof of Theorem 2.1 is a combination of the following two results (see Section 5.4).

**Proposition 2.7** (Free entropy lower bound). *Under hypotheses (H2)–(H5), we have for any  $M, N \geq 1$  that*

$$F_N(\lambda) \geq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda).$$

**Proposition 2.8** (Free entropy upper bound). *Under the assumptions of Theorem 2.1, we have that*

$$\limsup_{N \rightarrow \infty} F_N(\lambda) \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda). \quad (2.12)$$

Proposition 2.7 is proven in Section 4.1 using a standard application (apart from the use of Theorem 2.6) of Guerra’s interpolation method [30, 42, 43], while Proposition 2.8 requires a non-trivial adaptation of the cavity method. In the standard cavity computation for finite rank-models [43], one begins by Cesàro summing a telescoping series to obtain the upper bound

$$\begin{aligned} \limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda) &= \limsup_{N \rightarrow \infty} \frac{1}{NM} \sum_{n=0}^{N-1} (\mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0, \varepsilon}} \ln \tilde{Z}_{n+1, M} - \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0, \varepsilon}} \ln \tilde{Z}_{n, M}) \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{M} (\mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0, \varepsilon}} \ln \tilde{Z}_{N+1, M} - \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0, \varepsilon}} \ln \tilde{Z}_{N, M}), \end{aligned} \quad (2.13)$$

where  $\tilde{F}_N(\lambda)$  and  $\tilde{Z}_{N, M}$  are perturbations of the free entropy and partition function defined by equations (1.2) and (1.3), but with  $H_N(\mathbf{X})$  replaced by a *perturbed Hamiltonian*

$$\tilde{H}_{N, \varepsilon}(\mathbf{X}) := H_N(\mathbf{X}) + H_{N, \varepsilon}(\mathbf{X}) \quad (2.14)$$

and the expectation now containing an average over the perturbation parameter  $\varepsilon$  and any additional randomness due to the *perturbation Hamiltonian*  $H_{N, \varepsilon}(\mathbf{X})$ . This latter Hamiltonian is carefully chosen to induce concentration of the overlap matrix  $\mathbf{R}_{10}$

(see (2.8)) in the small  $\varepsilon$  regime; we show in Section 4.3 that for  $H_{N,\varepsilon}(X)$  defined by equation (2.22) forthcoming,  $\limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda)$  serves as an upper bound for the left-hand side of inequality (2.12), so bounding it by the right-hand side of said inequality completes the proof of Proposition 2.8. In the present setting, the rank  $M = M_N$  grows as a function of  $N$ , so inequality (2.13) becomes

$$\limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda) \leq \limsup_{N \rightarrow \infty} \frac{1}{M_N} (\mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{N+1, M_{N+1}} - \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{N, M_N}),$$

which is no longer amenable to the standard Aizenman–Sims–Starr approach [3, 43]: there are two growing, interacting scales that need to be considered jointly. This challenge can be overcome by putting the  $N$  and  $M$  parameters on equal footing and splitting the  $n$ - and  $m_n$ -dependencies within our telescoping sum (now letting  $0 \leq m_n \leq M_N$  be a rank parameter scaling with  $0 \leq n \leq N$  as  $M_N$  does with  $N$ ) across two separate sums. Thus, we proceed by defining

$$\begin{aligned} \tilde{\Delta}_N(n) &:= \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{n+1, m_{n+1}} - \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{n, m_{n+1}}, \\ \tilde{\Delta}_M(n) &:= \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{n, m_{n+1}} - \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{n, m_n}, \end{aligned}$$

and observing that

$$\tilde{F}_N(\lambda) = \frac{1}{NM_N} \left( \sum_{n=0}^{N-1} \tilde{\Delta}_N(n) + \sum_{n=0}^{N-1} \tilde{\Delta}_M(n) \right); \quad (2.15)$$

the upshot here is that each of the summands in the above is a finite difference of two terms that differ only in one index. Simplifying the right-hand side of equation (2.15), as detailed in Section 5.1, then leads to the following identity bounding the limit supremum of the right-hand side of equation (2.15) by a tractable convex combination.

**Theorem 2.9** (Multiscale Aizenman–Sims–Starr scheme). *Letting  $M = M_N \in \mathbb{N}$  be an increasing sequence, there exist some  $0 \leq \alpha, \beta \leq 1$  such that*

$$\limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda) \leq \alpha \limsup_{N \rightarrow \infty} \frac{\Delta_N}{M} + (1 - \alpha) \limsup_{N \rightarrow \infty} \frac{\Delta_M}{N}, \quad (2.16)$$

$$\liminf_{N \rightarrow \infty} \tilde{F}_N(\lambda) \geq \beta \liminf_{N \rightarrow \infty} \frac{\Delta_N}{M} + (1 - \beta) \liminf_{N \rightarrow \infty} \frac{\Delta_M}{N}, \quad (2.17)$$

where

$$\Delta_N := \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{N+1, M_{N+1}} - \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{N, M_{N+1}}, \quad (2.18)$$

$$\Delta_M := \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{N, M_{N+1}} - \mathbb{E}_{\mathbf{Z}, \mathbf{X}_{0,\varepsilon}} \ln \tilde{Z}_{N, M_N}. \quad (2.19)$$

Furthermore, if  $M = \lfloor f(N) \rfloor$  for some differentiable, strictly increasing function (e.g., a monotonic cubic interpolation of the strictly increasing subsequence of  $M_N$

such that  $M_N = M_{N-1} + 1$ )  $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , then we have

$$\liminf_{N \rightarrow \infty} \frac{f(N-1)}{f(N) + Nf'(N)} \leq \beta \leq \alpha \leq \limsup_{N \rightarrow \infty} \frac{f(N)}{f(N) + Nf'(N)}. \quad (2.20)$$

**Remark 2.10.** (1) For  $\gamma \in (0, 1]$ , the bounds of Theorem 2.9 hold with  $\alpha = \beta = (1 + \gamma)^{-1}$  for sublinear polynomial growth  $f(N) = N^\gamma$ , while if  $f(N) = (\ln N)^\gamma$ , we instead have  $\alpha = \beta = 1$ . Taking  $\gamma \rightarrow 0$  recovers the usual Aizenman–Sims–Starr scheme.

(2) Our interest is primarily in the bound (2.16), but inequality (2.17) shows that these results are sharp in the sense that if the limits on the right-hand sides exist and  $\alpha = \beta$ , then we have the equality

$$\lim_{N \rightarrow \infty} \tilde{F}_N(\lambda) = \alpha \lim_{N \rightarrow \infty} \frac{\Delta_N}{M} + (1 - \alpha) \lim_{N \rightarrow \infty} \frac{\Delta_M}{N}.$$

(3) In contrast to the Bayes-optimal setting, in the usual mixed  $p$ -spin models (see, for example, [56]) or mismatched inference problems [31], the upper bound is typically proved by interpolation and the lower bound is proved using the cavity method. In particular, inequality (2.17) will be the relevant bound for the cavity computations when studying growing rank versions of these models.

We have stated here the multiscale Aizenman–Sims–Starr scheme in terms of the (perturbed) free entropy for convenience, but it actually holds for *any* generic sequence with two growing indices. Indeed, it is a general fact about multiscale sequences and uses no properties of the log partition function, as can be seen from the proof presented in Section 5.1. The multiscale cavity method is a powerful tool because the limit of a two-indices sequence with non-linear asymptotic growth rate can be computed by finding the limits of differences where one of the coordinates is *fixed*, instead of having to deal with increments in both coordinates simultaneously. In other words, the  $\Delta_N$  and  $\Delta_M$  can be computed independently of one another to isolate the effects of adding one spin or one rank coordinate. In statistical physics parlance, this reduces the problem to those of computing a cavity in the spins for fixed rank dimension and a cavity in the rank for fixed spin dimension in the thermodynamic limit. The  $\Delta_N$  is the standard cavity studied in classical models, while the  $\Delta_M$  term only appears when examining multiscale models.

With Theorem 2.9 in hand, it remains to bound the limit suprema in the right-hand side of inequality (2.16) and compute their convex combination. In Section 5.2, we confirm via the standard fixed-rank cavity computations therein that

$$\limsup_{N \rightarrow \infty} \frac{\Delta_N}{M}$$

is bounded above by the right-hand side of inequality (2.12). Remarkably, thanks to the rank-one reduction described in Theorem 2.6, it turns out that the novel  $\Delta_M$  cavity term also obeys the same bound, hence so too does the right-hand side of inequality (2.16), as desired – we expect this observation to extend to the full sublinear-rank regime  $M = o(N)$ . This drastic simplification is however not expected to hold in the regime of extensive rank  $M = \Theta(N)$ . In that setting, Theorem 2.6 can no longer be used, so the right-hand sides of Propositions 2.7 and 2.8 become more complicated and cease to match. Thus, more refined tools are needed to derive extensive-rank counterparts of said propositions.

In the present sublinear-rank setting, the rank-one equivalence of Theorem 2.6 and a bootstrapping argument suggests to write the heuristic

$$\begin{aligned} \frac{\Delta_M}{N} &= \frac{1}{N} \mathbb{E}_{\mathbf{Z}, \mathbf{X}_0, \varepsilon} \ln \tilde{Z}_{N, M_{N+1}} - \frac{1}{N} \mathbb{E}_{\mathbf{Z}, \mathbf{X}_0, \varepsilon} \ln \tilde{Z}_{N, M_N} \\ &\approx (M_N + 1) \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) - M_N \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) \\ &= \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) \end{aligned} \quad (2.21)$$

for  $N$  such that  $M_{N+1} = M_N + 1$ . In Section 5.3, we present some manipulations involving Proposition 2.7 and the fixed-rank  $\Delta_N$  cavity bound to prove that the intuition behind approximation (2.21) is the correct one. As mentioned above, said  $\Delta_N$  cavity bound is obtained in Section 5.2 using standard fixed-rank arguments ([3], [43, Sections 4.3 and 6.2.5]). The idea, then, is to compare the rank  $M$ , perturbed, spiked Wigner model of size  $N + 1$  to that of size  $N$ , where we now fix our choice of perturbation to be

$$H_{N, \varepsilon}(\mathbf{X}) := \text{Tr} \left( \sqrt{\varepsilon} \tilde{\mathbf{Z}} \mathbf{X}^\top + \varepsilon \mathbf{X}_0 \mathbf{X}^\top - \frac{\varepsilon}{2} \mathbf{X} \mathbf{X}^\top \right), \quad \varepsilon \in [s_N, 2s_N], \quad (2.22)$$

with  $\tilde{\mathbf{Z}} \in \mathbb{R}^{N \times M}$  a standard Ginibre matrix consisting of i.i.d. standard Gaussian entries independent of everything else and  $(s_N)_{N \geq 1}$  a sequence of numbers in  $(0, 1)$  that decrease slowly to zero as  $N \rightarrow \infty$ . For the size  $N + 1$  model, denoting the  $(N + 1)$ th row of the signal by  $\boldsymbol{\eta}_0 \sim \mathbb{P}_{\mathbf{X}, M} = \mathbb{P}_{\mathbf{X}}^{\otimes M}$  and that of the noise by  $(\boldsymbol{\xi}, \xi)$  with  $\boldsymbol{\xi} \sim \mathcal{N}(0, I_N)$  and  $\xi \sim \mathcal{N}(0, 1)$  (the first  $N$  rows being  $\mathbf{X}_0$  and  $(\mathbf{Z}, \boldsymbol{\xi}^\top)$ , respectively) shows that

$$\begin{aligned} \tilde{H}_{N+1, \varepsilon_{N+1}}(\mathbf{X}, \boldsymbol{\eta}) &= H'_N(\mathbf{X}) + H_{N, \varepsilon_{N+1}}(\mathbf{X}) \\ &\quad + \boldsymbol{\eta} z(\mathbf{X}) + \boldsymbol{\eta} s(\mathbf{X}) \boldsymbol{\eta}^\top + m(\boldsymbol{\eta}) + r_{\varepsilon_{N+1}}(\boldsymbol{\eta}), \end{aligned} \quad (2.23)$$

$$\begin{aligned} H'_N(\mathbf{X}) &:= \frac{1}{2} \text{Tr} \left( \sqrt{\frac{\lambda}{N+1}} \mathbf{Z} \mathbf{X} \mathbf{X}^\top + \frac{\lambda}{N+1} \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X} \mathbf{X}^\top \right. \\ &\quad \left. - \frac{\lambda}{2(N+1)} \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \right), \end{aligned}$$

$$\begin{aligned}
 z(\mathbf{X}) &:= \sqrt{\frac{\lambda}{N+1}} \mathbf{X}^\top \boldsymbol{\xi}^\top + \frac{\lambda}{N+1} \mathbf{X}^\top \mathbf{X}_0 \boldsymbol{\eta}_0^\top, \\
 s(\mathbf{X}) &:= -\frac{\lambda}{2(N+1)} \mathbf{X}^\top \mathbf{X}, \\
 m(\boldsymbol{\eta}) &:= \frac{1}{2} \left( \sqrt{\frac{\lambda}{N+1}} \boldsymbol{\xi} + \frac{\lambda}{N+1} \boldsymbol{\eta}_0 \boldsymbol{\eta}_0^\top - \frac{\lambda}{2(N+1)} \boldsymbol{\eta} \boldsymbol{\eta}^\top \right) \boldsymbol{\eta} \boldsymbol{\eta}^\top, \\
 r_{\varepsilon_{N+1}}(\boldsymbol{\eta}) &:= \sqrt{\varepsilon_{N+1}} \tilde{\boldsymbol{\xi}} \boldsymbol{\eta}^\top + \varepsilon_{N+1} \boldsymbol{\eta}_0 \boldsymbol{\eta}^\top - \frac{\varepsilon_{N+1}}{2} \boldsymbol{\eta} \boldsymbol{\eta}^\top,
 \end{aligned}$$

where we have made explicit the  $N$ -dependence of  $\varepsilon = \varepsilon_N$  and denoted the  $(N+1)$ th row of the perturbation noise in  $H_{N+1, \varepsilon_{N+1}}(\mathbf{X}, \boldsymbol{\eta})$  by  $\tilde{\boldsymbol{\xi}}$ . The point of decomposing  $\tilde{H}_{N+1, \varepsilon_{N+1}}(\mathbf{X}, \boldsymbol{\eta})$  into the form (2.23) is that one may similarly observe for

$$\hat{y}(\mathbf{X}) := \frac{1}{2} \text{Tr} \left( \frac{\sqrt{\lambda}}{N} \hat{\mathbf{Z}} \mathbf{X} \mathbf{X}^\top + \frac{\lambda}{N^2} \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X} \mathbf{X}^\top - \frac{\lambda}{2N^2} \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \right),$$

with  $\hat{\mathbf{Z}}$  an independent copy of  $\mathbf{Z}$ , that  $\tilde{H}_{N, \varepsilon_N}(\mathbf{X})$  is statistically equivalent (up to some lower order terms) to

$$H'_N(\mathbf{X}) + H_{N, \varepsilon_N}(\mathbf{X}) + \hat{y}(\mathbf{X})$$

in the sense that the associated partition functions are equal in distribution. Thus, the difference  $\Delta_N$  (see (2.18)) can be expressed in terms of averages of exponentials of the cavity fields  $z(\mathbf{X})$ ,  $s(\mathbf{X})$ ,  $m(\boldsymbol{\eta})$ ,  $r_{\varepsilon_{N+1}}(\boldsymbol{\eta})$ ,  $\hat{y}(\mathbf{X})$ , and  $H_{N, \varepsilon_{N+1}}(\mathbf{X}) - H_{N, \varepsilon_N}(\mathbf{X})$  with respect to the Gibbs (posterior) average (writing  $\varepsilon_N$  as  $\varepsilon$ )

$$\langle \cdot \rangle'_{N, \varepsilon} := \frac{1}{Z'_{N, \varepsilon}} \int_{\mathbb{R}^{N \times M}} \cdot e^{H'_N(\mathbf{X}) + H_{N, \varepsilon}(\mathbf{X})} d\mathbb{P}_{X, N, M}(\mathbf{X}), \quad (2.24)$$

$$Z'_{N, \varepsilon} := \int_{\mathbb{R}^{N \times M}} e^{H'_N(\mathbf{X}) + H_{N, \varepsilon}(\mathbf{X})} d\mathbb{P}_{X, N, M}(\mathbf{X}). \quad (2.25)$$

In Section 5.2, we exploit the relative tractability of our cavity fields and the fact that the noise within them is independent of the noise in  $H'_N(\mathbf{X}) + H_{N, \varepsilon_N}(\mathbf{X})$  to control certain rates of convergence and ultimately deduce that

$$\limsup_{N \rightarrow \infty} \frac{\Delta_N}{M_N} \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda),$$

as desired. We emphasize here that although the techniques used in Section 5.2 are already in the literature, our contribution lies in our tracking the  $M$ -dependence of rates of convergence that arise throughout the computations. These become consequential in the thermodynamic limit when working in the sublinear-rank regime.

The final key ingredient for carrying out the cavity computation is the thermal concentration of the overlap matrix  $\mathbf{R}_{10}$  (see (2.8)). This result allows us to replace  $\mathbf{R}_{10}$  by the better behaved quantity  $\langle \mathbf{R}_{10} \rangle'_{N,\varepsilon}$  throughout the computations of Section 5.2. While it is sufficient for our purposes to prove thermal concentration of the overlap matrix with respect to the perturbed Hamiltonian  $H'_N(\mathbf{X}) + H_{N,\varepsilon}(\mathbf{X})$ , we are able to do so for a larger class of Hamiltonians and so believe our result to be a valuable tool for future studies of Bayes-optimal inference problems with factorized priors and growing rank.

**Theorem 2.11** (Thermal concentration of the overlap matrix). *Let  $\mathbb{P}_X$  be a centered distribution with bounded fourth moment, assume hypothesis (H2), let  $H''_N(\mathbf{X})$  denote the log-likelihood of a fully factorized channel*

$$\begin{aligned} \mathbf{Y}'' &\sim \mathbb{P}_{\text{out}}(\cdot \mid \mathbf{X}_0 \mathbf{X}_0^\top), \\ \mathbb{P}_{\text{out}}(\mathbf{Y}'' \mid \mathbf{X}_0 \mathbf{X}_0^\top) &= \prod_{i=1}^N \mathbb{P}_{\text{out}}^{(i=j)}(\mathbf{Y}''_{ii} \mid \mathbf{x}_{0,i}^\top \mathbf{x}_{0,i}) \\ &\quad \times \prod_{1 \leq i < j \leq N} \mathbb{P}_{\text{out}}^{(i < j)}(\mathbf{Y}''_{ij} \mid \mathbf{x}_{0,i}^\top \mathbf{x}_{0,j}), \end{aligned} \quad (2.26)$$

with general conditional probability distributions  $\mathbb{P}_{\text{out}}^{(i=j)}$ ,  $\mathbb{P}_{\text{out}}^{(i < j)}$  sufficiently integrable, and define  $\langle \cdot \rangle'_{N,\varepsilon}$  to be the Gibbs average corresponding to the perturbed Hamiltonian

$$\tilde{H}''_{N,\varepsilon}(\mathbf{X}) := H''_N(\mathbf{X}) + H_{N,\varepsilon}(\mathbf{X}), \quad (2.27)$$

à la equation (2.24). Then, there exists a finite positive constant  $C$  independent of  $M$ ,  $N$  and depending only on properties of  $\mathbb{P}_X$  such that

$$\frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E} \langle \|\mathbf{R}_{10} - \langle \mathbf{R}_{10} \rangle'_{N,\varepsilon}\|_{\mathbb{F}}^2 \rangle'_{N,\varepsilon} d\varepsilon \leq \Gamma(N, M) := \frac{CM^2}{\sqrt{Ns_N}},$$

where the expectation  $\mathbb{E}[\cdot]$  is taken over  $\tilde{\mathbf{Z}}, \mathbf{X}_0$  and the randomness in  $\mathbb{P}_{\text{out}}(\cdot \mid \mathbf{X}_0 \mathbf{X}_0^\top)$ .

The proof of this theorem, given in Section 4.2, relies on the presence of the perturbation Hamiltonian  $H_{N,\varepsilon}(\mathbf{X})$  (see (2.22)) in the definition of  $\langle \cdot \rangle'_{N,\varepsilon}$ . This perturbation corresponds to side information coming from the matrix Gaussian channel  $\sqrt{\varepsilon} \mathbf{X}_0 + \tilde{\mathbf{Z}}$ , which serves to temper singularities of  $\mathbb{E} \langle \|\mathbf{R}_{10} - \langle \mathbf{R}_{10} \rangle'_{N,\varepsilon}\|_{\mathbb{F}}^2 \rangle'_{N,\varepsilon}$  that otherwise prevent concentration. The necessity of this side information and the novelty of the proof of Theorem 2.11 is discussed in further detail in Section 4.2. A tool used throughout said proof and indeed this entire paper is the *Nishimori identity* or the tower rule for conditional expectation (see, e.g., [43]), which in our setting states that for continuous, bounded functions  $g$ ,

$$\mathbb{E}_{\mathbf{Z}, \mathbf{X}_0} \langle g(\mathbf{Y}, \mathbf{X}_1, \dots, \mathbf{X}_k) \rangle = \mathbb{E}_{\mathbf{Z}, \mathbf{X}_0} \langle g(\mathbf{Y}, \mathbf{X}_0, \mathbf{X}_2, \dots, \mathbf{X}_k) \rangle, \quad k \in \mathbb{N},$$

where  $X_1, \dots, X_k$  are i.i.d. samples from the posterior  $\mathbb{P}_{X|Y,N,M}(X | Y(X_0, \mathbf{Z}))$  and  $\langle \cdot \rangle$  denotes the Gibbs average with respect to the product conditional distribution

$$\prod_{i=1}^k \mathbb{P}_{X|Y,N,M}(X_i | Y(X_0, \mathbf{Z})).$$

Henceforth, when dealing with multiple conditionally independent samples from a Bayes-posterior distribution, we will write  $\mathbf{X} = X_1$  for the first sample and the Gibbs average will refer to an average over all samples.

### 3. Relating the rank- $M$ variational formula to its rank-one analog

This section concerns only the replica symmetric potential  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  (see (2.2)), so there is no  $N$  parameter nor large  $N$  limit to consider. Hence,  $M$  is a fixed, finite parameter throughout this section and, consequently, we have access to the literature on finite- $M$  analysis. In particular, we are able to benefit from characterizations of maximizers  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  given in [43]. It turns out that the growing rank considered in this paper only comes into play when we link the free entropy  $F_N(\lambda)$  (see (1.2)) of the sublinear-rank spiked Wigner model (1.1) to the replica symmetric potential. The upshot here is that since the supremum of the rank- $M$  replica symmetric potential is equivalent to its rank-one analog, many complications arising from the growing rank can eventually be circumvented.

#### 3.1. Information-theoretic inequalities on worst Gaussian noise

We now present the proofs of Lemma 2.3 and Corollary 2.4 given in Section 2.2. The first follows from the definitions and standard properties of mutual information and conditional (differential) entropy (see, e.g., [21]), the chain rule for mutual information, the fact that the signal  $\mathbf{x}_0$  and noise  $\mathbf{z}$  have i.i.d. entries, and from the rotational invariance of standard Gaussian vectors. The corollary follows from a convexity property [66] of the mutual information whose use is enabled by our prior having  $D$ -bounded support.

*Proof of Lemma 2.3.* We recall that we take  $\mathbb{P}_{X,M}$  to have factorized form  $\mathbb{P}_X^{\otimes M}$ ,  $\mathbf{x}_0 \sim \mathbb{P}_{X,M}$ ,  $x_0 \sim \mathbb{P}_X$ ,  $\mathbf{z} \sim \mathcal{N}(0, I_M)$ ,  $z \sim \mathcal{N}(0, 1)$ , and  $\Sigma \in \mathcal{S}_M$ . Applying the chain rule for mutual information twice shows that

$$\begin{aligned} I(\mathbf{x}_0; \mathbf{x}_0 + \Sigma^{1/2}\mathbf{z}) &= \sum_{i=1}^M I(\mathbf{x}_{0,i}; \mathbf{x}_0 + \Sigma^{1/2}\mathbf{z} | \{\mathbf{x}_{0,k}\}_{k < i}) \\ &= \sum_{i,j=1}^M I(\mathbf{x}_{0,i}; \mathbf{x}_{0,j} + (\Sigma^{1/2}\mathbf{z})_j | \{\mathbf{x}_{0,k}\}_{k < i}, \{\mathbf{x}_{0,k'} + (\Sigma^{1/2}\mathbf{z})_{k'}\}_{k' < j}). \end{aligned}$$

Thus, by the non-negativity of mutual information, we can discard all terms for which  $i \neq j$  to see that

$$\begin{aligned} I(\mathbf{x}_0; \mathbf{x}_0 + \Sigma^{1/2} \mathbf{z}) &\geq \sum_{i=1}^M I(\mathbf{x}_{0,i}; \mathbf{x}_{0,i} + (\Sigma^{1/2} \mathbf{z})_i \mid \{\mathbf{x}_{0,k} \mathbf{x}_{0,k} + (\Sigma^{1/2} \mathbf{z})_k\}_{k < i}) \\ &= \sum_{i=1}^M I(\mathbf{x}_{0,i}; \mathbf{x}_{0,i} + (\Sigma^{1/2} \mathbf{z})_i \mid \{(\Sigma^{1/2} \mathbf{z})_k\}_{k < i}). \end{aligned} \quad (3.1)$$

The second line follows from the fact that knowledge of  $\mathbf{x}_{0,k}$  and  $\mathbf{x}_{0,k} + (\Sigma^{1/2} \mathbf{z})_k$  is equivalent to that of  $\mathbf{x}_{0,k}$  and  $(\Sigma^{1/2} \mathbf{z})_k$ , but also that the mutual information between  $\mathbf{x}_{0,i}$  and  $\mathbf{x}_{0,i} + (\Sigma^{1/2} \mathbf{z})_i$  is independent of  $\{\mathbf{x}_{0,k}\}_{k < i}$  due to  $\mathbf{x}_0$  having i.i.d. entries.

Now, notice that for a generic random variable  $X$  independent of  $Z_1, Z_2$ , but with  $Z_1$  and  $Z_2$  possibly depending on each other in an arbitrary way,

$$\begin{aligned} I(X; X + Z_1 \mid Z_2) &= H(X \mid Z_2) - H(X \mid X + Z_1, Z_2) \\ &\geq H(X) - H(X \mid X + Z_1) \\ &= I(X; X + Z_1). \end{aligned}$$

Here,  $H(X)$  denotes the (differential) entropy of  $X$  and we have used the facts that  $H(X \mid Z_2) = H(X)$  and  $H(X \mid X + Z_1, Z_2) \leq H(X \mid X + Z_1)$ . Applying this inequality with  $X = \mathbf{x}_{0,i}$ ,  $Z_1 = (\Sigma^{1/2} \mathbf{z})_i$ , and  $Z_2 = \{(\Sigma^{1/2} \mathbf{z})_k\}_{k < i}$  to the summand in the right-hand side of inequality (3.1) then shows that

$$I(\mathbf{x}_0; \mathbf{x}_0 + \Sigma^{1/2} \mathbf{z}) \geq \sum_{i=1}^M I(\mathbf{x}_{0,i}; \mathbf{x}_{0,i} + (\Sigma^{1/2} \mathbf{z})_i). \quad (3.2)$$

Since  $\Sigma \in \mathcal{S}_M$ , we may diagonalize it as  $\Sigma = \mathbf{O} \boldsymbol{\sigma} \mathbf{O}^\top$  with  $\boldsymbol{\sigma} = \text{diag}(\sigma_1, \dots, \sigma_M)$  being the diagonal matrix of eigenvalues of  $\Sigma$  and  $\mathbf{O}$  being the corresponding real orthogonal matrix of eigenvectors. Then, due to properties of standard Gaussian vectors, we have in law that

$$(\Sigma^{1/2} \mathbf{z})_i = (\mathbf{O} \boldsymbol{\sigma}^{1/2} \mathbf{O}^\top \mathbf{z})_i \stackrel{d}{=} (\mathbf{O} \boldsymbol{\sigma}^{1/2} \mathbf{z})_i \stackrel{d}{=} \left( \sum_{j=1}^M \mathbf{O}_{ij}^2 \sigma_j \right)^{1/2} z = \sqrt{\Sigma_{ii}} z.$$

Observing also that  $\mathbf{x}_{0,i} \stackrel{d}{=} x_0$ , inequality (3.2) reduces to the sought form (2.10).

As  $\mathbf{z}, \mathbf{x}_0$  have i.i.d. entries, it is immediate from properties of mutual information that inequality (2.10) turns into an equality upon setting

$$\Sigma = \text{diag}(\Sigma_{11}, \dots, \Sigma_{MM}). \quad \blacksquare$$

*Proof of Corollary 2.4.* We remain in the setting of the above proof, but further recall that  $\mathbb{P}_X$  has  $D$ -bounded support and that  $\Sigma \in \mathcal{S}_M$  is such that  $\Sigma_{ii} \geq D^2$  for each  $1 \leq i \leq M$ . Then, setting  $t = \Sigma_{ii}$  in [66, Theorem 2] shows that  $I(x_0; x_0 + \sqrt{\Sigma_{ii}}z)$  is a convex function of  $\Sigma_{ii}$  in the  $\Sigma_{ii} \geq D^2$  regime. Thus, applying Jensen's inequality to the right-hand side of inequality (2.10) yields

$$\begin{aligned} I(x_0; x_0 + \Sigma^{1/2}z) &\geq \sum_{i=1}^M I(x_0; x_0 + \sqrt{\Sigma_{ii}}z) \\ &\geq MI \left( x_0; x_0 + \left( \frac{1}{M} \sum_{i=1}^M \Sigma_{ii} \right)^{1/2} z \right), \end{aligned}$$

which is seen to be equivalent to inequality (2.11) upon recalling that  $\sigma$  denotes the normalized trace  $\text{Tr } \Sigma / M$ .

It is again immediate that inequality (2.11) is an equality when  $\Sigma = \sigma I_M$ .  $\blacksquare$

### 3.2. Properties of maximizers of the replica symmetric potential

We now give a series of lemmas that amount to Proposition 2.5. We recall that this proposition lists properties of maximizers  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  (see (2.2)) that will be useful in the following subsection. To begin, it is essential to observe that we are allowed to restrict our search for maximizers to solutions of a certain criticality condition; note that although it is necessary for  $\mathbf{Q}^*(\lambda)$  to satisfy the following condition, said condition is not sufficient for completely characterizing  $\mathbf{Q}^*(\lambda)$ .

**Lemma 3.1** (A criticality condition). *Diagonalize  $\mathbf{Q} \in \mathcal{S}_M$  as  $\mathbf{Q} = \mathbf{O}\mathbf{q}\mathbf{O}^\top$  with  $\mathbf{q} = \text{diag}(q_1, \dots, q_M)$  being the diagonal matrix of eigenvalues of  $\mathbf{Q}$  and  $\mathbf{O}$  being the corresponding real orthogonal matrix of eigenvectors. Then, any maximizer  $\mathbf{Q}$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  must be such that the following system of equations, solved in particular by all critical points of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$ , is satisfied:*

$$\sqrt{q_i} (\mathbf{O}_i^\top \mathbb{E}_{z, x_0} \langle \mathbf{x} \mathbf{x}_0^\top \rangle_{\text{RS}} \mathbf{O}_i - q_i) = 0, \quad 1 \leq i \leq M, \quad (3.3)$$

where  $\mathbf{O}_i$  is the  $i$ th column of  $\mathbf{O}$  and we define the Gibbs average associated with the replica symmetric potential (2.2) by

$$\langle \cdot \rangle_{\text{RS}} := \frac{1}{Z_M^{\text{RS}}} \int \cdot e^{\sqrt{\lambda} \mathbf{x}^\top \sqrt{\mathbf{Q}} z + \lambda \mathbf{x}_0^\top \mathbf{Q} \mathbf{x} - \frac{\lambda}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x}} d\mathbb{P}_{X, M}(\mathbf{x}). \quad (3.4)$$

In particular, if  $q_1, \dots, q_M > 0$ , this system simplifies to the fixed-point equation

$$\mathbf{Q} = \mathbb{E}_{z, x_0} \langle \mathbf{x} \mathbf{x}_0^\top \rangle_{\text{RS}}.$$

*Proof.* Let  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_M)$  and interpret  $F_M^{\text{RS}}(\boldsymbol{Q}, \lambda)$  as a function of  $\sqrt{\boldsymbol{q}}$  at fixed eigenvectors. Then, defining  $\boldsymbol{Q}_{\boldsymbol{\gamma}}$  such that  $\sqrt{\boldsymbol{Q}_{\boldsymbol{\gamma}}} = \boldsymbol{O}(\sqrt{\boldsymbol{q}} + \text{diag } \boldsymbol{\gamma})\boldsymbol{O}^\top$ , we have that the  $\boldsymbol{\gamma}$ -directional  $\sqrt{\boldsymbol{q}}$ -derivative of the replica symmetric potential is given by

$$\begin{aligned} \nabla_{\boldsymbol{\gamma}} F_M^{\text{RS}}(\boldsymbol{Q}, \lambda) &= \sum_{i=1}^M \gamma_i \lim_{\boldsymbol{\gamma} \rightarrow \mathbf{0}} \frac{\partial}{\partial \gamma_i} F_M^{\text{RS}}(\boldsymbol{Q}_{\boldsymbol{\gamma}}, \lambda) \\ &= \frac{1}{M} \sum_{i=1}^M \gamma_i \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle 2\lambda \sqrt{q_i} \boldsymbol{x}_0^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x} - \lambda \sqrt{q_i} \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x} \rangle_{\text{RS}} \\ &\quad + \frac{1}{M} \sum_{i=1}^M \gamma_i (\mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle \sqrt{\lambda} \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{z} \rangle_{\text{RS}} - \lambda q_i^{3/2}). \end{aligned} \quad (3.5)$$

Gaussian integration by parts shows that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle \sqrt{\lambda} \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{z} \rangle_{\text{RS}} &= \sum_{j=1}^M \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \frac{\partial}{\partial z_j} \langle \sqrt{\lambda} (\boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x})_j \rangle_{\text{RS}} \\ &= \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} (\langle \lambda \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \sqrt{\boldsymbol{Q}} \boldsymbol{x} \rangle_{\text{RS}} - \langle \lambda \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \sqrt{\boldsymbol{Q}} \rangle_{\text{RS}} \langle \boldsymbol{x} \rangle_{\text{RS}}) \\ &= \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle \lambda \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \sqrt{\boldsymbol{Q}} \boldsymbol{x} - \lambda \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \sqrt{\boldsymbol{Q}} \boldsymbol{x}_2 \rangle_{\text{RS}} \\ &= \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle \lambda \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \sqrt{\boldsymbol{Q}} \boldsymbol{x} - \lambda \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \sqrt{\boldsymbol{Q}} \boldsymbol{x}_0 \rangle_{\text{RS}} \\ &= \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle \lambda \sqrt{q_i} \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x} - \lambda \sqrt{q_i} \boldsymbol{x}^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x}_0 \rangle_{\text{RS}}, \end{aligned}$$

where in the third line we have written  $\boldsymbol{x}_2$  for a sample, conditionally independent of  $\boldsymbol{x}$ , from the posterior distribution of the Gibbs average (3.4) and redefined said Gibbs average to be over both samples, while in the fourth line we have used the Nishimori identity to replace  $\boldsymbol{x}_2$  with  $\boldsymbol{x}_0$ ; the fifth line follows from recalling that  $\boldsymbol{O}_i$  is an eigenvector of  $\sqrt{\boldsymbol{Q}}$  with eigenvalue  $\sqrt{q_i}$ . Substituting this into the right-hand side of equation (3.5) then shows that

$$\begin{aligned} \nabla_{\boldsymbol{\gamma}} F_M^{\text{RS}}(\boldsymbol{Q}, \lambda) &= \frac{1}{M} \sum_{i=1}^M \gamma_i (\mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle \lambda \sqrt{q_i} \boldsymbol{x}_0^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x} \rangle_{\text{RS}} - \lambda q_i^{3/2}) \\ &= \frac{\lambda}{M} \sum_{i=1}^M \gamma_i \sqrt{q_i} (\boldsymbol{O}_i^\top \mathbb{E}_{\boldsymbol{z}, \boldsymbol{x}_0} \langle \boldsymbol{x} \boldsymbol{x}_0^\top \rangle_{\text{RS}} \boldsymbol{O}_i - q_i), \end{aligned} \quad (3.6)$$

where we have used the fact that  $\boldsymbol{x}_0^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x} = \text{Tr } \boldsymbol{x}_0^\top \boldsymbol{O}_i \boldsymbol{O}_i^\top \boldsymbol{x}$  along with the cyclic property of the trace. Equation (3.3) follows from noting that all critical points  $\boldsymbol{Q}$  of  $F_M^{\text{RS}}(\boldsymbol{Q}, \lambda)$  must be such that  $\nabla_{\boldsymbol{\gamma}} F_M^{\text{RS}}(\boldsymbol{Q}, \lambda) = 0$  for all choices of  $\boldsymbol{\gamma}$ , including those for which all but one  $\gamma_i$  are zero.

To discard the possibility of maximizers on the boundary of  $\mathcal{S}_M$  that do not satisfy the system of equations (3.3), note that being a boundary point means that at least one

of the  $q_i$ , say  $q_1$ , must vanish or tend to infinity. In the latter case, one may check from the form (2.9) of the replica symmetric potential that  $\lim_{q_1 \rightarrow \infty} F_M^{\text{RS}}(\mathbf{Q}, \lambda) = -\infty$ . In the former case, we assume for the sake of contradiction that there is at least one  $q_i \neq 0$  with  $i > 1$ , say  $q_2$ , that does not satisfy equation (3.3). Then, we note from equation (3.6) that  $\nabla_{(1,1,0,\dots,0)} F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  and  $\nabla_{(1,-1,0,\dots,0)} F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  are non-zero and disagree in sign and, therefore, such a  $\mathbf{Q}$  is at best a saddle point and cannot be a maximizer. Thus, for  $\mathbf{Q}$  to be a maximizer of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  with some eigenvalues vanishing, the other eigenvalues must satisfy equation (3.3). ■

With this criticality condition in hand, we may now readily address the properties of the maximizers  $\mathbf{Q}^*(\lambda)$  listed in Proposition 2.5.

**Lemma 3.2** (Eigenvalues of  $\mathbf{Q}^*(\lambda)$  are bounded). *Assume hypothesis (H2), let  $\mathbb{P}_X$  have bounded second moment  $\rho = \mathbb{E}_{\mathbb{P}_X} X^2$ , and let  $q_1^*, \dots, q_M^*$  be the eigenvalues of a maximizer  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$ . Then, for all  $\lambda \geq 0$ ,*

$$0 \leq q_1^*, \dots, q_M^* \leq \rho.$$

*Proof.* As a result of Lemma 3.1, each eigenvalue  $q_i$  ( $1 \leq i \leq M$ ) of a maximizer  $\mathbf{Q}$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  must either be zero or equal  $\mathbf{O}_i^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \langle \mathbf{x} \mathbf{x}_0^\top \rangle_{\text{RS}} \mathbf{O}_i$ , where  $\mathbf{O}_i$  is the eigenvector of  $\mathbf{Q}$  corresponding to  $q_i$ . In the latter case, we see that

$$\begin{aligned} q_i &= \mathbf{O}_i^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \langle \mathbf{x} \mathbf{x}_0^\top \rangle_{\text{RS}} \mathbf{O}_i \\ &= \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \langle \mathbf{O}_i^\top \mathbf{x} (\mathbf{O}_i^\top \mathbf{x}_0) \rangle_{\text{RS}} \\ &= \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \langle \mathbf{O}_i^\top \mathbf{x} \rangle_{\text{RS}}^2 \\ &\leq \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \langle (\mathbf{O}_i^\top \mathbf{x})^2 \rangle_{\text{RS}} \\ &= \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} (\mathbf{O}_i^\top \mathbf{x}_0)^2 \\ &= \mathbf{O}_i^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} [\mathbf{x}_0 \mathbf{x}_0^\top] \mathbf{O}_i \\ &= \rho. \end{aligned}$$

Here, the second and sixth lines follow from the fact that  $\mathbf{x}_0^\top \mathbf{O}_i = \mathbf{O}_i^\top \mathbf{x}_0$ , the third and fifth lines are due to the Nishimori identity, the fourth line follows from applying Jensen's inequality, and the final line follows by noting that since  $\mathbf{x}_0$  has i.i.d. entries of second moment  $\rho$ , we must have that  $\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} [\mathbf{x}_0 \mathbf{x}_0^\top] = \rho \mathbf{I}_M$ . Since  $\mathbf{Q}$  is positive semidefinite, its eigenvalues are bounded below by zero, so we are done. ■

**Lemma 3.3** (Continuity of  $\|\mathbf{Q}^*(\lambda)\|_{\text{F}}$ ). *Assume hypothesis (H4). Then, any maximizer  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  is such that  $\|\mathbf{Q}^*(\lambda)\|_{\text{F}}$  is continuous on  $[0, \infty) \setminus \{\lambda_c\}$ .*

*Proof.* It is immediate that the proof of the rank-one result [43, Proposition 17] extends to the rank- $M$  case to yield

$$\|\mathbf{Q}^*(\lambda)\|_{\text{F}}^2 = 4\phi'_M(\lambda),$$

where we recall that  $\phi_M(\lambda) = \sup_{\mathbf{Q} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{Q}, \lambda)$ . By hypothesis,  $\phi_M(\lambda)$  is real analytic on  $[0, \infty) \setminus \{\lambda_c\}$ , so  $\phi'_M(\lambda)$  and  $\|\mathbf{Q}^*(\lambda)\|_{\text{F}} = \sqrt{4\phi'_M(\lambda)}$  are likewise. ■

**Lemma 3.4** (Eigenvalues of  $\mathbf{Q}^*(\lambda)$  at low SNR). *Assume hypotheses (H2) and (H4), fix  $\rho'_S \in (0, \rho)$ , let  $\mathbb{P}_X$  be centered, and let  $q_1^*, \dots, q_M^*$  be the eigenvalues of a maximizer  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$ . Then, there exists  $0 < \lambda'_S < \lambda_c$  such that whenever  $0 \leq \lambda < \lambda'_S$ , we have*

$$0 \leq q_1^*, \dots, q_M^* < \rho'_S.$$

*Proof.* At  $\lambda = 0$ , there is no coupling between  $\mathbf{x}$  and  $\mathbf{x}_0$  in the definition (3.4) of  $\langle \cdot \rangle_{\text{RS}}$ , so one sees that

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \langle \mathbf{x} \mathbf{x}_0^\top \rangle_{\text{RS}} = \langle \mathbf{x} \rangle_{\text{RS}} \Big|_{\lambda=0} \mathbb{E}_{\mathbf{x}_0} \mathbf{x}_0^\top = \mathbf{0},$$

where we have used the fact that  $\mathbb{P}_X$  is centered. Thus, by the criticality condition of Lemma 3.1, we have that  $\mathbf{Q}^*(0)$ , hence  $\|\mathbf{Q}^*(0)\|_{\text{F}}$ , must be zero. Then, the continuity of  $\|\mathbf{Q}^*(\lambda)\|_{\text{F}}$  due to Lemma 3.3 may be used to establish existence of  $0 < \lambda'_S < \lambda_c$  such that  $0 \leq \|\mathbf{Q}^*(\lambda)\|_{\text{F}} < \rho'_S$  for all  $0 \leq \lambda < \lambda'_S$ . In terms of eigenvalues, this shows that

$$0 \leq (q_1^*)^2 + \dots + (q_M^*)^2 < (\rho'_S)^2,$$

so the largest value an eigenvalue can attain occurs when the remaining eigenvalues are zero and this value is bounded above by  $\rho'_S$ . ■

**Lemma 3.5** (Eigenvalues of  $\mathbf{Q}^*(\lambda)$  at high SNR). *Let  $\mathbb{P}_{X, M}$  be as in Proposition 2.5, fix  $\rho'_L \in (0, \rho)$ , let  $q_1^*, \dots, q_M^*$  be the eigenvalues of a maximizer  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$ . Then, there exists  $\lambda'_L > \lambda_c$  such that whenever  $\lambda > \lambda'_L$ ,*

$$\rho'_L < q_1^*, \dots, q_M^* \leq \rho.$$

*Proof.* We read from [43, Proposition 43] that for all  $\lambda \neq \lambda_c$ , the limiting minimum mean-square error (2.7) of the spiked Wigner model of fixed rank  $M$  is given by

$$\lim_{N \rightarrow \infty} \text{MMSE}_{N, M}(\lambda) = M\rho^2 - \|\mathbf{Q}^*(\lambda)\|_{\text{F}}^2.$$

It is well known that the left-hand side of this equation vanishes in the  $\lambda \rightarrow \infty$  limit, so we have that  $\lim_{\lambda \rightarrow \infty} \|\mathbf{Q}^*(\lambda)\|_{\text{F}}^2 = M\rho^2$ . Then, the continuity of  $\|\mathbf{Q}^*(\lambda)\|_{\text{F}}$  due to Lemma 3.3 says that there must exist  $\lambda'_L > \lambda_c$  such that

$$\|\mathbf{Q}^*(\lambda)\|_{\text{F}}^2 > (M-1)\rho^2 + (\rho'_L)^2$$

whenever  $\lambda > \lambda'_L$ . In terms of eigenvalues, this translates to

$$(q_1^*)^2 + \dots + (q_M^*)^2 > (M-1)\rho^2 + (\rho'_L)^2.$$

From Lemma 3.2, we see that the largest value an eigenvalue can take is  $\rho$ . When  $M - 1$  of the eigenvalues take this value, the above inequality shows that the remaining eigenvalue must still be bounded below by  $\rho'_L$ . ■

Combining Lemmas 3.2–3.5 produces Proposition 2.5, which enables us to constrain the eigenvalues  $q_1^*, \dots, q_M^*$  of suitable maximizers  $\mathbf{Q}^*(\lambda)$  of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  depending on whether we are working in the high or low SNR regime. Thus, we may now proceed to the proof of Theorem 2.6.

### 3.3. Rank-one reduction for high, low, and all SNR

Since  $\mathbf{Q} \in \mathcal{S}_M$ , so too is  $(\lambda \mathbf{Q})^{-1}$ , which we denote as  $\mathbf{\Sigma}$  from here on out. Rewriting equation (2.9) in terms of  $\mathbf{\Sigma}$  shows that

$$F_M^{\text{RS}}(\mathbf{Q}, \lambda) = -\frac{1}{M} I(\mathbf{x}_0; \mathbf{x}_0 + \mathbf{\Sigma}^{1/2} \mathbf{z}) - \frac{\lambda}{4M} \|\mathbf{\Sigma}^{-1}/\lambda - \rho I_M\|_{\text{F}}^2 + \frac{\lambda \rho^2}{4}. \quad (3.7)$$

Maximizing this expression over  $\mathbf{Q} \in \mathcal{S}_M$  is equivalent to the problem of minimizing

$$\tilde{F}_M^{\text{RS}}(\mathbf{\Sigma}, \lambda) := I(\mathbf{x}_0; \mathbf{x}_0 + \mathbf{\Sigma}^{1/2} \mathbf{z}) + \frac{\lambda}{4} \|\mathbf{\Sigma}^{-1}/\lambda - \rho I_M\|_{\text{F}}^2 \quad (3.8)$$

over  $\mathbf{\Sigma} \in \mathcal{S}_M$ . It is immediate that Lemma 2.3 and Corollary 2.4 induce decoupling in this expression when  $\mathbf{\Sigma}$  is diagonal or, more specifically, has the form  $\sigma I_M$ . Our goal now is to show that this is indeed the form of  $\mathbf{\Sigma}$  for which  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  is maximized. We first do so in the low and high SNR regimes separately. The separation of our task into these two regimes corresponds, via Proposition 2.5, to the tasks of maximizing over  $\mathbf{Q}$  with eigenvalues in, respectively,  $[0, \rho']$  or  $[\rho', \rho]$ , where we now set  $\rho' := 2\rho/3$ . This choice of  $\rho'$  is of crucial interest because the squared Frobenius norm in expression (3.8) has first and second order partial derivatives with respect to the eigenvalues  $\sigma_i$  ( $1 \leq i \leq M$ ) of  $\mathbf{\Sigma}$  given by

$$\frac{\partial}{\partial \sigma_i} \left\| \frac{\mathbf{\Sigma}^{-1}}{\lambda} - \rho I_M \right\|_{\text{F}}^2 = \frac{2(\lambda \rho \sigma_i - 1)}{\lambda^2 \sigma_i^3}, \quad (3.9)$$

$$\frac{\partial^2}{\partial \sigma_i^2} \left\| \frac{\mathbf{\Sigma}^{-1}}{\lambda} - \rho I_M \right\|_{\text{F}}^2 = \frac{2(3 - 2\lambda \rho \sigma_i)}{\lambda^2 \sigma_i^4}, \quad (3.10)$$

so one can see that the second order partial derivatives vanish exactly when the eigenvalues  $q_i = (\lambda \sigma_i)^{-1}$  of  $\mathbf{Q}$  equal  $\rho' = 2\rho/3$ . Thus, the gradient  $\nabla \|\mathbf{\Sigma}^{-1}/\lambda - \rho I_M\|_{\text{F}}^2$  is a bijective function of  $\sigma_1, \dots, \sigma_M$  on the domain  $\{\sigma_1, \dots, \sigma_M > (\lambda \rho')^{-1}\}$ , while on  $\{(\lambda \rho)^{-1} \leq \sigma_1, \dots, \sigma_M < (\lambda \rho')^{-1}\}$ , it is convex – the significance of these properties will soon become apparent.

Let us now give the partial statement of Theorem 2.6 valid for the low SNR regime.

**Lemma 3.6** (Low SNR rank-one reduction). *Assume hypotheses (H2)–(H4) and fix  $\rho' = 2\rho/3$ . Then, there exists  $0 < \lambda_S < \lambda_c$  such that whenever  $0 \leq \lambda < \lambda_S$ , we have*

$$\sup_{\mathbf{Q} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{Q}, \lambda) = \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) = \sup_{q \in [0, \rho']} F_1^{\text{RS}}(q, \lambda). \quad (3.11)$$

*Proof.* Let  $\mathbf{Q}^*(\lambda) \in \mathcal{S}_M$  be a maximizer of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  with eigenvalues  $q_1^*, \dots, q_M^*$ . Then, by Proposition 2.5, there exists  $0 < \lambda'_S < \lambda_c$  such that for all  $0 \leq \lambda < \lambda'_S$ ,

$$0 \leq q_1^*, \dots, q_M^* < \rho'.$$

Thus, the eigenvalues  $\sigma_i = (\lambda q_i^*)^{-1}$  ( $1 \leq i \leq M$ ) of the minimizer  $\mathbf{\Sigma} = (\lambda \mathbf{Q}^*(\lambda))^{-1}$  of  $\tilde{F}_M^{\text{RS}}(\mathbf{\Sigma}, \lambda)$  (see (3.8)) are such that

$$\sigma_1, \dots, \sigma_M > \frac{1}{\lambda \rho'} \quad (3.12)$$

whenever  $0 \leq \lambda < \lambda'_S$ . Letting  $\mathbf{O}_i$  be the eigenvector of  $\mathbf{\Sigma}$  corresponding to  $\sigma_i$  and setting  $\lambda_S := \min\{\lambda'_S, D^{-2}/\rho\}$ , we then see that taking  $0 \leq \lambda < \lambda_S \leq D^{-2}/\rho$  enforces the following lower bound on the diagonal entries of  $\mathbf{\Sigma}$ :

$$\Sigma_{ii} = \sum_{j=1}^M \mathbf{O}_{ij}^2 \sigma_j > \frac{1}{\lambda \rho'} \sum_{j=1}^M \mathbf{O}_{ij}^2 = \frac{1}{\lambda \rho'} > \frac{D^2 \rho}{\rho'} > D^2, \quad 1 \leq i \leq M.$$

Hence, by Corollary 2.4, the mutual information term in equation (3.8) satisfies

$$I(\mathbf{x}_0; \mathbf{x}_0 + \mathbf{\Sigma}^{1/2} \mathbf{z}) \geq MI(\mathbf{x}_0; \mathbf{x}_0 + \sqrt{\sigma} \mathbf{z}),$$

where we recall that  $\mathbf{x}_0 \sim \mathbb{P}_X$ ,  $\mathbf{z} \sim \mathcal{N}(0, 1)$ , and  $\sigma = \text{Tr } \mathbf{\Sigma} / M$ . Therefore, we have that

$$\begin{aligned} \inf_{\mathbf{\Sigma} \in \mathcal{S}_M} \tilde{F}_M^{\text{RS}}(\mathbf{\Sigma}, \lambda) &\geq \inf_{\sigma_1, \dots, \sigma_M \geq 0} \tilde{F}_{M, \text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda), \\ \tilde{F}_{M, \text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda) &:= MI(\mathbf{x}_0; \mathbf{x}_0 + \sqrt{\sigma} \mathbf{z}) + \frac{\lambda}{4} \sum_{i=1}^M \left( \frac{1}{\lambda \sigma_i} - \rho \right)^2, \end{aligned} \quad (3.13)$$

where  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_M)$ . Further writing  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_M)$  and using equation (3.9) with the chain rule shows that the  $\boldsymbol{\gamma}$ -directional  $\boldsymbol{\sigma}$ -derivative of  $\tilde{F}_{M, \text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda)$  is

$$\nabla_{\boldsymbol{\gamma}} \tilde{F}_{M, \text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda) = \sum_{i=1}^M \gamma_i (\iota(\boldsymbol{\sigma}) + \nu(\sigma_i)),$$

where we define

$$\iota(\boldsymbol{\sigma}) := \frac{d}{d\boldsymbol{\sigma}} I(\mathbf{x}_0; \mathbf{x}_0 + \sqrt{\sigma} \mathbf{z}), \quad (3.14)$$

$$\nu(\sigma_i) := \frac{\lambda \rho \sigma_i - 1}{2\lambda \sigma_i^3}. \quad (3.15)$$

Hence, the critical points of  $\tilde{F}_{M,\text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda)$  occur at

$$\sigma_i = \nu^{-1} \circ \iota(\sigma), \quad 1 \leq i \leq M, \quad (3.16)$$

where we note that  $\nu(\sigma_i)$  is a monotone, bijective function when  $0 \leq \lambda < \lambda_S$ , since then we are constrained to the regime (3.12) on which  $\nu'(\sigma_i)$  is strictly negative due to equation (3.10). The system of equations (3.16) is thus satisfied only if  $\sigma_1 = \dots = \sigma_M$ , i.e., when  $\boldsymbol{\Sigma}$  has the form  $\boldsymbol{\Sigma} = \sigma I_M$ .

Now, a minimizer  $\boldsymbol{\sigma}$  of  $\tilde{F}_{M,\text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda)$  must either be a critical point or lie on the boundary of  $[0, \infty)^M$ . For  $\boldsymbol{\sigma}$  to be a boundary point, it must be such that at least one of the  $\sigma_i$ , say  $\sigma_1$ , vanishes or tends to infinity. If  $\sigma_1 = 0$ ,  $\boldsymbol{\sigma}$  cannot be a minimizer of  $\tilde{F}_{M,\text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda)$  in the  $0 \leq \lambda < \lambda_S$  regime due to the bound (3.12). On the other hand, if  $\sigma_1 \rightarrow \infty$ , we have by the I-MMSE relation for scalar Gaussian channels [33] that

$$\iota(\sigma) = -\frac{1}{2\sigma^2} \text{mmse}\left(\frac{1}{\sigma}\right) \quad (3.17)$$

(recall the definition (2.5) of the MMSE) and thus the bounds  $0 \leq \text{mmse}(t) \leq \rho^2$  valid for all  $t \geq 0$  (see, e.g., [43]) imply that

$$\lim_{\sigma_1 \rightarrow \infty} \nabla_{\boldsymbol{\gamma}} \tilde{F}_{M,\text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda) = \sum_{i=2}^M \gamma_i \frac{\lambda \rho \sigma_i - 1}{2\lambda \sigma_i^3}.$$

Thus, for  $\boldsymbol{\sigma}$  to be a minimizer of  $\tilde{F}_{M,\text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda)$  when  $\sigma_1 \rightarrow \infty$ , we require that all of the remaining  $\sigma_i$  also tend to infinity or equal  $(\lambda\rho)^{-1}$  so that the above gradient vanishes and we avoid saddle points of the type described in the proof of Lemma 3.1. Since  $\sigma_i = (\lambda\rho)^{-1} < (\lambda\rho')^{-1}$  is in contradiction with the bound (3.12), we must have that  $\sigma_2, \dots, \sigma_M \rightarrow \infty$  and we again see that this corresponds to the form  $\boldsymbol{\Sigma} = \sigma I_M$ .

We have established that a minimizer  $\boldsymbol{\sigma}$  of  $\tilde{F}_{M,\text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda)$ , regardless of whether it is a critical or boundary point, must have equal components when we are constrained to the  $0 \leq \lambda < \lambda_S$  regime. Our infimum of interest then decouples over the  $\sigma_1, \dots, \sigma_M$  according to

$$\inf_{\sigma_1, \dots, \sigma_M \geq 0} \tilde{F}_{M,\text{low}}^{\text{RS}}(\boldsymbol{\sigma}, \lambda) = M \inf_{\sigma \geq 0} \left\{ I(x_0; x_0 + \sqrt{\sigma}z) + \frac{\lambda}{4} \left( \frac{1}{\lambda\sigma} - \rho \right)^2 \right\}.$$

Since  $\tilde{F}_M^{\text{RS}}(\boldsymbol{\Sigma}, \lambda)$  decouples in a similar fashion upon setting  $\boldsymbol{\Sigma} = \sigma I_M$ , we see that inequality (3.13) is an equality with the right-hand side given by the above expression. Comparing equations (3.7) and (3.8) then shows that  $F_M^{\text{RS}}(\boldsymbol{Q}, \lambda)$  is maximized when  $\boldsymbol{Q}$  has the form  $\boldsymbol{Q} = q I_M$ . Substituting this into the left-hand side of equation (3.11) finally yields

$$\sup_{q \geq 0} F_M^{\text{RS}}(q I_M, \lambda) = \sup_{q \geq 0} F_1^{\text{RS}}(q, \lambda),$$

which reduces to the middle and right-hand side of equation (3.11) upon realizing that the eigenvalues of  $\mathbf{Q}$  are bounded due to Lemma 3.1 and our choice of  $\lambda'_S$ . ■

We comment on a few features of the above proof. First, we used the fact that the mutual information  $I(x_0; x_0 + \sqrt{\Sigma_{ii}z})$  is a convex function of  $\Sigma_{ii}$  in the low SNR regime, which allowed us to use Jensen's inequality to replace

$$\sum_{i=1}^M I(x_0; x_0 + \sqrt{\Sigma_{ii}z}) \quad \text{by} \quad MI(x_0; x_0 + \sqrt{\sigma}z),$$

through Corollary 2.4. Thus, the problem of minimizing  $\tilde{F}_M^{\text{RS}}(\Sigma, \lambda)$  over  $\Sigma$  became a problem of minimizing  $\tilde{F}_{M,\text{low}}^{\text{RS}}(\Sigma, \lambda)$  over the eigenvalues of  $\Sigma$ . We then showed that the critical points and indeed all minimizers of  $\tilde{F}_{M,\text{low}}^{\text{RS}}(\Sigma, \lambda)$  are of the form  $\Sigma = \sigma I_M$ , which relied crucially on the monotonicity of the gradient of the squared Frobenius norm  $\|\Sigma^{-1}/\lambda - \rho I_M\|_{\text{F}}^2$ .

In contrast, the upcoming proof of the high SNR counterpart of Lemma 3.6 is in some sense dual to the above, where we first use the convexity of  $\|\Sigma^{-1}/\lambda - \rho I_M\|_{\text{F}}^2$  in terms of the eigenvalues of  $\Sigma$  to replace it by a function of  $\sigma = \text{Tr } \Sigma/M$  via Jensen's inequality. This will then supplant the problem of minimizing  $\tilde{F}_M^{\text{RS}}(\Sigma, \lambda)$  over  $\Sigma$  by a minimization problem in terms of the diagonal entries of  $\Sigma$ . We will again show that said minimization problem is solved when  $\Sigma$  is proportional to the identity – recall that the eigenvalues and diagonal entries of  $\Sigma$  coincide when  $\Sigma$  is diagonal, thus explaining how the low and high SNR regimes will eventually be reconciled. Showing that our minimizers have the form  $\Sigma = \sigma I_M$  will require us to use monotonicity of the derivative of the mutual information  $I(x_0; x_0 + \sqrt{\Sigma_{ii}z})$  with respect to  $\Sigma_{ii}$ , which is guaranteed by hypothesis (H5) for small  $\Sigma_{ii}$  (i.e., high SNR, due to Proposition 2.5).

**Lemma 3.7** (High SNR rank-one reduction). *Assume hypotheses (H2), (H4), (H5). Let  $\mathbb{P}_X$  be centered with bounded second moment  $\rho = \mathbb{E}_{\mathbb{P}_X} X^2$  and fix  $\rho' = 2\rho/3$ . Then, there exists  $\lambda_L > \lambda_c$  such that whenever  $\lambda > \lambda_L$ ,*

$$\sup_{\mathbf{Q} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{Q}, \lambda) = \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) = \sup_{q \in [\rho', \rho]} F_1^{\text{RS}}(q, \lambda).$$

*Proof.* We are once again tasked with minimizing  $\tilde{F}_M^{\text{RS}}(\Sigma, \lambda)$  (see (3.8)) over  $\Sigma \in \mathcal{S}_M$ . In the high SNR setting, we lose the bound (3.12) on the eigenvalues  $\sigma_1, \dots, \sigma_M$  of minimizers  $\Sigma$  and hence Corollary 2.4 no longer applies. Instead, we begin by observing that due to Proposition 2.5 and the relation  $\Sigma = (\lambda \mathbf{Q}^*(\lambda))^{-1}$ , there exists  $\lambda'_L > \lambda_c$  such that

$$\frac{1}{\lambda\rho} < \sigma_1, \dots, \sigma_M < \frac{1}{\lambda\rho'} \quad (3.18)$$

whenever  $\lambda > \lambda'_L$ . Then, since the Frobenius norm term in equation (3.8) is convex on the domain (3.18), we may use Jensen's inequality to see that

$$\left\| \frac{\Sigma^{-1}}{\lambda} - \rho I_M \right\|_F^2 \geq M \left( \frac{1}{\lambda\sigma} - \rho \right)^2,$$

where we recall that  $\sigma = \text{Tr } \Sigma / M$  denotes the normalized trace of  $\Sigma$ . Inserting this into equation (3.8) and using Lemma 2.3 then shows that

$$\begin{aligned} \inf_{\Sigma \in \mathcal{S}_M} \tilde{F}_M^{\text{RS}}(\Sigma, \lambda) &\geq \inf_{\Sigma_{11}, \dots, \Sigma_{MM} \geq 0} \tilde{F}_{M, \text{high}}^{\text{RS}}(\text{diag } \Sigma, \lambda), \\ \tilde{F}_{M, \text{high}}^{\text{RS}}(\text{diag } \Sigma, \lambda) &:= \sum_{i=1}^M I(x_0; x_0 + \sqrt{\Sigma_{ii}}z) + \frac{\lambda M}{4} \left( \frac{1}{\lambda\sigma} - \rho \right)^2, \end{aligned} \quad (3.19)$$

where we further recall that  $x_0 \sim \mathbb{P}_X$ ,  $z \sim \mathcal{N}(0, 1)$  and write  $\text{diag } \Sigma$  for the diagonal of  $\Sigma$ . Writing  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_M)$ , the  $\boldsymbol{\gamma}$ -directional derivative of  $\tilde{F}_{M, \text{high}}^{\text{RS}}(\text{diag } \Sigma, \lambda)$  is

$$\nabla_{\boldsymbol{\gamma}} \tilde{F}_{M, \text{high}}^{\text{RS}}(\text{diag } \Sigma, \lambda) = \sum_{i=1}^M \gamma_i (\iota(\Sigma_{ii}) + \nu(\sigma)),$$

where  $\iota(\Sigma_{ii})$  and  $\nu(\sigma)$  are specified by equations (3.14) and (3.15), respectively. Now, by hypothesis (H5), there exists  $\lambda''_L > 0$  such that  $\lambda^2 \text{mmse}(\lambda)$  is monotone on  $[\lambda''_L, \infty)$ . Then, by equation (3.17),  $\iota(\Sigma_{ii})$  is monotone for  $\Sigma_{ii} < (\lambda''_L)^{-1}$ . This condition is satisfied when  $\lambda > \lambda'_L, \lambda''_L/\rho'$ , since then

$$\Sigma_{ii} = \sum_{j=1}^M \mathbf{O}_{ij}^2 \sigma_j < \frac{1}{\lambda\rho'} \sum_{j=1}^M \mathbf{O}_{ij}^2 = \frac{1}{\lambda\rho'} < \frac{1}{\lambda''_L}, \quad 1 \leq i \leq M,$$

where  $\mathbf{O}_i$  is the eigenvector of  $\Sigma$  corresponding to  $\sigma_i$ , the first inequality follows from the bound (3.18) valid for  $\lambda > \lambda'_L$ , and the second inequality is due to the constraint  $\lambda > \lambda''_L/\rho'$ . Hence, setting  $\lambda_L := \max\{\lambda'_L, \lambda''_L/\rho'\}$ ,  $\iota(\Sigma_{ii})$  is a monotone, bijective function for  $\lambda > \lambda_L$  and the critical points of  $\tilde{F}_{M, \text{high}}^{\text{RS}}(\text{diag } \Sigma, \lambda)$  occur at

$$\Sigma_{ii} = \iota^{-1} \circ \nu(\sigma), \quad 1 \leq i \leq M,$$

which is satisfied only if  $\Sigma_{11} = \dots = \Sigma_{MM}$ .

For  $\tilde{F}_{M, \text{high}}^{\text{RS}}(\text{diag } \Sigma, \lambda)$  to be minimized on the boundary of  $[0, \infty)^M$ , the proposed minimizer  $\text{diag } \Sigma$  must be such that at least one of the  $\Sigma_{ii}$  vanishes or tends to infinity. This is not possible for finite  $\lambda > \lambda_L$  since the bound (3.18) and the relation  $\Sigma = \sum_{j=1}^M \mathbf{O}_{ij}^2 \sigma_j$  implies the constraint  $(\lambda\rho)^{-1} < \Sigma_{ii} < (\lambda\rho')^{-1}$ . When  $\lambda \rightarrow \infty$ , this constraint suggests that  $\Sigma_{11} = \dots = \Sigma_{MM} \rightarrow 0$ . As in the proof of Lemma 3.6, we thus conclude that the right-hand side of inequality (3.19) reduces to

$$\inf_{\Sigma_{11}, \dots, \Sigma_{MM} \geq 0} \tilde{F}_{M, \text{high}}^{\text{RS}}(\text{diag } \Sigma, \lambda) = M \inf_{\sigma \geq 0} \left\{ I(x_0; x_0 + \sqrt{\sigma}z) + \frac{\lambda}{4} \left( \frac{1}{\lambda\sigma} - \rho \right)^2 \right\}.$$

Since substituting  $\Sigma = \sigma I_M$  into the left-hand side of inequality (3.19) yields the above, the remainder of this proof follows in the same manner as in the proof of Lemma 3.6. ■

In order to complete the proof of Theorem 2.6, it remains only to use analytic continuation to combine Lemmas 3.6 and 3.7.

*Proof of Theorem 2.6.* Using the notation of equation (2.4), we have by Lemmas 3.6 and 3.7 that there exist  $0 < \lambda_S < \lambda_c$  and  $\lambda_L > \lambda_c$  such that

$$\phi_1(\lambda) = \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) \quad \text{and} \quad \phi_M(\lambda) = \sup_{\mathbf{Q} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{Q}, \lambda)$$

agree on  $[0, \lambda_S) \cup (\lambda_L, \infty)$ . By hypothesis (H4),  $\phi_1(\lambda)$  and  $\phi_M(\lambda)$  are real analytic on  $[0, \infty) \setminus \{\lambda_c\}$ , so by the identity theorem for real analytic functions, the aforementioned agreement between said functions extends to the domain  $[0, \infty) \setminus \{\lambda_c\}$ . Since  $\phi_1(\lambda)$  and  $\phi_M(\lambda)$  are also continuous functions of  $\lambda$  (recall Remark 2.2), they must then agree at  $\lambda = \lambda_c$ , as well. ■

#### 4. Standard prerequisites: Interpolation and overlap concentration

With Theorem 2.6 established, we would now like to move on to the next stage of the proof of Theorem 2.1. We recall that said proof amounts to combining the lower and upper bounds on the limiting free entropy given by Propositions 2.7 and 2.8. The brunt of the argument henceforth lies within the proof of Proposition 2.8.

We present the key idea of the proof, which is a multiscale application of the cavity method, in Section 5 and use this section to lay out some standard prerequisite results that will be needed therein. In Section 4.1, we use Guerra's interpolation method [30] to prove Proposition 2.7, which is both a counterpart to Proposition 2.8 and an ingredient of its proof. In Section 4.2, we prove thermal concentration of the overlap matrix  $\mathbf{R}_{10}$  (see (2.8)) with respect to the perturbed Hamiltonian  $\tilde{H}_{N,\varepsilon}(\mathbf{X})$  (see (2.14)), as described in Theorem 2.11, along with some straightforward corollaries. Finally, in Section 4.3, we prove negligibility of the difference between the limiting free entropies

$$\lim_{N \rightarrow \infty} F_N(\lambda) \quad \text{and} \quad \lim_{N \rightarrow \infty} \tilde{F}_N(\lambda)$$

corresponding respectively to the original and perturbed Hamiltonians  $H_N(\mathbf{X})$  given in (1.4) and  $\tilde{H}_{N,\varepsilon}(\mathbf{X})$ .

#### 4.1. Free entropy lower bound: Guerra interpolation

Proposition 2.7 follows from the standard (finite-rank) interpolation argument seen in, e.g., [43] in combination with the rank-one reduction of Theorem 2.6. The main idea is to introduce a Hamiltonian that interpolates between  $H_N(\mathbf{X})$  and one corresponding to  $N$  copies of the Hamiltonian associated with the Gibbs average  $\langle \cdot \rangle_{\text{RS}}$  (see (3.4)). Then, it is simply a matter of bounding the derivative of the associated interpolating free entropy. We present the details for the sake of completeness.

*Proof of Proposition 2.7.* Introduce the interpolating Hamiltonian

$$H_t(\mathbf{X}) := \frac{1}{2} \text{Tr} \left( \sqrt{\frac{\lambda t}{N}} \mathbf{Z} \mathbf{X} \mathbf{X}^\top + \frac{\lambda t}{N} \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X} \mathbf{X}^\top - \frac{\lambda t}{2N} \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \right) \\ + \sum_{i=1}^N \left( \sqrt{\lambda(1-t)} \mathbf{x}_i^\top \sqrt{\mathbf{Q}} \tilde{\mathbf{z}}_i + (1-t) \lambda \mathbf{x}_{0,i}^\top \mathbf{Q} \mathbf{x}_i - \frac{(1-t)\lambda}{2} \mathbf{x}_i^\top \mathbf{Q} \mathbf{x}_i \right),$$

with  $\mathbf{Z}, \mathbf{X}_0$  as in equation (1.1), each  $\mathbf{x}_i \in \mathbb{R}^M$  the  $i$ th row of  $\mathbf{X}$ , each  $\mathbf{x}_{0,i} \in \mathbb{R}^M$  the  $i$ th row of  $\mathbf{X}_0$ , and  $\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_N \sim \mathcal{N}(0, I_M)$  i.i.d. standard Gaussian vectors independent of  $\mathbf{Z}$ . Define the associated interpolating free entropy as

$$\varphi(t) := \frac{1}{NM} \mathbb{E} \ln \int_{\mathbb{R}^{N \times M}} e^{H_t(\mathbf{X})} d\mathbb{P}_{\mathbf{X}, N, M}(\mathbf{X}),$$

with the expectation taken over  $\mathbf{Z}, \mathbf{X}_0, \tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_N$  here and throughout the remainder of this proof. It follows that

$$\begin{aligned} \varphi'(t) &= \frac{1}{NM} \mathbb{E} \left\langle \frac{d}{dt} H_t(\mathbf{X}) \right\rangle_t \\ &= \frac{1}{2NM} \text{Tr} \mathbb{E} \left\langle \frac{1}{2} \sqrt{\frac{\lambda}{tN}} \mathbf{Z} \mathbf{X} \mathbf{X}^\top + \frac{\lambda}{N} \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X} \mathbf{X}^\top - \frac{\lambda}{2N} \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \right\rangle_t \\ &\quad - \frac{1}{NM} \mathbb{E} \left\langle \sum_{i=1}^N \frac{1}{2} \sqrt{\frac{\lambda}{1-t}} \mathbf{x}_i^\top \sqrt{\mathbf{Q}} \tilde{\mathbf{z}}_i + \lambda \mathbf{x}_{0,i}^\top \mathbf{Q} \mathbf{x}_i - \frac{\lambda}{2} \mathbf{x}_i^\top \mathbf{Q} \mathbf{x}_i \right\rangle_t, \end{aligned} \quad (4.1)$$

where we write

$$\langle \cdot \rangle_t := \frac{1}{\hat{\mathcal{Z}}_{N,t}} \int_{\mathbb{R}^{N \times M}} \cdot e^{H_t(\mathbf{X})} d\mathbb{P}_{\mathbf{X}, N, M}, \quad \hat{\mathcal{Z}}_{N,t} := \int_{\mathbb{R}^{N \times M}} e^{H_t(\mathbf{X})} d\mathbb{P}_{\mathbf{X}, N, M}$$

for the Gibbs average with respect to the interpolating Hamiltonian  $H_t(\mathbf{X})$  and the corresponding partition function. By Gaussian integration by parts with respect to the (symmetric) Wigner matrix  $\mathbf{Z}$ , we have that

$$\begin{aligned} \text{Tr} \mathbb{E} \langle \mathbf{Z} \mathbf{X} \mathbf{X}^\top \rangle_t &= 2 \sum_{1 \leq i \leq j \leq N} \mathbb{E} \left[ \frac{\partial}{\partial \mathbf{Z}_{ij}} \langle (\mathbf{X} \mathbf{X}^\top)_{ji} \rangle_t \right] \\ &= 2 \sum_{1 \leq i \leq j \leq N} \mathbb{E} \left[ \left\langle (\mathbf{X} \mathbf{X}^\top)_{ji} \frac{\partial H_t(\mathbf{X})}{\partial \mathbf{Z}_{ij}} \right\rangle_t - \frac{1}{\hat{\mathcal{Z}}_{N,t}} \langle (\mathbf{X} \mathbf{X}^\top)_{ji} \rangle_t \frac{\partial \hat{\mathcal{Z}}_{N,t}}{\partial \mathbf{Z}_{ij}} \right] \end{aligned}$$

$$\begin{aligned}
 &= 2 \sum_{1 \leq i \leq j \leq N} \mathbb{E} \left[ \left\langle (X X^\top)_{ji} \frac{\partial H_t(X)}{\partial \mathbf{Z}_{ij}} \right\rangle_t - \langle (X X^\top)_{ji} \rangle_t \left\langle \frac{\partial H_t(X)}{\partial \mathbf{Z}_{ij}} \right\rangle_t \right] \\
 &= \sqrt{\frac{\lambda t}{N}} \operatorname{Tr} \mathbb{E} \langle X_1 X_1^\top X_1 X_1^\top - X_1 X_1^\top X_2 X_2^\top \rangle_t,
 \end{aligned}$$

where we have written  $X_1$  for  $X$ , defined  $X_2$  to be a sample, conditionally independent of  $X_1$ , from the interpolating posterior distribution, and have redefined  $\langle \cdot \rangle_t$  to be the Gibbs average over both samples  $X_1$  and  $X_2$ , as per the convention described at the end of Section 2.3. In a similar fashion, Gaussian integration by parts with respect to  $\tilde{z}_i$  ( $1 \leq i \leq N$ ) shows that

$$\mathbb{E} \langle \mathbf{x}_i^\top \sqrt{\mathbf{Q}} \tilde{\mathbf{z}}_i \rangle_t = \sqrt{\lambda(1-t)} \mathbb{E} \langle \mathbf{x}_{1,i}^\top \mathbf{Q} \mathbf{x}_{1,i} - \mathbf{x}_{1,i}^\top \mathbf{Q} \mathbf{x}_{2,i} \rangle_t.$$

Thus, using the cyclic property of the trace and defining

$$\mathbf{R}_{12} := \frac{1}{N} X_1^\top X_2 = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_{1,i} \mathbf{x}_{2,i}^\top \quad (4.2)$$

in analogy with equation (2.8), we may rewrite equation (4.1) as

$$\begin{aligned}
 \varphi'(t) &= \frac{\lambda}{2M} \operatorname{Tr} \mathbb{E} \left\langle \mathbf{R}_{10}^\top \mathbf{R}_{10} - \frac{1}{2} \mathbf{R}_{12}^\top \mathbf{R}_{12} \right\rangle_t \\
 &\quad - \frac{\lambda}{NM} \sum_{i=1}^N \operatorname{Tr} \mathbb{E} \left\langle \mathbf{Q} \mathbf{x}_{1,i} \mathbf{x}_{0,i}^\top - \frac{1}{2} \mathbf{Q} \mathbf{x}_{1,i} \mathbf{x}_{2,i}^\top \right\rangle_t \\
 &= \frac{\lambda}{2M} \operatorname{Tr} \mathbb{E} \left\langle \mathbf{R}_{10}^\top \mathbf{R}_{10} - \frac{1}{2} \mathbf{R}_{12}^\top \mathbf{R}_{12} \right\rangle_t - \frac{\lambda}{M} \operatorname{Tr} \mathbb{E} \left\langle \mathbf{Q} \mathbf{R}_{10} - \frac{1}{2} \mathbf{Q} \mathbf{R}_{12} \right\rangle_t.
 \end{aligned}$$

Using the Nishimori identity to replace  $X_2$  by  $X_0$ , this further simplifies to

$$\varphi'(t) = \frac{\lambda}{4M} \operatorname{Tr} \mathbb{E} \langle \mathbf{R}_{10}^\top \mathbf{R}_{10} - 2\mathbf{Q} \mathbf{R}_{10} \rangle_t \geq -\frac{\lambda}{4M} \operatorname{Tr} \mathbf{Q}^2,$$

with the inequality following from completing the square. Combining this with the definitions (1.2) and (2.2) of the free entropy and replica symmetric potential shows that

$$\begin{aligned}
 F_N(\lambda) - F_M^{\text{RS}}(\mathbf{Q}, \lambda) &= \varphi(1) - \varphi(0) + \frac{\lambda}{4M} \operatorname{Tr} \mathbf{Q}^2 \\
 &= \int_0^1 \varphi'(t) dt + \frac{\lambda}{4M} \operatorname{Tr} \mathbf{Q}^2 \geq 0.
 \end{aligned}$$

As this holds for all  $\mathbf{Q} \in \mathcal{S}_M$ , we then have that

$$F_N(\lambda) \geq \sup_{\mathbf{Q} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{Q}, \lambda).$$

We use Theorem 2.6 to replace the right-hand side of this identity, completing the proof.  $\blacksquare$

## 4.2. Thermal concentration of the overlap matrix

The cavity computations of Section 5.2 rely on a form of thermal concentration of the overlap matrix  $\mathbf{R}_{10}$  (see (2.8)). Taken at face value, one might expect this to mean that

$$\mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, \mathbf{X}_0} \left( \|\mathbf{R}_{10} - \langle \mathbf{R}_{10} \rangle'_{N,0}\|_{\mathbb{F}}^2 \right) \xrightarrow{N \rightarrow \infty} 0,$$

with  $\langle \cdot \rangle'_{N,0}$  specified by equation (2.24), so that one may replace  $\mathbf{R}_{10}$  by the more amenable quantity  $\langle \mathbf{R}_{10} \rangle'_{N,0}$  within the calculations of Section 5.2. However, it is known that the concentration result proposed above does not hold in generality, with “non-physical” singularities arising at particular model-dependent values of SNR  $\lambda$  (see [10] and references therein). Nonetheless, due to the symmetries of our model, we are able to add infinitesimal side information (without affecting the limiting free entropy, as discussed in Section 4.3) to induce thermal concentration of the overlap matrix up to an average over a perturbation parameter. This is achieved by replacing  $\langle \cdot \rangle'_{N,0}$  in the above with the Gibbs average  $\langle \cdot \rangle'_{N,\varepsilon}$  (see (2.24)), and then integrating the left-hand side over  $\varepsilon$  as described in Theorem 2.11, with the integration step serving to smooth out the aforementioned singularities. Similar perturbations have been commonly used in other spin glass models, where regularizing perturbations are introduced in order to enforce stability properties of the high-dimensional probability measures [2, 29] that in turn induce asymptotically nice structural phenomena such as ultrametricity [55] and synchronization [57, 58].

It turns out that the driving mechanisms behind our proof of thermal concentration of the overlap matrix are the presence of the perturbation Hamiltonian  $H_{N,\varepsilon}(\mathbf{X})$  (see (2.22)) and the Nishimori identity, which applies due to our working in the Bayes-optimal setting. Indeed, our proof holds in more general settings, so long as the model under consideration exhibits the same symmetry properties as the spiked Wigner model (1.1). Thus, we now consider the fully factorized channel (2.26)

$$\mathbb{P}_{\text{out}}(\mathbf{Y}'' \mid \mathbf{X}_0 \mathbf{X}_0^\top) = \prod_{i=1}^N \mathbb{P}_{\text{out}}^{(i=j)}(\mathbf{Y}''_{ii} \mid \mathbf{x}_{0,i}^\top \mathbf{x}_{0,i}) \prod_{1 \leq i < j \leq N} \mathbb{P}_{\text{out}}^{(i < j)}(\mathbf{Y}''_{ij} \mid \mathbf{x}_{0,i}^\top \mathbf{x}_{0,j})$$

and study the inference problem

$$\begin{cases} \mathbf{Y}'' \sim \mathbb{P}_{\text{out}}(\cdot \mid \mathbf{X}_0 \mathbf{X}_0^\top), \\ \mathbf{Y}^{(\varepsilon)} = \sqrt{\varepsilon} \mathbf{X}_0 + \tilde{\mathbf{Z}}, \end{cases}$$

with  $\tilde{\mathbf{Z}} \in \mathbb{R}^{N \times M}$  a standard Ginibre matrix of i.i.d. standard Gaussian entries independent of everything else,  $\varepsilon \in [s_N, 2s_N]$  a perturbation parameter, and  $(s_N)_{N \geq 1}$  a sequence of numbers in  $(0, 1)$  that decrease slowly to zero as  $N \rightarrow \infty$ . Denoting the log-likelihood of the first channel by  $H_N''(\mathbf{X})$  and recalling that the analogous term for

the second channel is  $H_{N,\varepsilon}(\mathbf{X})$  (see (2.22)), we see that the Hamiltonian corresponding to the full inference problem is  $\tilde{H}_{N,\varepsilon}''(\mathbf{X}) = H_N''(\mathbf{X}) + H_{N,\varepsilon}(\mathbf{X})$  (see (2.27)). As stated above, we observe that this Hamiltonian corresponds to the log-likelihood of an inference problem in the Bayes-optimal setting and the Nishimori identity holds, which induces concentration-of-measure [6, 10].

We now prove thermal concentration of  $\mathbf{R}_{10}$  (see (2.8)), with respect to the general perturbed Hamiltonian  $\tilde{H}_{N,\varepsilon}''(\mathbf{X})$ .

*Proof of Theorem 2.11.* To simplify notation throughout this proof, we write  $x$  for  $\mathbf{X}$ ,  $X^*$  for  $\mathbf{X}_0$ ,  $\mathbf{R}$  for  $\mathbf{R}_{10}$ ,  $C$  a generic positive constant independent of  $M, N$  and depending only on the properties of  $\mathbb{P}_X$ ,  $\mathbb{E}[\cdot]$  for expectation over  $\tilde{\mathbf{Z}}, X^*$  and the randomness in  $\mathbb{P}_{\text{out}}(\cdot | X^* X^{*\top})$ ,  $h_N$  for  $\tilde{H}_{N,\varepsilon}''(x)$ , and  $\langle \cdot \rangle$  for the Gibbs average

$$\begin{aligned} \langle \cdot \rangle_{N,\varepsilon}'' &:= \frac{1}{Z_{N,\varepsilon}''} \int_{\mathbb{R}^{N \times M}} \cdot e^{h_N} d\mathbb{P}_{X,N,M}(x), \\ Z_{N,\varepsilon}'' &:= \int_{\mathbb{R}^{N \times M}} e^{h_N} d\mathbb{P}_{X,N,M}(x). \end{aligned}$$

Following [6, 47], introduce the auxiliary function

$$\mathcal{L} := -\frac{1}{N} \frac{\partial h_N}{\partial \varepsilon} = -\frac{1}{N} \text{Tr} \left( \frac{1}{2\sqrt{\varepsilon}} \tilde{\mathbf{Z}} x^\top + X^* x^\top - \frac{1}{2} x x^\top \right).$$

Then, one sees from Gaussian integration by parts with respect to  $\tilde{\mathbf{Z}}$  and the Nishimori identity that

$$\frac{1}{N} \mathbb{E} \langle \mathcal{L} \rangle = -\frac{1}{2N} \text{Tr} \mathbb{E} \langle \mathbf{R} \rangle. \quad (4.3)$$

Differentiating both sides of this equation with respect to  $\varepsilon$  and simplifying the left-hand side shows that

$$\frac{1}{N} \mathbb{E} \left\langle \frac{\partial \mathcal{L}}{\partial \varepsilon} \right\rangle - \mathbb{E} \langle (\mathcal{L} - \langle \mathcal{L} \rangle)^2 \rangle = -\frac{1}{2N} \frac{\partial}{\partial \varepsilon} \text{Tr} \mathbb{E} \langle \mathbf{R} \rangle. \quad (4.4)$$

Another application of Gaussian integration by parts with respect to  $\tilde{\mathbf{Z}}$ , along with the Nishimori identity, shows that the first term in the above simplifies as

$$\frac{1}{N} \mathbb{E} \left\langle \frac{\partial \mathcal{L}}{\partial \varepsilon} \right\rangle = \frac{1}{N} \mathbb{E} \left\langle \frac{1}{4N\varepsilon^{3/2}} \text{Tr} \tilde{\mathbf{Z}} x^\top \right\rangle = \frac{1}{4\varepsilon N} \text{Tr} \mathbb{E} \left\langle \frac{1}{N} X^* X^{*\top} - \langle \mathbf{R} \rangle \right\rangle. \quad (4.5)$$

By the positive semidefiniteness of  $\langle x \rangle^\top \langle x \rangle$ , the Nishimori identity, Jensen's inequality, hypothesis (H2), and the boundedness of the second moment of  $\mathbb{P}_X$ , we have that

$$0 \leq \text{Tr} \mathbb{E} \langle \mathbf{R} \rangle = \frac{1}{N} \text{Tr} \mathbb{E} \langle x \rangle^\top \langle x \rangle \leq \frac{1}{N} \text{Tr} \mathbb{E} \langle x^\top x \rangle = \frac{1}{N} \text{Tr} \mathbb{E} X^{*\top} X^* \leq CM. \quad (4.6)$$

Hence, the quantity (4.5) is such that

$$0 \leq \frac{1}{N} \mathbb{E} \left\langle \frac{\partial \mathcal{L}}{\partial \varepsilon} \right\rangle \leq \frac{CM}{4\varepsilon N}, \quad (4.7)$$

and we read from equation (4.4) that

$$\frac{1}{2N} \frac{\partial}{\partial \varepsilon} \text{Tr} \mathbb{E} \langle \mathbf{R} \rangle \leq \mathbb{E} \langle (\mathcal{L} - \langle \mathcal{L} \rangle)^2 \rangle. \quad (4.8)$$

It remains now to show how the left-hand side of this inequality relates to the difference  $\mathbb{E} \langle (\mathbf{R}_{\ell\ell'} - \langle \mathbf{R}_{\ell\ell'} \rangle)^2 \rangle$ , which sums over  $1 \leq \ell, \ell' \leq M$  to give  $\mathbb{E} \langle \|\mathbf{R} - \langle \mathbf{R} \rangle\|_{\text{F}}^2 \rangle$ , and to prove concentration of the right-hand side.

Carrying out the differentiation on the left-hand side of inequality (4.8) shows that

$$\begin{aligned} \frac{1}{2N} \frac{\partial}{\partial \varepsilon} \text{Tr} \mathbb{E} \langle \mathbf{R} \rangle &= \frac{1}{2} \text{Tr} \mathbb{E} [\langle \mathbf{R} \rangle \langle \mathcal{L} \rangle - \langle \mathbf{R} \mathcal{L} \rangle] \\ &= \frac{1}{2N^2} \sum_{i,j=1}^N \sum_{\ell,\ell'=1}^M \mathbb{E} \left[ \frac{1}{2} X_{i\ell}^* \langle x_{i\ell} \rangle \langle x_{j\ell'}^2 \rangle - X_{i\ell}^* X_{j\ell'}^* \langle x_{i\ell} \rangle \langle x_{j\ell'} \rangle \right. \\ &\quad \left. - \frac{1}{2\sqrt{\varepsilon}} X_{i\ell}^* \langle x_{i\ell} \rangle \langle x_{j\ell'} \rangle \tilde{\mathbf{Z}}_{j\ell'} - \frac{1}{2} X_{i\ell}^* \langle x_{i\ell} x_{j\ell'}^2 \rangle \right. \\ &\quad \left. + X_{i\ell}^* X_{j\ell'}^* \langle x_{i\ell} x_{j\ell'} \rangle + \frac{1}{2\sqrt{\varepsilon}} X_{i\ell}^* \langle x_{i\ell} x_{j\ell'} \rangle \tilde{\mathbf{Z}}_{j\ell'} \right]. \end{aligned} \quad (4.9)$$

Gaussian integration by parts with respect to  $\tilde{\mathbf{Z}}_{j\ell'}$  yields

$$\begin{aligned} &\frac{1}{2\sqrt{\varepsilon}} \mathbb{E} [X_{i\ell}^* \langle x_{i\ell} x_{j\ell'} \rangle \tilde{\mathbf{Z}}_{j\ell'} - X_{i\ell}^* \langle x_{i\ell} \rangle \langle x_{j\ell'} \rangle \tilde{\mathbf{Z}}_{j\ell'}] \\ &= \mathbb{E} \left[ X_{i\ell}^* \langle x_{i\ell} \rangle \langle x_{j\ell'} \rangle^2 - X_{i\ell}^* \langle x_{i\ell} x_{j\ell'} \rangle \langle x_{j\ell'} \rangle \right. \\ &\quad \left. + \frac{1}{2} X_{i\ell}^* \langle x_{i\ell} x_{j\ell'}^2 \rangle - \frac{1}{2} X_{i\ell}^* \langle x_{i\ell} \rangle \langle x_{j\ell'}^2 \rangle \right], \end{aligned}$$

so that, after an application of the Nishimori identity, equation (4.9) reduces to

$$\frac{1}{2N} \frac{\partial}{\partial \varepsilon} \text{Tr} \mathbb{E} \langle \mathbf{R} \rangle = \frac{1}{2N^2} \sum_{i,j=1}^N \sum_{\ell,\ell'=1}^M \mathbb{E} [(\langle x_{i\ell} x_{j\ell'} \rangle - \langle x_{i\ell} \rangle \langle x_{j\ell'} \rangle)^2].$$

Substituting this into inequality (4.8) then gives

$$\frac{1}{2N^2} \sum_{i,j=1}^N \sum_{\ell,\ell'=1}^M \mathbb{E} [(\langle x_{i\ell} x_{j\ell'} \rangle - \langle x_{i\ell} \rangle \langle x_{j\ell'} \rangle)^2] \leq \mathbb{E} \langle (\mathcal{L} - \langle \mathcal{L} \rangle)^2 \rangle. \quad (4.10)$$

Our hypothesis on  $\mathbb{P}_{\text{out}}(\cdot \mid X_0 X_0^T)$  being fully factorized implies that the term below is unchanged upon exchanging indices  $k \leftrightarrow k'$ , so we see for any  $1 \leq k \leq M$  that

$$\begin{aligned}
 & \frac{1}{2N^2} \sum_{i,j=1}^N \mathbb{E}[(\langle x_{ik} x_{jk} \rangle - \langle x_{ik} \rangle \langle x_{jk} \rangle)^2] \\
 &= \frac{1}{2MN^2} \sum_{i,j=1}^N \sum_{\ell=1}^M \mathbb{E}[(\langle x_{i\ell} x_{j\ell} \rangle - \langle x_{i\ell} \rangle \langle x_{j\ell} \rangle)^2] \\
 &\leq \frac{1}{2MN^2} \sum_{i,j=1}^N \sum_{\ell,\ell'=1}^M \mathbb{E}[(\langle x_{i\ell} x_{j\ell'} \rangle - \langle x_{i\ell} \rangle \langle x_{j\ell'} \rangle)^2] \\
 &\leq \frac{1}{M} \mathbb{E}[(\mathcal{L} - \langle \mathcal{L} \rangle)^2];
 \end{aligned}$$

the first inequality is due to the trivial fact that we are adding expectations of squares, while the second inequality is an application of inequality (4.10). Combining this with the Cauchy–Schwarz inequality shows that

$$\begin{aligned}
 & \frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E}[(\mathbf{R}_{\ell\ell'} - \langle \mathbf{R}_{\ell\ell'} \rangle)^2] d\varepsilon \\
 &\leq \sqrt{\frac{1}{N^2} \sum_{i,j=1}^N \mathbb{E}[(X_{i\ell'}^* X_{j\ell'}^*)^2]} \\
 &\quad \times \sqrt{\frac{1}{N^2 s_N} \sum_{i,j=1}^N \int_{s_N}^{2s_N} \mathbb{E}[(\langle x_{i\ell} x_{j\ell} \rangle - \langle x_{i\ell} \rangle \langle x_{j\ell} \rangle)^2] d\varepsilon} \\
 &\leq C \sqrt{\frac{2}{M s_N} \int_{s_N}^{2s_N} \mathbb{E}[(\mathcal{L} - \langle \mathcal{L} \rangle)^2] d\varepsilon}. \tag{4.11}
 \end{aligned}$$

Moving on to the thermal concentration of  $\mathcal{L}$ , first note that

$$\mathbb{E}[(\mathcal{L} - \langle \mathcal{L} \rangle)^2] = \frac{M}{N} \frac{\partial^2 f_N}{\partial \varepsilon^2} + \frac{1}{N} \mathbb{E} \left\langle \frac{\partial \mathcal{L}}{\partial \varepsilon} \right\rangle \leq \frac{M}{N} \frac{\partial^2 f_N}{\partial \varepsilon^2} + \frac{CM}{4\varepsilon N},$$

where we have written

$$f_N := \frac{1}{NM} \mathbb{E} \ln Z''_{N,\varepsilon}$$

and used the bound (4.7). Integrating this quantity over  $\varepsilon$  then shows that

$$\begin{aligned}
 & \frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E}[(\mathcal{L} - \langle \mathcal{L} \rangle)^2] d\varepsilon \\
 &\leq \frac{M}{N s_N} \left( \frac{\partial f_N}{\partial \varepsilon} \Big|_{\varepsilon=2s_N} - \frac{\partial f_N}{\partial \varepsilon} \Big|_{\varepsilon=s_N} \right) + \frac{CM \ln 2}{4N s_N}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{Ns_N} (\mathbb{E}\langle \mathcal{L} \rangle|_{\varepsilon=s_N} - \mathbb{E}\langle \mathcal{L} \rangle|_{\varepsilon=2s_N}) + \frac{CM \ln 2}{4Ns_N} \\
 &= \frac{1}{2Ns_N} (\text{Tr} \mathbb{E}\langle \mathbf{R} \rangle|_{\varepsilon=2s_N} - \text{Tr} \mathbb{E}\langle \mathbf{R} \rangle|_{\varepsilon=s_N}) + \frac{CM \ln 2}{4Ns_N} \\
 &\leq \frac{CM}{Ns_N}.
 \end{aligned}$$

Here, the first line follows from the fundamental theorem of calculus, the second from comparing the definitions of  $\mathcal{L}$  and  $f_N$ , the third from equation (4.3), and the fourth from the fact that the bounds (4.6) hold uniformly over  $\varepsilon$ .

Substituting the above bound into inequality (4.11) shows that

$$\frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E} \langle (\mathbf{R}_{\ell\ell'} - \langle \mathbf{R}_{\ell\ell'} \rangle)^2 \rangle d\varepsilon \leq \frac{C}{\sqrt{Ns_N}}.$$

Summing this expression over  $1 \leq \ell, \ell' \leq M$  completes the proof.  $\blacksquare$

Note the particularly simple form of side information used here compared to, e.g., [6, 43, 47, 61], wherein the side channel is Bernoulli distributed or is a matrix Gaussian channel like ours, but with perturbation parameter  $\varepsilon$  of matrix form rather than scalar. Our choice of perturbation is effective due to the fact that the cavity method only requires concentration with respect to the Gibbs average  $\langle \cdot \rangle''_{N,\varepsilon}$  instead of the stronger statements proved in the aforementioned references that include concentration with respect to the quenched Gibbs average  $\mathbb{E}\langle \cdot \rangle''_{N,\varepsilon}$  – the latter corresponds to analogs of Theorem 2.11 involving the quantity  $\mathbb{E}\langle \|\mathbf{R}_{10} - \mathbb{E}\langle \mathbf{R}_{10} \rangle''_{N,\varepsilon} \|_{\mathbb{F}}^2 \rangle''_{N,\varepsilon}$ . Another key point of our simple perturbation is that it maintains the exchangeability of the variables  $(\mathbf{x}_{ij})$  under the quenched Gibbs average: no statistical symmetries present in the original model are broken by the perturbation apart from the trivial sign ambiguity  $\mathbf{X} \rightarrow -\mathbf{X}$ . These two points allow us to obtain concentration for higher regimes of rank than would be possible using more complicated perturbations. As a concrete consequence, Theorem 2.11 gives thermal concentration of the overlap matrix for  $M = o(N^{1/4})$ . Yet, there is room for improvement in order to reach  $M = o(N)$ , the regime for which we believe Theorem 2.1 to hold.

For later convenience, we present a corollary of Theorem 2.11 regarding concentration of the overlap matrix  $\mathbf{R}_{12}$  (see (4.2)).

**Corollary 4.1** (Thermal concentration of  $\mathbf{R}_{12}$ ). *Working in the setting of Theorem 2.11 and recalling that  $\mathbf{R}_{12} = \frac{1}{N} \mathbf{X}_1^\top \mathbf{X}_2$ , there exists a finite positive constant  $C$  independent of  $M, N$  and depending only on properties of  $\mathbb{P}_X$  such that*

$$\frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E} \langle \|\mathbf{R}_{12} - \langle \mathbf{R}_{12} \rangle''_{N,\varepsilon} \|_{\mathbb{F}}^2 \rangle''_{N,\varepsilon} d\varepsilon \leq \Gamma(N, M), \quad (4.12)$$

$$\frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E} \langle \|\mathbf{R}_{10} - \langle \mathbf{R}_{12} \rangle''_{N,\varepsilon} \|_{\mathbb{F}}^2 \rangle''_{N,\varepsilon} d\varepsilon \leq \Gamma(N, M). \quad (4.13)$$

*Proof.* We reuse the notation introduced in the proof of Theorem 2.11, with the Gibbs average  $\langle \cdot \rangle$  now being taken over  $\mathbf{X}_1, \mathbf{X}_2$ , and further write  $\tilde{\mathbf{R}}$  for  $\mathbf{R}_{12}$ . In analogy with the bound (4.11), we first compute through the Cauchy–Schwarz inequality, Jensen’s inequality, and the Nishimori identity that

$$\begin{aligned}
 & \frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E}[\langle \mathbf{R}_{\ell\ell'} \rangle^2 - \langle \tilde{\mathbf{R}}_{\ell\ell'} \rangle^2] d\varepsilon \\
 & \leq \sqrt{\frac{1}{N^2 s_N} \sum_{i,j=1}^N \int_{s_N}^{2s_N} \mathbb{E}[\langle x_{i\ell'} \rangle^2 \langle x_{j\ell'} \rangle^2] d\varepsilon} \\
 & \quad \times \sqrt{\frac{1}{N^2 s_N} \sum_{i,j=1}^N \int_{s_N}^{2s_N} \mathbb{E}[(\langle x_{i\ell} x_{j\ell} \rangle - \langle x_{i\ell} \rangle \langle x_{j\ell} \rangle)^2] d\varepsilon} \\
 & \leq \sqrt{\frac{1}{N^2} \sum_{i,j=1}^N \sqrt{\mathbb{E}(X_{i\ell'}^*)^4} \sqrt{\mathbb{E}(X_{j\ell'}^*)^4}} \\
 & \quad \times \sqrt{\frac{1}{N^2 s_N} \sum_{i,j=1}^N \int_{s_N}^{2s_N} \mathbb{E}[(\langle x_{i\ell} x_{j\ell} \rangle - \langle x_{i\ell} \rangle \langle x_{j\ell} \rangle)^2] d\varepsilon} \\
 & \leq C \sqrt{\frac{2}{M s_N} \int_{s_N}^{2s_N} \mathbb{E}(\langle \mathcal{L} - \langle \mathcal{L} \rangle)^2) d\varepsilon}.
 \end{aligned}$$

Thus, we may conclude, as in the proof of Theorem 2.11, that

$$\frac{1}{s_N} \int_{s_N}^{2s_N} \mathbb{E}[\|\langle \mathbf{R} \rangle\|_{\mathbb{F}}^2 - \|\langle \tilde{\mathbf{R}} \rangle\|_{\mathbb{F}}^2] d\varepsilon \leq \Gamma(N, M). \quad (4.14)$$

Next, we see that

$$\begin{aligned}
 \mathbb{E}\langle \|\tilde{\mathbf{R}} - \langle \tilde{\mathbf{R}} \rangle\|_{\mathbb{F}}^2 \rangle &= \mathbb{E}\langle \tilde{\mathbf{R}}^{\top} \tilde{\mathbf{R}} \rangle - \mathbb{E}\langle \tilde{\mathbf{R}}^{\top} \rangle \langle \tilde{\mathbf{R}} \rangle \\
 &= \mathbb{E}\langle \mathbf{R}^{\top} \mathbf{R} \rangle - \mathbb{E}\langle \tilde{\mathbf{R}}^{\top} \rangle \langle \tilde{\mathbf{R}} \rangle \\
 &= \mathbb{E}\langle \|\mathbf{R} - \langle \mathbf{R} \rangle\|_{\mathbb{F}}^2 \rangle + \mathbb{E}[\|\langle \mathbf{R} \rangle\|_{\mathbb{F}}^2 - \|\langle \tilde{\mathbf{R}} \rangle\|_{\mathbb{F}}^2],
 \end{aligned}$$

where the second line follows from using the Nishimori identity on the first term and the final line follows from adding and subtracting  $\mathbb{E}\|\langle \mathbf{R} \rangle\|_{\mathbb{F}}^2$ . Integrating this over  $\varepsilon$  and using Theorem 2.11 together with inequality (4.14) yields inequality (4.12). To obtain inequality (4.13) from (4.12), we need only observe that due to the Nishimori identity,

$$\mathbb{E}\langle \|\mathbf{R} - \langle \tilde{\mathbf{R}} \rangle\|_{\mathbb{F}}^2 \rangle = \mathbb{E}\langle \|\tilde{\mathbf{R}} - \langle \tilde{\mathbf{R}} \rangle\|_{\mathbb{F}}^2 \rangle. \quad \blacksquare$$

### 4.3. Negligibility of the side information

In the previous subsection, we showed that upon introduction of the perturbation Hamiltonian  $H_{N,\varepsilon}(X)$  (see (2.22)), the overlap matrix  $R_{10}$  (see (2.8)) can be shown to concentrate with respect to the general Gibbs average  $\langle \cdot \rangle'_{N,\varepsilon}$  for  $M = o(N^{1/4})$ . Our interest lies in the particular Gibbs average  $\langle \cdot \rangle'_{N,\varepsilon}$  specified by equation (2.24), which relates via the cavity method described in Section 5 to the perturbed Hamiltonian  $\tilde{H}_{N,\varepsilon}(X)$  (see (2.14)) and perturbed free entropy

$$\tilde{F}_N(\lambda) := \frac{1}{NM} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0, \varepsilon} \ln \int_{\mathbb{R}^{N \times M}} e^{\tilde{H}_{N,\varepsilon}(X)} d\mathbb{P}_{X,N,M}(X),$$

where the average over  $\varepsilon$  is with respect to the uniform measure on  $[s_N, 2s_N]$ . We have thus far claimed that upper bounding  $\limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda)$  by  $\sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda)$  is sufficient for proving Proposition 2.8 since said quantity is itself an upper bound on the limit supremum of the free entropy  $F_N(\lambda)$  of the spiked Wigner model (1.1). We now present a more precise finite  $N$  result that implies this claim.

**Lemma 4.2** (Negligibility of perturbation). *Let  $(s_N)_{N \geq 1}$  be a sequence of numbers in  $(0, 1)$  that decrease slowly to zero as  $N \rightarrow \infty$  and let  $\varepsilon = \varepsilon_N \in [s_N, 2s_N]$ . Recall the definition (1.3) of the partition function  $Z_{N,M}$  and that the perturbed partition function is given by*

$$\tilde{Z}_{N,M} = \int_{\mathbb{R}^{N \times M}} e^{\tilde{H}_{N,\varepsilon}(X)} d\mathbb{P}_{X,N,M}(X)$$

with perturbed Hamiltonian  $\tilde{H}_{N,\varepsilon}$  specified by equations (2.14) and (2.22). Then, for each  $N \geq 1$ , we have that

$$\frac{1}{NM} \mathbb{E}_{\mathbf{Z}, X_0} \ln Z_{N,M} \leq \frac{1}{NM} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \ln \tilde{Z}_{N,M}. \quad (4.15)$$

*Proof.* We employ an interpolation argument akin to that of Section 4.1. Hence, define an interpolating Hamiltonian

$$\begin{aligned} H_t(X) &:= \frac{1}{2} \text{Tr} \left( \sqrt{\frac{\lambda}{N}} \mathbf{Z} X X^\top + \frac{\lambda}{N} X_0 X_0^\top X X^\top - \frac{\lambda}{2N} X X^\top X X^\top \right) \\ &\quad + \text{Tr} \left( \sqrt{\varepsilon t} \tilde{\mathbf{Z}} X X^\top + \varepsilon t X_0 X_0^\top - \frac{\varepsilon t}{2} X X^\top \right), \end{aligned}$$

with  $\mathbf{Z}, X_0$  as in equation (1.1) and  $\tilde{\mathbf{Z}}$  as in equation (2.22). The associated interpolating free entropy is

$$\varphi(t) := \frac{1}{NM} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \ln \int_{\mathbb{R}^{N \times M}} e^{H_t(X)} d\mathbb{P}_{X,N,M}(X).$$

Writing

$$\langle \cdot \rangle_t := \frac{1}{\widehat{Z}_{N,t}} \int_{\mathbb{R}^{N \times M}} \cdot e^{H_t(X)} d\mathbb{P}_{X,N,M}, \quad \widehat{Z}_{N,t} := \int_{\mathbb{R}^{N \times M}} e^{H_t(X)} d\mathbb{P}_{X,N,M},$$

we see that

$$\begin{aligned} \varphi'(t) &= \frac{1}{NM} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \left\langle \frac{d}{dt} H_t(X) \right\rangle_t \\ &= \frac{1}{NM} \text{Tr} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \left\langle \frac{1}{2} \sqrt{\frac{\varepsilon}{t}} \tilde{\mathbf{Z}} \mathbf{X}^\top + \varepsilon \mathbf{X}_0 \mathbf{X}^\top - \frac{\varepsilon}{2} \mathbf{X} \mathbf{X}^\top \right\rangle_t \\ &= \frac{\varepsilon}{2M} \text{Tr} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \langle \mathbf{R}_{10} \rangle_t \\ &\geq 0, \end{aligned}$$

with the last two lines following from Gaussian integration by parts with respect to  $\tilde{\mathbf{Z}}$ , the Nishimori identity, and the positive semidefiniteness of

$$\mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \langle \mathbf{R}_{10} \rangle_t = \frac{1}{N} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \langle \mathbf{X}^\top \rangle_t \langle \mathbf{X} \rangle_t.$$

By the definition of  $\varphi(t)$ , we then have that

$$\frac{1}{NM} \mathbb{E}_{\mathbf{Z}, \tilde{\mathbf{Z}}, X_0} \ln \tilde{Z}_{N,M} - \frac{1}{NM} \mathbb{E}_{\mathbf{Z}, X_0} \ln Z_{N,M} = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t) dt \geq 0,$$

as desired.  $\blacksquare$

Since the left-hand side of equation (4.15) is independent of  $\varepsilon$ , it is immediate from integrating both sides over  $\varepsilon \in [s_N, 2s_N]$  and then taking the limit supremum that

$$\limsup_{N \rightarrow \infty} F_N(\lambda) \leq \limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda). \quad (4.16)$$

## 5. Treatment of growing rank: The multiscale cavity method

The main methodological novelty begins now. In this section, we prove that the limit supremum of the free entropy  $\tilde{F}_N(\lambda)$  of the perturbed model corresponding to the Hamiltonian  $\tilde{H}_{N,\varepsilon}(X)$  (see (2.14)) is bounded above by the supremum of the rank-one replica symmetric potential  $F_1^{\text{RS}}(q, \lambda)$  (see (2.3)). By Lemma 4.2 above, this is sufficient for proving Proposition 2.8, which combines with Proposition 2.7 to produce Theorem 2.1, as described in Section 2.3 and detailed in Section 5.4.

For convenience, we repeat definitions of the involved Hamiltonians and cavity fields here, taking care to highlight now-relevant dependencies on SNR and various

sources of noise. First, we recall that we take  $(s_N)_{N \geq 1}$  to be a sequence of numbers in  $(0, 1)$  that decrease slowly to zero as  $N \rightarrow \infty$  and that for each  $N \geq 1$ , we have  $\varepsilon_N \in [s_N, 2s_N]$ . Then, the Hamiltonian of the size  $N$  perturbed model corresponding to the inference problem

$$\begin{cases} \mathbf{Y} = \sqrt{\frac{\lambda}{N}} \mathbf{X}_0 \mathbf{X}_0^\top + \mathbf{Z}, \\ \mathbf{Y}^{(\varepsilon_N)} = \sqrt{\varepsilon_N} \mathbf{X}_0 + \tilde{\mathbf{Z}}, \end{cases}$$

with  $\mathbf{X}_0 \sim \mathbb{P}_{X,N,M}$ ,  $\mathbf{Z} \in \mathbb{R}^{N \times N}$  Wigner, and  $\tilde{\mathbf{Z}} \in \mathbb{R}^{N \times M}$  Ginibre is

$$\tilde{H}_{N,\varepsilon_N}(\mathbf{X}; \mathbf{Z}, \tilde{\mathbf{Z}}, \lambda) = H_N(\mathbf{X}; \mathbf{Z}, \lambda) + H_{N,\varepsilon_N}(\mathbf{X}; \tilde{\mathbf{Z}}), \quad (5.1)$$

where the Hamiltonians corresponding to each channel are respectively

$$\begin{aligned} H_N(\mathbf{X}; \mathbf{Z}, \lambda) &= \frac{1}{2} \text{Tr} \left( \sqrt{\frac{\lambda}{N}} \mathbf{Z} \mathbf{X} \mathbf{X}^\top + \frac{\lambda}{N} \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X} \mathbf{X}^\top - \frac{\lambda}{2N} \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \right), \\ H_{N,\varepsilon_N}(\mathbf{X}; \tilde{\mathbf{Z}}) &= \text{Tr} \left( \sqrt{\varepsilon_N} \tilde{\mathbf{Z}} \mathbf{X}^\top + \varepsilon_N \mathbf{X}_0 \mathbf{X}^\top - \frac{\varepsilon_N}{2} \mathbf{X} \mathbf{X}^\top \right). \end{aligned} \quad (5.2)$$

The related perturbed free entropy and partition function are then

$$\tilde{F}_N(\lambda) = \frac{1}{NM} \mathbb{E} \ln \tilde{Z}_{N,M}, \quad (5.3)$$

$$\tilde{Z}_{N,M} = \int_{\mathbb{R}^{N \times M}} e^{\tilde{H}_{N,\varepsilon_N}(\mathbf{X}; \mathbf{Z}, \tilde{\mathbf{Z}}, \lambda)} d\mathbb{P}_{X,N,M}(\mathbf{X}), \quad (5.4)$$

with the expectation here and henceforth being taken over  $\mathbf{Z}, \tilde{\mathbf{Z}}, \mathbf{X}_0$ , and uniformly over  $\varepsilon_N \in [s_N, 2s_N]$ . Letting  $\boldsymbol{\eta} \in \mathbb{R}^M$ ,  $\boldsymbol{\eta}_0 \sim \mathbb{P}_{X,M}$ ,  $\boldsymbol{\xi} \sim \mathcal{N}(0, I_N)$ ,  $\tilde{\boldsymbol{\xi}} \sim \mathcal{N}(0, I_M)$ , and  $\xi \sim \mathcal{N}(0, 1)$  account for the  $(N+1)$ th coordinate, we have that the size  $N+1$  equivalent of the Hamiltonian  $\tilde{H}_{N,\varepsilon_N}(\mathbf{X}; \mathbf{Z}, \tilde{\mathbf{Z}}, \lambda)$  is given by

$$\begin{aligned} \tilde{H}_{N+1,\varepsilon_{N+1}}(\mathbf{X}, \boldsymbol{\eta}; \mathbf{Z}, \tilde{\mathbf{Z}}, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \xi, \lambda) &= H'_N(\mathbf{X}; \mathbf{Z}, \lambda) + H_{N,\varepsilon_{N+1}}(\mathbf{X}; \tilde{\mathbf{Z}}) \\ &\quad + \boldsymbol{\eta} z(\mathbf{X}; \boldsymbol{\xi}, \lambda) + \boldsymbol{\eta} s(\mathbf{X}; \lambda) \boldsymbol{\eta}^\top + m(\boldsymbol{\eta}; \xi, \lambda) + r_{\varepsilon_{N+1}}(\boldsymbol{\eta}; \tilde{\boldsymbol{\xi}}), \end{aligned} \quad (5.5)$$

with cavity fields

$$\begin{aligned} H'_N(\mathbf{X}; \mathbf{Z}, \lambda) &= \frac{1}{2} \text{Tr} \left( \sqrt{\frac{\lambda}{N+1}} \mathbf{Z} \mathbf{X} \mathbf{X}^\top + \frac{\lambda}{N+1} \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X} \mathbf{X}^\top \right. \\ &\quad \left. - \frac{\lambda}{2(N+1)} \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \right), \end{aligned} \quad (5.6)$$

$$z(\mathbf{X}; \boldsymbol{\xi}, \lambda) = \sqrt{\frac{\lambda}{N+1}} \mathbf{X}^\top \boldsymbol{\xi}^\top + \frac{\lambda}{N+1} \mathbf{X}^\top \mathbf{X}_0 \boldsymbol{\eta}_0^\top, \quad (5.7)$$

$$s(\mathbf{X}; \lambda) = -\frac{\lambda}{2(N+1)} \mathbf{X}^\top \mathbf{X}, \quad (5.8)$$

$$m(\boldsymbol{\eta}; \boldsymbol{\xi}, \lambda) = \frac{1}{2} \left( \sqrt{\frac{\lambda}{N+1}} \boldsymbol{\xi} + \frac{\lambda}{N+1} \boldsymbol{\eta}_0 \boldsymbol{\eta}_0^\top - \frac{\lambda}{2(N+1)} \boldsymbol{\eta} \boldsymbol{\eta}^\top \right) \boldsymbol{\eta} \boldsymbol{\eta}^\top, \quad (5.9)$$

$$r_{\varepsilon_{N+1}}(\boldsymbol{\eta}; \tilde{\boldsymbol{\xi}}) = \sqrt{\varepsilon_{N+1}} \tilde{\boldsymbol{\xi}} \boldsymbol{\eta}^\top + \varepsilon_{N+1} \boldsymbol{\eta}_0 \boldsymbol{\eta}^\top - \frac{\varepsilon_{N+1}}{2} \boldsymbol{\eta} \boldsymbol{\eta}^\top. \quad (5.10)$$

The perturbed free entropy  $\tilde{F}_{N+1}(\lambda)$  and partition function  $\tilde{Z}_{N+1, M}$  are defined as their size  $N$  analogs according to equations (5.3) and (5.4), but with the expectation  $\mathbb{E}[\cdot]$  now being taken over  $\varepsilon_{N+1} \in [s_{N+1}, 2s_{N+1}]$  instead of  $\varepsilon_N$  and over the additional variables  $\boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\xi}, \boldsymbol{\eta}_0$ .

In Section 5.1, we prove Theorem 2.9, which shows that  $\limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda)$  is bounded above by a convex combination of the limit suprema of  $\Delta_N/M$  and  $\Delta_M/N$ , where

$$\Delta_N = \mathbb{E} \ln \tilde{Z}_{N+1, M_{N+1}} - \mathbb{E} \ln \tilde{Z}_{N, M_{N+1}}, \quad (5.11)$$

$$\Delta_M = \mathbb{E} \ln \tilde{Z}_{N, M_{N+1}} - \mathbb{E} \ln \tilde{Z}_{N, M_N}. \quad (5.12)$$

As  $\Delta_N$  corresponds to a cavity in  $N$  of fixed rank  $M_{N+1}$  and vice versa for  $\Delta_M$ , this enables us to decompose the problem of dealing with two growing indices  $M, N$  into two separate single-index cavity computations. It turns out that said convex combination has the desired upper bound  $\sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda)$  due to both of the involved limit suprema being bounded above by this quantity. This is shown in Section 5.2 and Section 5.3. In the former, we give a detailed presentation of the standard fixed-rank Aizenman–Sims–Starr scheme ([3], [43, Sections 4.3 and 6.2.5]), while in the latter, we prove the heuristic (valid for  $N$  such that  $M_{N+1} = M_N + 1$ )

$$\begin{aligned} \frac{\Delta_M}{N} &= \frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M_{N+1}} - \frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M_N} \\ &\approx (M_N + 1) \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) - M_N \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) = \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda). \end{aligned}$$

### 5.1. Free entropy upper bound: Multiscale Aizenman–Sims–Starr scheme

In this subsection, we allow the matrix size  $N$  and rank  $M$  to grow generically with each other. Our primary goal is to show how the standard Cesàro sum extends to the multiscale setting to give

$$\limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda) \leq \alpha \limsup_{N \rightarrow \infty} \frac{\Delta_N}{M} + (1 - \alpha) \limsup_{N \rightarrow \infty} \frac{\Delta_M}{N}$$

for some  $0 \leq \alpha \leq 1$ .

*Proof of Theorem 2.9.* We focus on proving the upper bound (2.16), since the analogous lower bound (2.17) will follow from obvious modifications. Let  $\epsilon > 0$ , and for  $1 \leq n \leq N$ , define  $m_n = M_n$  and its generalized inverse

$$n_m = \inf\{n \mid m_n \geq m\}$$

in order to denote the  $m$  index as a function of  $n$  and vice versa; these dependencies are the same as between  $M_N$  and  $N$ . By the definition of the limit supremum, there exists  $T = T(\epsilon)$  independent of  $N$  such that for all  $n \geq T$ ,

$$\frac{\tilde{\Delta}_N(n)}{m_n} \leq \limsup_{N \rightarrow \infty} \frac{\Delta_N}{M} + \epsilon \quad (5.13)$$

and (considering only  $n \geq T$  for which  $m_{n+1} \neq m_n$ )

$$\frac{\tilde{\Delta}_M(n)}{n} \leq \limsup_{N \rightarrow \infty} \frac{\Delta_M}{N} + \epsilon. \quad (5.14)$$

With  $T$  in hand, we may now proceed as in the standard proof of the Aizenman–Sims–Starr scheme. Thus, express the log partition function as the telescoping sum

$$\begin{aligned} \tilde{F}_N(\lambda) &= \frac{1}{NM} \sum_{n=T}^{N-1} (\mathbb{E} \ln \tilde{Z}_{n+1, m_{n+1}} - \mathbb{E} \ln \tilde{Z}_{n, m_n}) + o_N(T) \\ &= \frac{1}{NM} \left( \sum_{n=T}^{N-1} \tilde{\Delta}_N(n) + \sum_{n=T}^{N-1} \tilde{\Delta}_M(n) \right) + o_N(T), \end{aligned} \quad (5.15)$$

where  $\tilde{Z}_{0,0} \equiv 1$  by convention, we write

$$\begin{aligned} \tilde{\Delta}_N(n) &:= \mathbb{E} \ln \tilde{Z}_{n+1, m_{n+1}} - \mathbb{E} \ln \tilde{Z}_{n, m_{n+1}}, \\ \tilde{\Delta}_M(n) &:= \mathbb{E} \ln \tilde{Z}_{n, m_{n+1}} - \mathbb{E} \ln \tilde{Z}_{n, m_n}, \end{aligned}$$

and the error term is the truncated sum

$$o_N(T) = \frac{1}{NM} \sum_{n=0}^{T-1} (\tilde{\Delta}_N(n) + \tilde{\Delta}_M(n)),$$

which vanishes as  $N \rightarrow \infty$ :  $T$  is independent of  $N$  and for each  $0 \leq n \leq T-1$ ,  $\tilde{\Delta}_N(n)$  and  $\tilde{\Delta}_M(n)$  are finite  $(nm_n)$ -dimensional integrals that are also independent of  $N$ .

Notice that  $\tilde{\Delta}_M(n) = 0$  if  $m_{n+1} = m_n$ , so keeping only the terms for which the  $m_n$  coordinate increases, we read from the bound (5.14) that

$$\sum_{n=T}^{N-1} \tilde{\Delta}_M(n) \leq \sum_{m=M_T}^{M_{N-1}} n_m \left( \limsup_{N \rightarrow \infty} \frac{\Delta_M}{N} + \epsilon \right).$$

Substituting this and the analogous statement deriving from the bound (5.13) into equation (5.15) gives the upper bound

$$\begin{aligned} \tilde{F}_N(\lambda) \leq & \left( \frac{1}{NM} \sum_{n=T}^{N-1} m_n \right) \left( \limsup_{N \rightarrow \infty} \frac{\Delta_N}{M} + \epsilon \right) \\ & + \left( \frac{1}{NM} \sum_{m=M_T}^{M_{N-1}} n_m \right) \left( \limsup_{N \rightarrow \infty} \frac{\Delta_M}{N} + \epsilon \right) + o_N(T). \end{aligned} \quad (5.16)$$

To obtain agreement with the bound (2.16), we let

$$\alpha = \limsup_{N \rightarrow \infty} \frac{1}{NM} \sum_{n=T}^{N-1} m_n,$$

so that the proof of inequality (2.16) is completed upon showing that  $0 \leq \alpha \leq 1$  and that

$$\limsup_{N \rightarrow \infty} \frac{1}{NM} \sum_{m=M_T}^{M_{N-1}} n_m = 1 - \alpha. \quad (5.17)$$

As  $1 \leq m_n \leq M$  for each  $1 \leq n \leq N$ , it is immediate that  $1 \leq \sum_{n=T}^{N-1} m_n \leq NM$ , so  $0 \leq \alpha \leq 1$ . To prove the second claim, we first recall Young's inequality for increasing functions [68]:

$$ab \leq \int_0^a f(x) dx + \int_0^b f^{-1}(x) dx, \quad a, b > 0,$$

where  $f$  is differentiable, strictly increasing, and such that  $f(0) = 0$ . When  $b = f(a)$ , this inequality is an equality that we may discretize and adopt to our setting to see that

$$\sum_{n=T}^{N-1} m_n + \sum_{m=M_T}^{M_{N-1}} n_m = NM_{N-1} - T(M_T - 1).$$

Normalizing by  $NM$  and taking the limit supremum then shows that

$$\begin{aligned} \alpha + \limsup_{N \rightarrow \infty} \frac{1}{NM} \sum_{m=M_T}^{M_{N-1}} n_m &= \limsup_{N \rightarrow \infty} \frac{1}{NM} \left( \sum_{n=T}^{N-1} m_n + \sum_{m=M_T}^{M_{N-1}} n_m \right) \\ &= \limsup_{N \rightarrow \infty} \frac{NM_{N-1} - T(M_T - 1)}{NM}, \end{aligned}$$

which equals one, hence equation (5.17) is verified. Taking the limit supremum of both sides of inequality (5.16) and then taking  $\epsilon \rightarrow 0$  completes the proof of the

upper bound. The lower bound follows from a similar argument to that just presented with

$$\beta = \liminf_{N \rightarrow \infty} \frac{1}{NM} \sum_{n=T}^{N-1} m_n$$

in place of  $\alpha$ .

We now show that  $\alpha, \beta$  satisfies the relation (2.20) in the case when  $M = \lfloor f(N) \rfloor$ . Without loss of generality, we modify  $f$  so that  $f(N) \in \mathbb{N}$  to simplify notation. Noting that  $m_n = \lfloor f(n) \rfloor$ , using L'Hôpital's rule to simplify the integral upper bound of the normalized sum shows that

$$\alpha = \limsup_{N \rightarrow \infty} \frac{1}{NM} \sum_{n=T}^{N-1} m_n \leq \limsup_{N \rightarrow \infty} \int_T^N \frac{f(x)}{Nf(N)} dx \leq \limsup_{x \rightarrow \infty} \frac{f(x)}{f(x) + xf'(x)}.$$

Similarly, consideration of the integral lower bound of the same sum gives

$$\beta = \liminf_{N \rightarrow \infty} \frac{1}{NM} \sum_{n=T}^{N-1} m_n \geq \liminf_{N \rightarrow \infty} \int_{T-1}^{N-1} \frac{f(x)}{Nf(N)} dx \geq \liminf_{x \rightarrow \infty} \frac{f(x-1)}{f(x) + xf'(x)},$$

concluding the proof. ■

As mentioned before, it can be observed that the above proof applies to generic sequences with two growing indices and that no properties of the log partition function are used. If  $\tilde{F}_N(\lambda)$  remains finite in the thermodynamic limit, then  $\Delta_N$  is an order  $M$  quantity and  $\Delta_M$  is an order  $N$  quantity, so the normalized differences in the spin and rank coordinates are both of constant order and both have non-negligible contribution to the limiting value when  $\alpha \neq 0, 1$ . In other words, even if  $M_N \ll N$  so that adding a column occurs much less frequently than rows when the system size grows, the free entropy difference due to column addition is much larger than the one due to row addition because columns are of length  $N$ , while rows are of length  $M$ . These two effects perfectly compensate each other so that both row and column addition contributes the same order in the multiscale Aizenman–Sims–Starr scheme.

As Theorem 2.9 shows that the effects of incrementing  $N$  or  $M$  can be isolated to the  $\Delta_N$  and  $\Delta_M$  cavities, we now move on to the task of showing that the limit suprema of both of these are bounded above by the right-hand side of inequality (2.12).

## 5.2. The standard row cavity bound

As  $\Delta_N$  (see (5.11)) is a difference of two terms that correspond to the same rank  $M_{N+1}$ , it suffices to bound  $\limsup_{N \rightarrow \infty} \Delta_N/M$  using the standard fixed-rank arguments of [3, 43] while tracking any  $M$ -dependencies that could become problematic

in the sublinear-rank regime. Thus, we begin by showing that  $\Delta_N/M$  is bounded above by a simpler analog that is obtained by showing that in the large  $N$  limit:

- the cavity fields  $m(\boldsymbol{\eta})$  (see (5.9)),  $r_{\varepsilon_{N+1}}(\boldsymbol{\eta})$  (see (5.10)),  $H_{N,\varepsilon_{N+1}}(\mathbf{X}) - H_{N,\varepsilon_N}(\mathbf{X})$  (see (5.2)) become negligible;
- the cavity fields  $z(\mathbf{X})$  (see (5.7)) and  $s(\mathbf{X})$  (see (5.8)) can be replaced by

$$\begin{aligned}\widehat{z}(\mathbf{X}; \boldsymbol{\xi}, \lambda) &:= \sqrt{\frac{\lambda}{N}} \mathbf{X}^\top \boldsymbol{\xi}^\top + \lambda \mathbf{R}_{10} \boldsymbol{\eta}_0^\top, \\ \widehat{s}(\mathbf{X}; \lambda) &:= -\frac{\lambda}{2} \mathbf{R}_{11},\end{aligned}$$

where we recall that  $\mathbf{R}_{10} = \frac{1}{N} \mathbf{X}^\top \mathbf{X}_0$  (see (2.8)) and write  $\mathbf{R}_{11} = \frac{1}{N} \mathbf{X}_1^\top \mathbf{X}_1 = \frac{1}{N} \mathbf{X}^\top \mathbf{X}$  in accordance with equation (4.2);

- and the Hamiltonian  $\widetilde{H}_{N,\varepsilon_N}(\mathbf{X})$  is statistically equivalent to

$$H_N(\mathbf{X}) + H_{N,\varepsilon_N}(\mathbf{X}) + \widehat{y}(\mathbf{X}),$$

where for  $\widehat{\mathbf{Z}}$ , an independent copy of  $\mathbf{Z}$ ,

$$\widehat{y}(\mathbf{X}; \mathbf{X}_0, \widehat{\mathbf{Z}}, \lambda) = \frac{1}{2} \text{Tr} \left( \frac{\sqrt{\lambda}}{N} \widehat{\mathbf{Z}} \mathbf{X} \mathbf{X}^\top + \lambda \mathbf{R}_{10}^\top \mathbf{R}_{10} - \frac{\lambda}{2} \mathbf{R}_{11}^\top \mathbf{R}_{11} \right). \quad (5.18)$$

Then, we use the thermal concentration results of Theorem 2.11 and Corollary 4.1 to show that said upper bound can be simplified even further to be written in terms of the replica symmetric potential  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  (see (2.2)).

**Lemma 5.1** (Standard Aizenman–Sims–Starr scheme). *Assume that hypothesis (H2) holds. Let  $\mathbb{P}_X$  be a centered distribution with bounded fourth moment, let  $\boldsymbol{\eta}_0 \sim \mathbb{P}_{X,M}$ , let  $\widetilde{\mathbf{Z}}_{N,M}$  be specified by equation (5.4), and let  $\widehat{z}(\mathbf{X})$ ,  $\widehat{s}(\mathbf{X})$ , and  $\widehat{y}(\mathbf{X})$  be as above. Then,*

$$\begin{aligned}& \frac{\mathbb{E}_{\boldsymbol{\xi}, \widetilde{\boldsymbol{\xi}}, \boldsymbol{\xi}, \boldsymbol{\eta}_0} \mathbb{E} \ln \widetilde{\mathbf{Z}}_{N+1,M} - \mathbb{E} \ln \widetilde{\mathbf{Z}}_{N,M}}{M} \\ & \leq \frac{1}{M} \mathbb{E}_{\boldsymbol{\xi}, \boldsymbol{\eta}_0} \mathbb{E} \ln \left\langle \int_{\mathbb{R}^M} e^{\boldsymbol{\eta} \widehat{z}(\mathbf{X}; \boldsymbol{\xi}, \lambda) + \boldsymbol{\eta} \widehat{s}(\mathbf{X}; \lambda) \boldsymbol{\eta}^\top} d\mathbb{P}_{X,M}(\boldsymbol{\eta}) \right\rangle'_{N,\varepsilon_N} \\ & \quad - \frac{1}{M} \mathbb{E}_{\widehat{\mathbf{Z}}} \mathbb{E} \ln \left\langle e^{\widehat{y}(\mathbf{X}; \mathbf{X}_0, \widehat{\mathbf{Z}}, \lambda)} \right\rangle'_{N,\varepsilon_N} + \Gamma'(N, M),\end{aligned} \quad (5.19)$$

where the unspecified expectation  $\mathbb{E}[\cdot]$  is as in (5.3), the Gibbs average  $\langle \cdot \rangle'_{N,\varepsilon_N}$  is specified by equation (2.24), and we define, with  $C$  some finite positive constant independent of  $M, N$ ,

$$\Gamma'(N, M) := C \left( \frac{M - N}{N + 1} + 1 + s_N \right),$$

which is  $o_N(1)$  for any  $M = o(N)$ .

*Proof.* We break the proof into two parts. First, we quantify the difference

$$\frac{\mathbb{E}_{\xi, \tilde{\xi}, \xi, \eta_0} \mathbb{E} \ln \tilde{Z}_{N+1, M}}{M} - \frac{1}{M} \mathbb{E}_{\xi, \eta_0} \mathbb{E} \ln \left[ Z'_{N, \varepsilon_N} \left\langle \int_{\mathbb{R}^M} e^{\eta \hat{z}(X; \xi, \lambda) + \eta \hat{s}(X; \lambda) \eta^\top} d\mathbb{P}_{X, M}(\eta) \right\rangle'_{N, \varepsilon_N} \right], \quad (5.20)$$

where we recall the definition (2.25) of the partition function  $Z'_{N, \varepsilon_N}$ . To this end, we introduce the interpolating Hamiltonian

$$H_{1,t}(X, \eta) := \tilde{H}_{N+1, t s_{N+1} \hat{\varepsilon}}(X, \eta; \mathbf{Z}, \tilde{\mathbf{Z}}, \xi, \tilde{\xi}, \xi, t\lambda) + H'_N(X; U, (1-t)\lambda) \\ + H_{N, (1-t) s_N \hat{\varepsilon}}(X; \tilde{U}) + \eta \hat{z}(X; \mu, (1-t)\lambda) + \eta \hat{s}(X; (1-t)\lambda) \eta^\top,$$

with  $U, \tilde{U}, \mu$  being independent copies of  $\mathbf{Z}, \tilde{\mathbf{Z}}, \xi$  and  $\hat{\varepsilon}$  a perturbation parameter distributed uniformly on  $[1, 2]$  so that  $\varepsilon_N = s_N \hat{\varepsilon}$  in distribution. Letting  $\mathbb{E}_1[\cdot]$  denote expectation with respect to the signal variables  $X_0, \eta_0$ , the perturbation parameter  $\hat{\varepsilon}$ , and the noise variables  $\mathbf{Z}, \tilde{\mathbf{Z}}, U, \tilde{U}, \xi, \tilde{\xi}, \xi, \mu$ , the corresponding free entropy

$$\varphi_1(t) := \frac{1}{M} \mathbb{E}_1 \ln \int_{\mathbb{R}^{(N+1) \times M}} e^{H_{1,t}(X, \eta)} d\mathbb{P}_{X, N+1, M}(X, \eta)$$

is such that  $\varphi_1(1)$  and  $\varphi_1(0)$ , respectively, equal the first and second terms of the difference (5.20). As usual, we proceed by computing the derivative of this free entropy and see upon substituting in equation (5.5) and writing  $\langle \cdot \rangle_{1,t}$  for the Gibbs average with respect to the Hamiltonian  $H_{1,t}(X, \eta)$  that

$$\varphi'_1(t) = \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} H_{1,t}(X, \eta) \right\rangle_{1,t} \\ = \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} [H'_N(X; \mathbf{Z}, t\lambda) + H'_N(X; U, (1-t)\lambda)] \right\rangle_{1,t} \\ + \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} [H_{N, t s_{N+1} \hat{\varepsilon}}(X; \tilde{\mathbf{Z}}) + H_{N, (1-t) s_N \hat{\varepsilon}}(X; \tilde{U})] \right\rangle_{1,t} \\ + \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} \eta [z(X; \xi, t\lambda) + \hat{z}(X; \mu, (1-t)\lambda)] \right\rangle_{1,t} \\ + \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} \eta [s(X; t\lambda) + \hat{s}(X; (1-t)\lambda)] \eta^\top \right\rangle_{1,t} \\ + \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} m(\eta; \xi, t\lambda) \right\rangle_{1,t} + \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} r_{t s_{N+1} \hat{\varepsilon}}(\eta; \tilde{\xi}) \right\rangle_{1,t}.$$

By Gaussian integration by parts and the Nishimori identity, we see that

$$\frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} [H'_N(X; \mathbf{Z}, t\lambda) + H'_N(X; U, (1-t)\lambda)] \right\rangle_{1,t} = 0,$$

$$\begin{aligned}
 & \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} [H_{N,tS_{N+1}} \hat{\varepsilon}(\mathbf{X}; \tilde{\mathbf{Z}}) + H_{N,(1-t)S_N} \hat{\varepsilon}(\mathbf{X}; \tilde{\mathbf{U}})] \right\rangle_{1,t} \\
 &= \frac{N(S_{N+1} - S_N)}{2M} \text{Tr} \mathbb{E}_1 \langle \hat{\varepsilon} \mathbf{R}_{10} \rangle_{1,t}, \\
 & \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} \eta [z(\mathbf{X}; \xi, t\lambda) + \hat{z}(\mathbf{X}; \mu, (1-t)\lambda)] \right\rangle_{1,t} \\
 &= -\frac{\lambda}{2M(N+1)} \mathbb{E}_1 \langle \eta \mathbf{R}_{10} \eta_0^\top + \eta \mathbf{R}_{11} \eta^\top \rangle_{1,t}, \\
 & \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} \eta [s(\mathbf{X}; t\lambda) + \hat{s}(\mathbf{X}; (1-t)\lambda)] \eta^\top \right\rangle_{1,t} = \frac{1}{2M(N+1)} \mathbb{E}_1 \langle \eta \mathbf{R}_{11} \eta^\top \rangle_{1,t}, \\
 & \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} m(\eta; \xi, t\lambda) \right\rangle_{1,t} = \frac{\lambda}{4M(N+1)} \mathbb{E}_1 \langle \eta_0 \eta_0^\top \eta \eta^\top \rangle_{1,t}, \\
 & \frac{1}{M} \mathbb{E}_1 \left\langle \frac{d}{dt} r_{tS_{N+1}} \hat{\varepsilon}(\eta; \tilde{\xi}) \right\rangle_{1,t} = \frac{S_{N+1}}{2M} \mathbb{E}_1 \langle \hat{\varepsilon} \eta_0 \eta^\top \rangle_{1,t}.
 \end{aligned}$$

Let  $\mathbb{E}'_1[\cdot]$  denote the same expectation as  $\mathbb{E}_1[\cdot]$  but without the average over  $\hat{\varepsilon}$  so that  $\mathbb{E}_1[\cdot] = \mathbb{E}_{\hat{\varepsilon}} \mathbb{E}'_1[\cdot]$ . Due to the same reasoning as behind the series of inequalities (4.6), we have that

$$0 \leq \text{Tr} \mathbb{E}'_1 \langle \mathbf{R}_{10} \rangle_{1,t}, \mathbb{E}_1 \langle \eta \mathbf{R}_{10} \eta_0^\top \rangle_{1,t}, \frac{1}{M} \mathbb{E}_1 \langle \eta_0 \eta_0^\top \eta \eta^\top \rangle_{1,t}, \mathbb{E}'_1 \langle \eta_0 \eta^\top \rangle_{1,t} \leq CM$$

for some finite positive constant  $C$  independent of  $M, N$  and depending only on properties of  $\mathbb{P}_X$ . Then, since  $s_N$  is a decreasing sequence, we see that

$$\frac{N(S_{N+1} - S_N)}{2M} \text{Tr} \mathbb{E}_1 \langle \hat{\varepsilon} \mathbf{R}_{10} \rangle_{1,t} = \frac{N(S_{N+1} - S_N)}{2M} \mathbb{E}_{\hat{\varepsilon}} [\hat{\varepsilon} \text{Tr} \mathbb{E}'_1 \langle \mathbf{R}_{10} \rangle_{1,t}] \leq 0,$$

as we have the product of the expectation of a positive trace and the negative quantity  $N(S_{N+1} - S_N)/(2M)$ . Likewise, we have that the other terms in  $\varphi'_1(t)$  can be bounded according to

$$\begin{aligned}
 & -\frac{\lambda}{2M(N+1)} \mathbb{E}_1 \langle \eta \mathbf{R}_{10} \eta_0^\top + \eta \mathbf{R}_{11} \eta^\top \rangle_{1,t} + \frac{1}{2M(N+1)} \mathbb{E}_1 \langle \eta \mathbf{R}_{11} \eta^\top \rangle_{1,t} \\
 &= -\frac{\lambda}{2M(N+1)} \mathbb{E}_1 \langle \eta \mathbf{R}_{10} \eta_0^\top \rangle_{1,t} \leq 0, \\
 & \frac{\lambda}{4M(N+1)} \mathbb{E}_1 \langle \eta_0 \eta_0^\top \eta \eta^\top \rangle_{1,t} \leq \frac{C\lambda M}{4(N+1)}, \\
 & \frac{S_{N+1}}{2M} \mathbb{E}_1 \langle \hat{\varepsilon} \eta_0 \eta^\top \rangle_{1,t} = \frac{S_{N+1}}{2M} \mathbb{E}_{\hat{\varepsilon}} [\hat{\varepsilon} \mathbb{E}'_1 \langle \eta_0 \eta^\top \rangle_{1,t}] \leq \frac{C S_{N+1}}{2} \mathbb{E}_{\hat{\varepsilon}} \hat{\varepsilon} \leq 3C S_N.
 \end{aligned}$$

Combining all of this together shows that

$$\varphi'_1(t) \leq \frac{C\lambda M}{4(N+1)} + 3C S_N.$$

Since this is independent of  $t$ , we have that

$$\varphi_1(1) - \varphi_1(0) = \int_0^1 \varphi_1'(t) dt \leq \frac{C\lambda M}{4(N+1)} + 3C_{SN}, \quad (5.21)$$

so the difference (5.20) obeys the same upper bound.

We now move on to quantifying the difference

$$\frac{1}{M} \mathbb{E}_{\hat{\mathbf{Z}}} \mathbb{E} \ln [Z'_{N,\varepsilon_N} \langle e^{\hat{\gamma}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, \lambda)} \rangle'_{N,\varepsilon_N}] - \frac{1}{M} \mathbb{E} \ln \tilde{Z}_{N,M}. \quad (5.22)$$

The relevant interpolating Hamiltonian and free entropy are given by

$$\begin{aligned} H_{2,t}(\mathbf{X}) &:= H'_N(\mathbf{X}; \mathbf{Z}, t\lambda) + H_{N,t\varepsilon_N}(\mathbf{X}; \tilde{\mathbf{Z}}) + \hat{\gamma}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, t\lambda) \\ &\quad + \tilde{H}_{N,(1-t)\varepsilon_N}(\mathbf{X}; \mathbf{U}, \tilde{\mathbf{U}}, (1-t)\lambda), \\ \varphi_2(t) &:= \frac{1}{M} \mathbb{E}_2 \ln \int_{\mathbb{R}^{N \times M}} e^{H_{2,t}(\mathbf{X})} d\mathbb{P}_{X,N,M}(\mathbf{X}), \end{aligned}$$

where  $\mathbf{U}, \tilde{\mathbf{U}}$  are independent copies of  $\mathbf{Z}, \tilde{\mathbf{Z}}$ , as before, and  $\mathbb{E}_2[\cdot]$  denotes an expectation with respect to  $\mathbf{X}_0, \mathbf{Z}, \tilde{\mathbf{Z}}, \hat{\mathbf{Z}}, \mathbf{U}, \tilde{\mathbf{U}}$ , and  $\varepsilon_N$ . Observe that  $\varphi_2(1)$  and  $\varphi_2(0)$ , respectively, equal the first and second terms of the difference (5.22). Writing  $\langle \cdot \rangle_{2,t}$  for the Gibbs average with respect to  $H_{2,t}(\mathbf{X})$  and using equation (5.1) shows that

$$\begin{aligned} \varphi_2'(t) &= \frac{1}{M} \mathbb{E}_2 \left\langle \frac{d}{dt} H_{2,t}(\mathbf{X}) \right\rangle_{2,t} \\ &= \frac{1}{M} \mathbb{E}_2 \left\langle \frac{d}{dt} [H'_N(\mathbf{X}; \mathbf{Z}, t\lambda) + H_N(\mathbf{X}; \mathbf{U}, (1-t)\lambda)] \right\rangle_{2,t} \\ &\quad + \frac{1}{M} \mathbb{E}_2 \left\langle \frac{d}{dt} [H_{N,t\varepsilon_N}(\mathbf{X}; \tilde{\mathbf{Z}}) + H_{N,(1-t)\varepsilon_N}(\mathbf{X}; \tilde{\mathbf{U}})] \right\rangle_{2,t} \\ &\quad + \frac{1}{M} \mathbb{E}_2 \left\langle \frac{d}{dt} \hat{\gamma}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, t\lambda) \right\rangle_{2,t}. \end{aligned}$$

By Gaussian integration by parts and the Nishimori identity, we see that

$$\begin{aligned} &\frac{1}{M} \mathbb{E}_2 \left\langle \frac{d}{dt} [H'_N(\mathbf{X}; \mathbf{Z}, t\lambda) + H_N(\mathbf{X}; \mathbf{U}, (1-t)\lambda)] \right\rangle_{2,t} \\ &= -\frac{\lambda N}{4(N+1)M} \text{Tr} \mathbb{E}_2 \langle \mathbf{R}_{10}^\top \mathbf{R}_{10} \rangle_{2,t}, \\ &\frac{1}{M} \mathbb{E}_2 \left\langle \frac{d}{dt} [H_{N,t\varepsilon_N}(\mathbf{X}; \tilde{\mathbf{Z}}) + H_{N,(1-t)\varepsilon_N}(\mathbf{X}; \tilde{\mathbf{U}})] \right\rangle_{2,t} = 0, \\ &\frac{1}{M} \mathbb{E}_2 \left\langle \frac{d}{dt} \hat{\gamma}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, t\lambda) \right\rangle_{2,t} = \frac{\lambda}{4M} \text{Tr} \mathbb{E}_2 \langle \mathbf{R}_{10}^\top \mathbf{R}_{10} \rangle_{2,t}. \end{aligned}$$

Combining these equalities shows that

$$\varphi_2'(t) = \frac{\lambda}{4M} \left( 1 - \frac{N}{N+1} \right) \text{Tr} \mathbb{E}_2 \langle \mathbf{R}_{10}^\top \mathbf{R}_{10} \rangle_{2,t}.$$

Extending the logic behind the bounds (4.6) shows that

$$\begin{aligned} 0 &\leq \text{Tr} \mathbb{E}_2 \langle \mathbf{R}_{10}^\top \mathbf{R}_{10} \rangle_{2,t} \leq \frac{1}{N^2} \text{Tr} \mathbb{E}_2 [\mathbf{X}_0^\top \mathbf{X}_0 \mathbf{X}_0^\top \mathbf{X}_0] \\ &= \frac{1}{N^2} \sum_{i,j=1}^N \sum_{\ell,\ell'=1}^M \mathbb{E}_2 [\mathbf{x}_{0,i\ell} \mathbf{x}_{0,i\ell'} \mathbf{x}_{0,j\ell} \mathbf{x}_{0,j\ell'}] \leq C \left( M + \frac{M^2}{N} \right), \end{aligned}$$

where the last inequality follows from the fact that, since the entries of  $\mathbf{X}_0$  are i.i.d. with bounded fourth moments,  $\mathbb{E}_2 [\mathbf{x}_{0,i\ell} \mathbf{x}_{0,i\ell'} \mathbf{x}_{0,j\ell} \mathbf{x}_{0,j\ell'}]$  is given by some finite positive constant  $C$  if either  $i = j$  or  $\ell = \ell'$ , and is otherwise zero; the summand is non-zero for  $\mathfrak{o}(MN^2 + M^2N)$  combinations of  $i, j, \ell, \ell'$ . As  $M \leq N$ , we hence see that

$$\varphi'_2(t) \leq \frac{C\lambda}{4} \left( 1 - \frac{N}{N+1} \right)$$

and integrating this over  $t$  shows that the difference (5.22) also obeys this upper bound. Summing this with the upper bound (5.21) on the difference (5.20) produces  $\Gamma'(N, M)$ . Thus, we are done upon noting that  $\mathbb{E}_{\xi, \eta_0} \mathbb{E} \ln Z'_{N, \varepsilon N} = \mathbb{E}_{\hat{\mathbf{z}}} \mathbb{E} \ln Z'_{N, \varepsilon N}$ . ■

Lemma 5.1 displays a similarity in structure to the replica symmetric potential  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$  (see (2.2)). Our goal now is to refine this observation. Writing

$$\begin{aligned} \hat{H}_{M,N}(\mathbf{A}, \mathbf{B}, \mathbf{x}; \mathbf{z}^{(N)}, \lambda) &:= \sqrt{\frac{\lambda}{N}} \mathbf{x}^\top \mathbf{A}^\top \mathbf{z}^{(N)} \\ &\quad + \frac{\lambda}{N} \mathbf{x}^\top \mathbf{A}^\top \mathbf{B} \mathbf{x}_0 - \frac{\lambda}{2N} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x} \end{aligned} \quad (5.23)$$

for  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times M}$ ,  $\mathbf{x} \in \mathbb{R}^M$ ,  $\mathbf{z}^{(N)} \sim \mathcal{N}(0, I_N)$ , and  $\mathbf{x}_0 \sim \mathbb{P}_{X,M}$ , one observes upon replacing  $\xi^\top, \eta^\top, \eta_0^\top$ , respectively, by  $\mathbf{z}^{(N)}, \mathbf{x}, \mathbf{x}_0$  that the first term in the right-hand side of inequality (5.19) is

$$\frac{1}{M} \mathbb{E}_{\mathbf{z}^{(N)}, \mathbf{x}_0} \mathbb{E} \ln \left\langle \int_{\mathbb{R}^M} e^{\hat{H}_{M,N}(\mathbf{X}, \mathbf{X}_0, \mathbf{x}; \mathbf{z}^{(N)}, \lambda)} d\mathbb{P}_{X,M}(\mathbf{x}) \right\rangle'_{N,\varepsilon}.$$

On the other hand, further replacing  $\mathbf{X}$  and  $\mathbf{X}_0$  by  $\langle \mathbf{X} \rangle'_{N,\varepsilon}$  in the integrand produces

$$\begin{aligned} &\frac{1}{M} \mathbb{E}_{\mathbf{z}^{(N)}, \mathbf{x}_0} \mathbb{E} \ln \left\langle \int_{\mathbb{R}^M} e^{\hat{H}_{M,N}(\langle \mathbf{X} \rangle'_{N,\varepsilon}, \langle \mathbf{X} \rangle'_{N,\varepsilon}, \mathbf{x}; \mathbf{z}^{(N)}, \lambda)} d\mathbb{P}_{X,M}(\mathbf{x}) \right\rangle'_{N,\varepsilon} \\ &= \frac{1}{M} \mathbb{E}_{\mathbf{z}^{(N)}, \mathbf{x}_0} \mathbb{E} \ln \int_{\mathbb{R}^M} e^{\hat{H}_{M,N}(\langle \mathbf{X} \rangle'_{N,\varepsilon}, \langle \mathbf{X} \rangle'_{N,\varepsilon}, \mathbf{x}; \mathbf{z}^{(N)}, \lambda)} d\mathbb{P}_{X,M}(\mathbf{x}) \end{aligned} \quad (5.24)$$

$$= \frac{1}{M} \mathbb{E}_{\mathbf{z}^{(M)}, \mathbf{x}_0} \mathbb{E} \ln \int_{\mathbb{R}^M} e^{\hat{H}_{M,M}(\sqrt{\mathbf{T}}, \sqrt{\mathbf{T}}, \mathbf{x}; \mathbf{z}^{(M)}, \lambda_N)} d\mathbb{P}_{X,M}(\mathbf{x}), \quad (5.25)$$

where we have defined

$$\mathbf{T} := \langle \mathbf{R}_{12} \rangle'_{N,\varepsilon} = \frac{1}{N} \langle \mathbf{X}^\top \rangle'_{N,\varepsilon} \langle \mathbf{X} \rangle'_{N,\varepsilon} = \frac{1}{N} \sum_{i=1}^N \langle \mathbf{x}_i \rangle'_{N,\varepsilon} \langle \mathbf{x}_i^\top \rangle'_{N,\varepsilon}. \quad (5.26)$$

Here, equation (5.24) is due to the fact that the integral is independent of  $\mathbf{X}$ , which is what the Gibbs average  $\langle \cdot \rangle'_{N,\varepsilon}$  is over, while equation (5.25) follows from the fact that the Gaussian vectors  $\langle \mathbf{X}^\top \rangle'_{N,\varepsilon} \mathbf{z}^{(N)}$  and  $\sqrt{N\mathbf{T}} \mathbf{z}^{(M)}$  are equal in distribution, since  $\mathbf{z}^{(N)}$ ,  $\mathbf{z}^{(M)}$  are independent of  $\langle \mathbf{X} \rangle'_{N,\varepsilon}$ .

Similarly, substituting  $\mathbf{X}$  and  $\mathbf{X}_0$  by  $\langle \mathbf{X} \rangle'_{N,\varepsilon}$  in the exponent of the second term in the right-hand side of inequality (5.19) yields the simplification

$$\begin{aligned} \frac{1}{M} \mathbb{E}_{\hat{\mathbf{Z}}} \mathbb{E} \ln \langle e^{\hat{\gamma}(\langle \mathbf{X} \rangle'_{N,\varepsilon}; \langle \mathbf{X} \rangle'_{N,\varepsilon}, \hat{\mathbf{Z}}, \lambda)} \rangle'_{N,\varepsilon} &= \frac{1}{M} \mathbb{E}_{\hat{\mathbf{Z}}} \mathbb{E} [\hat{\gamma}(\langle \mathbf{X} \rangle'_{N,\varepsilon}; \langle \mathbf{X} \rangle'_{N,\varepsilon}, \hat{\mathbf{Z}}, \lambda)] \\ &= \frac{\lambda}{4M} \text{Tr} \mathbb{E} \mathbf{T}^2, \end{aligned}$$

where the first equality follows as above, while the second is due to the independence of  $\hat{\mathbf{Z}}$  and  $\langle \mathbf{X} \rangle'_{N,\varepsilon}$ . Combining this with equation (5.25) shows that replacing  $\mathbf{X}$  and  $\mathbf{X}_0$  by  $\langle \mathbf{X} \rangle'_{N,\varepsilon}$  within the averages in the right-hand side of inequality (5.19) produces the averaged replica symmetric potential  $\mathbb{E} F_M^{\text{RS}}(\mathbf{T}, \lambda)$ . Connection with Theorem 2.1 is made upon noting that  $\mathbf{T}$  is symmetric positive semidefinite, replacing the average with a supremum, and using Theorem 2.6 to establish that

$$\mathbb{E} F_M^{\text{RS}}(\mathbf{T}, \lambda) \leq \sup_{\mathbf{Q} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{Q}, \lambda) = \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda).$$

The driving mechanism behind the above is the action of replacing  $\mathbf{X}$  and  $\mathbf{X}_0$  by  $\langle \mathbf{X} \rangle'_{N,\varepsilon}$ . This is allowed, with a negligible cost, due to the concentration results proved in Section 4.2. We now make this precise through a slight rewriting of the arguments of [43].

**Lemma 5.2** (First application of overlap concentration). *Assume hypotheses (H2) and (H3). Define*

$$\begin{aligned} U_1 &:= \left\langle \int_{\mathbb{R}^M} e^{\hat{H}_{M,N}(\mathbf{X}, \mathbf{X}_0, \mathbf{x}; \mathbf{z}, \lambda)} d\mathbb{P}_{\mathbf{X}, M}(\mathbf{x}) \right\rangle'_{N,\varepsilon}, \\ V_1 &:= \int_{\mathbb{R}^M} e^{\hat{H}_{M,N}(\langle \mathbf{X} \rangle'_{N,\varepsilon}, \langle \mathbf{X} \rangle'_{N,\varepsilon}, \mathbf{x}; \mathbf{z}, \lambda)} d\mathbb{P}_{\mathbf{X}, M}(\mathbf{x}), \end{aligned}$$

with  $\mathbf{z} \sim \mathcal{N}(0, I_N)$ ,  $\hat{H}_{M,N}(\mathbf{A}, \mathbf{B}, \mathbf{x}; \mathbf{z}, \lambda)$  as specified by equation (5.23), and the Gibbs average  $\langle \cdot \rangle'_{N,\varepsilon}$  as in equation (2.24). Then, there is a positive constant  $C$  independent of  $M, N$  such that

$$|\mathbb{E}'_{\mathbf{z}} \ln U_1 - \mathbb{E}'_{\mathbf{z}} \ln V_1| \leq C \sqrt{M} e^{CM^2} \Gamma(N, M)^{1/4},$$

where  $\Gamma(N, M) = CM^2/\sqrt{Ns_N}$  is the rate of concentration given in Theorem 2.11 and we write  $\mathbb{E}'[\cdot]$  for the expectation over  $\mathbf{Z}$ ,  $\tilde{\mathbf{Z}}$ ,  $\mathbf{X}_0$ ,  $\mathbf{x}_0$ , and uniformly over  $\varepsilon \in [s_N, 2s_N]$ .

*Proof.* First observe that since  $|\ln(x) - \ln(y)| \leq \max\{x^{-1}, y^{-1}\} |x - y|$ , we have by Jensen's inequality and the Cauchy–Schwarz inequality that

$$\begin{aligned} & |\mathbb{E}'\mathbb{E}_{\mathbf{z}} \ln U_1 - \mathbb{E}'\mathbb{E}_{\mathbf{z}} \ln V_1| \\ & \leq \mathbb{E}'\mathbb{E}_{\mathbf{z}} |\ln U_1 - \ln V_1| \\ & \leq \mathbb{E}'\sqrt{\mathbb{E}_{\mathbf{z}}[U_1^{-2} + V_1^{-2}]} \sqrt{\mathbb{E}_{\mathbf{z}}[(U_1 - V_1)^2]} \\ & \leq \mathbb{E}'\sqrt{\mathbb{E}_{\mathbf{z}}[U_1^{-2} + V_1^{-2}]} \sqrt{2|\mathbb{E}_{\mathbf{z}}U_1^2 - \mathbb{E}_{\mathbf{z}}V_1^2| + 2|\mathbb{E}_{\mathbf{z}}[U_1V_1] - \mathbb{E}_{\mathbf{z}}V_1^2|}. \end{aligned} \quad (5.27)$$

Thus, we bound each of the terms

$$\mathbb{E}_{\mathbf{z}}[U_1^{-2} + V_1^{-2}], \quad |\mathbb{E}_{\mathbf{z}}U_1^2 - \mathbb{E}_{\mathbf{z}}V_1^2|, \quad |\mathbb{E}_{\mathbf{z}}[U_1V_1] - \mathbb{E}_{\mathbf{z}}V_1^2|$$

in a series of three steps and then simplify the above using Corollary 4.1 of Theorem 2.11.

*Step 1.* Applying Jensen's inequality shows that

$$U_1 \geq \int_{\mathbb{R}^M} e^{\langle \hat{H}_{M,N}(\mathbf{X}, \mathbf{X}_0, \mathbf{x}; \mathbf{z}, \lambda) \rangle'_{N,\varepsilon}} d\mathbb{P}_{\mathbf{X},M}(\mathbf{x}).$$

Thus, using Jensen's inequality again shows that

$$U_1^{-2} \leq \int_{\mathbb{R}^M} e^{-2\langle \hat{H}_{M,N}(\mathbf{X}, \mathbf{X}_0, \mathbf{x}; \mathbf{z}, \lambda) \rangle'_{N,\varepsilon}} d\mathbb{P}_{\mathbf{X},M}(\mathbf{x}).$$

Now, using the fact that  $\mathbb{E}_z e^{az} = e^{a^2/2}$  for  $z \sim \mathcal{N}(0, 1)$  and  $a$  independent of  $z$  gives

$$\mathbb{E}_{\mathbf{z}} U_1^{-2} \leq \int_{\mathbb{R}^M} e^{2\lambda \mathbf{x}^\top \mathbf{T} \mathbf{x} - \lambda(2\mathbf{x}^\top \mathbf{R}_{10} \mathbf{x}_0 - \mathbf{x}^\top \mathbf{R}_{11} \mathbf{x})}'_{N,\varepsilon} d\mathbb{P}_{\mathbf{X},M}(\mathbf{x}),$$

where we recall that  $\mathbf{T} = \frac{1}{N} \langle \mathbf{X}^\top \rangle'_{N,\varepsilon} \langle \mathbf{X} \rangle'_{N,\varepsilon}$  (see (5.26)). Since  $\mathbb{P}_{\mathbf{X}}$  has  $D$ -bounded support, the entries of  $\mathbf{X}$ ,  $\mathbf{X}_0$  lie in  $[-D, D]$ , so

$$|\mathbf{x}^\top \mathbf{T} \mathbf{x}|, |\mathbf{x}^\top \mathbf{R}_{10} \mathbf{x}_0|, |\mathbf{x}^\top \mathbf{R}_{11} \mathbf{x}| \leq M^2 D^4,$$

and we obtain the upper bound  $\mathbb{E}_{\mathbf{z}} U_1^{-2} \leq e^{5\lambda M^2 D^4}$ . The exact same argument shows that we also have  $\mathbb{E}_{\mathbf{z}} V_1^{-2} \leq e^{5\lambda M^2 D^4}$ , hence

$$\mathbb{E}_{\mathbf{z}} [U_1^{-2} + V_1^{-2}] \leq 2e^{5\lambda M^2 D^4}. \quad (5.28)$$

*Step 2.* Letting  $\langle \cdot \rangle'_{N,\varepsilon}$  now refer to a Gibbs average with respect to the replicas  $\mathbf{X}_1, \mathbf{X}_2$  of  $\mathbf{X}$ , we have that

$$\begin{aligned} \mathbb{E}_z U_1^2 &= \mathbb{E}_z \left\langle \int_{\mathbb{R}^M \times \mathbb{R}^M} e^{\hat{H}_{M,N}(\mathbf{X}_1, \mathbf{X}_0, \mathbf{x}_1; z, \lambda) + \hat{H}_{M,N}(\mathbf{X}_2, \mathbf{X}_0, \mathbf{x}_2; z, \lambda)} d\mathbb{P}_{X,M}^{\otimes 2}(\mathbf{x}_1, \mathbf{x}_2) \right\rangle'_{N,\varepsilon} \\ &= \left\langle \int_{\mathbb{R}^M \times \mathbb{R}^M} e^{\lambda(\mathbf{x}_1^\top \mathbf{R}_{10} \mathbf{x}_0 + \mathbf{x}_2^\top \mathbf{R}_{20} \mathbf{x}_0 + \mathbf{x}_1^\top \mathbf{R}_{12} \mathbf{x}_2)} d\mathbb{P}_{X,M}^{\otimes 2}(\mathbf{x}_1, \mathbf{x}_2) \right\rangle'_{N,\varepsilon}, \end{aligned}$$

where we have used the fact that

$$\mathbb{E}_z \exp\left(\sqrt{\frac{\lambda}{N}}(\mathbf{X}_1 \mathbf{x}_1 + \mathbf{X}_2 \mathbf{x}_2)^\top z\right) = \exp\left(\frac{\lambda}{2N} \|\mathbf{X}_1 \mathbf{x}_1 + \mathbf{X}_2 \mathbf{x}_2\|_{\mathbb{F}}^2\right).$$

Defining  $F_1: ([-D^2, D^2]^{M \times M})^3 \rightarrow \mathbb{R}$  by

$$F_1(\mathbf{A}, \mathbf{B}, \mathbf{C}) := \int_{\mathbb{R}^M \times \mathbb{R}^M} e^{\lambda(\mathbf{x}_1^\top \mathbf{A} \mathbf{x}_0 + \mathbf{x}_2^\top \mathbf{B} \mathbf{x}_0 + \mathbf{x}_1^\top \mathbf{C} \mathbf{x}_2)} d\mathbb{P}_{X,M}^{\otimes 2}(\mathbf{x}_1, \mathbf{x}_2),$$

we observe that

$$\mathbb{E}_z U_1^2 = \langle F_1(\mathbf{R}_{10}, \mathbf{R}_{20}, \mathbf{R}_{12}) \rangle'_{N,\varepsilon}.$$

Likewise, we have that

$$\mathbb{E}_z V_1^2 = F_1(\mathbf{T}, \mathbf{T}, \mathbf{T}).$$

As  $\mathbb{P}_X$  has  $D$ -bounded support, we compute for all  $\mathbf{A}, \mathbf{B}, \mathbf{C} \in [-D^2, D^2]^{M \times M}$  that

$$\|\nabla F_1(\mathbf{A}, \mathbf{B}, \mathbf{C})\|_{\mathbb{F}} \leq \sqrt{3} \lambda M D^2 e^{3\lambda M^2 D^4}.$$

Thus,  $F_1(\mathbf{A}, \mathbf{B}, \mathbf{C})$  is  $(C_1 M e^{C_2 M^2})$ -Lipschitz for some constants  $C_1, C_2$  independent of  $M, N$ . Hence, using Jensen's inequality and our Lipschitz bound shows that

$$\begin{aligned} &|\mathbb{E}_z U_1^2 - \mathbb{E}_z V_1^2| \\ &= |\langle F_1(\mathbf{R}_{10}, \mathbf{R}_{20}, \mathbf{R}_{12}) \rangle'_{N,\varepsilon} - F_1(\mathbf{T}, \mathbf{T}, \mathbf{T})| \\ &\leq \langle |F_1(\mathbf{R}_{10}, \mathbf{R}_{20}, \mathbf{R}_{12}) - F_1(\mathbf{T}, \mathbf{T}, \mathbf{T})| \rangle'_{N,\varepsilon} \\ &\leq C_1 M e^{C_2 M^2} \left\langle \sqrt{\|\mathbf{R}_{10} - \mathbf{T}\|_{\mathbb{F}}^2 + \|\mathbf{R}_{20} - \mathbf{T}\|_{\mathbb{F}}^2 + \|\mathbf{R}_{12} - \mathbf{T}\|_{\mathbb{F}}^2} \right\rangle'_{N,\varepsilon}. \quad (5.29) \end{aligned}$$

*Step 3.* From the same reasoning as in the above step, we see that

$$\begin{aligned} &\mathbb{E}_z [U_1 V_1] \\ &= \mathbb{E}_z \left\langle \int_{\mathbb{R}^M \times \mathbb{R}^M} e^{\hat{H}_{M,N}(\mathbf{X}, \mathbf{X}_0, \mathbf{x}_1; z, \lambda) + \hat{H}_{M,N}(\langle \mathbf{X} \rangle'_{N,\varepsilon}, \langle \mathbf{X} \rangle'_{N,\varepsilon}, \mathbf{x}_2; z, \lambda)} d\mathbb{P}_{X,M}^{\otimes 2}(\mathbf{x}_1, \mathbf{x}_2) \right\rangle'_{N,\varepsilon} \\ &= \langle F_1(\mathbf{R}_{10}, \mathbf{T}, \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon}) \rangle'_{N,\varepsilon}. \end{aligned}$$

Hence, using Jensen's inequality and our Lipschitz bound on  $F_1(\mathbf{A}, \mathbf{B}, \mathbf{C})$  yet again shows that

$$\begin{aligned}
 & |\mathbb{E}_z[U_1 V_1] - \mathbb{E}_z V_1^2| \\
 &= |\langle F_1(\mathbf{R}_{10}, \mathbf{T}, \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon})'_{N,\varepsilon} - F_1(\mathbf{T}, \mathbf{T}, \mathbf{T}) | \\
 &\leq \langle |F_1(\mathbf{R}_{10}, \mathbf{T}, \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon}) - F_1(\mathbf{T}, \mathbf{T}, \mathbf{T})|'_{N,\varepsilon} \\
 &\leq C_1 M e^{C_2 M^2} \left\langle \sqrt{\|\mathbf{R}_{10} - \mathbf{T}\|_{\mathbb{F}}^2 + \left\| \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon} - \mathbf{T} \right\|_{\mathbb{F}}^2} \right\rangle'_{N,\varepsilon}. \quad (5.30)
 \end{aligned}$$

*Step 4.* Applying Jensen's inequality and Corollary 4.1 to inequality (5.29) shows that

$$\begin{aligned}
 & \mathbb{E}' |\mathbb{E}_z U_1^2 - \mathbb{E}_z V_1^2| \\
 &\leq C_1 M e^{C_2 M^2} \sqrt{\mathbb{E}' \langle \|\mathbf{R}_{10} - \mathbf{T}\|_{\mathbb{F}}^2 + \|\mathbf{R}_{20} - \mathbf{T}\|_{\mathbb{F}}^2 + \|\mathbf{R}_{12} - \mathbf{T}\|_{\mathbb{F}}^2 \rangle'_{N,\varepsilon}} \\
 &\leq C_1 M e^{C_2 M^2} \sqrt{3\Gamma(N, M)}, \quad (5.31)
 \end{aligned}$$

where we have used the Nishimori identity to replace  $\mathbf{R}_{20}$  by  $\mathbf{R}_{10}$ . Doing the same to inequality (5.30) shows that

$$\begin{aligned}
 \mathbb{E}' |\mathbb{E}_z[U_1 V_1] - \mathbb{E}_z V_1^2| &\leq C_1 M e^{C_2 M^2} \sqrt{\Gamma(N, M) + \mathbb{E}' \left\langle \left\| \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon} - \mathbf{T} \right\|_{\mathbb{F}}^2 \right\rangle'_{N,\varepsilon}} \\
 &\leq C_1 M e^{C_2 M^2} \sqrt{2\Gamma(N, M)}, \quad (5.32)
 \end{aligned}$$

where in the last line we have used the fact that, if we let  $\langle \cdot \rangle$  denote the Gibbs average  $\langle \cdot \rangle'_{N,\varepsilon}$  with respect to only  $\mathbf{X}_2$ , we have

$$\begin{aligned}
 \mathbb{E}' \left\langle \left\| \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon} - \mathbf{T} \right\|_{\mathbb{F}}^2 \right\rangle'_{N,\varepsilon} &= \mathbb{E}' \left\langle \left\| \frac{1}{N} \mathbf{X}_1^\top \mathbf{X}_2 - \mathbf{T} \right\|_{\mathbb{F}}^2 \right\rangle'_{N,\varepsilon} \\
 &\leq \mathbb{E}' \langle \|\mathbf{R}_{12} - \mathbf{T}\|_{\mathbb{F}}^2 \rangle'_{N,\varepsilon} \\
 &= \mathbb{E}' \langle \|\mathbf{R}_{12} - \mathbf{T}\|_{\mathbb{F}}^2 \rangle'_{N,\varepsilon} \leq \Gamma(N, M).
 \end{aligned}$$

Returning to inequality (5.27), inserting inequality (5.28), using Jensen's inequality, and then finally using inequalities (5.31) and (5.32), we see that there is a finite positive constant  $C$  independent of  $M, N$  such that

$$\begin{aligned}
 |\mathbb{E}' \mathbb{E}_z \ln U_1 - \mathbb{E}' \mathbb{E}_z \ln V_1| &\leq C e^{C M^2} \mathbb{E}' \sqrt{2|\mathbb{E}_z U_1^2 - \mathbb{E}_z V_1^2| + 2|\mathbb{E}_z[U_1 V_1] - \mathbb{E}_z V_1^2|} \\
 &\leq C e^{C M^2} \sqrt{\mathbb{E}' |\mathbb{E}_z U_1^2 - \mathbb{E}_z V_1^2| + \mathbb{E}' |\mathbb{E}_z[U_1 V_1] - \mathbb{E}_z V_1^2|} \\
 &\leq C \sqrt{M} e^{C M^2} \Gamma(N, M)^{1/4},
 \end{aligned}$$

as desired. ■

**Lemma 5.3** (Second application of overlap concentration). *Assume hypotheses (H2) and (H3). Define*

$$U_2 := \left\langle e^{\hat{y}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, \lambda)} \right\rangle'_{N, \varepsilon}, \quad V_2 := e^{\hat{y}(\langle \mathbf{X} \rangle'_{N, \varepsilon}; \langle \mathbf{X} \rangle'_{N, \varepsilon}, \hat{\mathbf{Z}}, \lambda)}$$

with  $\hat{y}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, \lambda)$  as specified by equation (5.18) and the Gibbs average  $\langle \cdot \rangle'_{N, \varepsilon}$  as in equation (2.24). Then, there is a positive constant  $C$  independent of  $M, N$  such that

$$|\mathbb{E}' \mathbb{E}_{\hat{\mathbf{Z}}} \ln U_2 - \mathbb{E}' \mathbb{E}_{\hat{\mathbf{Z}}} \ln V_2| \leq C \sqrt{M} e^{CM^2} \Gamma(N, M)^{1/4},$$

where  $\Gamma(N, M) = CM^2 / \sqrt{Ns_N}$  is the rate of concentration given in Theorem 2.11 and  $\mathbb{E}'[\cdot]$  retains its meaning from Lemma 5.2.

*Proof.* The structure of the proof is identical to that of Lemma 5.2. Indeed, as before, we have that

$$\begin{aligned} & |\mathbb{E}' \mathbb{E}_{\hat{\mathbf{Z}}} \ln U_2 - \mathbb{E}' \mathbb{E}_{\hat{\mathbf{Z}}} \ln V_2| \\ & \leq \mathbb{E}' \sqrt{\mathbb{E}_{\hat{\mathbf{Z}}} [U_2^{-2} + V_2^{-2}]} \sqrt{2|\mathbb{E}_{\hat{\mathbf{Z}}} U_2^2 - \mathbb{E}_{\hat{\mathbf{Z}}} V_2^2| + 2|\mathbb{E}_{\hat{\mathbf{Z}}} [U_2 V_2] - \mathbb{E}_{\hat{\mathbf{Z}}} V_2^2|}, \end{aligned} \quad (5.33)$$

and we proceed by bounding each of the terms

$$\mathbb{E}_{\hat{\mathbf{Z}}} [U_2^{-2} + V_2^{-2}], \quad |\mathbb{E}_{\hat{\mathbf{Z}}} U_2^2 - \mathbb{E}_{\hat{\mathbf{Z}}} V_2^2|, \quad |\mathbb{E}_{\hat{\mathbf{Z}}} [U_2 V_2] - \mathbb{E}_{\hat{\mathbf{Z}}} V_2^2|$$

in three steps analogous to those of the proof of Lemma 5.2, before applying Corollary 4.1.

*Step 1.* Using Jensen's inequality shows that

$$U_2^{-2} \leq e^{-2\langle \hat{y}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, \lambda) \rangle'_{N, \varepsilon}}, \quad V_2^{-2} \leq e^{-2\langle \hat{y}(\langle \mathbf{X} \rangle'_{N, \varepsilon}; \langle \mathbf{X} \rangle'_{N, \varepsilon}, \hat{\mathbf{Z}}, \lambda) \rangle'_{N, \varepsilon}}.$$

Then, since

$$\mathbb{E}_{\hat{\mathbf{Z}}} e^{\text{Tr}(\hat{\mathbf{Z}} \mathbf{A})} = e^{\text{Tr} \mathbf{A}^2} \quad (5.34)$$

for symmetric  $\mathbf{A} \in \mathbb{R}^{N \times N}$  independent of  $\hat{\mathbf{Z}}$ , we have that

$$\begin{aligned} \mathbb{E}_{\hat{\mathbf{Z}}} U_2^{-2} & \leq e^{\lambda \text{Tr}(\langle \mathbf{R}_{11}^T \rangle'_{N, \varepsilon} \langle \mathbf{R}_{11} \rangle'_{N, \varepsilon} - \langle \mathbf{R}_{10}^T \mathbf{R}_{10} \rangle'_{N, \varepsilon} + \frac{1}{2} \langle \mathbf{R}_{11}^T \mathbf{R}_{11} \rangle'_{N, \varepsilon})}, \\ \mathbb{E}_{\hat{\mathbf{Z}}} V_2^{-2} & \leq e^{\frac{\lambda}{2} \text{Tr} \mathbf{T}^2}, \end{aligned}$$

where we recall that  $\mathbf{T} = \frac{1}{N} \langle \mathbf{X}^T \rangle'_{N, \varepsilon} \langle \mathbf{X} \rangle'_{N, \varepsilon}$  (5.26). As the entries of  $\mathbf{X}, \mathbf{X}_0$  lie in  $[-D, D]$ , we have that

$$|\text{Tr}(\langle \mathbf{R}_{11}^T \rangle'_{N, \varepsilon} \langle \mathbf{R}_{11} \rangle'_{N, \varepsilon})|, |\text{Tr}(\mathbf{R}_{10}^T \mathbf{R}_{10})|, |\text{Tr}(\mathbf{R}_{11}^T \mathbf{R}_{11})|, |\text{Tr} \mathbf{T}^2| \leq M^2 D^4,$$

and thus

$$\mathbb{E}_{\hat{\mathbf{Z}}} [U_2^{-2} + V_2^{-2}] \leq 2e^{\frac{5}{2} \lambda M^2 D^4}. \quad (5.35)$$

*Step 2.* Letting  $\langle \cdot \rangle'_{N,\varepsilon}$  denote the Gibbs average with respect to the replicas  $\mathbf{X}_1, \mathbf{X}_2$  of  $\mathbf{X}$ , we have

$$\mathbb{E}_{\hat{\mathbf{Z}}} U_2^2 = \mathbb{E}_{\hat{\mathbf{Z}}} \langle e^{\hat{\gamma}(\mathbf{X}_1; \mathbf{X}_0, \hat{\mathbf{Z}}, \lambda) + \hat{\gamma}(\mathbf{X}_2; \mathbf{X}_0, \hat{\mathbf{Z}}, \lambda)} \rangle'_{N,\varepsilon} = \langle F_2(\mathbf{R}_{10}, \mathbf{R}_{20}, \mathbf{R}_{12}) \rangle'_{N,\varepsilon},$$

where we have again used identity (5.34) and defined  $F_2: ([-D, D]^{M \times M})^3 \rightarrow \mathbb{R}$  by

$$F_2(\mathbf{A}, \mathbf{B}, \mathbf{C}) := e^{\frac{\lambda}{2} \text{Tr}(\mathbf{A}^\top \mathbf{A} + \mathbf{B}^\top \mathbf{B} + \mathbf{C}^\top \mathbf{C})}.$$

Note also that

$$\mathbb{E}_{\hat{\mathbf{Z}}} V_2^2 = \mathbb{E}_{\hat{\mathbf{Z}}} e^{2\hat{\gamma}(\langle \mathbf{X} \rangle'_{N,\varepsilon}; \langle \mathbf{X} \rangle'_{N,\varepsilon}, \hat{\mathbf{Z}}, \lambda)} = e^{\frac{3\lambda}{2} \text{Tr} \mathbf{T}^\top \mathbf{T}} = F_2(\mathbf{T}, \mathbf{T}, \mathbf{T}).$$

Now, on the domain of  $F_2$ , we have that

$$\|\nabla F_2(\mathbf{A}, \mathbf{B}, \mathbf{C})\|_{\text{F}} \leq \frac{\sqrt{3}}{2} \lambda M D^2 e^{\frac{3}{2} \lambda M^2 D^4}.$$

Thus,  $F_2(\mathbf{A}, \mathbf{B}, \mathbf{C})$  is  $(C_1 M e^{C_2 M^2})$ -Lipschitz for some constants  $C_1, C_2$  independent of  $M, N$ . Hence, using Jensen's inequality shows that

$$\begin{aligned} & |\mathbb{E}_{\hat{\mathbf{Z}}} U_2^2 - \mathbb{E}_{\hat{\mathbf{Z}}} V_2^2| \\ &= |\langle F_2(\mathbf{R}_{10}, \mathbf{R}_{20}, \mathbf{R}_{12}) \rangle'_{N,\varepsilon} - F_2(\mathbf{T}, \mathbf{T}, \mathbf{T})| \\ &\leq \langle |F_2(\mathbf{R}_{10}, \mathbf{R}_{20}, \mathbf{R}_{12}) - F_2(\mathbf{T}, \mathbf{T}, \mathbf{T})| \rangle'_{N,\varepsilon} \\ &\leq C_1 M e^{C_2 M^2} \left\langle \sqrt{\|\mathbf{R}_{10} - \mathbf{T}\|_{\text{F}}^2 + \|\mathbf{R}_{20} - \mathbf{T}\|_{\text{F}}^2 + \|\mathbf{R}_{12} - \mathbf{T}\|_{\text{F}}^2} \right\rangle'_{N,\varepsilon}. \end{aligned} \quad (5.36)$$

*Step 3.* Similarly to above, we see that

$$\begin{aligned} \mathbb{E}_{\hat{\mathbf{Z}}} [U_2 V_2] &= \mathbb{E}_{\hat{\mathbf{Z}}} \langle e^{\hat{\gamma}(\mathbf{X}; \mathbf{X}_0, \hat{\mathbf{Z}}, \lambda) + \hat{\gamma}(\langle \mathbf{X} \rangle'_{N,\varepsilon}; \langle \mathbf{X} \rangle'_{N,\varepsilon}, \hat{\mathbf{Z}}, \lambda)} \rangle'_{N,\varepsilon} \\ &= \left\langle F_2\left(\mathbf{R}_{10}, \mathbf{T}, \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon}\right) \right\rangle'_{N,\varepsilon}. \end{aligned}$$

Hence, using Jensen's inequality and inserting our Lipschitz constant shows that

$$\begin{aligned} & |\mathbb{E}_{\hat{\mathbf{Z}}} [U_2 V_2] - \mathbb{E}_{\hat{\mathbf{Z}}} V_2^2| \\ &= \left| \left\langle F_2\left(\mathbf{R}_{10}, \mathbf{T}, \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon}\right) \right\rangle'_{N,\varepsilon} - F_2(\mathbf{T}, \mathbf{T}, \mathbf{T}) \right| \\ &\leq \left\langle \left| F_2\left(\mathbf{R}_{10}, \mathbf{T}, \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon}\right) - F_2(\mathbf{T}, \mathbf{T}, \mathbf{T}) \right| \right\rangle'_{N,\varepsilon} \\ &\leq C_1 M e^{C_2 M^2} \left\langle \sqrt{\|\mathbf{R}_{10} - \mathbf{T}\|_{\text{F}}^2 + \left\| \frac{1}{N} \mathbf{X}^\top \langle \mathbf{X} \rangle'_{N,\varepsilon} - \mathbf{T} \right\|_{\text{F}}^2} \right\rangle'_{N,\varepsilon}. \end{aligned} \quad (5.37)$$

*Step 4.* As inequalities (5.36), (5.37) are equivalent to inequalities (5.29), (5.30), we may read off from Step 4 of the proof of Lemma 5.2 that

$$\begin{aligned}\mathbb{E}'|\mathbb{E}_{\hat{\mathbf{z}}}[U_2^2] - \mathbb{E}_{\hat{\mathbf{z}}}[V_2^2]| &\leq C_1 M e^{C_2 M^2} \sqrt{3\Gamma(N, M)}, \\ \mathbb{E}'|\mathbb{E}_{\hat{\mathbf{z}}}[U_2 V_2] - \mathbb{E}_{\hat{\mathbf{z}}}[V_2^2]| &\leq C_1 M e^{C_2 M^2} \sqrt{2\Gamma(N, M)}.\end{aligned}$$

Therefore, inserting inequality (5.35) into inequality (5.33), using Jensen's inequality, and then finally using the above two inequalities, we see that there is a finite positive constant  $C$  independent of  $M, N$  such that

$$\begin{aligned}|\mathbb{E}'\mathbb{E}_{\hat{\mathbf{z}}}\ln U_2 - \mathbb{E}'\mathbb{E}_{\hat{\mathbf{z}}}\ln V_2| &\leq C e^{C M^2} \mathbb{E}'\sqrt{2|\mathbb{E}_{\hat{\mathbf{z}}}[U_2^2] - \mathbb{E}_{\hat{\mathbf{z}}}[V_2^2]| + 2|\mathbb{E}_{\hat{\mathbf{z}}}[U_2 V_2] - \mathbb{E}_{\hat{\mathbf{z}}}[V_2^2]|} \\ &\leq C e^{C M^2} \sqrt{\mathbb{E}'|\mathbb{E}_{\hat{\mathbf{z}}}[U_2^2] - \mathbb{E}_{\hat{\mathbf{z}}}[V_2^2]| + \mathbb{E}'|\mathbb{E}_{\hat{\mathbf{z}}}[U_2 V_2] - \mathbb{E}_{\hat{\mathbf{z}}}[V_2^2]|} \\ &\leq C \sqrt{M} e^{C M^2} \Gamma(N, M)^{1/4}. \quad \blacksquare\end{aligned}$$

To conclude this subsection, it remains now to combine Lemmas 5.2 and 5.3 with Lemma 5.1. We use the equality in the distribution of  $\langle \mathbf{X}^\top \rangle'_{N,\varepsilon} \mathbf{z}^{(N)}$  and  $\sqrt{N\mathbf{T}} \mathbf{z}^{(M)}$  for  $\mathbf{z}^{(N)} \sim \mathcal{N}(0, I_N)$  and  $\mathbf{z}^{(M)} \sim \mathcal{N}(0, I_M)$  discussed earlier, replace the expectation over  $\mathbf{T}$  with a supremum, and finally apply Theorem 2.6.

**Proposition 5.4** ( $\Delta_N/M$  cavity upper bound). *Assume all of the hypotheses of Theorem 2.1, let  $M = M_N$ , and recall (5.11) that*

$$\Delta_N = \mathbb{E} \ln \tilde{\mathbf{Z}}_{N+1, M_{N+1}} - \mathbb{E} \ln \tilde{\mathbf{Z}}_{N, M_{N+1}},$$

where  $\mathbb{E}[\cdot]$  denotes an expectation with respect to  $\mathbf{Z}, \tilde{\mathbf{Z}}, \mathbf{X}_0, \xi, \tilde{\xi}, \xi, \eta_0, \varepsilon_N, \varepsilon_{N+1}$ . There exists a finite positive constant  $C$  independent of  $M, N$  such that enforcing the constraint  $s_N \geq \min\{0.9, N^{-1/9}(C e^{C M^2})^{8/9}\}$ , with  $s_N$  still vanishing at large  $N$ , ensures that

$$\limsup_{N \rightarrow \infty} \frac{\Delta_N}{M} \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda). \quad (5.38)$$

*Proof.* First, we set  $M = M_{N+1}$  instead of  $M = M_N$ , recalling that Lemma 5.1 holds for any  $M$ . Rewriting said lemma in the notation of Lemmas 5.2 and 5.3 then yields

$$\frac{\Delta_N}{M} \leq \frac{\mathbb{E} \ln U_1 - \mathbb{E}_{\hat{\mathbf{z}}} \mathbb{E} \ln U_2}{M} + \Gamma'(N, M),$$

where we have replaced  $\xi^\top, \eta^\top, \eta_0^\top$  by  $\mathbf{z}, \mathbf{x}, \mathbf{x}_0$ . By Lemmas 5.2 and 5.3, we then have

$$\frac{\Delta_N}{M} \leq \frac{\mathbb{E} \ln V_1 - \mathbb{E}_{\hat{\mathbf{z}}} \mathbb{E} \ln V_2}{M} + \frac{C}{\sqrt{M}} e^{C M^2} \Gamma(N, M)^{1/4} + \Gamma'(N, M) \quad (5.39)$$

for some finite positive constant  $C$  independent of  $M, N$ . Now, writing  $\mathbf{z}' \sim \mathcal{N}(0, I_M)$ , we observe that  $\langle \mathbf{X}^\top \rangle'_{N, \varepsilon_N} \mathbf{z}$  and  $\sqrt{N} \mathbf{T} \mathbf{z}'$  are equal in distribution, so

$$\frac{1}{M} \mathbb{E} \ln V_1 = \frac{1}{M} \mathbb{E}_{\mathbf{z}'} \mathbb{E} \ln \int_{\mathbb{R}^M} e^{\sqrt{\lambda} \mathbf{x}^\top \sqrt{\mathbf{T}} \mathbf{z}' + \mathbf{x}^\top \mathbf{T} \mathbf{x}_0 - \frac{\lambda}{2} \mathbf{x}^\top \mathbf{T} \mathbf{x}} d\mathbb{P}_{\mathbf{X}, M}(\mathbf{x}).$$

On the other hand, using the independence of  $\widehat{\mathbf{Z}}$  and  $\langle \mathbf{X} \rangle'_{N, \varepsilon_N}$  shows that

$$\frac{1}{M} \mathbb{E}_{\widehat{\mathbf{Z}}} \mathbb{E} \ln V_2 = \frac{1}{2M} \text{Tr} \mathbb{E}_{\widehat{\mathbf{Z}}} \mathbb{E} \left[ \sqrt{\lambda} \widehat{\mathbf{Z}} \mathbf{T} + \frac{\lambda}{2} \mathbf{T}^2 \right] = \frac{\lambda}{4M} \text{Tr} \mathbb{E} \mathbf{T}^2.$$

Inserting these two expressions back into inequality (5.39) gives

$$\begin{aligned} \frac{\Delta_N}{M} &\leq \frac{1}{M} \mathbb{E}_{\mathbf{z}'} \mathbb{E} \ln \int_{\mathbb{R}^M} e^{\sqrt{\lambda} \mathbf{x}^\top \sqrt{\mathbf{T}} \mathbf{z}' + \mathbf{x}^\top \mathbf{T} \mathbf{x}_0 - \frac{\lambda}{2} \mathbf{x}^\top \mathbf{T} \mathbf{x}} d\mathbb{P}_{\mathbf{X}, M}(\mathbf{x}) \\ &\quad - \frac{\lambda}{4M} \text{Tr} \mathbb{E} \mathbf{T}^2 + \frac{C}{\sqrt{M}} e^{CM^2} \Gamma(N, M)^{1/4} + \Gamma'(N, M). \end{aligned}$$

As this integral depends on all of the involved random variables through only  $\mathbf{T}, \mathbf{z}', \mathbf{x}_0$ , we recognize from equation (2.2) that the above simplifies to

$$\frac{\Delta_N}{M} \leq \mathbb{E}_{\mathbf{T}} F_M^{\text{RS}}(\mathbf{T}, \lambda) + \frac{C}{\sqrt{M}} e^{CM^2} \Gamma(N, M)^{1/4} + \Gamma'(N, M).$$

Noting that  $\mathbf{T}$  is symmetric positive semidefinite, we circumvent the intractability of the expectation over  $\mathbf{T}$  by replacing it with a supremum to obtain

$$\begin{aligned} \frac{\Delta_N}{M} &\leq \sup_{\mathbf{T} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{T}, \lambda) + \frac{C}{\sqrt{M}} e^{CM^2} \Gamma(N, M)^{1/4} + \Gamma'(N, M) \\ &= \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + \frac{C}{\sqrt{M}} e^{CM^2} \Gamma(N, M)^{1/4} + \Gamma'(N, M), \end{aligned} \quad (5.40)$$

with the second line following from using Theorem 2.6 to replace the  $M$ -dependent supremum  $\sup_{\mathbf{T} \in \mathcal{S}_M} F_M^{\text{RS}}(\mathbf{T}, \lambda)$  by its  $M$ -independent analog. To make connection with inequality (5.38), we multiply both sides by  $M_{N+1}/M_N$  to obtain

$$\begin{aligned} \frac{\Delta_N}{M_N} &\leq \frac{M_{N+1}}{M_N} \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) \\ &\quad + \frac{C \sqrt{M_{N+1}}}{M_N} e^{CM_{N+1}^2} \Gamma(N, M_{N+1})^{1/4} + \frac{M_{N+1}}{M_N} \Gamma'(N, M_{N+1}). \end{aligned}$$

By hypothesis, either  $M_{N+1} = M_N$  or  $M_{N+1} = M_N + 1$ . In either case, the result follows due to our choice of  $s_N$  inducing the vanishing of all error terms when taking the limit supremum; the second term in the right-hand side vanishes because

$$M_N = o(\sqrt{\ln N}) \quad \text{and} \quad s_N \geq N^{-1/9} (C e^{CM_N^2})^{8/9},$$

while the final term vanishes because  $s_N \rightarrow 0$ . ■

### 5.3. The novel column cavity bound

In this subsection, we formalize the intuition contained within the approximation given in (2.21) to prove the counterpart to Proposition 5.4 bounding the limit supremum of  $\Delta_M/N$ . This is a novel term that only comes into play when examining multiscale models and arises here due to our working in the sublinear rank regime  $M = o(\sqrt{\ln N})$ . As per the approximation (2.21), our approach is to bootstrap the finite  $M, N$  cavity bounds through a second application of the cavity method.

The strategy then is to fix  $N$  and  $M = M_N$  at a value such that  $M_{N+1} = M_N + 1$  and bound each of the terms in (5.12):

$$\frac{\Delta_M}{N} = \frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M+1} - \frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M}$$

separately. The upper bound on the second term, being negative, can be obtained simply from the free entropy lower bound of Proposition 2.7; note that this proposition holds for any  $M, N \geq 1$ . To obtain an upper bound on  $\frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M+1}$ , we proceed by introducing the auxiliary difference

$$\hat{\Delta}_N(n) := \mathbb{E} \ln \tilde{Z}_{n+1, M+1} - \mathbb{E} \ln \tilde{Z}_{n, M+1}, \quad (5.41)$$

so that we have the telescoping series

$$\frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M+1} = \frac{1}{N} \sum_{n=0}^{N-1} \hat{\Delta}_N(n).$$

Then, we truncate this sum at a truncation parameter  $n = T$  and show that the truncated sum can be bounded by a standard application of the cavity method, while the remainder is bounded with  $T$ -dependent error due to Proposition 5.4. Finally, we take the limit supremum, choosing  $T$  such that all error terms vanish at large  $N$ .

**Lemma 5.5** (Bounding the truncated sum of  $\hat{\Delta}_N(n)$ ). *Let  $N, M, T \geq 1$ , assume hypotheses (H2) and (H3), and let  $\hat{\Delta}_N(n)$  be defined by equation (5.41). Then, there exists some positive constant  $C$  independent of  $M, N, T$  such that*

$$\sum_{n=0}^{T-1} \hat{\Delta}_N(n) \leq CTM^2.$$

*Proof.* First, note that

$$\sum_{n=0}^{T-1} \hat{\Delta}_N(n) \leq T \sup_{n \leq T} \hat{\Delta}_N(n). \quad (5.42)$$

Then, replacing  $N$  by  $n$  and  $M$  by  $M + 1$  in Lemma 5.1 shows that for  $C$  a generic finite positive constant independent of  $N, M, T$  here on out,

$$\begin{aligned} \widehat{\Delta}_N(n) &\leq \mathbb{E}_{\xi', \eta_0} \mathbb{E} \ln \left\langle \int_{\mathbb{R}^{M+1}} e^{\eta \widehat{z}(X'; \xi', \lambda) + \eta \widehat{s}(X'; \lambda) \eta^\top} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n} \\ &\quad - \mathbb{E}_{\widehat{Z}'} \mathbb{E} \ln \langle e^{\widehat{y}(X'; X'_0, \widehat{Z}', \lambda)} \rangle'_{n, \varepsilon_n} + (M + 1) \Gamma'(n, M + 1), \end{aligned} \quad (5.43)$$

where we define  $X' \in \mathbb{R}^{n \times (M+1)}$  to be the integration variable of the Gibbs average  $\langle \cdot \rangle'_{n, \varepsilon_n}$ , let  $X'_0 \in \mathbb{R}^{n \times (M+1)}$  have i.i.d. entries drawn from  $\mathbb{P}_X$ , let  $\eta_0 \sim \mathbb{P}_{X, M+1}$ , let  $\widehat{Z}' \in \mathbb{R}^{n \times n}$  be a standard Wigner matrix, and let  $\xi' \sim \mathcal{N}(0, I_n)$ .

Substituting  $\widehat{z}(X')$  (see (5.7)) and  $\widehat{s}(X')$  (see (5.8)) into the first term of the right-hand side of inequality (5.43), using Jensen's inequality, and then using the identity  $\mathbb{E}_z e^{az} = e^{a^2/2}$  for  $z \sim \mathcal{N}(0, 1)$  and  $a$  independent of  $z$  shows that

$$\begin{aligned} &\mathbb{E}_{\xi', \eta_0} \mathbb{E} \ln \left\langle \int_{\mathbb{R}^{M+1}} e^{\eta \widehat{z}(X'; \xi', \lambda) + \eta \widehat{s}(X'; \lambda) \eta^\top} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n} \\ &= \mathbb{E}_{\xi', \eta_0} \mathbb{E} \ln \left\langle \int_{\mathbb{R}^{M+1}} e^{\sqrt{\frac{\lambda}{n}} \eta X'^\top \xi'^\top + \frac{\lambda}{n} \eta X'^\top X'_0 \eta_0 - \frac{\lambda}{2n} \eta X'^\top X' \eta^\top} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n} \\ &\leq \ln \mathbb{E}_{\eta_0} \mathbb{E} \left\langle \int_{\mathbb{R}^{M+1}} \mathbb{E}_{\xi'} e^{\sqrt{\frac{\lambda}{n}} \eta X'^\top \xi'^\top + \frac{\lambda}{n} \eta X'^\top X'_0 \eta_0 - \frac{\lambda}{2n} \eta X'^\top X' \eta^\top} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n} \\ &= \ln \mathbb{E}_{\eta_0} \mathbb{E} \left\langle \int_{\mathbb{R}^{M+1}} e^{\frac{\lambda}{n} \eta X'^\top X'_0 \eta_0} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n}. \end{aligned}$$

As  $\mathbb{P}_X$  has  $D$ -bounded support, we have that  $|\eta X'^\top X'_0 \eta_0| \leq n(M + 1)^2 D^4$ , and therefore

$$\begin{aligned} &\mathbb{E}_{\xi', \eta_0} \mathbb{E} \ln \left\langle \int_{\mathbb{R}^{M+1}} e^{\eta \widehat{z}(X'; \xi', \lambda) + \eta \widehat{s}(X'; \lambda) \eta^\top} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n} \\ &\leq \ln \mathbb{E}_{\eta_0} \mathbb{E} \left\langle \int_{\mathbb{R}^{M+1}} e^{\frac{\lambda}{n} \eta X'^\top X'_0 \eta_0} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n} \\ &\leq \ln \mathbb{E}_{\eta_0} \mathbb{E} \left\langle \int_{\mathbb{R}^{M+1}} e^{\lambda(M+1)^2 D^4} d\mathbb{P}_{X, M+1}(\eta) \right\rangle'_{n, \varepsilon_n} \\ &= \lambda(M + 1)^2 D^4. \end{aligned} \quad (5.44)$$

Likewise, substituting  $\widehat{y}(X')$  (see (5.18)) into the second term of the right-hand side of inequality (5.43), using Jensen's inequality, using the independence of  $\widehat{Z}'$  and  $X'$ , and then using the  $D$ -boundedness of  $\text{supp } \mathbb{P}_X$  shows that

$$\begin{aligned} &-\mathbb{E}_{\widehat{Z}'} \mathbb{E} \ln \langle e^{\widehat{y}(X'; X'_0, \widehat{Z}', \lambda)} \rangle'_{n, \varepsilon_n} \\ &= -\mathbb{E}_{\widehat{Z}'} \mathbb{E} \ln \left\langle e^{\frac{1}{2} \text{Tr} \left( \frac{\sqrt{\lambda}}{n} \widehat{Z}' X' X'^\top + \frac{\lambda}{n^2} X'_0{}^\top X' X'^\top X'_0 - \frac{\lambda}{2n^2} X'^\top X' X'^\top X' \right)} \right\rangle'_{n, \varepsilon_n} \end{aligned}$$

$$\begin{aligned}
 &\leq -\mathbb{E}_{\tilde{\mathbf{Z}}'} \mathbb{E} \left\langle \frac{1}{2} \operatorname{Tr} \left( \frac{\sqrt{\lambda}}{n} \tilde{\mathbf{Z}}' \mathbf{X}' \mathbf{X}'^\top + \frac{\lambda}{n^2} \mathbf{X}'^\top \mathbf{X}' \mathbf{X}'^\top \mathbf{X}'_0 - \frac{\lambda}{2n^2} \mathbf{X}'^\top \mathbf{X}' \mathbf{X}'^\top \mathbf{X}' \right) \right\rangle'_{n, \varepsilon_n} \\
 &= -\mathbb{E} \left\langle \frac{1}{2} \operatorname{Tr} \left( \frac{\lambda}{n^2} \mathbf{X}'^\top \mathbf{X}' \mathbf{X}'^\top \mathbf{X}'_0 - \frac{\lambda}{2n^2} \mathbf{X}'^\top \mathbf{X}' \mathbf{X}'^\top \mathbf{X}' \right) \right\rangle'_{n, \varepsilon_n} \\
 &\leq \frac{3\lambda(M+1)^2 D^4}{4}. \tag{5.45}
 \end{aligned}$$

Combining inequalities (5.44) and (5.45) with the fact that the remaining error term  $(M+1)\Gamma'(n, M+1)$  in inequality (5.43) is at most quadratic in  $M$  for all choices of  $n$ , said inequality simplifies to give

$$\hat{\Delta}_N(n) \leq CM^2.$$

Substituting this into inequality (5.42) completes the proof.  $\blacksquare$

The next proposition allows us to deduce the  $\Delta_M$  cavity bound from the  $\Delta_N$  cavity bound at a slight loss of the rate, provided that we have a rank-one equivalence formula.

**Proposition 5.6** ( $\Delta_M/N$  cavity upper bound). *Assume hypotheses (H2)–(H5) and recall (5.12) that for  $N$  and  $M = M_N$  such that  $M_{N+1} = M+1$ , we have*

$$\Delta_M = \mathbb{E} \ln \tilde{Z}_{N, M+1} - \mathbb{E} \ln \tilde{Z}_{N, M},$$

where  $\mathbb{E}[\cdot]$  denotes an expectation with respect to  $\mathbf{Z}, \tilde{\mathbf{Z}}, \mathbf{X}_0, \varepsilon_N$ . Furthermore, suppose that

$$\frac{\Delta_N}{M+1} \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + \Phi(N, M) \tag{5.46}$$

for some error function  $\Phi(N, M): \mathbb{N}^2 \rightarrow \mathbb{R}_+$  decreasing in  $N$ . Then, for any  $T \leq N$ , there exists a finite positive constant  $C$  independent of  $M, N, T$  such that

$$\frac{\Delta_M}{N} \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + (M+1)\Phi(T, M) + \frac{CTM^2}{N}. \tag{5.47}$$

*Proof.* By Lemma 4.2 and Proposition 2.7, we have for each  $N, M \geq 1$  that

$$\frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M} \geq MF_N(\lambda) \geq M \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda).$$

Hence, it is immediate that

$$\frac{\Delta_M}{N} \leq \frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M+1} - M \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda). \tag{5.48}$$

Recalling that  $\hat{\Delta}_N(n) = \mathbb{E} \ln \tilde{Z}_{n+1, M+1} - \mathbb{E} \ln \tilde{Z}_{n, M+1}$  (5.41) and letting  $1 \leq T \leq N - 1$ , we have the telescoping series

$$\begin{aligned} \frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M+1} &= \frac{1}{N} \sum_{n=0}^{N-1} \hat{\Delta}_N(n) \\ &= \frac{1}{N} \sum_{n=0}^{T-1} \hat{\Delta}_N(n) + \frac{1}{N} \sum_{n=T}^{N-1} \hat{\Delta}_N(n) \\ &\leq \left(1 - \frac{T}{N}\right) \sup_{n \geq T} \hat{\Delta}_N(n) + \frac{CTM^2}{N}, \end{aligned} \quad (5.49)$$

where we have used Lemma 5.5. Replacing  $N$  by  $n$  in inequality (5.46) shows that for all  $n \geq T$ ,

$$\hat{\Delta}_N(n) \leq (M+1) \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + (M+1)\Phi(T, M).$$

As  $T \leq N - 1$ , substituting this into inequality (5.49) produces

$$\frac{1}{N} \mathbb{E} \ln \tilde{Z}_{N, M+1} \leq (M+1) \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + (M+1)\Phi(T, M) + \frac{CTM^2}{N},$$

which further combines with inequality (5.48) to yield

$$\frac{\Delta_M}{N} \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + (M+1)\Phi(T, M) + \frac{CTM^2}{N},$$

as required. ■

A key property of this upper bound is that it is sharp when one is able to take  $T$  such that both  $T = o(NM_N^{-2})$  and  $M_N\Phi(T, M_N) \rightarrow 0$  at large  $N$ . Indeed, a sharp bound is made possible upon taking  $M_N = o(\sqrt{\ln N})$  and setting

$$\Phi(N, M) = \frac{C}{\sqrt{M+1}} e^{C(M+1)^2} \Gamma(N, M+1)^{1/4} + \Gamma'(N, M+1), \quad (5.50)$$

as prescribed by inequality (5.40) (recalling that  $M$  therein stands for  $M_{N+1}$ ).

**Corollary 5.7.** *Assume the hypotheses of Theorem 2.1 and recall (5.12) that for  $N$  and  $M = M_N$  such that  $M_{N+1} \neq M$ , we have*

$$\Delta_M = \mathbb{E} \ln \tilde{Z}_{N, M+1} - \mathbb{E} \ln \tilde{Z}_{N, M},$$

where  $\mathbb{E}[\cdot]$  denotes an expectation with respect to  $\mathbf{Z}$ ,  $\tilde{\mathbf{Z}}$ ,  $\mathbf{X}_0$ ,  $\varepsilon_N$ . There exists a finite positive constant  $C$  independent of  $M, N$  such that choosing  $s_N$  to decay to zero while satisfying  $s_N \geq \min\{0.9, (C^2 \ln N / (8C + 9))^4 N^{-1/(8C+9)}\}$  ensures that

$$\limsup_{N \rightarrow \infty} \frac{\Delta_M}{N} \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda).$$

*Proof.* Inequality (5.40) stems from the same hypotheses of Proposition 5.4 as those being assumed here, so we read from said inequality that

$$\frac{\Delta_N}{M+1} \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + \frac{C}{\sqrt{M+1}} e^{C(M+1)^2} \Gamma(N, M+1)^{1/4} + \Gamma'(N, M+1).$$

Thus, we may take  $\Phi(N, M)$  within Proposition 5.6 to be as in equation (5.50) so that

$$\begin{aligned} \frac{\Delta_M}{N} &\leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) + C \sqrt{M+1} e^{C(M+1)^2} \Gamma(T, M+1)^{1/4} \\ &\quad + \Gamma'(T, M+1) + \frac{CTM^2}{N} \end{aligned} \quad (5.51)$$

for any  $T \leq N$ .

It remains now to choose  $T$  such that the error terms in the above vanish in the large  $N$  limit. Hence, let  $\epsilon = 1/(8C + 10)$  and set  $T = N^{1-\epsilon}$ . Since  $M = o(\sqrt{\ln N})$  by hypothesis (H1), we have that for  $N$  large enough,  $M < M+1 \leq \sqrt{\ln N}^\epsilon$  and thus

$$\begin{aligned} C \sqrt{M+1} e^{C(M+1)^2} \Gamma(T, M+1)^{1/4} &= e^{C(M+1)^2} \frac{C(M+1)}{T^{1/8} s_T^{1/8}} \\ &\leq C \sqrt{\epsilon} \sqrt{\ln N} N^{(C+1/8)\epsilon-1/8} s_T^{-1/8}, \\ (M+1) \Gamma'(T, M+1) &= C \left( \frac{M+1-T}{T+1} + 1 + s_T \right) \\ &\leq C \left( \frac{\sqrt{\epsilon} \sqrt{\ln N} - T}{T+1} + 1 + s_T \right), \\ \frac{CTM^2}{N} &\leq C \epsilon N^{-\epsilon} \ln N. \end{aligned}$$

By our choice of  $s_N$ , we have that  $s_T \geq (C^2(\epsilon/(1-\epsilon)) \ln T)^4 T^{\epsilon/(\epsilon-1)}$  in the thermodynamic limit. Thus, inserting  $T = N^{1-\epsilon}$  into the first error term above shows that it is bounded above by  $N^{-\epsilon}$ , and hence vanishes at large  $N$ . As  $s_T \rightarrow 0$  and  $\sqrt{\ln N} \ll T = N^{1-\epsilon}$ , the second error term also vanishes at large  $N$ . Likewise for the final error term. Therefore, taking the limit supremum in inequality (5.51) produces the stated result.  $\blacksquare$

We conclude this subsection with a demonstration of how the arguments underlying Proposition 5.6 can be extended to the case where  $\Phi(N, M)$  vanishes in the  $M = o(N^{\gamma/2})$  regime for some  $0 < \gamma \leq 1$  (in line with the rate of concentration in Theorem 2.11), highlighting the fact that the bottleneck posed by hypothesis (H1) is purely due to the method of proof used in Section 5.2.

**Lemma 5.8.** *If there exists  $0 < \gamma \leq 1$  such that*

$$\Phi(N, M) = \frac{M^2}{N^\gamma},$$

*then for  $M = o(N^{\gamma/(2\gamma+3)})$ , there exists  $T$  a sequence in  $N$  such that*

$$\max\left\{M\Phi(T, M), \frac{TM^2}{N}\right\} \rightarrow 0.$$

*Proof.* We show that  $M = o(N^{\gamma/(2\gamma+3)})$  is the optimal rate for which there exists  $T$  such that the error terms in question vanish at large  $N$ . We set  $M = N^x$  and  $T = N^y$  for some yet to be determined constants  $0 \leq x, y \leq 1$ . We have that

$$M\Phi(T, M) = \frac{M^3}{N^\gamma} \rightarrow 0 \iff x < \frac{\gamma y}{3}$$

and

$$\frac{TM^2}{N} \rightarrow 0 \iff x < \frac{1-y}{2}.$$

The curves bounding these regions intersect precisely when

$$y = \frac{3}{2\gamma + 3}.$$

Therefore, for any  $\epsilon > 0$  sufficiently small, both error terms limit to zero upon setting  $x = \gamma/(2\gamma + 3) - \epsilon$  and  $y = 3/(2\gamma + 3) - \epsilon$ . ■

**Remark 5.9.** Notice that even upon assuming that a rate of  $M = o(N^{\gamma/2})$  is sufficient for the error term  $\Phi(N, M)$  in the  $\Delta_N$  cavity bound (5.46) to vanish at large  $N$ , one requires the slower rate  $M = o(N^{\gamma/(2\gamma+3)})$  for the equivalent error term of the accompanying  $\Delta_M$  cavity bound (5.47) to also vanish as  $N \rightarrow \infty$ .

#### 5.4. Proof of Proposition 2.8 and Theorem 2.1

We now conclude this paper by giving a brief demonstration of how the above results constitute a proof of Theorem 2.1. The majority of the argument is in fact a proof of Proposition 2.8, which combines with Proposition 2.7, proven in Section 4.1, to give Theorem 2.1.

*Proof of Theorem 2.1.* Throughout this paper,  $C$  has stood for a generic finite positive constant independent of  $N, M$ , and the truncation parameter  $T$  of Lemma 5.5. Noting that each inequality of this paper involving  $C$  remains true upon arbitrarily increasing  $C$ , we now take  $C$  to be large enough such that every such inequality is true for the same choice of  $C$ . Then, setting

$$s_N = \min\left\{0.9, \max\left\{N^{-1/9}(Ce^{CM^2})^{8/9}, (C^2 \ln N / (8C + 9))^4 N^{-1/(8C+9)}\right\}\right\}$$

in order to satisfy the hypotheses of Proposition 5.4 and Corollary 5.7, we have by said proposition and corollary that

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{\Delta_N}{M} &\leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda), \\ \limsup_{N \rightarrow \infty} \frac{\Delta_M}{N} &\leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda). \end{aligned}$$

Thus, by Theorem 2.9, there exists some  $0 \leq \alpha \leq 1$  such that

$$\begin{aligned} \limsup_{N \rightarrow \infty} \tilde{F}_N(\lambda) &\leq \alpha \limsup_{N \rightarrow \infty} \frac{\Delta_N}{M} + (1 - \alpha) \limsup_{N \rightarrow \infty} \frac{\Delta_M}{N} \\ &\leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda). \end{aligned}$$

Proposition 2.8 hence follows from inequality (4.16), which we recall is a straightforward consequence of Lemma 4.2. Combining Proposition 2.8 with the result of taking the limit infimum in Proposition 2.7 finally shows that

$$\limsup_{N \rightarrow \infty} F_N(\lambda) \leq \sup_{q \in [0, \rho]} F_1^{\text{RS}}(q, \lambda) \leq \liminf_{N \rightarrow \infty} F_N(\lambda).$$

This is equivalent to Theorem 2.1. ■

## A. The replica symmetric potential in terms of mutual information

Consider the vector Gaussian channel

$$\mathbf{y} = \sqrt{\lambda} \mathbf{Q} \mathbf{x}_0 + \mathbf{z} \tag{A.1}$$

with  $\mathbf{z} \in \mathbb{R}^M$  a standard Gaussian vector and  $\mathbf{x}_0 \sim \mathbb{P}_{X, M}$ . Inspection of the replica symmetric potential (2.2) reveals that it is closely related to the free entropy of this vector channel. Indeed, the intuition behind the replica method is that the size  $N$ , rank  $M$  spiked Wigner model (1.1) is, up to the regularization term  $-(\lambda/4M) \text{Tr} \mathbf{Q}^2$ , statistically equivalent to the mean field model corresponding to  $N$  replicas of the channel (A.1); the quadratic interactions within the  $\mathbf{X}_0 \mathbf{X}_0^\top$  term of (1.1) are accounted for by the order parameter  $\mathbf{Q}$ , which is eventually seen to play the role of the averaged overlap matrix  $\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \langle \mathbf{x} \mathbf{x}_0^\top \rangle_{\text{RS}}$  (see Lemma 3.1).

To rewrite the replica symmetric potential in the form (2.9), we study the vector channel (A.1) from an information theoretic viewpoint. Thus, begin by recalling that the *differential entropy* [21] of a random vector  $\mathbf{X} \sim p_X(\mathbf{X})$  is defined as

$$H(\mathbf{X}) := - \int_{\mathbb{R}^M} \ln p_X(\mathbf{X}) \, d p_X(\mathbf{X}).$$

Then, the mutual information between  $\mathbf{x}_0$  and the output  $\mathbf{y}$  of the channel (A.1) is

$$I(\mathbf{x}_0; \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y} | \mathbf{x}_0) = H(\sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z}) - H(\mathbf{z}); \quad (\text{A.2})$$

it is a standard result that  $H(\mathbf{z}) = M \ln(2\pi e)/2$ . By Bayes' rule, the first term in the right-hand side of equation (A.2) reads

$$H(\sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z}) = H(\mathbf{y}) = -\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \ln \int_{\mathbb{R}^M} \mathbb{P}_{Y|X}(\mathbf{y}(\mathbf{x}_0, \mathbf{z}) | \mathbf{x}) d\mathbb{P}_{X, M}(\mathbf{x}).$$

Conditionally on  $\mathbf{x}_0$ , the data  $\mathbf{y}$  is Gaussian distributed with mean  $\sqrt{\lambda \mathbf{Q}} \mathbf{x}_0$ , so the above expression for the entropy reduces further to

$$\begin{aligned} H(\sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z}) &= -\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \ln \int_{\mathbb{R}^M} \exp\left(-\frac{1}{2} \|\sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z} - \sqrt{\lambda \mathbf{Q}} \mathbf{x}\|_{\mathbb{F}}^2\right) d\mathbb{P}_{X, M}(\mathbf{x}) \\ &\quad + \frac{M}{2} \ln(2\pi) \\ &= -\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \ln \int_{\mathbb{R}^M} \exp\left(\sqrt{\lambda} \mathbf{x}^\top \sqrt{\mathbf{Q}} \mathbf{z} + \lambda \mathbf{x}_0^\top \mathbf{Q} \mathbf{x} - \frac{\lambda}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x}\right) d\mathbb{P}_{X, M}(\mathbf{x}) \\ &\quad + \frac{1}{2} \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \|\sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z}\|_{\mathbb{F}}^2 + \frac{M}{2} \ln(2\pi). \end{aligned}$$

Since  $\mathbf{x}_0$  and  $\mathbf{z}$  are independent and  $\mathbf{z}$  has identity covariance, we have that

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \|\sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z}\|_{\mathbb{F}}^2 = \lambda \mathbb{E}_{\mathbf{z}, \mathbf{x}_0} [\mathbf{x}_0^\top \mathbf{Q} \mathbf{x}_0] + M.$$

Moreover, we have by hypothesis (H2) that our prior distribution on  $\mathbf{x}_0$  factorizes as  $\mathbb{P}_{X, M} = \mathbb{P}_X^{\otimes M}$  according to (2.1), so

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} [\mathbf{x}_0^\top \mathbf{Q} \mathbf{x}_0] = \lambda \operatorname{Tr} \mathbf{Q} \mathbb{E}[\mathbf{x}_0 \mathbf{x}_0^\top] = \lambda \rho \operatorname{Tr} \mathbf{Q},$$

where we recall that  $\rho$  denotes the variance of  $\mathbb{P}_X$ . Thus, returning to equation (A.2), we find that

$$I(\mathbf{x}_0; \mathbf{y}) = -\mathbb{E}_{\mathbf{z}, \mathbf{x}_0} \ln \int_{\mathbb{R}^M} e^{\sqrt{\lambda} \mathbf{x}^\top \sqrt{\mathbf{Q}} \mathbf{z} + \lambda \mathbf{x}_0^\top \mathbf{Q} \mathbf{x} - \frac{\lambda}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x}} d\mathbb{P}_{X, M}(\mathbf{x}) + \frac{\lambda \rho}{2} \operatorname{Tr} \mathbf{Q}.$$

Comparing this expression to the definition (2.2) of  $F_M^{\text{RS}}(\mathbf{Q}, \lambda)$ , and completing the square to write  $\operatorname{Tr} \mathbf{Q}^2 - 2\rho \operatorname{Tr} \mathbf{Q} = \|\mathbf{Q} - \rho I_M\|_{\mathbb{F}}^2 - M\rho^2$  finally shows that

$$\begin{aligned} F_M^{\text{RS}}(\mathbf{Q}, \lambda) &= -\frac{1}{M} I(\mathbf{x}_0; \sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z}) + \frac{\lambda \rho}{2M} \operatorname{Tr} \mathbf{Q} - \frac{\lambda}{4M} \operatorname{Tr} \mathbf{Q}^2 \\ &= -\frac{1}{M} I(\mathbf{x}_0; \sqrt{\lambda \mathbf{Q}} \mathbf{x}_0 + \mathbf{z}) - \frac{\lambda}{4M} \|\mathbf{Q} - \rho I_M\|_{\mathbb{F}}^2 + \frac{\lambda \rho^2}{4}. \end{aligned}$$

**Acknowledgments.** The authors thank the anonymous referees for careful reading of the manuscript and helpful suggestions. The authors acknowledge crucial discussions with Francesco Camilli and Koki Okajima which led to the joint development of the multiscale cavity method. J. Barbier wishes to thank Nicolas Macris for the uncountable discussions concerning this problem and many others over the years.

**Funding.** J. Barbier and A. A. Rahman were funded by the European Union (ERC, CHORAL, project number 101039794). J. Ko was funded by the European Union (ERC, LDRAM, project number 884584) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-04597). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## References

- [1] E. Abbe, Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.* **18** (2017), article no. 177 Zbl 1403.62110 MR 3827065
- [2] M. Aizenman and P. Contucci, On the stability of the quenched state in mean-field spin-glass models. *J. Statist. Phys.* **92** (1998), no. 5-6, 765–783 Zbl 0963.82045 MR 1657840
- [3] M. Aizenman, R. Sims, and S. L. Starr, Extended variational principle for the Sherrington–Kirkpatrick spin-glass model. *Phys. Rev. B* **68** (2003), article no. 214403
- [4] A. Auffinger and Y. Zhou, On properties of the spherical mixed vector  $p$ -spin model. *Stochastic Process. Appl.* **146** (2022), 382–413 Zbl 1484.60102 MR 4383209
- [5] J. Baik, G. Ben Arous, and S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** (2005), no. 5, 1643–1697 Zbl 1086.15022 MR 2165575
- [6] J. Barbier, Overlap matrix concentration in optimal Bayesian inference. *Inf. Inference* **10** (2021), no. 2, 597–623 Zbl 1475.62119 MR 4270761
- [7] J. Barbier, F. Camilli, J. Ko, and K. Okajima, Phase diagram of extensive-rank symmetric matrix denoising beyond rotational invariance. *Phys. Rev. X* **15** (2025), article no. 021085
- [8] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 424–432, Curran Associates, Red Hook, NY, 2016
- [9] J. Barbier and N. Macris, Statistical limits of dictionary learning: Random matrix theory and the spectral replica method. *Phys. Rev. E* **106** (2022), no. 2, article no. 024136 MR 4489098

- [10] J. Barbier and D. Panchenko, [Strong replica symmetry in high-dimensional optimal Bayesian inference](#). *Comm. Math. Phys.* **393** (2022), no. 3, 1199–1239 Zbl 1494.82020 MR 4453233
- [11] E. Bates and Y. Sohn, [Parisi formula for balanced Potts spin glass](#). *Comm. Math. Phys.* **405** (2024), no. 10, article no. 228 Zbl 07923631 MR 4797739
- [12] D. Belius, L. Fröber, and J. Ko, [TAP variational principle for the constrained overlap multiple spherical Sherrington–Kirkpatrick model](#). 2023, arXiv:2304.04031v1
- [13] A. Bodin and N. Macris, [Gradient flow on extensive-rank positive semi-definite matrix denoising](#). In *2023 IEEE Information Theory Workshop (ITW)*, pp. 365–370, IEEE, 2023
- [14] F. Camilli and M. Mézard, [Matrix factorization with neural networks](#). *Phys. Rev. E* **107** (2023), no. 6, article no. 064308 MR 4618735
- [15] F. Camilli and M. Mézard, [The decimation scheme for symmetric matrix factorization](#). *J. Phys. A* **57** (2024), no. 8, article no. 085002 Zbl 1548.15030 MR 4705338
- [16] H.-B. Chen, [On Parisi measures of Potts spin glasses with correction](#). *Electron. Commun. Probab.* **29** (2024), article no. 35 Zbl 1551.82024 MR 4781870
- [17] H.-B. Chen, [On the self-overlap in vector spin glasses](#). *J. Math. Phys.* **65** (2024), no. 3, article no. 031901 Zbl 1535.82024 MR 4713500
- [18] H.-B. Chen, [Parisi PDE and convexity for vector spins](#). *Stochastic Process. Appl.* **190** (2025), article no. 104746 Zbl 08111209 MR 4935983
- [19] H.-B. Chen and J.-C. Mourrat, [On the free energy of vector spin glasses with nonconvex interactions](#). *Probab. Math. Phys.* **6** (2025), no. 1, 1–80 Zbl 1563.82013 MR 4848391
- [20] Y. Chen and J. Xu, [Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices](#). *J. Mach. Learn. Res.* **17** (2016), article no. 27 Zbl 1360.62320 MR 3491121
- [21] T. M. Cover and J. A. Thomas, *Elements of information theory*. Second edn., Wiley-Interscience, Hoboken, NJ, 2006 Zbl 1140.94001 MR 2239987
- [22] Y. Deshpande, E. Abbe, and A. Montanari, [Asymptotic mutual information for the balanced binary stochastic block model](#). *Inf. Inference* **6** (2017), no. 2, 125–170 Zbl 1383.62021 MR 3671474
- [23] Y. Deshpande and A. Montanari, [Information-theoretically optimal sparse PCA](#). In *2014 IEEE International Symposium on Information Theory (ISIT)*, pp. 2197–2201, IEEE, 2014
- [24] S. N. Diggavi and T. M. Cover, [Is maximum entropy noise the worst?](#) In *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, p. 278, IEEE, 1997
- [25] S. N. Diggavi and T. M. Cover, [The worst additive noise under a covariance constraint](#). *IEEE Trans. Inform. Theory* **47** (2001), no. 7, 3072–3081 Zbl 1021.94515 MR 1872870
- [26] T. Dominguez, [The  \$\ell^p\$ -Gaussian–Grothendieck problem with vector spins](#). *Electron. J. Probab.* **27** (2022), article no. 70 Zbl 1490.82020 MR 4440068
- [27] D. L. Donoho and M. J. Feldman, [Optimal eigenvalue shrinkage in the semicircle limit](#). [v1] 2022, [v2] 2023, arXiv:2210.04488v2
- [28] M. J. Feldman, [Spiked singular values and vectors under extreme aspect ratios](#). *J. Multivariate Anal.* **196** (2023), article no. 105187 Zbl 1520.15018 MR 4575691

- [29] S. Ghirlanda and F. Guerra, [General properties of overlap probability distributions in disordered spin systems. Towards Parisi ultrametricity](#). *J. Phys. A* **31** (1998), no. 46, 9149–9155 Zbl [0953.82037](#) MR [1662161](#)
- [30] F. Guerra, [Broken replica symmetry bounds in the mean field spin glass model](#). *Comm. Math. Phys.* **233** (2003), no. 1, 1–12 Zbl [1013.82023](#) MR [1957729](#)
- [31] A. Guionnet, J. Ko, F. Krzakala, and L. Zdeborová, [Estimating rank-one matrices with mismatched prior and noise: Universality and large deviations](#). *Comm. Math. Phys.* **406** (2025), no. 1, article no. 9 Zbl [07963928](#) MR [4837935](#)
- [32] A. Guionnet, J. Ko, F. Krzakala, and L. Zdeborová, [Low-rank matrix estimation with inhomogeneous noise](#). *Inf. Inference* **14** (2025), no. 2, article no. iaaf010 Zbl [08113037](#) MR [4893427](#)
- [33] D. Guo, S. Shamai, and S. Verdú, [Mutual information and minimum mean-square error in Gaussian channels](#). *IEEE Trans. Inform. Theory* **51** (2005), no. 4, 1261–1282 Zbl [1309.94099](#) MR [2241490](#)
- [34] D. Guo, Y. Wu, S. Shamai, and S. Verdú, [Estimation in Gaussian noise: Properties of the minimum mean-square error](#). *IEEE Trans. Inform. Theory* **57** (2011), no. 4, 2371–2385 Zbl [1366.94144](#) MR [2809096](#)
- [35] B. Hajek, Y. Wu, and J. Xu, [Submatrix localization via message passing](#). *J. Mach. Learn. Res.* **18** (2017), article no. 186 Zbl [1468.68155](#) MR [3827074](#)
- [36] J. Huang, [Mesoscopic perturbations of large random matrices](#). *Random Matrices Theory Appl.* **7** (2018), no. 2, article no. 1850004 Zbl [1390.15121](#) MR [3786885](#)
- [37] J. Husson and J. Ko, [Spherical integrals of sublinear rank](#). *Probab. Theory Related Fields* **193** (2025), no. 1-2, 1–88 Zbl [08109723](#) MR [4970431](#)
- [38] I. M. Johnstone, [On the distribution of the largest eigenvalue in principal components analysis](#). *Ann. Statist.* **29** (2001), no. 2, 295–327 Zbl [1016.62078](#) MR [1863961](#)
- [39] J. Ko, [The Crisanti–Sommers formula for spherical spin glasses with vector spins](#). 2019, arXiv:[1911.04355v1](#)
- [40] J. Ko, [Free energy of multiple systems of spherical spin glasses with constrained overlaps](#). *Electron. J. Probab.* **25** (2020), article no. 28 Zbl [1437.60075](#) MR [4073689](#)
- [41] S. B. Korada and N. Macris, [Exact solution of the gauge symmetric  \$p\$ -spin glass model on a complete graph](#). *J. Stat. Phys.* **136** (2009), no. 2, 205–230 Zbl [1180.82200](#) MR [2525244](#)
- [42] F. Krzakala, J. Xu, and L. Zdeborová, [Mutual information in rank-one matrix estimation](#). In *2016 IEEE Information Theory Workshop (ITW)*, pp. 71–75, IEEE, 2016
- [43] M. Lelarge and L. Miolane, [Fundamental limits of symmetric low-rank matrix estimation](#). *Probab. Theory Related Fields* **173** (2019), no. 3-4, 859–929 Zbl [1411.60014](#) MR [3936148](#)
- [44] T. Lesieur, F. Krzakala, and L. Zdeborová, [MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel](#). In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing*, pp. 680–687, IEEE, 2015
- [45] T. Lesieur, F. Krzakala, and L. Zdeborová, [Phase transitions in sparse PCA](#). In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1635–1639, IEEE, 2015

- [46] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová, [Statistical and computational phase transitions in spiked tensor estimation](#). In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 511–515, IEEE, 2017
- [47] C. Luneau, J. Barbier, and N. Macris, [Mutual information for low-rank even-order symmetric tensor estimation](#). *Inf. Inference* **10** (2021), no. 4, 1167–1207 Zbl 1490.62034 MR 4353122
- [48] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, [Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising](#). *J. Stat. Mech. Theory Exp.* (2022), no. 8, article no. 083301 Zbl 1539.94007 MR 4504819
- [49] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond*. World Sci. Lecture Notes Phys. 9, World Scientific, Teaneck, NJ, 1987 Zbl 0992.82500 MR 1026102
- [50] L. Miolane, Fundamental limits of low-rank matrix estimation: The non-symmetric case. [v1] 2017, [v3] 2018, arXiv:1702.00473v3
- [51] L. Miolane, Phase transitions in spiked matrix estimation: Information-theoretic analysis. [v1] 2018, [v3] 2019, arXiv:1806.04343v3
- [52] A. Montanari and E. Richard, A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems 27*, pp. 2987–2905, Curran Associates, Red Hook, NY, 2014
- [53] A. Montanari and Y. Wu, [Fundamental limits of low-rank matrix estimation with diverging aspect ratios](#). *Ann. Statist.* **52** (2024), no. 4, 1460–1484 Zbl 07928784 MR 4804816
- [54] J.-C. Mourrat, [Free energy upper bound for mean-field vector spin glasses](#). *Ann. Inst. Henri Poincaré Probab. Stat.* **59** (2023), no. 3, 1143–1182 Zbl 1549.82038 MR 4635706
- [55] D. Panchenko, [The Parisi ultrametricity conjecture](#). *Ann. of Math. (2)* **177** (2013), no. 1, 383–393 Zbl 1270.60060 MR 2999044
- [56] D. Panchenko, *The Sherrington–Kirkpatrick model*. Springer Monogr. Math., Springer, New York, 2013 Zbl 1266.82005 MR 3052333
- [57] D. Panchenko, [Free energy in the mixed  \$p\$ -spin models with vector spins](#). *Ann. Probab.* **46** (2018), no. 2, 865–896 Zbl 1430.60032 MR 3773376
- [58] D. Panchenko, [Free energy in the Potts spin glass](#). *Ann. Probab.* **46** (2018), no. 2, 829–864 Zbl 1430.60088 MR 3773375
- [59] S. Péché, [The largest eigenvalue of small rank perturbations of Hermitian random matrices](#). *Probab. Theory Related Fields* **134** (2006), no. 1, 127–173 Zbl 1088.15025 MR 2221787
- [60] F. Pourkamali, J. Barbier, and N. Macris, [Matrix inference in growing rank regimes](#). *IEEE Trans. Inform. Theory* **70** (2024), no. 11, 8133–8163 Zbl 1564.94023 MR 4817652
- [61] G. Reeves, [Information-theoretic limits for the matrix tensor product](#). *IEEE J. Sel. Areas Inf. Theory* **1** (2020), no. 3, 777–798
- [62] G. Reeves, H. D. Pfister, and A. Dytso, [Mutual Information as a function of matrix SNR for linear Gaussian channels](#). In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1754–1758, IEEE, 2018

- [63] I. Shomorony and A. S. Avestimehr, [Is Gaussian noise the worst-case additive noise in wireless networks?](#) In *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 214–218, IEEE, 2012
- [64] M. Talagrand, *Mean field models for spin glasses. Volume 1*. *Ergeb. Math. Grenzgeb. (3)* 54, Springer, Berlin, 2011 Zbl [1214.82002](#) MR [2731561](#)
- [65] S. Tarmoun, G. Franca, B. D. Haeffele, and R. Vidal, Understanding the dynamics of gradient flow in overparameterized linear models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10153–10161, Proceedings of Machine Learning Research 139, PMLR, 2021
- [66] A. Wibisono and V. Jog, [Convexity of mutual information along the heat flow](#). In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1615–1619, IEEE, 2018
- [67] Y. Wu and S. Verdú, [MMSE dimension](#). *IEEE Trans. Inform. Theory* **57** (2011), no. 8, 4857–4879 Zbl [1365.62104](#) MR [2849092](#)
- [68] W. H. Young, [On classes of summable functions and their Fourier series](#). *Lond. Roy. Soc. Proc. (A)* **87** (1912), no. 594, 225–229 Zbl [43.1114.12](#)
- [69] L. Zdeborová and F. Krzakala, [Statistical physics of inference: Thresholds and algorithms](#). *Adv. Phys.* **65** (2016), no. 5, 453–552

Received 18 April 2025; revised 17 November 2025.

### **Jean Barbier**

Quantitative Life Sciences and Mathematics Sections, The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy; [jbarbier@ictp.it](mailto:jbarbier@ictp.it)

### **Justin Ko**

Unité de Mathématiques Pures at Appliquées, École Normale Supérieure de Lyon, 15 parvis René Descartes, 69342 Lyon, France; Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, N2L 3G1, Canada; [jko104@syr.edu](mailto:jko104@syr.edu)

### **Anas A. Rahman**

Quantitative Life Sciences and Mathematics Sections, The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy; [anas.rahman@live.com.au](mailto:anas.rahman@live.com.au)