PSEUDO-MAXIMUM LIKELIHOOD THEORY FOR HIGH-DIMENSION RANK ONE INFERENCE

CURTIS GRANT, AUKOSH JAGANNATH, AND JUSTIN KO

ABSTRACT. We develop a pseudo-likelihood theory for rank one matrix estimation problems in the high dimensional limit. We prove a variational principle for the limiting pseudo-likelihood's value, and show it is universal, depending only on four information parameters determined by the corresponding null model. Through this universality we establish equivalence for estimation problems, and in particular relate recovery in spiked matrix models with the community detection problem. We further give a complete description of the performance of the least-squares estimator for any rank one recovery problem.

1. INTRODUCTION

Suppose that we are given data in the form of a real, symmetric $N \times N$ matrix, $Y \in \mathbb{R}^{N\times N}$, whose entries are conditionally independent given an unknown vector $\mathbf{x}_0 \in \Omega^N \subseteq \mathbb{R}^N$ and where each entry of Y has conditional law

$$Y_{ij} \sim \mathbb{P}_Y\left(\cdot \,|\, \frac{\lambda}{\sqrt{N}} x_i^0 x_j^0 \right) \quad \text{for } i \leq j\,,$$

for some $\lambda > 0$, and $\Omega \subseteq \mathbb{R}$ is compact.[1] Our goal is to infer $\mathbf{x}_0$.

High-dimensional rank one estimation tasks with structure form one of the central classes of problems in high-dimensional statistics. This data model captures a broad range of problems that have received a tremendous amount of attention in recent years, such as Sparse PCA [83], $\mathbb{Z}_2$ Synchronization [48], submatrix localization [13, 38], matrix factorization [50], community detection [53], biclustering [59], and non-linear spiked matrix models [69] among many others.

From a statistical perspective, a substantial literature on these problems has emerged over the past decade, particularly from the perspective of hypothesis testing and Bayesian inference. The fundamental limits of hypothesis testing have been explored in [5]. The fundamental limits of Bayesian inference, specifically computing the mutual information of and characterizing the performance for the (matrix) minimum mean-squared error estimator, has been explored in [55, 58, 75, 20, 19, 25]. More generally, the setting of "mismatched" Bayesian inference was developed in [15, 8, 10, 73, 33, 37]. From an algorithmic perspective, various algorithms (along with performance guarantees) for specific problems and estimators—including the MMSE—have been introduced in recent years using the frameworks of Approximate Message Passing [27, 74, 24, 56], Spectral methods [70, 61], Semi-definite programs [79, 80, 49], Low-degree methods [63], and the sum-of-squares hierarchy [42, 41].

A natural question is to understanding the statistical performance of more general optimization based procedures, such as maximum likelihood estmation (MLE), maximum a posteriori (MAP) estimation, or best low rank approximations. The literature for these methods, however, is far

Date: January 14, 2025.

[1]The scaling assumption here in N matches the regime where non-trivial high-dimensional effects, such as the BBP phase transition, occur. It guarantees that both the operator norm of Y and $\frac{1}{\sqrt{N}}(\mathbf{x^0})(\mathbf{x^0})^\intercal$ are of the same order.

more sparse. To our knowledge, to date, there have been a sharp understanding only the case of the MLE in Sparse PCA [44] as well as variational inference for $\mathbb{Z}_2$-synchronization [30, 17].

We seek here to close this gap. To this end, observe that many popular optimization based estimators for such problems, such as those mentioned above, can be interpreted as pseudo-likelihood methods [34]. In this paper, we provide a unified analysis of the performance of pseudo-likelihood methods.

We develop a pseudo maximum likelihood theory for rank one inference tasks in the high-dimensional regime for when the latent vector, $\boldsymbol{x}^0$, is structured. We provide exact variational formulas for the asymptotic pseudo-likelihood and, as a direct consequence, obtain exact variational characterizations for the performance of the corresponding estimators. See Section 2.

We find that these problem exhibit "universal" behaviour in that these variational characterizations depend only on four scalar quantities, which we call the *information parameters*. These parameters encode certain Fisher-type information of the pseudolikelihood with respect to a "null" model and are reminiscent of the score parameters appearing in the classical regime [32].

Surprisingly, we find that if one of these information parameters, which we call the *score parameter*, is not zero, then it entirely dictates the effectiveness of our inference method, and the effect is typically catastrophic. We refer to such models as *ill-scored* models. We present here a data-driven approach to systematically correct for this effect and obtain a corresponding variational characterization for the performance of this *score-corrected* method. See Section 2.7.

Since a given inference tasks is entirely characterized by its information parameters, our analysis yields two general notions of equivalence of inference tasks, called strong and coarse equivalence. For example, we give a precise sense in which the problem of maximum likelihood estimation for certain spiked matrix models and the stochastic block model are equivalent. See Section 3.

We then illustrate our results with a broad range of examples. First, we present a complete analysis of the performance of the popular "Best rank 1 approximation" procedure [28]. We also provide a method to correct for some of these issues, by introducing the score-corrected least squares procedure. Surprisingly, however, we find that in natural problems, such as a sparse Rademacher matrices, the best rank one approximation and its score-corrected version are necessarily completely uninformative. Indeed, we provide a sufficient condition for the failure of such methods. Finally in Section 5, we illustrate how our approach can be used to analyze a broad range of problems and methods. Specifically, we study popular inference methods for spiked matrix models, $\mathbb{Z}_2$-synchronization, the Stochastic Block Model, sparse rademacher matrices, and non-linear transformations of spiked matrix models.

Let us pause here to discuss the technical tools involved in our work and how they compare to the above mentioned literature. Since the latent vector is structured, standard tools of high-dimensional statistics, such as concentration of measure or random matrix theory, are unable to yield a sharp understanding of these problems. To circumvent this, the recent progress in the past decade has used deep connections to statistical physics, specifically to the theory of spin glasses. In particular, the central insight is that Hypothesis Testing and Bayesian inference of matrix models are deeply connected to the Sherrington-Kirkpatrick model [78, 68, 81, 64] (and its relatives) in a special regime called the "Nishimori Line" as a consequence of Bayes theorem [43].

With this in mind, it is natural that optimization-based procedures have been less understood: on the "Nishimori line" the corresponding spin glass model is in the so-called "replica symmetric phase". While deeply challenging, this regime is comparatively simpler to understand as the corresponding variational problems reduce to optimizing functions of one real variable [55, 9]. To understand more general optimization methods, such as maximum likelihood estimation, one must the recently developed tool, called the *method of annealing* to understand the "zero-temperature" asymptotics of spin glasses [7, 47]. Here the corresponding model enters the so-called "replica symmetry breaking" phase and can exhibit deep and nontrivial structure [22, 82, 64].

The key technical insight is that one can view pseudolikelihood inference as a "zero temperature" asymptotic of mis-matched Bayesian inference. We can then combine the recent analysis of such problems developed by one of us and co-authors in [37] with the Γ-convergence based "method of annealing" approach developed by one of us and co authors in [45]. The combination of these works is non-trivial and several new techniques were utilized. To deal with ill-scored models we generalize the universality result of [37] and remove the simplifying assumption [37, Hypothesis 2.3] (see Appendix A) and prove the analogous universality statement for pseudo maximum likelihood estimation (see Appendix B). Ill-scored models introduce an additional mean parameter that has to be controlled, so we use the techniques developed in [37] and prove a generalized variational formula for ill-scored models. We also proved new regularity results for the variational formulas with respect to general reference measures, extending the results in [12], which were previously only done for the uniform measure on $\{\pm 1\}$ (see Appendix D and Appendix E). Lastly, we extended the work of [45] to allow for random initial conditions (See Appendix C).

2. Variational Characterization for Pseudo-Likelihood Estimation

2.1. **Data model and assumptions.** Suppose that we are given data in the form of a real, symmetric $N \times N$ matrix $Y \in \mathbb{R}^{N\times N}$, whose upper entries are conditionally independent given an unknown vector $\mathbf{x}^{0,N} \in \Omega_0^N \subseteq \mathbb{R}^N$ with law

$$Y_{ij} \sim \mathbb{P}_Y(\cdot | \frac{1}{\sqrt{N}} x_i^{0,N} x_j^{0,N}) \quad \text{for } i \leq j\,, \tag{2.1}$$

for some $\lambda > 0$. Here $\Omega_0 \subseteq \mathbb{R}$ is a compact set. Our goal is to infer $\mathbf{x}^{0,N}$.

We will assume throughout that the laws of Y_{ij} are jointly absolutely continuous with respect to either Lebesgue measure on $\mathbb{R}^{N\times N}$ or a product of counting measures on Ω_0, so that the conditional densities are well-defined. In the following, we denote to the underling Lebesgue or counting measure by dy. (The meaning of notation will be clear from context.) We denote the log-likelihood of a single coordinate as $g_0(y, w)$, i.e.,

$$g_0(y, w) = \log \mathbb{P}_Y(Y = y \mid w)\,,$$

and the log-likelihood of Y given $\mathbf{x} \in \mathbb{R}^N$ is

$$\mathcal{L}_N^{g_0}(Y, \mathbf{x}) = \sum_{i \leq j} g_0\Big(Y_{ij}, \frac{x_i x_j}{\sqrt{N}}\Big)\,.$$

We will further assume that there is a null model whose likelihood we denote by $g_0(y, 0)$, and we denote the corresponding null measure by $\mathbb{P}_0$. The null model corresponds to the case $\mathbf{x}^{0,N} = 0$. Under the assumptions above a maximum likelihood estimator is defined as:

$$\hat{\mathbf{x}}_{\text{MLE}} = \arg\max_{x \in \Omega_0^N} \mathcal{L}_N^{g_0}(Y, x)\,.$$

Note that, at this level of generality, this estimator may not be uniquely defined.

We are also interested in understanding the pseudo-likelihood. Here we allow for misspecification of both the likelihood function and the support of the unknown vector Ω_0. In this case we denote the pseudo-likelihood by g and the parameter set by Ω. Throughout we shall denote our pseudo maximum likelihood estimator by $\hat{\mathbf{x}}_{\text{PMLE}}$, and it is given by:

$$\hat{\mathbf{x}}_{\text{PMLE}} := \arg\max_{x \in \Omega^N} \sum_{i \leq j} g\Big(Y_{ij}, \frac{x_i x_j}{\sqrt{N}}\Big)\,, \tag{2.2}$$

which again may not be uniquely defined. We measure the performance of the estimator by its cosine similarity with the unknown vector, that is:

$$\lim_{N\to\infty} \mathbf{CS}(\hat{\mathbf{x}}, \mathbf{x}^0) \qquad \text{where} \qquad \mathbf{CS}(x,y) := \frac{x\cdot y}{\|x\|\,\|y\|}\,, \tag{2.3}$$

and its squared norm.

In order to develop a high-dimensional theory, we need certain basic assumptions on the data distribution, and the unknown vector. Note that as Ω_0 is compact, for any sequence $\mathbf{x}^{0,N}$, the sequence of empirical measures

$$\mu_{\mathbf{x}^{0,N}} = \frac{1}{N}\sum_i \delta_{x_i^{0,N}},$$

is always tight.

Definition 2.1. We say that a sequence $\mathbf{x}^N \in \Omega_0^N$ is *tame* if $\mu_{\mathbf{x}^{0,N}} \to \mathbb{Q}$ weakly for some probability measure $\mathbb{Q}$.

We work under the assumption that $\mathbf{x}^{0,N}$ is tame. This assumption is common in the high-dimensional statistics literature (see, e.g., [27, 29, 71]). Next we need some basic regularity assumptions on the (pseudo)-likelihood. To this end, we need the following function class

Definition 2.2. Let $\mathcal{F}_0(dy)$ denote the set of pairs of functions, $(f_1(y,w), f_2(y,w))$, with common domain $\mathbb{R}\times U \subseteq \mathbb{R}^2$, where U is an open neighborhood of 0, that are three times continuously differentiable in w for every y and satisfy the following four conditions:

$$\begin{aligned} \int_{\Omega_0} \exp(f_i(y,0))dy, \qquad & \int_{\Omega_0} [|\partial_w f_i(y,0)|^4] \exp(f_1(y,0))dy, \\ \|\partial_w^2 f_i(\cdot,0)\|_\infty, \qquad & \|\partial_w^3 f_i(\cdot,\cdot)\|_\infty < \infty. \end{aligned} \tag{2.4}$$

for each $i = 1,2$.

($\mathcal{F}_0$ in principle depends on the choice of U, whose choice is problem dependent. We suppress this dependence for the sake of exposition.) We will assume throughout the following that the pair of the likelihood, g_0, and pseudo-likelihood g are in this class, $(g_0, g) \in \mathcal{F}_0(dy)$. Since the pair (g_0, g) completely specify the underlying inference problem, we refer to this pair as an *inference task*.

2.2. **Information parameters.** One of our central results is that the performance of the (pseudo)-likelihood estimator in problems of this class is *entirely* determined by the *information parameters* of the pair (g_0, g), which are defined as follows

Definition 2.3. The *information parameters* of an inference task (g_0, g) are

$$\beta_1(g_0,g) = \mathbb{E}_{\mathbb{P}_0}\left[(\partial_w g(Y,0) - \mathbb{E}_0[\partial_w g(Y,0)])^2\right] \tag{2.5}$$

$$\beta_2(g_0,g) = \mathbb{E}_{\mathbb{P}_0}\left[\partial_w g(Y,0)\partial_w g_0(Y,0)\right] \tag{2.6}$$

$$\beta_3(g_0,g) = -\mathbb{E}_{\mathbb{P}_0}\left[\partial_w^2 g(Y,0)\right] \tag{2.7}$$

$$\beta_4 = \mathbb{E}_{\mathbb{P}_0}\left[\partial_w g(Y,0)\right]. \tag{2.8}$$

The information parameters measure the effect of the misspecification on the null model. Observe that when $g = g_0$, if we denote the null Fisher information by

$$\beta_* = \mathbb{E}_{\mathbb{P}_0}\left[(\partial_w g_0(Y,0))^2\right], \tag{2.9}$$

then by standard properties of score functions [77, Chapter 2.3], the information parameters satisfy the *Rao relation*

$$\beta_* = \beta_1 = \beta_2 = \beta_3.$$

2.3. **Well-scored v.s. ill-scored psuedolikelihoods.** A classical fact is that the expected score of g_0 is 0. As we are allowing for the case that $g \neq g_0$, however, this identity may no longer hold. As we shall see below, the failure of this identity has substantial repercussions for inference. To this end, it helps to introduce the following criterion.

Definition 2.4 (well-scored pseudo-likelihood)**.** We say that a pseudo-likelihood function $g(y, w)$ is *well-scored* if its score satisfies

$$\beta_4 = \mathbb{E}_{\mathbb{P}_0}\big[\partial_w g(Y, 0)\big] = 0. \tag{2.10}$$

Otherwise we call it *ill-scored*.

The case of well-scored models represents an ideal case for pseudo-maximum likelihood theory. On the contrary, in the ill-scored setting, the pseudo-maximum likelihood is heavily influenced by the sign of the score parameter, β_4, and can lead to complete failure of pseudo maximum likelihood estimation (see Section 2.6).

2.4. **Variational characterization of performance for well-scored PMLEs (and MLEs).** We are now in the position to state our main results. We begin by discussing the case of well-scored models. Our first main result is a variational formula for the asymptotic pseudo-maximum likelihood and corresponding characterization of the asymptotic performance of pseudo-maximum likelihood estimators.

To this end, we need to define a corresponding Parisi-type functional. Let $\mathcal{M}([0, S])$ denote the space of non-negative, finite measures on $[0, S]$ equipped with the weak-* topology, and let $\mathcal{A}_S \subseteq \mathcal{M}([0, S])$ be the subset

$$\mathcal{A}_S = \{\nu \in \mathcal{M}([0, S]) : m(s)ds + c\delta_S, m(s) \geq 0 \text{ non-decreasing}\}.$$

For each $\gamma = mds + c\delta_S \in \mathcal{A}_S$, let $\Phi_{\gamma,\lambda,\mu}(t, y)$ denote the weak solution to the Hamilton-Jacobi-Bellman equation,

$$\begin{cases} \partial_t \Phi_{\gamma,\lambda,\mu} = -\frac{\beta_1^2}{4}\Big[\partial_y^2 \Phi_{\gamma,\lambda,\mu} + m(t)(\partial_y \Phi_{\gamma,\lambda,\mu})^2\Big] & (t, y) \in [0, S] \times \mathbb{R} \\ \Phi_{\gamma,\lambda,\mu}(S, y) = \max_{x\in\Omega}\left(yx + \lambda x x^0 + (\mu + \frac{\beta_1^2}{2}c)x^2\right) & y \in \mathbb{R} \end{cases}$$

For the notion of weak solution for partial differential equations (PDEs) of this type see, e.g., [46] and for the existence, uniqueness and regularity of weak solutions to this PDE see [45, Appendix A].

Let us now define the functional $\psi_{\bar{\beta}}$ which will characterize the maximum of the asymptotic pseudo-likeihood when restricted to parameters with a prescribed cosine similarity, M and squared norm, S.

$$\psi_\beta(S, M) = \inf_{\mu,\lambda}\inf_{\gamma}\left(\mathbb{E}_{\mathbb{Q}}[\Phi_{\lambda,\mu,\gamma}(0, 0)] - \frac{\beta_1^2}{2}\int_0^S t d\gamma(t) - \mu S - \lambda M + \frac{\beta_2 M^2}{2} - \frac{\beta_3 S^2}{4}\right). \tag{2.11}$$

Observe that ψ_β is well-defined on $[0, (\max\Omega)^2] \times [\min\Omega, \max\Omega]$ and upper semicontinuous C.2 there, though it may take the value $-\infty$. The (effective) domain of ψ_β is the set

$$\mathcal{C} = \cap_{\rho,\tau\in[-1,1]^2}\{(S, M) : \mathbb{E}_{x^0\sim\mathbb{Q}}[\inf_{x\in\Omega}\{\rho x^2 + \tau x x^0\}] \leq \rho S + \tau M \leq \mathbb{E}_{x^0\sim\mathbb{Q}}[\sup_{x\in\Omega}\{\rho x^2 + \tau x x^0\}]\}. \tag{2.12}$$

Observe that the set $\mathcal{C}$ is convex and compact and depends implicitly on $\mathbb{Q}$. Let $\mathcal{C}_\beta$ denote the set of maximizers of ψ_β over $\mathcal{C}$, that is

$$\mathcal{C}_\beta = \operatorname*{argmax}_{(S,M)\in\mathcal{C}} \psi_\beta(S, M). \tag{2.13}$$

Finally, let $M_N(\mathbf{x})$ and $S_N(\mathbf{x})$ be

$$M_N(\mathbf{x}) = \frac{1}{N}\mathbf{x}\cdot\mathbf{x}^{0,N} \qquad \text{and} \qquad S_N(\mathbf{x}) = \frac{1}{N}\|\mathbf{x}\|^2$$

Observe that $\mathbf{CS}(\mathbf{x},\mathbf{x}^{0,N}) = M_N(\mathbf{x})/\sqrt{S_N(\mathbf{x})S_N(\mathbf{x}^{0,N})}$, and under the assumption that $\mathbf{x}^{0,N}$ is tame, that $S_N(\mathbf{x}^{0,N}) \to \mathbb{E}_{\mathbb{Q}}[x_0^2]$. With this in hand, we can now state our main technical result.

Theorem 2.1. *Suppose that $\mathbf{x}^{0,N}$ is tame and that g is a well-scored pseudo-likelihood. The maximum pseudo-likelihood satisfies*

$$\frac{1}{N}\Big[\max_{x\in\Omega^N}\mathcal{L}_N^g(Y,x) - \sum_{i\le j} g(Y_{ij},0)\Big] \to \sup_{(S,M)\in\mathcal{C}} \psi_{\bar\beta}(S,M) \qquad a.s. \tag{2.14}$$

Furthermore, for any sequence of choices of $\hat{\mathbf{x}}^N_{\text{PMLE}}$, the corresponding sequence of overlaps

$$(S_N(\hat{\mathbf{x}}_{\text{PMLE}}), M_N(\hat{\mathbf{x}}_{\text{PMLE}}))\,,$$

is tight, with limit points contained in $\mathcal{C}_\beta$.

To better interpret Theorem 2.1, it helps to observe that ψ_β and $\mathcal{C}_\beta$ have an intrinsic statistical meaning. ψ_β is the maximum of the pseudolikelihood when restrcited to that set of overlaps and squared norms and $\mathcal{C}_\beta$ is the set of overlaps and norms of near maxima of the (normalized) pseudo likelihood. To make this precise, let $\Omega^N_\varepsilon(S,M) = \{x\in\Omega^N : |M_N(x)-M|\le\varepsilon, |S_N(x)-S|\le\varepsilon\}$

$$\mathcal{L}_N^{g;\varepsilon}(S,M) = \max_{x\in\Omega^N_\varepsilon(S,M)}\mathcal{L}_N^g(Y,x).$$

We then have the following.

Theorem 2.2. *For every $(S,M)\in\mathcal{C}$, we have that*

$$\lim_{\varepsilon\to0}\lim_{N\to\infty}\frac{1}{N}\Big[\mathcal{L}_N^{g;\varepsilon}(S,M) - \sum_{i\le j} g(Y_{ij},0)\Big] = \psi_\beta(S,M)$$

almost surely.

Observe that in the above, we do not guarantee the convergence of cosine similarity. This is because, in some settings, $\mathcal{C}_\beta$ may contain several points. This is due to the existence of many near maximizers of the pseudo MLE. It is natural to ask under which regimes one has true convergence. A sufficient condition is if $\mathcal{C}_\beta$ consists of at most two points.

Assumption 1. Suppose that $\bar\beta$ is such that $\mathcal{C}_{\bar\beta}$ consists of at most two points. Furthermore, the coordinate associated with the m parameter is unique up to a sign.

Corollary 2.1. *Suppose that $\bar\beta$ satisfies Assumption 1, then for any sequence of pseudo-likelihood estimators, $(S_N(\hat{\mathbf{x}}^N_{\text{PMLE}}), |M_N(\hat{\mathbf{x}}^N_{\text{PMLE}})|) \to (s,m)$ almost surely. In particular, the absolute cosine similarity converges almost surely to $m/\sqrt{s\mathbb{E}_{\mathbb{Q}}x_0^2}$*

We note here the following remark regarding the centreing in (2.14).

Remark 2.1. The term $\sum_{i\le j} g(Y,0)$ does not depend on x, so it will not affect the pseudo-maximum likelihood estimator. However, these normalization terms need to be subtracted off for the PMLE to have a well-defined limit. For example, with data $\boldsymbol{Y}\sim\boldsymbol{G}+\frac{\mathbf{x}\mathbf{x}^\intercal}{\sqrt{N}}$ the likelihood,

$$\frac{1}{N}\mathcal{L}_N^{g_0}(Y,x) = \frac{1}{N}\sum_{ij} -\frac{1}{2}G_{ij}^2 + \frac{x_i^0x_j^0x_ix_j}{N} - \frac{x_i^2x_j^2}{2N}$$

diverges because $\frac{1}{N}\sum_{ij} g_0(Y,0) = \frac{1}{N}\sum_{ij}\frac{1}{2}G_{ij}^2$.

2.5. **Behavior of Ill-scored pseudolikelihoods.** Before turning to our variational characterization in the case of ill-scored models, let us pause here for a discussion of the key issue in this setting. Suppose that

$$\beta_4 = \mathbb{E}_{\mathbb{P}_0}[\partial_w g(Y,0)] = C \neq 0\,.$$

When $\beta_4 \neq 0$, the leading order behaviour of the pseudo-likelihood is dominated by the empirical mean of the parameter. Roughly speaking, in this regime one has the expansion

$$\max_{x\in\Omega_N} \mathcal{L}_N^g(Y,x) = N^{3/2} C \max_{x\in\Omega_N} \left(\frac{1}{N}\sum_{i=1}^N x_i\right)^2 + o\Big(N^{3/2}\Big).$$

Note that the leading order term here does *not* depend on the unknown parameter, x_0, and is an order of magnitude larger than in the well-scored setting (c.f. Example 2.1 and Lemma 6.1).

This has catastrophic consequences on inference which are best illustrated by way of example.

Example 2.1. Consider the data matrix generated from a spiked Gaussian matrix with non-zero mean in the null model,

$$Y_{ij} \sim \mathcal{N}\bigg(C + \frac{x_i^0 x_j^0}{\sqrt{N}}, 1\bigg).$$

To infer $\mathbf{x}_0$, we take a irregular psudeo likelihood from a centered Gaussian likelihood,

$$g(y,w) = -\frac{1}{2}(y-w)^2,$$

which represents a large misspecification of an order 1 parameter of the data model. It follows that

$$\beta_4 = \mathbb{E}_{\mathbb{P}_0}[\partial_w g(Y,0)] = \mathbb{E}_{\mathbb{P}_0}[Y] = C.$$

When $C > 0$, then the data distribution Y has a large positive eigenvalue of order N while the positive eigenvalue of the spike we want to infer $\frac{x_i^0 x_j^0}{\sqrt{N}}$ is of order $\sqrt{N}$. Conversely, $C < 0$, then the data distribution Y has a large negative eigenvalue of order N while the positive eigenvalue of the spike we want to infer $\frac{x_i^0 x_j^0}{\sqrt{N}}$ is of order $\sqrt{N}$. In either case, there is large but spurious shift in the likelihood that obscures the parameter we want to infer.

In the following sections, we begin by first stating the variational formula in the case of ill-scored pseudolikelihoods. Importantly, however, the statistician does not, a priori, have access to the underlingly null distribution $\mathbb{P}_0$. As such it is important to understand whether or not it is possible to determine if one is in the ill-scored scenario and, in particular, if it is possible to systematically correct for this effect. We present such an approach in the subsequent section.

2.6. **Variational Formula for Ill-scored pseudolikelihood.** We now state an extension of Theorem 2.1 to these ill-scored models. Due to the importance of the sign of β_4, the results will be separated into cases. We begin with the case of $\beta_4 > 0$.

Theorem 2.3 (Positive β_4)**.** *Suppose that $\mathbf{x}^0$ is tame and $\beta_4 > 0$. Let*

$$\tilde{x}_+ = \sup \Omega \qquad \tilde{x}_- = \inf \Omega$$

denote the respective largest point and smallest points in our parameter space. Let

$$x_+ = \begin{cases} \tilde{x}_+ & \text{if } |\tilde{x}_+| \geq |\tilde{x}_-| \\ \tilde{x}_- & \text{if } |\tilde{x}_+| < |\tilde{x}_-| \end{cases}$$

We have $\hat{\mathbf{x}}_{\mathrm{PMLE}} = x_+\mathbf{1}$ is a constant vector, so the set of limit points of $S_N(\hat{x}_{PMLE})$ and $M_N(\hat{x}_{\mathrm{PMLE}})$ is unique and given by $\mathcal{C}_{\bar{\beta}} = \{(x_+^2, x_+\mathbb{E}_{\mathbb{Q}}(x^0))\}$. In particular,

$$\mathbf{CS}(\hat{\mathbf{x}}_{\mathrm{PMLE}}, \mathbf{x}_0) \to \frac{\mathbb{E}_{\mathbb{Q}}(x^0)}{(\mathbb{E}_{\mathbb{Q}}(x^0)^2)^{1/2}}$$

$$\frac{1}{N}\|\hat{x}_{PMLE}\|^2 \to x_+^2.$$

Evidently if $\bar{x}_0 = 0$ then $\mathbf{CS}(\hat{\mathbf{x}}_{\text{PMLE}}, \mathbf{x}_0) = 0$ and the estimator is useless.

The case when $\beta_4 < 0$ is more delicate since there is the large spurious information induced by the misspecification is, in some sense, in the *opposite* direction of the vector we want to infer.

We have the following formula for the restricted ground state. Let

$$\begin{cases} \partial_t \Phi_{\gamma,\lambda,\mu,\rho} = -\frac{\beta_1^2}{4}\Big[\partial_y^2 \Phi_{\gamma,\lambda,\mu,\rho} + m(t)(\partial_y \Phi_{\gamma,\lambda,\mu,\rho})^2\Big] & (t,y) \in [0,S] \times \mathbb{R} \\ \Phi_{\gamma,\lambda,\mu,\rho}(S,y) = \max_{x\in\Omega}\left(yx + \lambda x x^0 + (\mu + \frac{\beta_1^2}{2}c)x^2 + \rho x\right) & y \in \mathbb{R} \end{cases}$$

and using a slight abuse of notation, we define

$$\psi_{\bar{\beta},-}(S,M,v) = \inf_{\mu,\lambda,\rho}\inf_{\gamma}\left(\mathbb{E}_{\mathbb{Q}}[\Phi_{\lambda,\mu,\gamma,\rho}(0,0)] - \frac{\beta_1^2}{2}\int_0^S t d\gamma(t) - \mu S - \lambda M - \rho v + \frac{\beta_2 M^2}{2} - \frac{\beta_3 S^2}{4}\right). \tag{2.15}$$

Notice that the $\psi_{\bar{\beta},-}$ defined here differs from (2.11) by an extra Lagrange multiplier term ρ. If it is clear from context which scenario we are in, we will sometimes exclude the $-$ in the subscript.

The domain of $\psi_{\bar{\beta},-}$ is the set

$$\mathcal{C}_- = \cap_{\rho,\tau,\eta\in[-1,1]^3}\{(S,M,v) : \mathbb{E}_{x^0\sim\mathbb{Q}}[\inf_{x\in\Omega}\{\rho x^2 + \tau x x^0 + x\eta\}] \le \rho S + \tau M + \eta v \le \mathbb{E}_{x^0\sim\mathbb{Q}}[\sup_{x\in\Omega}\{\rho x^2 + \tau x x^0 + x\eta\}]\}. \tag{2.16}$$

Furthermore, in the context of illscored models, we let $\mathcal{C}_\beta$ denote the set of maximizers of ψ_β given in (2.15) over the set $\mathcal{C}$ defined in (2.16) subject to a constraint on the third coordinate, that is

$$\mathcal{C}_{\bar{\beta}} = \operatorname*{argmax}_{(S,M,v)\in\mathcal{C}_-: v = x_-} \psi_{\beta,-}(S,M,x_-). \tag{2.17}$$

We use the same notation as the previous set of maximizers defined in (2.13), but it is understood that the tuple $\bar{\beta}$ has a negative fourth coordinate in (2.17), while in (2.13), the fourth coordinate is 0.

Theorem 2.4 (Negative β_4)**.** *Suppose that $\mathbf{x}^{0,N}$ is tame and $\beta_4 < 0$. Let $x_- = \min \operatorname{conv}(\Omega)$ denote the point in the convex hull of the parameter space closest to the origin. The maximum pseudo-likelihood satisfies*

$$\max_{x\in\Omega^N}\frac{1}{N}\left(\mathcal{L}_N^g(Y,x) - \sum_{i\le j} g(Y_{ij},0)\right) - \sqrt{N}(x_-)^2\beta_4 \to \sup_{(S,M)\in\mathcal{C}} \psi_{\bar{\beta},-}(S,M,x_-) \tag{2.18}$$

almost surely. Furthermore, for any sequence of choices of $\hat{\mathbf{x}}^N_{\text{PMLE}}$, the corresponding sequence $(S_N(\hat{\mathbf{x}}_{\text{PMLE}}), M_N(\hat{\mathbf{x}}_{\text{PMLE}}))$ is tight with limit points contained in $\mathcal{C}_{\bar{\beta}}$.

Similarly to the case with positive score, if $\bar{x}_0 = c\mathbf{1}$ and $x_- = 0$ then $\mathbf{CS}(\hat{\mathbf{x}}_{\text{PMLE}}, \mathbf{x}_0) = 0$ and the estimator is useless.

2.7. **The score-corrected pseudolikelihood.** As seen in the previous section, ill-scored pseudolikelihoods have behaviour dictated by the sign of β_4 which introduces a very large uniformative "spike" in the models, which can lead to a complete failure of the inferential procedure in certain scenarios. To this end, we propose a correction to the pseudo-likelihood estimator that resolves this issue.

A natural way to deal with the high order term when $\beta_4 \neq 0$ is to introduce an additional term to the pseudo-likelihood to centre the corresponding score by using an estimate of the score parameter,

β_4. A priori, the statistician does not have access to $\mathbb{P}_0$. That said, for any pseudo-likelihood, g, we can consider the estimator,

$$\hat{\beta}_4 = \frac{1}{N^2} \sum_{i \le j} \partial_w g(Y_{ij}, 0).$$

If we let $\bar{\mathbf{x}}_{\mathbf{0}} = \frac{1}{N} \sum x_i^0$, then by the law of large numbers, this quantity will concentrate around its expected value

$$\begin{aligned} \hat{\beta}_4 = \frac{1}{N^2} \sum_{i \le j} \partial_w g(Y_{ij}, 0) \approx \mathbb{E}_Y[\partial_w g(Y_{ij}, 0)] &= \mathbb{E}_{\mathbb{P}_0}[\partial_w g(Y_{ij}, 0)] + \frac{1}{N^2} \beta_2 \sum_{i \le j} \frac{x_i^0 x_j^0}{\sqrt{N}} + O(N^{-1}) \\ &= \beta_4 + \beta_2 \frac{\bar{\mathbf{x}}_{\mathbf{0}}^2}{\sqrt{N}} + O(N^{-1}) \end{aligned}$$

where the lower order terms $\beta_2 \frac{\bar{\mathbf{x}}_{\mathbf{0}}^2}{\sqrt{N}}$ come from the fact that we can estimate $\mathbb{E}_Y[\partial_w g(Y_{ij}, 0)]$ using the data distribution. (See Lemma F.2 for a precise statement.) While nominally, the second term is a lower order effect, this lower order term will have a nontrivial contribution when multiplied by $N^{3/2}$, i.e., the appropriate power of N to counteract the expected score. Thus $\mathbb{E}_{\mathbb{P}_0}[\partial_w g(Y_{ij}, 0)]$ unfortunately remains inaccessible.

To account for this, let us introduce a hyper-parameter $\alpha \in \mathbb{R}$ and define the corresponding score-corrected pseudo-likelihood by

$$\mathcal{L}^g_{N,\alpha}(Y, x) = \sum_{i \le j} g\Big(Y_{ij}, \frac{\lambda x_i x_j}{\sqrt{N}}\Big) - N^{\frac{3}{2}} \hat{\beta}_4 \bar{x}^2 + N\alpha \bar{x}^2. \tag{2.19}$$

The centering by $-N^{\frac{3}{2}} \hat{\beta}_4 (\bar{x})^2$ kills of the large effect induced by score parameter, while α is a ridge correction term to offset the lower order terms in the score approximation $\hat{\beta}_4$. If $\alpha = \beta_2 \mathbb{E}_{\mathbb{Q}}[x_0]^2$ then pseudo maximum likelihood estimation on the score-corrected likelihood is equivalent to optimizing the pseudo-likelihood

$$\mathcal{L}^g_{N,\alpha}(Y, x) = \sum_{i \le j} g\Big(Y_{ij}, \frac{\lambda x_i x_j}{\sqrt{N}}\Big) - N^{\frac{3}{2}} \beta_4 \bar{x}^2.$$

Remark 2.2. One might also consider a slightly generalized version of the score corrected pseudo likelihood,

$$\mathcal{L}^g_{N,\gamma}(Y, x) = \sum_{i \le j} g\Big(Y_{ij}, \frac{\lambda x_i x_j}{\sqrt{N}}\Big) - N^{\frac{3}{2}} \gamma \bar{x}^2.$$

If we take $\gamma = \beta_4$, then this will also remove the adverse effect caused by non-zero score. However, the scaling of the correction term is order $N^{3/2}$, so that γ must be calibrated to within $o(N^{\frac{1}{2}})$ of β_4 to avoid introducing lower order corrections.

We have the following variational formula for the score-corrected pseudo-maximum likleihood. Let

$$\psi_{\beta,\alpha}(S, M, v) = \psi_{\beta,-}(S, M, v) - \frac{\beta_2 [\mathbb{E}_{\mathbb{Q}} x_0]^2 v^2}{2} + \frac{\alpha v^2}{2}. \tag{2.20}$$

Note that the information parameters $\beta_1, \beta_2, \beta_3$ are defined with respect to g and not g_c. The domain of this function is $\mathcal{C}_-$ defined in (2.16). Let

$$\mathcal{C}_{\beta,\alpha} = \operatorname*{argmax}_{(S,M,v) \in \mathcal{C}} \psi_{\beta,\alpha}(S, M, v).$$

Theorem 2.5. *Suppose that* $\mathbf{x}^0$ *is tame. The maximum pseudo-likelihood satisfies*

$$\max_{x\in\Omega^N}(\mathcal{L}^g_{N,\alpha}(Y,x)-\sum_{i\le j}g(Y_{ij},0))\to\sup_{(S,M,v)}\psi_{\bar\beta,\alpha}(S,M,v) \tag{2.21}$$

almost surely. Furthermore, for any sequence of choices of $\hat{\mathbf{x}}^N_{\mathrm{PMLE}}$*, the corresponding sequence* $(S_N(\hat{\mathbf{x}}_{\mathrm{PMLE}}),M_N(\hat{\mathbf{x}}_{\mathrm{PMLE}}))$ *is tight with limit points contained in* $\mathcal{C}_{\beta,\alpha}$*.*

Remark 2.3. If $\alpha=\beta_2\mathbb{E}_{\mathbb{Q}}[x_0]^2$, then the variational formula is equivalent to a regular model with information parameters β_1,β_2,β_3.

3. Strong and coarse equivalence of inference tasks and a universal task

It is natural to ask if two pseudo likelihoods lead to estimators that are, from a statistical perspective, equivalent. For example, in the spiked matrix model, while the top eigenvector obtains a nontrivial cosine similarity with the ground truth, any other unit vector with the same cosine similarity has the same performance with respect to the underlying statistical task. It turns out that our results lead to an even deeper notion of equivalence between pseudolikelihood estimation *problems*. For example, there is a precise sense in which maximum likelihood estimation of the "spike" in spiked matrix models is "equivalent" to maximum likelihood estimation of the communities in stochastic block models!

Our first notion of equivalence is *strong equivalence*.

Definition 3.1. We say that two inference tasks are *strongly equivalent* if they have the same information parameters.

Evidently strong equivalence is an equivalence relation. Furthermore, there is a natural universal statistical task corresponding to given information parameters which is defined as follows.

Let g^1 denote a pseudo likelihood whose information parameters with respect to g_0 are given by $\bar\beta=(\beta_1,\beta_2,\beta_3,\beta_4)$. We consider the corresponding inference task with likelihoods given by

$$g^{\bar\beta}_{U,0}(y,w)=-\frac{1}{2\beta_1}(y-\beta_2w-\beta_4)^2-\frac{1}{2}\log(2\pi\beta_1)\,, \tag{3.1}$$

$$g^{\bar\beta}_{U,1}(y,w)=-\frac{1}{2}(y-w)^2-\frac{\beta_3-1}{2}w^2\,, \tag{3.2}$$

which corresponds to least squares estimation with a *correction*. The universal statistical corresponds to estimating the spike in the matrix

$$Y=G+\frac{\beta_2}{\sqrt{N}}x^0(x^0)^T\,,$$

where G has i.i.d $\mathcal{N}(\beta_4,\beta_1)$ entries, via the pseudo-likelihood $g^{\bar\beta}_{U,1}$.

Theorem 3.1. *Any inference task* (g_0,g_1) *with information parameters given by* $\bar\beta$ *is strongly equivalent to the inference task* $(g^{\bar\beta}_{U,0},g^{\bar\beta}_{U,1})$*.*

Remark 3.1. Theorem 3.1 simplifies greatly in the well scored case with $\beta_3>0$. In this case $g^{\bar\beta}_{U,1}$, may instead be taken to be $-\frac{1}{2}(y-\sqrt{\beta_3}w)^2$, with an appropriate normalization in $g^{\bar\beta}_{U,0}$.

An important consequence of our work is that there is in fact a substantially weaker notion of equivalence that captures the underlying statistical task. Recalling the statistical interpretation of $\mathcal{C}_\beta$ from Theorem 2.1, 2.3, 2.4, 2.5 as the set of near optimal overlaps of estimators in the respective problems, we are led to the following natural notion.

Definition 3.2. We say that two inference tasks (g_0^1, g^1) and (g_0^2, g^2) are *coarsely equivalent* if $\mathcal{C}_{\bar\beta^1} = \mathcal{C}_{\bar\beta^2}$ where $\bar\beta^i = \bar\beta(g_0^i, g^i)$ for $i = 1, 2$ are their corresponding information parameters.

Notice that if g_1 and g_2 have the same information parameters with respect to g_0^1, g_0^2 then they are coarsely equivalent. More generally, one has the following sufficient conditions for coarse equivalence of well-scored pseudolikelihoods.

Theorem 3.2. *Consider two well-scored inference tasks (g_0^1, g^1) and (g_0^2, g^2), with information parameters $\bar\beta(g^1) = \bar\beta(g_0^1, g^1)$ and $\bar\beta(g^2) = \bar\beta(g_0^2, g^2)$. Suppose that at least one of the following conditions are true*

- *The ratio of all the information parameters is constant*

$$\frac{\beta_1(g^1)}{\beta_1(g^2)} = \frac{\beta_2(g^1)}{\beta_2(g^2)} = \frac{\beta_3(g^1)}{\beta_3(g^2)}, \tag{3.3}$$

- *There exists a constant C such that the parameter space Ω satisfies $|\boldsymbol{x}| = C$ for every $\boldsymbol{x} \in \Omega$ and the first ratio of the two information parameters are equal*

$$\frac{\beta_1(g^1)}{\beta_1(g^2)} = \frac{\beta_2(g^1)}{\beta_2(g^2)}, \tag{3.4}$$

then (g_0^1, g^1) and (g_0^2, g^2) are coarsely equivalent.

In the case of ill-scored pseudolikelihoods one must further include a condition on the correction parameters used:

Theorem 3.3. *Consider two ill-scored inference tasks (g_0^1, g^1) and (g_0^2, g^2) with information parameters $\bar\beta(g^1) = \bar\beta(g_0^1, g^1)$ and $\bar\beta(g^2) = \bar\beta(g_0, g^2)$, and let α^1 and α^2 be the correction parameters for g^1 and g^2 respectively. Suppose that at least one of the following conditions are true*

- *The ratio of all the information and correction parameters are constant*

$$\frac{\beta_1(g^1)}{\beta_1(g^2)} = \frac{\beta_2(g^1)}{\beta_2(g^2)} = \frac{\beta_3(g^1)}{\beta_3(g^2)} = \frac{\beta_4(g^1)}{\beta_4(g^2)} = \frac{\alpha^1}{\alpha^2}, \tag{3.5}$$

- *There exists a constant C such that the parameter space Ω satisfies $|\boldsymbol{x}| = C$ for every $\boldsymbol{x} \in \Omega$ and the first ratio of the the information and correction parameters are equal*

$$\frac{\beta_1(g^1)}{\beta_1(g^2)} = \frac{\beta_2(g^1)}{\beta_2(g^2)} = \frac{\beta_4(g^1)}{\beta_4(g^2)} = \frac{\alpha^1}{\alpha^2}, \tag{3.6}$$

then (g_0^1, g_1) and (g_0^2, g_2) with correction parameters α^1 and α^2 are coarsely equivalent.

Remark 3.2. While Theorem 3.1 guarantees a measure of the performance of the pseudo likelihood g_1 in terms of the performance of a least squares problem, it does not mean that the performance of g_1 is equivalent to the performance of least-squares for the initial matrix Y. In fact, we show in Section 4 that the least square estimator can be completely uninformative regardless of the SNR used.

4. Application to Gaussian Pseudolikelihoods (a.k.a. the best rank 1 approximation)

A popular approach to tackling rank one estimation problems is to consider the *best rank 1 approximation.* That is, consider a vector, $x \in \Omega^N$, such that

$$\hat{\mathbf{x}}_{LS}(\lambda) = \arg\min_{x \in \Omega^N} \frac{1}{2} \left\| \boldsymbol{Y} - \frac{\lambda \mathbf{x}\mathbf{x}^\intercal}{\sqrt{N}} \right\|_F^2,$$

where $\lambda > 0$ is a scale hyper-parameter. Observe that this corresponds to pseudolikelihood estimation with a Gaussian likelihood $g(y, w) = -\frac{1}{2\sigma^2}(y - \lambda w)^2$ where $\sigma^2 = \mathbb{E}_0[Y^2]$. Let

$$\beta_{LS} = \frac{\mathbb{E}[Y \partial_w g_0(Y, 0)]}{\sigma}.$$

We then have the following.

Proposition 4.1. *The pair (g, g_0) has information parameters $(\lambda^2, \lambda\beta_{LS}, \lambda^2, \lambda\mathbb{E}_0 Y)$. In particular g is well-scored if and only if $\mathbb{E}_0 Y = 0$.*

Let us pause to consider the case that g is well-scored and $\beta_{LS} > 0$. In this case Theorem 2.1 applies. In particular, the corresponding overlap and squared norm have limit points lying in $\mathcal{C}_\beta$. If $\lambda = \sigma\sqrt{\beta_{LS}}$ then the information parameters satisfy the Rao relation. Note that by Cauchy-Schwarz, $\beta_{LS} \leq \sqrt{\beta_0}$. If, furthermore, $\beta_{LS} = \sqrt{\beta_0}$ then information parameters are equal to those of the log-likelihood.

In practice, we do not necessarily know that the data distribution under the null model has zero mean, and, as shown in Example 2.1 above, a seemingly innocuous misspecification can lead to substantial effects on inference. In order to counteract these potential effects, we introduce a score-corrected best rank 1 approximation (as in Section 2.7) by subtracting off the mean of the data distribution and adding a ridge term.
Let

$$\bar{\boldsymbol{Y}} = \frac{1}{N^2} \sum_{i \leq j} Y_{ij}$$

and consider the *score-corrected least squares estimator*

$$\hat{\mathbf{x}}_{LS,\alpha}(\lambda, \alpha) = \arg \min_{x \in \Omega^N} \frac{1}{2} \left\| \boldsymbol{Y} - \frac{\lambda \mathbf{x}\mathbf{x}^\intercal}{\sqrt{N}} \right\|_F^2 + N^{3/2} \bar{\boldsymbol{Y}} \bar{x}^2 - N\alpha\bar{x}^2,$$

where $\lambda > 0$ is a scale parameter. To offset the correction term, we set

$$\alpha = \beta_2 \mathbb{E}_{\mathbb{Q}}[x_0]^2 = \lambda\beta_{LS}\mathbb{E}_{\mathbb{Q}}[x_0]^2 \,,$$

and then Theorem 2.5 applies.

It is interesting to note that the above gives the following important negative result.

Proposition 4.2. *If $\beta_{LS} = 0$ and $\mathbf{x}^{0,N}$ is tame with $\mathbb{E}_{\mathbb{Q}} x^0 = 0$, then*

$$\mathbf{CS}(\hat{\mathbf{x}}_{\text{PMLE}}, \mathbf{x}) \to 0 \qquad a.s.$$

This occurs, for example, in the case of sparse Rademacher matrices. See Section 5.3 and Table 1 below.

5. Examples

In this section, we outline several explicit models which fall into our framework. We summarize some of the models we consider, and their corresponding likelihoods and information parameters in the table below.

5.1. **Spiked Matrices and $\mathbb{Z}_2$ synchronization.**

Suppose that we want to recover an unknown vector $\mathbf{x}^0$ that has been corrupted with additive Gaussian noise $G_{ij} \sim N(0, 1)$ at signal to noise ratio λ_0. That is,

$$Y = G + \frac{\lambda_0}{\sqrt{N}} x_0 x_0^\intercal \,.$$

In the case that $\boldsymbol{x}^0$ has $\{\pm 1\}$ valued entries, this is known as the $\mathbb{Z}_2$ synchronization problem. This special case has been studied extensively (see [57], [71], [62], [18], [11])

Model Type	Likelihood	β_1	β_2	β_3	β_4
Spiked Wigner with SNR λ_0	$-\frac{1}{2}(y-\lambda w+C)^2-\frac{1}{2}\log(2\pi)$	λ^2	$\lambda\lambda_0$	λ^2	C
Community Detection $(\frac{1}{2},\mu_0)$	$y\log(\frac{1}{2}+\mu w)+(1-y)\log(\frac{1}{2}-\mu w)$	$4\mu^2$	$4\mu\mu_0$	$4\mu^2$	0
Sparse Rademacher	$-\frac{1}{2}(y-\lambda w)^2-\frac{1}{2}\log(2\pi)$	λp	0	λ^2	0
Signs of Spiked Wigner Matrix with SNR λ_0	$\frac{(1-y)}{2}\log\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-\lambda w}e^{-\frac{x^2}{2}}dx$ $+\frac{(1+y)}{2}\log\frac{1}{\sqrt{2\pi}}\int_{-\lambda w}^{\infty}e^{-x^2/2}dx$	$\frac{2}{\pi}\lambda^2$	$\frac{2}{\pi}\lambda^2$	$\frac{2}{\pi}\lambda^2$	0

Table 1. This table lists the information parameters of several well studied inference problems that fall under our framework. These examples are described in detail in Section 5.

In this case, the log likelihood of any coordinate is given by:

$$g_0(y,w)=-\frac{1}{2}(y-\lambda_0 w)^2-\frac{1}{2}\log(2\pi)\,.$$

Suppose that we have misspecified the signal to noise ratio $\lambda\neq\lambda_0$, and we build statistical estimators from the following misspecified spiked matrix model

$$Y=G+\frac{\lambda}{\sqrt{N}}x_0x_0^\intercal\,, \tag{5.1}$$

in other words, we assume that the log likelihood is given by:

$$g_\lambda(y,w)=-\frac{(y-\lambda w)^2}{2}-\frac{1}{2}\log(2\pi).$$

The information parameters for are given by

$$\beta_0=\lambda_0^2\,,\ \beta_1(\lambda)=\lambda^2\,,\ \beta_2(\lambda)=\lambda\lambda_0\,,\ \beta_3(\lambda)=\lambda^2 \tag{5.2}$$

where we recall that β_0 is the true information parameter associated with the correctly specified model.

5.2. **Stochastic Block Model.** We now consider a community detection problem with two groups. We work with the Stochastic Block Model SBM$(n,\frac{1}{2}+\mu_0N^{-1/2},\frac{1}{2}-\mu_0N^{-1/2})$ on two communities. In this model we shall assume that our unknown signal $\boldsymbol{x}^0$ lies in $\{\pm1\}^N$, and serves as the index vector for the two communities. The corresponding data matrix is the adjacency matrix, and its entries have distribution given by:

$$\mathbb{P}\left(Y_{i,j}=1\mid\frac{x_ix_j}{\sqrt{N}}\right)=\frac{1}{2}+\mu_0\frac{x_ix_j}{\sqrt{N}}\qquad\text{and}\qquad\mathbb{P}\left(Y_{i,j}=0\mid\frac{x_ix_j}{\sqrt{N}}\right)=\frac{1}{2}-\mu_0\frac{x_ix_j}{\sqrt{N}}\,.$$

The parameter $\mu_0>0$ represents the difference between the probability of edges appearing within and outside of each group. Notice that when x_i,x_j take the same sign, the probability is higher, and when x_ix_j take different signs then the probability of connecting an edge is lower. The $\sqrt{N}$ scaling is such that the detection problem becomes non-trivial, and a phase transition on the weak recovery of the groups is observable (see [57]). More generally the Stochastic block model has been studied in a wide variety of regimes for the connection probabilities between communities (see [1] for a detailed overview of different regimes). There is a large collection of literature concerned with showing when different notions of recovery of the communities is possible, as well as when there

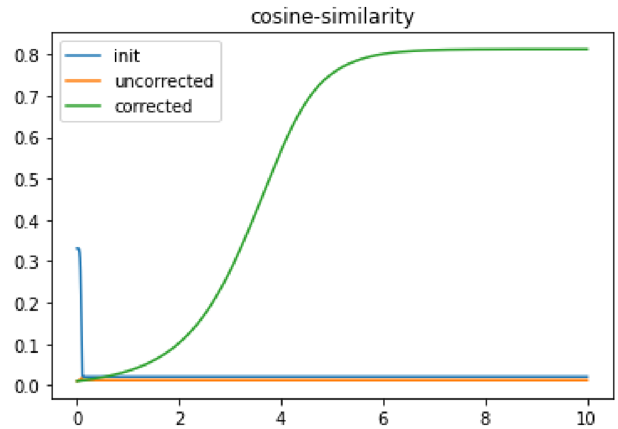

FIGURE 1. The cosine similarity in the spiked matrix problem with Rademacher latent variable and noise with mean 1. A data matrix of size 2500×2500 and the uncorrected and corrected likelihoods were optimized using gradient descent. It is clear that correcting the likelihood gets rid of the effect from the score parameter, and the corresponding PMLE estimator achieves a non-zero cosine similiarty.

are efficient algorithms for recovery. See [2, 40, 62, 39, 60, 26] and the references therein.
For the SBM with connection probabilities as above, the loglikelihood is given by

$$g_0(Y,w) = Y\ln(\frac{1}{2}+\mu_0 w) + (1-Y)\ln(\frac{1}{2}-\mu_0 w).$$

Suppose that a signal to noise ratio $\mu_0 \neq \mu$ is chosen, that is, we choose the pseudo-likelihood

$$g(Y,w) = Y\ln(\frac{1}{2}+\mu w) + (1-Y)\ln(\frac{1}{2}-\mu w)\,,$$

then the information parameters are given by

$$\beta_1 = 4\mu^2, \beta_2 = 4\mu\mu_0, \beta_3 = 4\mu^2, \beta_4 = 0\,,$$

and the Rao relation is not satisfied. We note however that $\beta_4 = 0$ and so our choice of pseudo-likelihood is well scored.

One method to introduce an ill-scored pseudo-likelihood is to work with an incorrect assumption on the null-model. If we suppose that null model corresponds to the adjacency matrix of a $G(n,p)$ matrix with $p \neq 1/2$, that is the pseudo-likelihood is given by:

$$g_p(y,w) = y\log(p+\mu w) + (1-y)\log(1-p-\mu w)\,,$$

and a direct computation yields:

$$\beta_4^{g_p} = \frac{\mu(1-2p)}{2p(1-p)}\,,$$

which is zero if and only if $p = 1/2$.

5.3. **Sparse Rademacher Matrices and Best Rank-1 approximation.** We now consider a class of sparse submatrix detection problems [14]. For this example, we suppose that our unknown vector lies in Ω^N where Ω is either an interval $[a,b]$ or a finite set.
Consider the case where Y is a sparse Rademacher matrix, i.e Y, conditionally on w, takes values in $\{-1,0,1\}$ with probabilities given by:

$$\mathbb{P}\left(Y_{ij} = \pm 1 | \frac{x_i x_j}{\sqrt{N}}\right) = \frac{p}{2} + \lambda\frac{x_i x_j}{\sqrt{N}}\,, \qquad \text{and} \qquad \mathbb{P}\left(Y = 0 | \frac{x_i x_j}{\sqrt{N}}\right) = 1 - p - 2\lambda\frac{x_i x_j}{\sqrt{N}}\,, \tag{5.3}$$

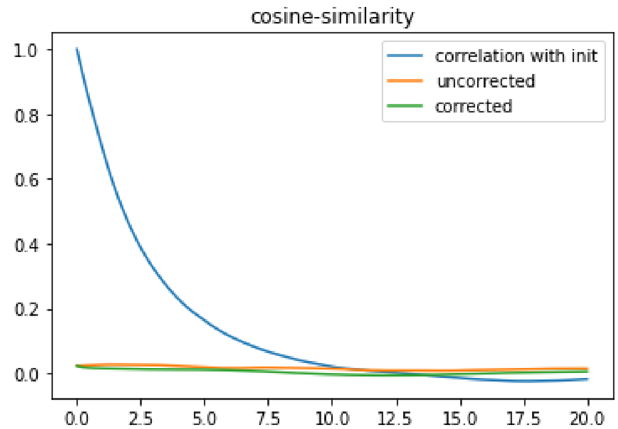

FIGURE 2. The cosine similarity in the sparse Rademacher problem. A data matrix of size 2500×2500 and corrected and uncorrected least squares was performed on the data matrix and optimized using gradient descent. It is clear that the least squares estimator was uninformative and always achieved a cosine-similarity of zero.

where throughout p is a fixed number in $(0,1)$. In this case the log-likelihood is given by:

$$g_0(Y,w) = (1-Y^2)\log(1-p-2\lambda w) + \frac{Y(Y-1)}{2}\log(p/2+\lambda w) - \frac{Y(1+Y)}{2}\log\left(\frac{p}{2}+\lambda w\right),$$

and the corresponding score parameters are given by:

$$\beta_1 = \beta_2 = \beta_3 = \frac{4\lambda^2}{1-p} + \frac{\lambda^2}{p}\,.$$

Suppose now that we try to infer the unknown vector via the best rank 1 approximation, that is, we try to minimize

$$\min_{x\in\Omega^N}\left\|Y - \lambda\frac{\mathbf{x}\mathbf{x}^T}{\sqrt{N}}\right\|_F^2,$$

then as discussed in Section 4, the corresponding estimator $\mathbf{x}_{LS}$ corresponds to a pseudo likelihood estimator $\hat{\mathbf{x}}_{\text{PMLE}}$ with pseudo-likelihood given by:

$$g(Y,w) = -\frac{1}{2}(Y-\lambda w)^2 - \log(2\pi)\,.$$

By Proposition 4.1 the model is well scored, and furthermore, an explicit computation shows the Fisher parameters for the proxy model are given by:

$$\beta_1 = \lambda p, \beta_2 = 0, \beta_3 = \lambda^2\,,$$

and consequently by Proposition 4.2 the least-square estimator is completely uninformative provided that the limiting empirical measure of $\mathbf{x}^{0,N}$ is balanced.

5.4. **Non-Linear transformations of rank 1 matrices.** Consider a data vector $x \in \{\pm 1\}^N$, and a spiked Wigner matrix W given by:

$$W = G + \frac{\lambda}{\sqrt{N}}xx^T\,,$$

where G is a symmetric matrix with i.i.d standard Gaussian entries. From W we consider the transformation taking each entry W_{ij} and sending them to $Y_{ij} = F(W_{ij})$ for some function F. Non-linear transformations of random matrices have applications in to kernel methods [76, 52, 51] and the spectra of one-layer neural networks [69, 72, 21]. The spectra of $F(W_{ij})$ was thoroughly analyzed in [36] and [31].

From the matrix Y we will study the behavior of maximum likelihood estimation for certain choices of F. We remark that some choices of F will lead to irregular likelihoods that do not fall into our framework. We provide an example in Section 5.4.2.

5.4.1. *Rounded Entries:* Suppose that $F(x) = \mathrm{sgn}(x)$ (with the convention that $\mathrm{sgn}(0) = 1$). This is the censored spiked matrix model that was studied recently in [54]. In this case the likelihood of the output matrix Y is given by:

$$g(y,w) = \frac{(1-y)}{2}\log\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-\lambda w} e^{-\frac{x^2}{2}}dx + \frac{(1+y)}{2}\log\frac{1}{\sqrt{2\pi}}\int_{-\lambda w}^{\infty} e^{-x^2/2}dx\,.$$

We may explicitly compute the β values in this case:

$$\beta_1 = \beta_2 = \beta_3 = \frac{2}{\pi}\lambda^2, \beta_4 = 0\,.$$

5.4.2. *Squaring Entries:* We now provide an example which does not fall into the class $\mathcal{F}_0$. Suppose we choose $F(x) = x^2$, then explicitly one computes the log-likelihood to be given by:

$$g(y,w) = -\frac{1}{2}\log(2\pi) - \log(2) - \frac{1}{2}\log(y) + \log\left[e^{-(\sqrt{y}-\lambda w)^2/2} + e^{-(\sqrt{y}+\lambda w)^2/2}\right]\,.$$

In particular, the second derivative of $g(y,w)$ at $w = 0$ is given by:

$$\partial_w^2 g(y,0) = \lambda^2(y-1)\,,$$

and consequently the bound $\|\partial_w^2 g(\cdot,0)\| < \infty$ fails.

6. Outline of Proofs

In this section, we will summarize the strategy to prove the main results. The proofs will be deferred to the relevant sections of the Appendix. To simplify the notation in this section, we only consider ill-scored scenario. The case for well-scored problems are simpler and the proof is essentially the same. The only difference is the constraint on the mean $\bar{x}$, which is unneeded.

6.1. **Universality.** We begin by showing that the limit of the (normalized) maximum pseudo-likelihood is equivalent to that obtained by a maximization of a Gaussian model parameterized by the information parameters. The Gaussian model is given by

$$\begin{aligned} H_N^{\bar{\beta}}(x) &= \frac{\sqrt{\beta_1}}{\sqrt{N}}\sum_{1\le i\le j\le N} g_{ij}x_ix_j + \frac{\beta_2}{N}\sum_{1\le i\le j\le N} x_i^0x_j^0x_ix_j - \frac{\beta_3}{2N}\sum_{1\le i\le j\le N} x_i^2x_j^2 + \beta_4\sum_{1\le i\le j\le N}\frac{x_ix_j}{\sqrt{N}} \\ &= \frac{\sqrt{\beta_1}}{\sqrt{N}}\sum_{ij} g_{ij}x_ix_j + \frac{N\beta_2}{2}R_{10}^2 - \frac{N\beta_3}{4}R_{11}^2 + \beta_4 N^{\frac{3}{2}}(\bar{x})^2 + o_N(1) \end{aligned} \tag{6.1}$$

where we for vectors $\mathbf{x}^1, \mathbf{x}^2 \in \mathbb{R}^N$, we denote the normalized inner product by

$$R(\mathbf{x}^1,\mathbf{x}^2) = \frac{1}{N}\sum_{i=1}^N x_i^1x_i^2\,,$$

and we set $R_{11} = R(\mathbf{x},\mathbf{x})$, $R_{10} = R(\mathbf{x},\mathbf{x}^0)$, $\bar{x} = R(\mathbf{x},\mathbf{1})$.

We will prove in Appendix A that the asymptotic MLE is equal to the one given by the maximum of the proxy model on average. Given $S, M, v \subseteq \mathbb{R}$, let $\Omega_\varepsilon(S,M,v)$ denote the set of points in Ω^N within ε of (S,M,v), i.e

$$\Omega_\varepsilon(S,M,v) := \{x \in \Omega^N : |R_{1,1} - S| \le \varepsilon, |R_{1,0} - M| \le \varepsilon, |\overline{x} - v| \le \varepsilon\}\,,$$

and let us define

$$\mathcal{L}_N^{g,\varepsilon}(S,M,v) = \mathbb{E} \max_{\Omega_\varepsilon(S,M,v)} \Big(\sum_{i<j} g\Big(Y_{ij}, \frac{x_i x_j}{\sqrt{N}}\Big) - \sum_{i<j} g(Y_{ij},0)\Big)\,,$$

and

$$\mathcal{L}_N^{\bar\beta,\varepsilon}(S,M,v) = \mathbb{E} \max_{\Omega_\varepsilon(S,M,v)} H_N^{\bar\beta}(x)\,,$$

to denote the restricted pseudo MLE and the proxy MLE respectively. We will prove in Section A that the pseudo MLE and the proxy MLE are equivalent in the following sense.

Lemma 6.1. *If $g, g_0 \in \mathcal{F}_0$, then for any $(S,M,v) \in \mathcal{C}_c$*

$$\lim_{N\to\infty} \frac{1}{N} |\mathcal{L}_N^{g,\varepsilon}(S,M,v) - \mathcal{L}_N^{\bar\beta,\varepsilon}(S,M,v)| = 0\,,$$

where $\bar\beta$ are the information and score parameters defined in (2.5), (2.6), (2.7), (2.8).

This is proved by showing equivalence for smooth approximations of the MLE. Let $\mathbb{P}_X(x)$ be the uniform measure on Ω. We approximate the pseudo MLE with log likelihood ratio of the posterior. We define the log-likelihood ratio associated with the pseudo likelihood

$$F_N(g,\varepsilon;S,M,v) := \frac{1}{N}\Big(\mathbb{E}_Y\Big(\log \int_{\Omega_\varepsilon(S,M,v)} e^{\sum_{i<j} g\big(Y_{ij},\frac{x_i x_j}{\sqrt{N}}\big)}\, d\,\mathbb{P}_X^{\otimes N}(\mathbf{x}) - \sum_{i<j} g(Y_{ij},0)\Big)\Big)\,, \tag{6.2}$$

where $\mathbb{E}_Y$ is with the average with respect to the conditional data distribution (2.1). On the other hand, we define the Gaussian log-likelihood ratio for $\bar\beta = (\beta_1,\beta_2,\beta_3,\beta_4) \in \mathbb{R}^4$ by

$$F_N(\bar\beta,\varepsilon;S,M,v) := \frac{1}{N}\mathbb{E}_Y \log \int_{\Omega_\varepsilon(S,M,v)} e^{H_N^{\bar\beta}(\mathbf{x})}\, d\,\mathbb{P}_X^{\otimes N}(\mathbf{x})\,, \tag{6.3}$$

where $H_N^{\bar\beta}(\mathbf{x})$ is as in (6.1). We define $F_N^{g,\varepsilon}$ and $F_N^{\bar\beta,\varepsilon}$ to be the equal to (6.2) and (6.3) without the constraints. These quantities approximate the pseudo MLE in the sense that for any $(S,M,v) \in \mathcal{C}_c$

$$\lim_{L\to\infty}\lim_{N\to\infty} |\frac{1}{L}F_N(Lg,\varepsilon;S,M,v) - \frac{1}{N}\mathcal{L}_N^{g,\varepsilon}(S,M,v)| = 0 \tag{6.4}$$

$$\lim_{L\to\infty}\lim_{N\to\infty} |\frac{1}{L}F_N(L\bar\beta,\varepsilon;S,M,v) - \frac{1}{N}\mathcal{L}_N^{\bar\beta,\varepsilon}(S,M,v)| = 0\,. \tag{6.5}$$

An analogous statement holds for the unconstrained versions. Universality for the PMLE is then proved in the following:

Lemma 6.2. *If $g, g_0 \in \mathcal{F}_0$, then for any $S,M,v \in \mathcal{C}_c$*

$$\lim_{L\to\infty}\lim_{N\to\infty} \left|\frac{1}{L}F_N(Lg,\varepsilon;S,M,v) - \frac{1}{L}F_N(L\bar\beta,\varepsilon;S,M,v)\right| = 0$$

and

$$\lim_{L\to\infty}\lim_{N\to\infty} \left|\frac{1}{L}F_N^{Lg,\varepsilon} - \frac{1}{L}F_N^{L\bar\beta,\varepsilon}\right| = 0$$

where $\bar\beta$ are the information and score parameters defined in (2.5), (2.6), (2.7), (2.8).

The key idea in this proof is that at the level of the log likelihood, we are able to use Taylor's theorem to expand around the likelihood in the exponent with respect to the $\frac{x_i x_j}{\sqrt{N}}$. The second order coefficients of the Taylor series will concentrate in the high dimensional limit while the first order term will be approximately Gaussian with a specific mean and variance given by the information coefficients.

The main consequence of Lemma 6.2 is that it suffices to compute the limit in the case of the Gaussian model instead of $F_N(\bar\beta; S, M, v)$ instead of the pseudo MLE. The computation of this limit is the focus of the following two sections.

6.2. **Derivation of the Variational Formula I.** In this section, we once again use the approximation of the likelihoods with the loglikelihood ratios and first compute the limit of the loglikelihood ratio. Our goal is to first define the variational formula for the loglikelihood ratios.

Notice that the term corresponding to $H_N^{\bar\beta}$ is of higher order, so this term must be corrected in order to have a well defined limit. To this end, we define

$$H_N^{\bar\beta,\alpha}(\mathbf{x}) = H_N^{\bar\beta}(\mathbf{x}) - \beta_4 N^{\frac{3}{2}}(\bar x)^2 + \alpha N(\bar x)^2\,,$$

which is the proxy model for (2.19). We define

$$F_{N,\alpha}(\bar\beta,\alpha,\varepsilon;S,M,v) := \frac{1}{N}\mathbb{E}_Y \log \int_{\Omega_\varepsilon(S,M,v)} e^{H_N^{\bar\beta,\alpha}(\mathbf{x})}\, d\,\mathbb{P}_X^{\otimes N}(\mathbf{x})\,,$$

and let $F_{N,\alpha}(\bar\beta,\alpha,\varepsilon)$ denotes its unconstrained version.

We now define the variational formula which will compute the limit. Let ζ be a probability measure, and let $\Phi_{\zeta,\mu,\lambda,\rho}(t,y)$ be the unique weak solution to the Parisi PDE

$$\begin{cases}\partial_t \Phi_{\zeta,\mu,\lambda,\rho} = -\frac{\beta_1}{4}(\partial_y^2\Phi_\zeta + \zeta([0,t])(\partial_y \Phi_\zeta)^2) & (t,y)\in(0,S)\times\mathbb{R} \\ \Phi_{\zeta,\mu,\lambda,\rho}(S,y;x^0) = \log\int e^{yx+\lambda x x^0 + \mu x^2 + \rho x}\, d\,\mathbb{P}_X(x) & \end{cases}. \tag{6.6}$$

See [46] for the notion of weak solutions for this PDE and the corresponding well-posedness. Define the corresponding Parisi functional by

$$\begin{aligned}&\varphi_{\bar\beta,\alpha}(S,M,v)\\ &= \inf_{\mu,\lambda,\rho,\zeta}\left(\mathbb{E}_{\mathbb{Q}}[\Phi_{\zeta,\mu,\lambda,\rho}(0,0;x^0)] - \frac{\beta_1}{2}\int_0^S t\zeta([0,t])\,dt - \mu S - \lambda M - \rho v + \frac{\beta_2 M^2}{2} - \frac{\beta_3 S^2}{4} + \frac{\alpha v^2}{2}\right).\end{aligned}$$

Furthermore, we see that $(R_{1,1}, R_{1,0}, \bar x)$ asymptotically live in the closed subset $\mathcal{C}_c$ of $[0,C^2]\times[-C^2,C^2]\times[-C,C]$ defined in (B.5). This is the domain of our functional. We will show in Appendix B that the limit of the loglikelihood is given by the Parisi functional.

Theorem 6.1. *For any β_1,β_2,β_3 and α and constraints $(S,M,v)\in\mathcal{C}$, we have*

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} F_{N,\alpha}(\bar\beta,\alpha,\varepsilon;S,M,v) = \varphi_{\beta,\alpha}(S,M,v)$$

and for the unconstrainted problem

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} F_{N,\alpha}(\bar\beta,\alpha,\varepsilon) = \sup_{S,M,v}\varphi_{\beta,\alpha}(S,M,v).$$

Remark 6.1. For regular models, the constraint on v can be completely removed and was proven in [37]. In such cases, the optimization is over the functional $\varphi_{\bar\beta}$ which is defined on only two parameters S and M.

The proof of this results relies heavily on the techniques first developed to study the Sherrington–Kirkpatrick model in spin glasses [35, 81, 4, 64]. By introducing small perturbation to the loglikelihood, we are able to show characterize the limiting behaviour of independent samples from the perturbed posterior measure and explicitly compute the limit. The proof also borrows techniques from large deviations to remove deal with the constraint on the overlaps.

Having computed the appropriate limit of $F_N(\bar\beta;S,M,v)$, we can apply Lemma 6.2 to recover the limit of the loglikihood ratio. By (6.4) if one can compute the limit of this quantity as $L\to\infty$, then we can recover the limiting formula for the pseudo MLE.

6.3. **Derivation of the Variational Formula II.** This variational formula holds for all $\beta_1, \beta_2, \beta_3$ and α, so it also holds when these parameters are scaled by L as in the smooth approximation. We will show in Appendix C that taking the limit as $L \to \infty$ of this variational formula will give the formula for the MLE, after an application of (6.4), which will give us a variational formula for the limit of pseudo MLE.

Lemma 6.3. *For any $\bar{\beta}$, in the constrained PMLE we have*

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathcal{L}_N^{g,\varepsilon}(S,M,v)=\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathcal{L}_N^{\bar{\beta},\varepsilon}(S,M,v)=\psi_{\bar{\beta},\alpha}(S,M,v)\,,$$

and in the unconstrained PMLE,

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathcal{L}_N^{g,\varepsilon}=\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathcal{L}_N^{\bar{\beta},\varepsilon}=\sup_{(S,M,v)\in\mathcal{C}_c}\psi_{\bar{\beta},\alpha}(S,M,v)\,,$$

where

$$\psi_{\bar{\beta},c}(S,M,v)=\psi_{\bar{\beta},-}(S,M,v)+\frac{\alpha v^2}{2}$$

and $\psi_{c,-}(S,M,v)$ is given by (2.15).

This proof uses the Γ limit of the solutions in [45] to the Parisi PDE (B.24) to identify $\psi_{\bar{\beta}}$ as the limit of $\frac{1}{L}\varphi_{\bar{\beta}}$. In the lemma above the constrained case is established in Appendix C. The limit formula in the unconstrained is established in Appendices D and E, by proving an equicontinuity statement for the family of functionals $\frac{1}{L}\varphi_{L\bar{\beta}}$.

Having understood the limiting variational formula, one can also show using (6.4), that this limiting variational formula characterizes the constrained pseudo MLE.

We can conclude that the limit of the pseudo MLE is a variational optimization over the parameters S, M, v. We will show in the next section that the maximizers of the variational problem encode the limiting performance of the maximum likelihood estimators.

6.4. **Characterization of the Maximizers.** In Appendix G we prove tightness of the overlaps as stated in Theorems 2.1 and 2.5. The tightness will follow from concentration properties satisfied by the proxy model (6.1), and the results proved in Appendix C.

Next, under the further assumption that $\psi_{\bar{\beta},\alpha}$ has a unique maximizer, we are able to prove the following characterization of performance:

Lemma 6.4. *For $g, g^0 \in \mathcal{F}_0$ let $\bar{\beta}$ denote the corresponding information and score paramters and suppose that $\psi_{\bar{\beta}}$ has a unique (up to the sign of m) maximizer $(s_{\bar{\beta}}, \pm m_{\bar{\beta}}, v_{\bar{\beta}})$. Then*

$$|\mathbf{CS}(\hat{\mathbf{x}}_{\mathrm{PMLE}},\mathbf{x}_0)|\to\frac{|m_{\bar{\beta}}|}{(s_{\bar{\beta}}\mathbb{E}_{\mathbb{Q}}(x^0)^2)^{1/2}}\qquad a.s \tag{6.7}$$

$$\frac{1}{N}\left\|\hat{x}_{PMLE}\right\|^2\to s_{\bar{\beta}}\qquad a.s. \tag{6.8}$$

6.5. **Coarse Equivalence of Estimators.** In Section H, we prove a sufficient condition for when two likelihoods are coarsely equivalent. Coarse equivalence will follow as a consequence of the universality result in Lemma 6.1. Given likelihoods g^1, g^2 which satisfy :

$$\frac{\sqrt{\beta_1(g^1)}}{\sqrt{\beta_1(g^2)}}=\frac{\beta_2(g^1)}{\beta_2(g^2)}=\frac{\beta_3(g^1)}{\beta_3(g^2)}=\frac{\beta_4(g^1)}{\beta_4(g^2)}=C, \tag{6.9}$$

the corresponding proxy-models will be a scalar multiple of each-other. Consequently the collection of near-maximizers will be the same for both problems, and the result will then follow from Theorems 2.1 and 2.5. We further prove theorem 3.1 in this section, it will be an immediate consequence of the universality established in Theorems 2.1 and 2.5.

Appendix A. Universality with Non-Zero Score

In this section, we prove universality of pseudo maximum likelihood estimation with possibly non-zero score parameters. This extends the universality result in [37] for the MLE and removes the zero score assumption in [37, Hypothesis 2.3].

Given the information and score parameters, we define the proxy log-likelihood function corresponding to $\bar{\beta}$ as

$$H_N^{\bar{\beta}}(x) = \frac{\beta_1}{\sqrt{N}} \sum_{1 \le i \le j \le N} g_{ij} x_i x_j + \frac{\beta_2}{N} \sum_{1 \le i \le j \le N} x_i^0 x_j^0 x_i x_j - \frac{\beta_3}{2N} \sum_{1 \le i \le j \le N} x_i^2 x_j^2 + \beta_4 \sum_{1 \le i \le j \le N} \frac{x_i x_j}{\sqrt{N}}$$
$$= \frac{\beta_1}{\sqrt{N}} \sum_{ij} g_{ij} x_i x_j + \frac{N\beta_2}{2} R_{10}^2 - \frac{N\beta_3}{4} R_{11}^2 + \beta_4 N^{\frac{3}{2}} (\bar{x})^2 + o_N(1)\,, \tag{A.1}$$

where we recall that the overlaps and magnetization are denoted by

$$R_{10} = \frac{1}{N}\sum_{i=1}^N x_i x_i^0, \quad R_{11} = \frac{1}{N}\sum_{i=1}^N x_i^2, \quad \bar{x} = \frac{1}{N}\sum_{i=1}^N x_i\,.$$

Given subsets A, B, C of $\mathbb{R}$, let

$$\mathcal{L}_N^g(A,B,C) := \frac{1}{N}\mathbb{E} \max_{R_{10}\in A, R_{11} \in B, \bar{x} \in C} \bigg(\sum_{i<j} g\Big(Y_{ij}, \frac{x_i x_j}{\sqrt{N}}\Big) - \sum_{i<j} g(Y_{ij}, 0) \bigg)$$
$$\mathcal{L}_N^{\bar{\beta}}(A,B,C) := \frac{1}{N}\mathbb{E} \max_{R_{10}\in A, R_{11} \in B, \bar{x} \in C} H_N^{\bar{\beta}}(x)$$

denote the normalized pseudo MLE and the proxy MLE respectively. The goal of this section is to show that these quantities converge to the same value.

Lemma A.1. *If $g, g_0 \in \mathcal{F}_0$, then for any $A, B, C \subset \mathbb{R}$*

$$\lim_{N\to\infty} |\mathcal{L}_N^g(A,B,C) - \mathcal{L}_N^{\bar{\beta}}(A,B,C)| = 0$$

where $\bar{\beta}$ are the information and score parameters defined in (2.5), (2.6), (2.7), (2.8).

The proof of the MLE formulas will follow from an extension of the universality for Bayesian models proven in [37]. In contrast to the Bayesian inference setting, we fix $\mathbf{x}^0$ and define the log-likelihood ratios

$$F_N(g; A, B, C)$$
$$= \frac{1}{N}\bigg(\mathbb{E}_Y \Big(\log \int \mathbb{1}(R_{10} \in A, R_{11} \in B, \bar{x} \in C) e^{\sum_{i<j} g\left(Y_{ij}, \frac{x_i x_j}{\sqrt{N}}\right)} \, d\, \mathbb{P}_X^{\otimes N}(\mathbf{x}) - \sum_{i<j} g(Y_{ij}, 0) \Big) \bigg) \tag{A.2}$$

where $\mathbb{E}_Y$ is with the average with respect to the conditional data distribution (2.1).

The dependence of F_N on $\mathbf{x}^0$ will remain implicit. We can interpret these log-likelihood as a smooth approximation of constrained Pseudo maximum likelihood estimation. We have standard bounds relating L_N and F_N given by:

$$\mathcal{L}_N^g(A,B,C) \le \frac{F_N(Lg; A,B,C)}{L} \le \mathcal{L}_N^g(A,B,C) + o_L(1)\,, \tag{A.3}$$

which are obtained by either replacing g in the exponent with its maximum value or by localizing around the maximum value. The regularity conditions on g are essential in this argument.

We also define the Gaussian log-likelihood ratio for $\bar{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) \in \mathbb{R}_+^3 \times \mathbb{R}$ by

$$F_N(\bar{\beta}; A, B, C) = \frac{1}{N}\mathbb{E}_Y \log \int \mathbb{1}(R_{10} \in A, R_{11} \in B, \bar{x} \in C) e^{H_N^{\bar{\beta}}(\mathbf{x})} \, d\, \mathbb{P}_X^{\otimes N}(\mathbf{x})\,,$$

where $H_N^{\bar{\beta}}(\mathbf{x})$ was defined in (A.1). In the case that Ω is discrete we let $\mathbb{P}_X$ denote counting measure, and in the case that Ω is an interval, we let $\mathbb{P}_X$ denote normalized Lebesgue measure. We start by proving universality for log-likelihood.

Proposition A.1 (Universality of Bayesian Models)**.** *If $g, g^0 \in \mathcal{F}_0$, then for any $A, B, C \subset \mathbb{R}$ there exists a constant $K > 0$ depending only on g, g^0 such that*

$$\left|F_N(g; A, B, C) - F_N(\bar{\beta}; A, B, C)\right| \leq \frac{K}{\sqrt{N}}$$

where $\bar{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ are the Fisher score parameters defined in (2.5), (2.6), (2.7), (2.8).

Proof. The proof is in Section 3 from [37]. We highlight the key steps. To simplify notation, we let

$$\Omega(A, B, C) := \{R_{10} \in A, R_{11} \in B, \bar{x} \in C\},$$

and we let K denote a universal constant that only depends on the supports Ω and Ω_0, but not on the dimension N.

Step 1 - Approximation by Third Order Terms: We first show that to leading order in N, it suffices to consider only a third order expansion of the loglikelihood around $w = 0$, define a proxy $F^{(1)}$ by:

$$F_N^{(1)}(g; A, B, C) = \frac{1}{N}\left(\mathbb{E}_Y\left(\log \int \mathbb{1}(\Omega(A, B, C)) e^{\sum_{i<j} \partial_w g(Y_{ij}, 0) w_{ij} + \frac{1}{2}\partial_w^{(2)} g(Y_{ij}, 0) w_{ij}^2} \, d\,\mathbb{P}_X^{\otimes N}(\mathbf{x})\right)\right).$$

By our regularity assumptions on g we may Taylor expand the log-likelihood. In particular, Taylor's theorem implies there is $\theta_{ij} \in [0, 1]$ such that

$$(g(Y_{ij}, w_{ij}) - g(Y_{ij}, 0)) = \partial_w g(Y_{ij}, 0) w_{ij} + \frac{1}{2}\partial_w^{(2)} g(Y_{ij}, 0) w_{ij}^2 + \frac{w_{ij}^3}{3!}\partial_w^{(3)} g(Y_{ij}, \theta_{ij} w_{ij}),$$

and since $\partial_w^{(3)} g(Y_{ij}, \theta_{ij} w_{ij})$ is uniformly bounded and $|w_{ij}| \leq \frac{C^2\lambda}{\sqrt{N}}$, we have

$$\left|F_N(g; A, B, C) - F_N^{(1)}(g; A, B, C)\right| \leq \frac{\|\partial_w^{(3)} g(Y_{ij}, \theta_{ij} w_{ij})\|_\infty K}{\sqrt{N}},$$

and thus, it suffices to compute the limit for $F_N^{(1)}$.

Step 2 - Control of the Second Order Terms: We now show that we can replace $\partial_w^{(2)} g(Y_{ij}, 0) w_{ij}^2$ in the exponent with its average. Define

$$F_N^{(2)}(g; A, B, C) = \frac{1}{N}\left(\mathbb{E}_Y\left(\log \int \mathbb{1}(\Omega(A, B, C)) e^{\sum_{i<j} \partial_w g(Y_{ij}, 0) w_{ij} + \frac{1}{2}\mathbb{E}_Y[\partial_w^{(2)} g(Y, 0)] w_{ij}^2} \, d\,\mathbb{P}_X^{\otimes N}(\mathbf{x})\right)\right),$$

then we may express the difference of $F_N^{(1)}$ and $F_N^{(2)}$ as follows:

$$F_N^{(1)}(g; A, B, C) - F_N^{(2)}(g; A, B, C) = \mathbb{E}_Y \frac{1}{N} \ln \left\langle e^{\frac{1}{2\sqrt{N}} \sum_{i<j} \frac{1}{\sqrt{N}} (\partial_w^{(2)} g(Y_{ij}, 0) - \mathbb{E}_Y[\partial_w^{(2)} g(Y_{ij}, 0)])(x_i x_j)^2)} \right\rangle,$$

where for a function $f : \Omega^N \to \mathbb{R}$, the average $\langle f \rangle$ is defined as:

$$\langle f \rangle := \frac{\int \mathbb{1}(\Omega(A, B, C)) f(\boldsymbol{x}) e^{\sum_{i<j} \partial_w g(Y_{ij}, 0) w_{ij} + \frac{1}{2}\mathbb{E}_0[\partial_w^{(2)} g(Y, 0)] w_{ij}^2} d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\int \mathbb{1}(\Omega(A, B, C)) e^{\sum_{i<j} \partial_w g(Y_{ij}, 0) w_{ij} + \frac{1}{2}\mathbb{E}_Y[\partial_w^{(2)} g(Y, 0)] w_{ij}^2} d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}.$$

Now if Z denotes the $N \times N$ symmetric matrix with entries

$$Z_{ij} := \frac{1}{2\sqrt{N}}(\partial_w^{(2)} g(Y_{ij}, 0) - \mathbb{E}_0[\partial_w^{(2)} g(Y_{ij}, 0)]),$$

then one may write:

$$\sum_{i<j} \frac{1}{2\sqrt{N}}(\partial_w^{(2)} g(Y_{ij},0) - \mathbb{E}_Y[\partial_w^{(2)} g(Y_{ij},0)])(\boldsymbol{x}_i\boldsymbol{x}_j)^2 = \mathrm{Tr}\left(Z(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^2\right) .$$

Note that under $\mathbb{P}_0$, Z is a random matrix with centered entries and covariance bounded by C/N. Using standard concentration inequalities for random matrices (see [6, Theorem 2.3.5]), there exists $T_0 > 0$ such that

$$\mathbb{P}_0\left(\|Z\|_{op} \geq 2T_0\right) \leq e^{-cN} , \tag{A.4}$$

for some $c > 0$. Now on the event that $\{\|Z\|_{op} \leq 2T_0\}$,

$$\left|\mathrm{Tr}(Z(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^2)\right| = \left|\sum_{i,j} Z_{ij}x_i^2x_j^2\right| \leq 2T_0\sum_{i=1}^N x_i^4 \leq 2CT_0N$$

for some finite constant C depending only the bound on the support of $\mathbb{P}_X$, so

$$\mathbb{E}_Y\, \mathbb{1}_{\|Z\|_{op}\leq 2T_0} \frac{1}{N}\ln\left\langle e^{\frac{1}{\sqrt{N}}\sum_{i\leq j}\frac{1}{2\sqrt{N}}(\partial_w^{(2)}g(Y_{ij},0)-\mathbb{E}_Y[\partial_w^{(2)}g(Y_{ij},0)])(x_ix_j)^2)}\right\rangle \leq \frac{2T_0}{\sqrt{N}}.$$

Similarly, on the event that $\{\|Z\|_{op} \geq 2T_0\}$ we have:

$$\mathbb{E}_Y\, \mathbb{1}_{\|Z\|_{op}\geq 2T_0} \frac{1}{N}\ln\left\langle e^{\frac{1}{\sqrt{N}}\sum_{i\leq j}\frac{1}{2\sqrt{N}}(\partial_w^{(2)}g(Y_{ij},0)-\mathbb{E}_Y[\partial_w^{(2)}g(Y_{ij},0)])(x_ix_j)^2)}\right\rangle \leq e^{-cN} ,$$

which follows as $\partial_w^2 g(Y_{ij},0)$ is assumed to be uniformly bounded over i,j. Combining, we see for N sufficiently large that there is some $K > 0$ such that:

$$\left|F_N^{(1)}(g:A,B,C) - F_N^{(2)}(g:A,B,C)\right| \leq \frac{K}{\sqrt{N}} .$$

Step 3 - Expansion of The First Order Term: We now show that the first order term can be approximated by a Gaussian random variable with non-zero mean. Define

$$F_N^{(3)}(g;A,B,C) := \frac{1}{N}\left(\mathbb{E}_Y\left(\log\int \mathbb{1}(\Omega(A,B,C))e^{\sum_{i<j}\beta_1 w_{ij}+\beta_2 w_{ij}^0 w_{ij}-\frac{1}{2}\beta_3 w_{ij}^2+\beta_4 w_{ij}}\, d\,\mathbb{P}_X^{\otimes N}(\mathbf{x})\right)\right),$$

and consider the following moments of the information parameter under the data distribution,

(1) $\mu_{ij} = \mathbb{E}_Y[\partial_w g(Y_{ij},0) \mid \boldsymbol{x}^0]$
(2) $\sigma_{ij}^2 = \mathbb{E}_Y[(\partial_w g(Y_{ij},0) - \mu_{ij})^2 \mid \boldsymbol{x}^0]$
(3) $\gamma_{ij} = \mathbb{E}_Y[\partial_w^{(2)} g(Y_{ij},0) \mid \boldsymbol{x}^0]$.

Using Taylor's theorem, these parameters may be expressed in terms of the information parameters under the null model,

(1) With $w_{ij}^0 = x_ix_j/\sqrt{N}$, and recalling the fact that $\|\partial_w g^0\|_\infty, \|\partial_w g^0\|_\infty < \infty$ for $g, g^0 \in \mathcal{F}_0$, we may compute:

$$\begin{aligned}
\mu_{ij} &= \mathbb{E}_Y[\partial_w g(Y_{ij},0)|\boldsymbol{x}^0] = \int \partial_w g(y,0)e^{g^0(y,w_{ij}^0)}\,dy \\
&= \int \partial_w g(y,0)\Big(1 + \partial_w g^0(y,0)w_{ij}^0 + ((\partial_w g^0(y,0))^2 + \partial_w^{(2)} g^0(y,0))\frac{(w_{ij}^0)^2}{2} + O(N^{-1})\Big)e^{g^0(Y,0)}\,dy \\
&= \mathbb{E}_0\partial_w g(Y,0) + \frac{x_i^0x_j^0}{\sqrt{N}}\mathbb{E}_0\partial_w g(Y,0)g_w^0(Y,0) + \frac{\|\partial_w^{(2)} g(Y,0)\|K}{N} \\
&= \beta_4 + \frac{x_i^0x_j^0}{\sqrt{N}}\beta_2 + \frac{\|\partial_w^{(2)} g^0(Y,0)\|K}{N}
\end{aligned}$$

(2) Similarly, expanding the density we see that

$$\begin{aligned}\sigma_{ij}^2 &= \mathbb{E}_Y[(\partial_w g(Y,0))^2 - \mu_{ij}^2 \mid \boldsymbol{x}^0] \\ &= \int (\partial_w g(Y,0))^2 e^{g^0(Y,0)}(1+O(N^{-1/2})\,dy - \mu_{ij}^2 \\ &= \mathbb{E}_0(\partial_w g(Y,0))^2 - (\mathbb{E}_0 \partial_w g(Y_{ij},0))^2 + \mathbb{E}_0[\partial_w g(Y,0)^2 \partial_w g^0(Y,0)] w_{ij}^0 \\ &= \beta_1 + O\Big(\frac{L\sqrt{\mathbb{E}_0[g(Y,0)^4]\mathbb{E}_0[g_0(Y,0)^2]}}{N^{1/2}}\Big),\end{aligned}$$

(3) Similarly, expanding the density we see that

$$\begin{aligned}\gamma_{ij} &= \mathbb{E}_Y[\partial_w^{(2)} g(Y,0) \mid \boldsymbol{x}^0] \\ &= \int \partial_w^{(2)} g(Y,0) e^{g^0(Y,0)} + O(N^{-1/2})\,dy \\ &= \mathbb{E}_0 \partial_w^{(2)} g(Y,0) + \frac{\|\partial_w^{(2)} g(Y,0)\| K}{N^{1/2}} \\ &= -\beta_3 + \frac{\|\partial_w^{(2)} g(Y,0)\| K}{N^{1/2}}\end{aligned}$$

Heuristically, one can expect that in the large N limit, the first disorder term in $F_N^{(2)}$ can be approximated with a Gaussian with matching mean and variance, we have that

$$g(Y_{ij},0) \approx \sigma_{ij} g_{ij} + \mu_{ij} \approx \sqrt{\beta_1} g_{ij} + \beta_4 + \frac{x_i^0 x_j^0}{\sqrt{N}} \beta_2.$$

Since we have assumed that $\mathbb{E}_{\mathbb{P}_0}[(\partial_w f(Y,0))^3]$ is finite, the substitution can be made precise using approximate Gaussian integration by parts as was applied to prove universality for the SK model (see [16]) for disorder with finite third moments to conclude that

$$|F_N^{(2)}(g;A,B,C) - F_N^{(3)}(g;A,B,C)| \le \frac{K\mathbb{E}_0[g(Y,0)^3]}{\sqrt{N}}.$$

Step 4 - Summary: We can use the triangle inequality and the estimates in steps 1 to steps 3, combined with the fact that $F_N^{(3)}(g;A,B,C) = F_N(\bar{\beta};A,B,C)$ to conclude the statement of the result. □

As a consequence of Proposition A.1, we obtain the following universality result for pseudo maximum likelihood estimation:

Proposition A.2 (Universality of the Ground State)**.** *If our model is well-scored, then*

$$|\mathcal{L}_N^g(A,B,C) - \mathcal{L}_N^{\bar{\beta}}(A,B,C)| = O_L(N^{-1/2}) + o_L(1)$$

where $O_L(N^{-1/2})$ is a term that goes to 0 at rate $N^{-\frac{1}{2}}$ for every fixed L, $o_L(1) \to 0$ uniformly over N and $\bar{\beta} = (\beta_1,\beta_2,\beta_3,\beta_4)$ are the information paramters defined in (2.5), (2.6), (2.7), (2.8).

Proof. This follows from a direct application of the universality at finite temperatures and a careful analysis of the dependencies of the error terms on the norms of g.

The proof is relatively simple consequence of Proposition A.1 since the Fisher score parameters (2.5), (2.6), (2.7), (2.8) all scale linearly with scalar multiplication of g by the definition. This implies that $H_N^{\bar{\beta}}$ is homogeneous in $\bar{\beta}$, i.e. $H_N^{L\bar{\beta}} = LH_N^{\bar{\beta}}$.

By Proposition A.1 we have

$$\left|F_N(Lg;A,B) - F_N(\bar{\beta}_L;A,B)\right| \le \frac{K(Lg,g^0)}{\sqrt{N}},$$

where K is a univeral constant that only depends on g and g^0 and $\bar\beta_L = L\bar\beta$. Then the bounds in (A.3) implies that

$$|\mathcal{L}^g_N(A,B) - \mathcal{L}^{\bar\beta}_N(A,B)| \le |\frac{1}{L}F_N(Lg) - \frac{1}{L}F_N(L\bar\beta)| + o_L(1) \le O_L(N^{-1/2}) + o_L(1)\,.$$

Taking N to infinity, followed by L to infinity then gives the desired result. □

Appendix B. Variational Formula with Constrained Sample Mean

We now extend the earlier result with constrained overlaps in [37, Theorem 2.6] with an additional sample mean constraint. However, in the maximum likelihood setting the signal $\mathbf{x}_0$ is fixed and not random, so the technical details in this proof are simplified despite the inclusion of an extra constraint. The case without a sample mean constraint, which will be required for regular models, is a direct consequence of [37, Theorem 2.6] and will be stated at the end of this section.

By the universality results in Section A, it suffices to work solely with the effective log likelihood given by the information parameters. We also introduce the ridge regression term which will appeared in the corrected models.

$$H_N^{\bar\beta,\alpha}(x) = \frac{\beta_1}{\sqrt{N}}\sum_{1\le i\le j\le N} g_{ij}x_ix_j + \frac{\beta_2}{N}\sum_{1\le i\le j\le N} x_i^0x_j^0x_ix_j - \frac{\beta_3}{2N}\sum_{1\le i\le j\le N} x_i^2x_j^2 + \frac{\alpha}{N}\sum_{1\le i\le j\le N} x_ix_j \tag{B.1}$$

$$= \frac{\beta_1}{\sqrt{N}}\sum_{ij} g_{ij}x_ix_j + \frac{N\beta_2}{2}R_{10}^2 - \frac{N\beta_3}{4}R_{11}^2 + \frac{N\alpha}{2}\bar{x}^2 + o_N(1). \tag{B.2}$$

Let $\varepsilon > 0$, we can define the associated constrained free entropy,

$$F_N(\bar\beta,\alpha,\varepsilon;S,M,v) = \frac{1}{N}\mathbb{E}_Y \log\int \mathbb{1}(\Omega_\varepsilon(S,M,v))e^{H_N^{\bar\beta,\alpha}(\mathbf{x})}\,d\,\mathbb{P}_X^{\otimes N}(\mathbf{x})$$

where we defined the sets

$$A_M = (M-\varepsilon, M+\varepsilon), B_S = (S-\varepsilon, S+\varepsilon), C_v = (v-\varepsilon, v+\varepsilon),$$

and

$$\Omega_\varepsilon(S,M,v) = \{R_{10}\in A_M, R_{11}\in B_S, \bar{x}\in C_v\} \tag{B.3}$$

which implicitly depends on $\mathbf{x}_0$ through the constraint on $R_{1,0}$. To simplify notation, we will often keep the dependence on ε implicit.

The goal of this section is to prove a variational formula for this restricted model. Let $\zeta(t)$ be a c.d.f, and let $\Phi_{\zeta,\mu,\lambda,\rho}(t,y)$ is the solution to Parisi's PDE

$$\begin{cases}\partial_t\Phi_{\zeta,\mu,\lambda,\rho} = -\frac{\beta_1^2}{4}(\partial_y^2\Phi_\zeta + \zeta([0,t])(\partial_y\Phi_\zeta)^2) & (t,y)\in(0,S)\times\mathbb{R}\\ \Phi_{\zeta,\mu,\lambda,\rho}(S,y;x^0) = \log\int e^{yx+\lambda xx^0+\mu x^2+\rho x}\,d\,\mathbb{P}_X(x)\end{cases}. \tag{B.4}$$

Define the Parisi functional

$$\varphi_{\bar\beta}(S,M,v) = \inf_{\mu,\lambda,\rho,\zeta}\left(\mathbb{E}_{\mathbb{Q}}[\Phi_{\zeta,\mu,\lambda,\rho}(0,0;x^0)] - \frac{\beta_1^2}{2}\int_0^S t\zeta(t)\,dt - \mu S - \lambda M - \rho v + \frac{\beta_2M^2}{2} - \frac{\beta_3S^2}{4} + \frac{\alpha v^2}{2}\right).$$

We will see that $(R_{1,1},R_{1,0})$ asymptotically live in the closed subset $\mathcal{C}$ of $[0,C^2]\times[-C^2,C^2]\times[-C,C]$ given by

$$\mathcal{C} = \cap_{a,b,c\in[-1,1]^3}\{(S,M,v) : \mathbb{E}_{\mathbb{Q}}[\mathrm{essinf}_x\{ax^2+bxx^0+cx\}] \le aS+bM+cv \le \mathbb{E}_{\mathbb{Q}}[\mathrm{esssup}_x\{ax^2+bxx^0+cx\}]\}. \tag{B.5}$$

where C is the maximal point in the support of $\mathbb{P}_X$.

Theorem B.1. *For any $\beta_1, \beta_2, \beta_3$ and α and constraints (S, M, v), we have*

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} F_N(\bar\beta, \alpha, \varepsilon; S, M, v) = \varphi_\beta(S, M, v).$$

The proof will be split into two parts. We first begin with a proof of the upper bound of the constrained free entropy. The following proofs are stated in terms of a quantity called the Ruelle probability cascades [65, Chapter 2]. A quick summary of the notation is provided for convenience in Appendix I.

Proposition B.1 (Large Deviation Upper Bound of the Free Energy)**.** *There exists a universal finite constant L such that for every $S, M, v \in \mathcal{C}$, and every real numbers μ, λ, ρ, we have*

$$F_N(\bar\beta : \Omega_\varepsilon(S, M, v)) \le -\lambda S - \mu M - \rho v + \frac{1}{N}\sum_{i=1}^N \Phi_{\lambda,\mu,\zeta}(0,0;x_i^0) - \frac{\beta_1^2}{2}\int_0^S t\zeta(t)\,dt \\ + \frac{\beta_2}{2}M^2 - \frac{\beta_3}{4}S^2 + \frac{\alpha}{2}v^2 + L\varepsilon(|\mu| + |\lambda|) + o_{N,\varepsilon}(1)$$

where $\Phi_{\lambda,\mu,\zeta}(0,0;x_i^0)$ solves the afforementioned PDE (B.24) at parameter x^0. Moreover $o_{N,\varepsilon}(1) = O(\varepsilon) + O(N^{-1})$ is independent of λ, μ.

Proof. This proof follows from the classical Guerra interpolation argument and holds verbatim as the one appearing in [37, Section 4]. There is an extra constraint parameter, but this is dealt by introducing Lagrange multipliers for the sum $\sum_{i=1}^N x_i$. One key difference is that the upper bound is written in terms of the Ruelle probability cascades with an extra Lagrange multiplier parameter $\sum_{i=1}^N \rho x_i$, but this representation is equivalent to (B.24) (see [65, Chapter 4]). □

We now claim that the upper bound is sharp in the sense that after one minimizes over the parameters μ, λ, ρ, the upper bound is equal to the constrained integral.

For $(\lambda, \mu, \rho) \in \mathbb{R}^3$, consider the annealed log Laplace transform

$$\Lambda(\lambda, \mu, \rho) := \int \left(\log \int e^{\lambda x^2 + \mu x x^0 + \rho x}\, d\,\mathbb{P}_X(x)\right) d\,\mathbb{Q}(x^0)$$

and consider the rate function on $\mathbb{R}^3$ given by

$$\mathcal{I}(S, M, v) = \sup_{(\lambda,\mu,\rho)\in\mathbb{R}^2} \{I_{S,M,v}(\lambda, \mu, \rho)\}, \text{ with } I_{S,M,v}(\lambda, \mu) = \lambda S + \mu M + \rho v - \Lambda(\lambda, \mu)\,. \tag{B.6}$$

This quantity gives the entropy of the set $\Omega_\varepsilon(S, M, v)$ under $\mathbb{P}_X$ by [37, Proposition 5.3].

Lemma B.1 (Sharp Lower Bound)**.** *For $(S, M, v) \in \mathcal{C}$ and any $\varepsilon, \delta > 0$ small enough,*

$$\varliminf_{N\to\infty} \frac{1}{N}\mathbb{E}_Z \log \sum_\alpha v_\alpha \int_{\Omega_\varepsilon(S,M,v)} e^{\sum_{i\le N}\beta_1 Z_i(\alpha) x_i}\, d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x}) \\ \ge \inf_{\mu,\lambda,\rho}\Big(-\lambda S - \mu M - \rho v + \mathbb{E}_{Z,\mathbb{Q}} \log \sum_\alpha v_\alpha \int e^{\beta_1 Z(\alpha) x + \lambda x^2 + \mu x x^0 + \rho x}\, d\,\mathbb{P}_X(\boldsymbol{x})\Big). \tag{B.7}$$

Moreover, the right hand side is equal to $-\infty$ if $\mathcal{I}(S, M, v) = \infty$. Furthermore, if S, M, v belong to the interior of $\mathcal{C}$, then the minimizer is attained at a unique μ and λ, such that $|\mu| + |\lambda| + |\rho| \le C(S, M, v)$ where the constant C only depends on the distance from (S, M, v) to the boundary.

Proof. A similar result is proved in [66, Section 7] and [37, Lemma 5.4]. We follow the proof of Gartner-Ellis theorem [23, Theorem 2.3.6], taking into account the random density depending on the Z_i's.

We first show that we can restrict ourselves to (S, M, v) with finite entropy because the lower bound in (B.7) is infinite otherwise. Indeed,

$$\mathbb{E}_Z \log \sum_\alpha v_\alpha \int e^{\beta_1 Z(\alpha)x+\lambda x^2+\mu x x^0+\rho x}\, d\,\mathbb{P}_X(\boldsymbol{x})$$

$$\leq \mathbb{E}_Z \log \sum_\alpha v_\alpha \int e^{\beta_1 |Z(\alpha)| C} + \mathbb{E}_{\mathbb{Q}} \log \int e^{\lambda x^2+\mu x x^0+\rho x}\, d\,\mathbb{P}_X(\boldsymbol{x})\,,$$

and $\mathbb{E}\log\sum_\alpha v_\alpha e^{\beta_1|Z(\alpha)|C}$ is bounded uniformly, so by the properties of the Ruelle probability cascades,

$$\mathbb{E} e^{|\sum_{k=1}^r (Q_k^2-Q_{k-1}^2)^{1/2} z_k| C} < \infty$$

using the moment generating function for folded normals. Therefore there exists a finite constant L such that

$$\inf_{\mu,\lambda}\Big(-\lambda S - \mu M - \rho v + \mathbb{E}_{Z,\mathbb{Q}} \log \sum_\alpha v_\alpha \int e^{\beta_1 Z(\alpha)x+\lambda x^2+\mu x x^0+\rho x}\, d\,\mathbb{P}_X(\boldsymbol{x})\Big) \leq -\mathcal{I}(S,M,v) + L.$$

Thus we may restrict to values of (S, M, v) with finite entropy.

We next adapt the Gartner-Ellis argument [23, Section 2.3] to our setting. It is based on a large deviation upper bound for certain titled measures. Namely let $\lambda, \mu, \rho \in \mathbb{R}^3$. We show for every $(S, M, v) \in [0, C^2] \times [-C^2, C^2] \times [-C, C]$,

$$\begin{aligned}\overline{\lim_{N\to\infty}} \frac{1}{N}\mathbb{E}_{Z,x^0} \log \frac{\sum_\alpha v_\alpha \int_{\Omega_\varepsilon(S,M,v)} e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\, d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha \int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\, d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})} \\ \leq -\Lambda^*_{\lambda,\mu}(S,M,v) + O(\varepsilon)\,,\end{aligned} \tag{B.8}$$

with

$$\Lambda^*_{\lambda,\mu}(S,M) = -\lambda S - \mu M - \rho v + \Lambda(\mu,\lambda) + \sup_{\lambda',\mu',\rho'}\{\lambda' S + \mu' M + \rho' v - \Lambda(\lambda',\mu',\rho')\}\,,$$

where

$$\Lambda(\lambda,\mu) = \mathbb{E}_{Z,\mathbb{Q}} \log \sum_\alpha v_\alpha \int e^{\beta_1 Z(\alpha)x+\lambda x^2+\mu x x^0+\rho x}\, d\,\mathbb{P}_X(\boldsymbol{x})\,.$$

We denote in short $\Lambda^* = \Lambda^*_{0,0,0}$. Indeed, (B.8) is a direct consequence of the fact that the v_α are non-negative, and almost surely we have

$$\int_{\Omega_\varepsilon(S,M)} e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\, d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})$$

$$\leq e^{N(\lambda-\lambda')S+N(\mu-\mu')M+N(\rho-\rho')v+NO(\varepsilon)} \int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda' x_i^2+\mu' x_i x_i^0+\rho' x_i)}\, d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})\,.$$

We next introduce the notion of exposed points: (S, M, v) is said to be exposed if and only if there exists (λ, μ, ρ) such that for every $(S', M', v') \neq (S, M, v)$ we have

$$\lambda S + \mu M + \rho v - \Lambda^*(S,M,v) > \lambda S' + \mu M' + \rho' v - \Lambda^*(S',M',v') = -\Lambda^*_{\lambda,\mu,\rho}(S',M',v') + \Lambda(0,0,0)\,. \tag{B.9}$$

The set (λ, μ, ρ) is called an exposing hyperplane. We first prove (B.8) for an exposed point (S, M, v) with exposing hyperplane (λ, μ, ρ) by showing that the associated tilted measure puts some mass on a neighborhood of (S, M, v), see (B.12). To see this, we first claim that for every $(S', M', v') \neq (S, M, v)$,

$$\begin{aligned}\Lambda^*_{\lambda,\mu,\rho}(S',M',v') &= \Lambda^*(S',M',v') - (\lambda S' + \mu M' + \rho v - \Lambda(\mu,\lambda,\rho) + \Lambda(0,0,0)) \\ &\geq \Lambda^*(S',M',v') - (\lambda(S'-S) + \mu(M'-M) + \rho(v-v') + \Lambda^*(S,M,v)) > 0\end{aligned}$$

Moreover, it is easy to see that $\Lambda^*_{\lambda,\mu}$ is a good rate function so that it achieves its minimum value on the closure $\bar\Omega_\varepsilon(S,M,v)^c$ of $\Omega_\varepsilon(S,M,v)^c$, hence $\inf_{\bar\Omega_\varepsilon(S,M,v)^c}\Lambda^*_{\lambda,\mu,\rho}\geq\kappa>0$. Moreover, we can cover $\bar\Omega_\varepsilon(S,M,v)^c$ by a union of finitely many balls $(B_j)_{j\leq K}$ so that for each $j\leq K$

$$\begin{aligned}&\varlimsup_{N\to\infty}\frac{1}{N}\mathbb{E}_Z\log\frac{\sum_\alpha v_\alpha\int_{(R_{1,1},R_{1,0},\bar x)\in B_j}e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}\\&\quad\leq-\kappa+O(\delta)\,.\end{aligned}\tag{B.10}$$

Therefore, there exists $\kappa=\kappa_\varepsilon>0$ such that

$$\begin{aligned}&\varlimsup_{N\to\infty}\frac{1}{N}\mathbb{E}_Z\log\frac{\sum_\alpha v_\alpha\int_{\bar\Omega_\varepsilon(S,M)^c}e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}\\&\leq\varlimsup_{N\to\infty}\frac{1}{N}\mathbb{E}_Z\log\sum_{j\leq K}\frac{\sum_\alpha v_\alpha\int_{B_j}e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}\\&\leq-\kappa\,,\end{aligned}\tag{B.11}$$

where in the last step we use Lemma I.3 to pull the sum outside of the logarithm. Applying Lemma I.3 again, we conclude that

$$\begin{aligned}0&=\varliminf_{N\to\infty}\frac{1}{N}\mathbb{E}_Z\log\frac{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}\\&\leq\max\left\{\varliminf_{N\to\infty}\frac{1}{N}\mathbb{E}_Z\log\frac{\sum_\alpha v_\alpha\int_{\Omega_\varepsilon(S,M)}e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})},-\kappa+\delta\right\}\end{aligned}$$

and therefore for δ small enough (depending on ε)

$$\varliminf_{N\to\infty}\frac{1}{N}\mathbb{E}_Z\log\frac{\sum_\alpha v_\alpha\int_{\Omega_\varepsilon(S,M,v)}e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}}\geq 0.\tag{B.12}$$

(B.7) then follows. Indeed, by Hölder's inequality

$$\begin{aligned}&\frac{1}{N}\mathbb{E}_Z\log\sum_\alpha v_\alpha\int_{\Omega_\varepsilon(S,M,v)}e^{\sum_{i\leq N}\beta_1 Z_i(\alpha)x_i}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})\\&\geq-\lambda S-\mu M-\rho v+\frac{1}{N}\mathbb{E}_Z\log\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})\qquad\text{(B.13)}\\&+\frac{1}{N}\mathbb{E}_Z\log\frac{\sum_\alpha v_\alpha\int_{\bar\Omega_\varepsilon(S,M)}e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})}{\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i)}\,d\,\mathbb{P}_X^{\otimes N}}+O(\varepsilon).\qquad\text{(B.14)}\end{aligned}$$

By weak convergence we also have that the absolute difference between

$$\frac{1}{N}\mathbb{E}_Z\log\sum_\alpha v_\alpha\int e^{\sum_{i\leq N}(\beta_1 Z_i(\alpha)x_i+\lambda x_i^2+\mu x_i x_i^0+\rho x_i}\,d\,\mathbb{P}_X^{\otimes N}(\boldsymbol{x})$$

$$\mathbb{E}_{Z,\mathbb{Q}}\log\sum_\alpha v_\alpha\int e^{(\beta_1 Z(\alpha)x+\lambda x^2+\mu x x^0+\rho x_i}\,d\,\mathbb{P}_X(\boldsymbol{x})$$

tends to zero as $N\to\infty$. Hence, letting N go to infinity, δ to zero and then ε to zero we arrive at the desired statement.

To conclude that the lower bound holds not only for exposed points we appeal to Rockafellar's lemma, see [23, Lemma 2.3.12], which shows that it suffices to prove that Λ is essentially smooth, lower semi-continuous and convex. This follows as $\mathbb{P}_X$ and $\mathbb{Q}$ are compactly supported. Consequently, the relative interior of the set of points where Λ^* is finite is included in the set of exposed points, and so by our earlier reduction to points with finite entropy, the Lemma is proven. □

The rest of the proof of the lower bound can be adapted from [37]. The main difference is that in our setting $\mathbf{x}_0$ is non-random, while in the Bayesian setting, there is a prior on $\mathbf{x_0}$. The current setting with non-random $\mathbf{x}_0$ is actually much simpler, and can the lower bound can be proved using the classical perturbations without localizing $\mathbf{x}_0$ around typical values. We sketch the key steps below.

Proposition B.2 (Lower Bound of the Free Energy)**.** *For any real numbers $\beta_1, \beta_2, \beta_3$, for any $(S, M, v) \in \mathcal{C}$, for any $\varepsilon > 0$, we have*

$$\varliminf_{N\to\infty} F_N(\bar\beta, \varepsilon; S, M, v) \geq \varphi_{\bar\beta}(S, M, v) + O(\varepsilon)$$

Proof. The key ideas of the proof is similar to the ones used to derive the lower bound of the Sherrington–Kirkpatrick model. The approximation techniques used to deal with the random constraint set in [37] is also not needed in this setting, since $\mathbf{x}_0$ is fixed and non-random. We summarize the key steps.

Step 1: We first introduce a perturbation of the likelihood function that will allow us to characterize its limiting distribution. To introduce the perturbed Hamiltonian let us first fix the self-overlap by setting

$$\hat{\boldsymbol{x}} = \frac{\sqrt{SN}}{\|\boldsymbol{x}\|_2}\boldsymbol{x} \tag{B.15}$$

The entries of $\hat{\boldsymbol{x}}$ are still uniformly bounded for $\boldsymbol{x}$ so that $R_{1,1} = \frac{1}{N}\|x\|_2^2$ is at ε distance of S, provided $\varepsilon < S/2$. Throughout D will denote such a uniform bound (which depends on S and C). For $p \geq 1$, consider

$$g_p(\hat{\boldsymbol{x}}) = \frac{1}{N^{p/2}} \sum_{i_1,\ldots,i_p} g_{i_1,\ldots,i_p} \hat{x}_{i_1} \cdots \hat{x}_{i_p} \,,$$

and the Gaussian process

$$g(\hat{\boldsymbol{x}}) = \sum_{p\geq 1} 2^{-p} D^{-p} t_p g_p(\hat{\boldsymbol{x}}) \,, \tag{B.16}$$

where the $g_{i_1,\ldots,i_p}$ are independent standard Gaussians and $(t_p)_{p\geq 1}$ is a sequence of parameters such that $t_p \in [0, 3]$ for all $p \geq 1$. Notice that the covariance is bounded

$$\mathbb{E} g(\hat{\boldsymbol{x}}^1) g(\hat{\boldsymbol{x}}^2) = \sum_{p\geq 1} 4^{-p} D^{-2p} t_p^2 (\frac{1}{N}\sum_{i=1}^N \hat{x}_i^1 \hat{x}_i^2)^p \leq \sum_{p\geq 1} 4^{-p} D^{-2p} t_p^2 D^{2p} \leq 3 \,, \tag{B.17}$$

where the first inequality uses $R_{1,2} = \frac{1}{N}\sum \hat{x}_i^1 \hat{x}_i^2 \leq C^2$. For $s > 0$, we define the interpolating Hamiltonian as

$$H_N^{\text{pert}}(\boldsymbol{x}) = H_N^{SK}(\boldsymbol{x}) + s g(\hat{\boldsymbol{x}}). \tag{B.18}$$

A key consequence is that under the perturbed likelihood function, samples from the posterior will satisfy a concentration inequality called the Ghirlanda–Guerra identities.

Theorem B.2 (Ghirlanda–Guerra Identities)**.** *Let* $\hat{R}_{k,\ell} = \frac{1}{N}\sum_{i=1}^N \hat{x}_i^k \hat{x}_i^\ell$*. If* $s = N^\gamma$ *for* $1/4 < \gamma < 1/2$*, then*

$$\lim_{N\to\infty} \mathbb{E}_u \left| \mathbb{E}\langle f \hat{R}^p_{1,n+1}\rangle - \frac{1}{n}\mathbb{E}\langle f\rangle \mathbb{E}\langle \hat{R}^p_{1,2}\rangle - \frac{1}{n}\sum_{\ell=2}^n \mathbb{E}\langle f \hat{R}^p_{1,\ell}\rangle \right| = 0\,,$$

for any $p \geq 1$*,* $n \geq 2$ *and bounded measurable function* f *of the* $n \times n$ *sub array of the overlaps.*

Step 2: We now compute the limit by showing that the limit can be expressed as functions of samples from the posterior, which we hav a limiting characterization of. This is commonly known as the Aizenman–Sims–Starr scheme or cavity computations in statistical physics. We have for every $n \geq 1$,

$$\varliminf_{N\to\infty} F_N(\bar{\beta}, \varepsilon; S, M, v) \geq \varliminf_{N\to\infty} \frac{1}{n}\mathbb{E}\log Z_{N+n} - \frac{1}{N}\mathbb{E}\log Z_N\,,$$

where

$$Z_N = \int \mathbb{1}(\Omega_\varepsilon(S,M,v)) e^{H_N^{\bar{\beta},\alpha}(\mathbf{x})}\, d\,\mathbb{P}_X^{\otimes N}(\mathbf{x})\,.$$

Let $(\mathbf{x}, \boldsymbol{y}) \in \mathbb{R}^{N+n}$. We decompose H_{N+1} into terms that depend on the cavity coordinate x and its bulk terms. Consider the following cavity fields defined with respect to the modified coordinates $\hat{x}_i = \sqrt{(N+n)S}x_i/\|x\|_2$ (see (B.15)):

$$H^{\text{pert}}_{N,n}(\boldsymbol{x}) := \sum_{1\leq i<j\leq N} \beta \frac{W_{ij}}{\sqrt{(N+n)}} x_i x_j + s g_N(\hat{\boldsymbol{x}}), \tag{B.19}$$

$$z_i(\hat{\boldsymbol{x}}) = \frac{\beta}{\sqrt{N}} \sum_{j=1}^N W_{j,N+i}\hat{x}_j, \tag{B.20}$$

$$y(\hat{\boldsymbol{x}}) = \frac{\sqrt{n}\beta}{N} \sum_{1\leq i<j\leq N} W_{ij}\hat{x}_i\hat{x}_j\,. \tag{B.21}$$

By matching the covariances, we can see that

$$H_{N+n}(\boldsymbol{x},\boldsymbol{y}) \approx \sum_{i=1}^N z_i(\hat{\boldsymbol{x}})\hat{y}_i + H^{\text{pert}}_{N,n}(\boldsymbol{x})\,,$$

and

$$H_N(\boldsymbol{x},\boldsymbol{y}) \approx y(\hat{\boldsymbol{x}}) + H^{\text{pert}}_{N,n}(\boldsymbol{x})\,,$$

so

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\log Z_{N+n} - \frac{1}{N}\mathbb{E}\log Z_N \\
&\geq \frac{1}{n}\Bigg(\mathbb{E}\log\Bigg\langle \int \mathbb{1}(|R_{1,1}(\boldsymbol{y}) - S| \leq \varepsilon, |R_{1,0}(\boldsymbol{y}) - M| \leq \varepsilon, |\bar{\boldsymbol{y}} - v| \leq \varepsilon) e^{\sum_{i=1}^n z_i(\hat{\boldsymbol{x}})y_i}\, d\,\mathbb{P}_X^{\otimes n}(\boldsymbol{y})\Bigg\rangle^{\text{pert}}_{N,n} \\
&- \mathbb{E}\log\Big\langle e^{y(\hat{\boldsymbol{x}})}\Big\rangle^{\text{pert}}_{N,n}\Bigg) + o(1)\,,
\end{aligned}$$

where $\langle\cdot\rangle^{\text{pert}}_{N,n}$ denotes the average with respect to $H^{\text{pert}}_{N,n}(\boldsymbol{x})$. This lower bound can be approximated by a continuous function of finitely many samples from $\langle\rangle^{\text{pert}}_{N,n}$.

Lemma B.2 (Continuity of the Lower Bound with Respect to the Overlaps)**.** *Let* $\langle\cdot\rangle$ *be the average with respect to some non-random Gibbs measure* $\mathbb{G}$ *on the sphere with radius* $\sqrt{S}$ *in some*

Hilbert space H. Consider the Gaussian processes $Z(\boldsymbol{\sigma})$ and $Y(\boldsymbol{\sigma})$ indexed by points $\boldsymbol{\sigma}$ in H with covariances

$$\mathbb{E}Z(\boldsymbol{\sigma}^1)Z(\boldsymbol{\sigma}^2) = \langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle \qquad \mathbb{E}Y(\boldsymbol{\sigma}^1)Y(\boldsymbol{\sigma}^2) = \frac{\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle}{2}$$

Let n be a fixed integer number and (S, M) with finite entropy I so that there exists a finite constant c independent of n and ε such that for n large enough, $\mathbb{P}_X^{\otimes n}(\Omega_\varepsilon(S, M)) \geq e^{-cn}$ uniformly for all $\boldsymbol{y}_0$ with limiting empirical distribution $\mathbb{Q}$. Then the functionals

$$f_n^Z(S, M) = \frac{1}{n}\mathbb{E}_Z \log \left\langle \in \mathbb{1}(|R_{1,1}(\boldsymbol{y}) - S| \leq \varepsilon, |R_{1,0}(\boldsymbol{y}) - M| \leq \varepsilon, |\bar{\boldsymbol{y}} - v| \leq \varepsilon) e^{\sum_{i=1}^n Z_i(\boldsymbol{\sigma}) y_i} \, d\,\mathbb{P}_X^{\otimes n}(\boldsymbol{y}) \right\rangle$$

where Z_i are independent copies of Z and

$$f_n^Y = \frac{1}{n}\mathbb{E}_Z \log \left\langle e^{\sqrt{n}\beta Y(\boldsymbol{\sigma})} \right\rangle,$$

are continuous functionals of the distribution of the overlap array $(\boldsymbol{x}^\ell \cdot \boldsymbol{x}^{\ell'})_{\ell,\ell' \geq 1}$ under $\mathbb{G}^{\otimes\infty}$. In particular, for any $\eta > 0$ there exists a finite integer number $K(\eta)$ so that these functionals can be approximated by a continuous function of the finite array $(\boldsymbol{x}^\ell \cdot \boldsymbol{x}^{\ell'})_{1 \leq \ell,\ell' \leq K(\eta)}$ uniformly over all possible choices of Gibbs measures $\mathbb{G}$ and all $\boldsymbol{y}^0$ limiting empirical distribution $\mathbb{Q}$.

Step 3: We now identify the limit of this lower bound. Since R^∞ satisfies the Ghirlanda–Guerra identities, the distribution of the entire array is determined by $\zeta(t) = \mathbb{P}(R_{1,2}^\infty \leq t)$ [65, Theorem 2.13 and Theorem 2.17]. We can approximate $\zeta(t)$ in L^1 with a piecewise constant function $\mu(t)$, so that

$$\int |\zeta(t) - \mu(t)| \, dt < \varepsilon.$$

The density function μ of a measure can be encoded by the parameters

$$\zeta_{-1} = 0 < \zeta_0 < \cdots < \zeta_{r-1}, \tag{B.22}$$

and sequence

$$0 = Q_0 \leq Q_1 \leq \cdots \leq Q_{r-1} \leq Q_r = S. \tag{B.23}$$

That is, these sequences define the density function

$$\mu(Q) = \zeta_k \qquad \text{for} \qquad Q_k \leq Q < Q_{k+1}.$$

Let v_α denote the weights of the Ruelle probability cascades corresponding to the sequence (B.22). If $(\alpha^\ell)_{\ell \geq 1}$ are samples from the Ruelle probability cascades, then $\mathbb{P}(\alpha^1 \wedge \alpha^2 \leq t) = \mu(t)$ by construction. This gives us an explicit way to construct the off-diagonal entries of the overlap array in the limit. We define Gaussian processes $Z(\alpha)$ and $Y(\alpha)$ with covariance

$$\mathbb{E}Z(\alpha^1)Z(\alpha^2) = Q_{\alpha^1 \wedge \alpha^2} \quad \mathbb{E}Y(\alpha^1)Y(\alpha^2) = \frac{1}{2}Q^2_{\alpha^1 \wedge \alpha^2}$$

and let Z_i for $1 \leq i \leq n$ denote independent copies of Z. The functionals

$$f_n^Z(\mu) = \frac{1}{n}\mathbb{E}\log\sum_\alpha v_\alpha \int \mathbb{1}(|R_{1,1}(\boldsymbol{y}) - S| \leq \varepsilon, |R_{1,0}(\boldsymbol{y}) - M| \leq \varepsilon, |\bar{\boldsymbol{y}} - v| \leq \varepsilon) e^{\sum_{i \leq n} \beta Z_i(\alpha) y_i} \, d\,\mathbb{P}_X^{\otimes n}(\boldsymbol{y})$$

and

$$f_n^Y(\mu) = \frac{1}{n}\mathbb{E}\log\sum_\alpha v_\alpha e^{\sqrt{n}\beta Y(\alpha)},$$

are of the same form as the functionals in Lemma B.2 because they depend on the overlap array in exactly the same way. Furthermore, one can show that they are Lipschitz continuous [65, Lemma 4.1].

Step 4: To remove the constraint and identity the limit with its matching upper bound, we can apply Lemma B.1 to finish the proof. □

B.1. **Simplification when** $\alpha = 0$**.** When $\alpha = 0$, which corresponds to regular models, then the variational formula can be expressed in a simpler form. When $\alpha = 0$, the constraint on v is unnecessary and we instead define

$$\begin{cases} \partial_t \Phi_{\zeta,\mu,\lambda,0} = -\frac{\beta_1^2}{4}(\partial_y^2 \Phi_\zeta + \zeta([0,t])(\partial_y \Phi_\zeta)^2) & (t,y) \in (0,S) \times \mathbb{R} \\ \Phi_{\zeta,\mu,\lambda,0}(S,y;x^0) = \log \int e^{yx+\lambda x x^0 + \mu x^2} \, d\,\mathbb{P}_X(x) \end{cases}. \tag{B.24}$$

Define the Parisi functional

$$\varphi_{\bar{\beta}}(S,M) = \inf_{\mu,\lambda,\zeta} \left(\mathbb{E}_{\mathbb{Q}}[\Phi_{\zeta,\mu,\lambda,0}(0,0;x^0)] - \frac{\beta_1^2}{2} \int_0^S t\zeta(t)\,dt - \mu S - \lambda M + \frac{\beta_2 M^2}{2} - \frac{\beta_3 S^2}{4} \right). \tag{B.25}$$

With a slight abuse of notation, we notice that $\varphi_{\bar{\beta}}(S,M)$ and $\varphi_{\bar{\beta}}(S,M,v)$ are almost identical, but the former no longer depends on α, v, ρ. The next theorem shows that the maximum for regular models converges to $\sup \varphi_\beta(S,M)$.

Theorem B.3. *For any* $\beta_1, \beta_2, \beta_3$ *and* α *and constraints* (S,M,v)*, we have*

$$\lim_{\varepsilon \to 0} \lim_{N \to \infty} F_N(\bar{\beta}, \alpha, \varepsilon; S, M, v) = \varphi_\beta(S,M,v).$$

If $\alpha = 0$*, then for any constraints* $(S,M) \in \mathcal{C}$

$$\lim_{\varepsilon \to 0} \lim_{N \to \infty} F_N(\bar{\beta}, 0, \varepsilon; S, M, v) = \varphi_\beta(S,M)$$

Proof. This is a direct consequence of [37, Theorem 2.6]. One slight difference is that in the setting of MLE, $\mathbf{x}_0$ is taken to be non-random while in the Bayesian setting $\mathbf{x}_0$ is drawn from some prior $\mathbb{P}_*^{\otimes N}(\mathbf{x}_0)$. However, this is not an issue because the proof of [37] holds conditionally on a realization of $\mathbf{x}_0$, and we can simply view $\mathbf{x}_0$ as a realization of a sample from the limiting measure $\mathbb{Q}$. □

Appendix C. Gamma Convergence of Local Free Energies

In this section we show that the local quantities computed by taking the limit as N tends to infinity of $\frac{1}{NL} F_N^L(L\bar{\beta}, \varepsilon; S, M, v)$, are Γ convergent as L tends to infinity to $\psi_{\bar{\beta}}(S,M,v)$. We prove this result in the case when $\beta_4 = 0$ to simplify notation, but note that in the case where $\beta_4 \neq 0$, the modification is simple. We point out where the modifications are necessary as we go along.

We recall the following result in [37]. Let $\mathbb{P}_X$ denote either normalized Lebesgue measure on counting measure depending on if Ω is an interval or discrete. We consider the finite temperature free energy given by:

$$F_N(\overline{\beta}) = \lim_{N \to \infty} \frac{1}{N} \mathbb{E} \log \int_{\Omega^N} \exp\left(\frac{\beta_1}{\sqrt{N}} \sum_{ij} g_{ij} x_i x_j + \frac{\beta_2}{N} \sum_{ij} x_i^0 x_j^0 x_i x_j - \frac{\beta_3}{2N} \sum_{ij} x_i^2 x_j^2 \right) d\,\mathbb{P}_X^{\otimes N},$$

where x_i^0 are the entries of our rank one signal. Then $F(\overline{\beta})$ can be computed by solving the variational problem in Theorem B.3. defined by

$$\lim_{N \to \infty} F_N(\bar{\beta}) = F(\overline{\beta}) = \sup_{(S,M) \in \mathcal{C}} \varphi_{\bar{\beta}}(S,M)$$

where φ is defined in (B.25). In order to compute the limit of the pseudo MLE we must compute the quantity:

$$\lim_{L \to \infty} \mathbb{E}_{x^0 \sim \mathbb{Q}} \frac{1}{L} F(L\overline{\beta}),$$

and we shall do so by means of Γ convergence. For fixed $0 \le t \le S$, $h, y \in \mathbb{R}$ we define functionals $F_L(\zeta,\lambda,\mu)$ by:

$$\mathcal{F}_{L,S}(\zeta,\mu,\lambda;t,y,h) = \begin{cases} \Phi^L_{\zeta,\lambda,\mu}(t,y) & \text{if } \zeta = L\rho(t)dt \\ +\infty & \text{otherwise} \end{cases},$$

where $\Phi^L_{\zeta,\lambda,\mu}$ is the weak solution to the Parisi PDE:

$$\begin{cases} \partial_t\Phi + \frac{\beta_1}{4}(\Delta\Phi + L\rho(s)(\partial_y\Phi)^2) = 0 \\ \Phi(S,y) = \frac{1}{L}\log\int e^{L(yx+\lambda xh+\mu x^2)}d\,\mathbb{P}_X(x) =: f_L(y,\lambda,\mu) \end{cases}. \tag{C.1}$$

In [45] , the authors showed the following theorem:

Theorem C.1. *Fix t, y, h, then the sequence F_L is Γ-convergent to the functional F. In particular the following hold:*

(1) $(\Gamma - \underline{\lim})$ *For any sequence $(\zeta_L,\lambda_L,\mu_L) \to (\zeta,\lambda,\mu)$ we have the inequality:*

$$\underline{\lim}_{L\to\infty} F_L(\zeta_L,\lambda_L,\mu_L;t,y,h) \ge F(\zeta,\lambda,\mu;t,y,h)\,.$$

(2) $(\Gamma - \overline{\lim})$ *For any (ζ,λ,μ) there is a recovery sequence, i.e, there is $(\zeta_L,\lambda_L,\mu_L) \to (\zeta,\lambda,\mu)$ such that:*

$$\lim_{L\to\infty} F_L(\zeta_L,\lambda_L,\mu_L;t,y,h) = F(\zeta,\lambda,\mu;t,y,h)\,,$$

and furthermore the recovery sequence $(\zeta_L,\lambda_L,\mu_L)$ can be taken as (ζ_L,λ,μ) with ζ_L independent of the choice of t, y, h.

Remark C.1. Theorem C.1 remains true if the additional Lagrange multiplier corresponding to fixed magnetization v is added. Additionally the recovery sequence in the Γ-limsup condition can still be taken to be $(\zeta_L,\lambda,\mu,\eta)$ with ζ_L independent of the choice of h.

With this theorem in hand we may complete the proof of Γ-convergence to show $\mathbb{E}_{h\sim\mathbb{Q}}\mathcal{F}_L \to \mathbb{E}_{h\sim\mathbb{Q}}\mathcal{F}$.

Lemma C.1. *The functionals defined by $\mathbb{E}\mathcal{F}_L$ are Γ convergent to $\mathbb{E}\mathcal{F}$.*

Proof. To prove the $\Gamma\ \underline{\lim}$ inequality note that Theorem C.1 implies for every sequence $(\zeta_L,\lambda_L,\mu_L) \to (\zeta,\lambda,\mu)$ one has $\mathbb{Q}$ almost surely that

$$\underline{\lim}_{L}\,\Phi_{\zeta_L,\lambda_L,\mu_L}(t,y) \ge \Phi_{\zeta,\lambda,\mu}(t,y)\,,$$

and hence by Fatou's lemma one has:

$$\begin{aligned}\mathbb{E}F(\zeta,\lambda,\mu,t,y) = \mathbb{E}_{\mathbb{Q}}\Phi_{\zeta,\lambda,\mu)} &\le \mathbb{E}_{\mathbb{Q}}\,\underline{\lim}_{L}\,\Phi_{\zeta_L,\lambda_L,\mu_L} \\ &\le \underline{\lim}_{L}\,\mathbb{E}\Phi^L_{\zeta_L,\lambda_L,\mu_L}(t,y) \\ &= \underline{\lim}_{L}\,\mathbb{E}_{h\sim\mathbb{Q}}\;\mathcal{F}_L(\zeta_L,\lambda_L,\mu_L,t,y,h)\,.\end{aligned}$$

To prove the $\Gamma\ \overline{\lim}$ inequality we note that part (b) of Theorem C.1 implies the recovery sequence can be taken to be (ζ_L,λ,μ) with ζ_L independent of the realization of $\mathbb{Q}$. Lastly, we note that initial condition of (C.1) at finite L satisfies a uniform upper bound

$$f_L(y,\lambda,\mu,c) \le Cy + D\,,$$

for some constants $C, D > 0$. Consequently if we define $\hat{\Phi}^L_{\zeta_L,\lambda_L,\mu_L}(t,y)$ to be the solution to the Parisi PDE with initial condition given by:

$$\hat{f}(y) = \max_{y\in\Omega}(Cy + D)\,,$$

then $\mathbb{Q}$ almost surely we have the pointwise bound:

$$\Phi^L_{\zeta_L,\lambda,\mu}(t,y) \leq \hat{\Phi}^L_{\zeta_L,\lambda,\mu}(t,y)\,.$$

Indeed, the Parisi PDE is monotonic in the initial condition as the difference to any two solutions with different initial conditions solves a linear heat equation with non-negative first order term, and solutions to the heat equation with non-negative initial values are non-negative for all time. It is immediate to check the convergence

$$\hat{\Phi}^L_{\zeta_L,\lambda,\mu}(t,y) \to \hat{\Phi}_{\zeta,\lambda,\mu}\,,$$

where $\hat{\Phi}_{\zeta,\lambda,\mu}$ solves the zero temperature Parisi PDE with initial condition given by $\hat{f}(y)$. Consequently the generalized dominated convergence theorem implies that

$$\lim_{L\to\infty} \mathbb{E}_{\mathbb{Q}}\Phi^L_{\zeta_L,\lambda,\mu}(t,y) = \Phi_{\zeta,\lambda,\mu}(t,y)\,,$$

which completes the proof. □

As a consequence of the Γ-convergence in Lemma C.1, we have that ψ^L converges to ψ point wise on $\mathcal{C}$. We will establish in Sections D and E that these quantities in fact converge uniformly.

We conclude this section by proving an appropriate upper semi-continuity statement

Lemma C.2. *Suppose that $(S_L, M_L, v_L) \to (S, M, v)$, with $\Psi(S, M, v) > -\infty$. Then one has*

$$\overline{\lim}\, \Psi_L(S_L, M_L, v_L) \leq \Psi(S, M, v)\,,$$

and furthermore:

$$\overline{\lim}\, \Psi(S_L, M_L, v_L) \leq \Psi(S, M, v)\,.$$

Proof. For notation write $A_L = (S_L, M_L, v_L)$, $A = (S, M, v)$. By the assumption that $\Psi(A) > -\infty$, one has for any $\varepsilon > 0$ a pair $(\zeta, \lambda, \mu, \rho) \in \mathcal{A}_S \times \mathbb{R}^3$ such that:

$$\Psi(A) + \varepsilon \geq \mathcal{P}_A(\zeta, \lambda, \mu, \rho)\,.$$

(Here $\mathcal{P}_A$ is the zero temp parisi functional, we will write $\mathcal{P}_{A_L}$ for the finite temperature one evaluated at A_L.).

In order to prove the result it will suffice to show the following, there is a sequence $(\zeta_L, \lambda_L, \mu_L, \rho_L) \in \mathcal{A}_{S_L} \times \mathbb{R}^3$, such that

$$\lim_{L\to\infty} \mathcal{P}_{A_L}(\zeta_L, \lambda_L, \mu_L, \rho_L) = \mathcal{P}_A(\zeta, \lambda, \mu, \rho)\,.$$

We may take $\lambda_L = \lambda, \mu_L = \mu$ and $\rho_L = \rho$, so it will simply suffice to construct a recovery sequence $\zeta_L \to \zeta$. Write $\zeta = m(t)dt + c\delta_S$, then we have the following construction from [47, Lemma 2.1.2]

Lemma C.3. *Define c_L by:*

$$c_L = \begin{cases} c & \text{if } \zeta(\{S\}) = c > 0 \\ \frac{1}{L} & \text{otherwise} \end{cases}$$

For L sufficiently large, there is $q_L \in (0, \min(S_L, S))$ such that the following hold:

(1) $\int_{q_L}^{\min(S_L,S)} m(t)dt + c_L = L(\min(S_L, S) - q_L)$.
(2) $q_L \to S$.
(3) $m(q_L)/L \leq 1$.

The proof is the same as the one in [47].

With the sequence q_L defined, we can now define ζ_L as follows:

$$\zeta_L = \begin{cases} \frac{m(t)}{L} & 0 \leq t \leq q_L \\ 1 & q_L < t < S_L \end{cases}\,,$$

and immediately one has $L\zeta_L \to \zeta$. Now let Φ and Φ^L denote solutions to the initial value problems:

$$\begin{cases} \partial_t \Phi^L + \frac{\beta_1^2}{4}(\Phi^L_{yy} + L\zeta_L(s)(\Phi^L_y)^2) = 0 & (t,y) \in [0,S_L] \times \mathbb{R} \\ \Phi^L(S_L, y) = \frac{1}{L}\log \int e^{L(yx+\lambda x h+\mu x^2+\rho x)} d\,\mathbb{P}_X(x) \end{cases}$$

$$\begin{cases} \partial_t \Phi + \frac{\beta_1^2}{4}(\Phi_{yy} + m(t)(\Phi_y)^2) = 0 & (t,y) \in [0,S] \times \mathbb{R} \\ \Phi(S, y) = \max_{x\in\Omega}(yx+\lambda x h+\mu x^2+\rho x) \end{cases}$$

We claim for any $h \in \mathbb{R}$ one has $\Phi^L(0,0) \to \Phi(0,0)$. The claim is proven by following [45] Lemma 3.3 with a simple modification.

To finish the proof we proceed as follows. By construction we have that:

$$\mathcal{P}_{A_L}(\zeta_L, \lambda, \mu, \rho) \to \mathcal{P}_A(\zeta, \lambda, \mu, \rho)\,,$$

and hence by definition of Φ_L one has:

$$\varlimsup_{L\to\infty} \Phi_L(S_L, M_L, v_L) \le \varlimsup \mathcal{P}_{A_L}(\zeta_L, \lambda, \mu, \rho) = \mathcal{P}_A(\zeta, \lambda, \mu, \rho) \le \Psi(S, M, v) + \varepsilon\,.$$

Since the inequality holds for any $\varepsilon > 0$ the result follows. □

Appendix D. Proof of Limit Formulas for Discrete Parameter Spaces

Throughout this section, we assume that the parameter space $\Omega = \Omega_K$ is discrete and supported on exactly $K \ge 1$ points. We define

$$\mathcal{C}_P = \{(S, M, v) \in \mathcal{C} \mid \varphi_0(S, M, v) \ge -P\}\,,$$

to denote the sets with entropy bounded below by P, here ψ_0 is given by:

$$\begin{aligned}\varphi_0(S, M, v) &= \lim_{\varepsilon\to 0}\lim_{N\to\infty} \frac{1}{N}\log \int \mathbb{1}(|R_{11} - S| \le \varepsilon)\,\mathbb{1}(|R_{10} - M| \le \varepsilon)\,\mathbb{1}(|\bar{x} - v| \le \varepsilon)\, d\,\mathbb{P}_K(x) \\ &= \inf_{\lambda,\mu,\rho}\left(\mathbb{E}_{\mathbb{Q}} \log \int e^{\lambda x^0 x + \mu x^2 + \rho x}\, d\,\mathbb{P}_K(x) - \lambda M - \mu S - \rho v\right),\end{aligned}$$

where $\mathbb{P}_X$ is the uniform measure on Ω_K. We show Hölder continuity and equicontinuity for discretized priors.

Lemma D.1. *Suppose that $\Omega = \Omega_K$ is a finite collection distinct points. We have that for every $(S, M, v) \in \mathcal{C}$ there exists a constant C, independent of N and L, such that for every $L \ge 1$,*

$$|F^L_N(S, M, v; \varepsilon) - F^L_N(S, M, v; \eta)| \le C\sqrt{\varepsilon - \eta} + C(\varepsilon - \eta).$$

In particular, the family of functions $(F^L_N(S, M, v))_L$ is uniformly Hölder-$\frac{1}{2}$ on $\mathcal{C}_P$.

Proof. This proof follows the argument of [12, Lemma 7.3]. Furthermore, it suffices to prove the result for the case that $\beta_2 = \beta_3 = \beta_4 = 0$ since these terms are within an ε neighbourhood of M, S, v respectively on Ω_ε. The deviations in these terms by changing ε is clearly Lipschitz.

Fix $\varepsilon > \eta > 0$ and let $\pi^{\mathbf{x}_0} : \Omega_\varepsilon(\mathbf{x_0}) \to \Omega_\eta(\mathbf{x_0})$ be the map that takes $\mathbf{x}$ to $\pi^{\mathbf{x_0}}(\mathbf{x}) \in \Omega_\eta(\mathbf{x}_0)$ such that the Euclidean distance, $d(\pi^{\mathbf{x_0}}(x), x)$, is minimized. As Ω is finite, this map is well-defined. Furthermore, we can choose $\pi(\mathbf{x})$ so that $d(\pi^{\mathbf{x_0}}(x), x) \le C\sqrt{N}(\varepsilon - \eta)$.

By Dudley's entropy inequality, for any $\delta > 0$ and constant there exists a constant C such that

$$\mathbb{E} \sup_{\substack{d(\mathbf{x}_1,\mathbf{x}_2)\le\delta\sqrt{N} \\ \|\mathbf{x}_1\|,\|\mathbf{x}_2\| \le |\operatorname{supp}\Omega|\sqrt{N}}} |H^{\bar\beta}_N(\mathbf{x}_1) - H^{\bar\beta}_N(\mathbf{x}_2)| \le CN\delta\,.$$

We have the chain of inequalities,

$$\mathbb{E}F^L_N(S, M, v; \varepsilon) = \frac{1}{NL}\mathbb{E}\log\int_{\Omega_\varepsilon} e^{LH^{\bar\beta}_N(\mathbf{x})}\, d\,\mathbb{P}_K(\mathbf{x}) \le \frac{1}{NL}\mathbb{E}\log\int_{\Omega_\varepsilon} e^{LH^{\bar\beta}_N(\pi^{\mathbf{x_0}}(\mathbf{x}))}\, d\,\mathbb{P}_K(\mathbf{x}) + C(\varepsilon - \eta)$$

$$\le \frac{1}{NL}\mathbb{E}\log\int_{\Omega_\eta} e^{LH_N^{\bar\beta}(\mathbf{x})}|(\pi^{\mathbf{x_0}})^{-1}(\mathbf{x})|\,d\,\mathbb{P}_K(\mathbf{x}) + C(\varepsilon-\eta)$$
$$\le \frac{1}{NL}\mathbb{E}\log\int_{\Omega_\eta} e^{LH_N^{\bar\beta}(\mathbf{x})}|B_{R(x,x),R(x,x);\varepsilon-\eta}(\mathbf{x})|\,d\,\mathbb{P}_K(\mathbf{x}) + C(\varepsilon-\eta)\,.$$

To get from the second to third inequality, notice that for every $\mathbf{x}\in\Omega_\eta(\mathbf{x_0})$, we have

$$(\pi^{\mathbf{x_0}})^{-1}(\mathbf{x}) = \{\mathbf{y}\in\Omega_\varepsilon \mid \pi^{\mathbf{x_0}}(\mathbf{y}) = \mathbf{x}\} \subseteq B_{R(x,x),R(x,x;C(\varepsilon-\eta)}(\mathbf{x})$$

where for parameters S, M

$$B_{S,M;\varepsilon-\eta}(\mathbf{x}) = \{\mathbf{y}\in\Omega \mid R(\mathbf{y},\mathbf{y}) \approx_{\varepsilon-\eta} S, R(\mathbf{x},\mathbf{y}) \approx_{\varepsilon-\eta} M\}\,.$$

This follows because π was constructed so that $d(\pi(\mathbf{y}),\mathbf{x}) \le C\sqrt{N}(\varepsilon-\eta)$ for points $\mathbf{y}\in\Omega_\varepsilon$ and $\mathbf{x}\in\Omega_\eta$ and $B_{\|\mathbf{y}\|,\|\mathbf{y}\|;C(\varepsilon-\eta)}(\mathbf{x})$ contains all points within $C\sqrt{N}(\varepsilon-\eta)$ of $\mathbf{y}$. More precisely, for $\{\mathbf{y}\in\Omega_\varepsilon \mid \pi^{\mathbf{x_0}}(\mathbf{y}) = \mathbf{x}\}$,

$$|R(x,x)-R(y,y)| \le \frac{1}{N}\langle x-y, x+y\rangle \le \frac{1}{N}\sqrt{\|x-y\|\|x+y\|} \le C(\varepsilon-\eta)\,,$$

and

$$|R(x,y)-R(y,y)| = \Big|-\frac{1}{2}R(x,x)+R(x,y)-\frac{1}{2}R(y,y)+\frac{1}{2}R(x,x)-\frac{1}{2}R(y,y)\Big|$$
$$\le \frac{1}{N}\|x-y\|^2 + \frac{1}{N}|\|x\|^2-\|y\|^2| \le C(\varepsilon-\eta)\,.$$

We conclude that

$$\mathbb{E}F_N^L(S,M,v;\varepsilon) - \mathbb{E}F_N^L(S,M,v;\eta) \le \sup_y\left(\frac{1}{NL}\log|B_{R(x,x),R(x,x);\varepsilon-\eta}(\mathbf{x})|\right) + C(\varepsilon-\eta)\,.$$

It remains to find a uniform bound on the number of points in $B_{R(y,y),R(y,y);C(\varepsilon-\eta)}(\mathbf{y})$. Intuitively this set is small since it requires the $\mathbf{x}$ and $\mathbf{y}$ to be almost perfectly correlated, which means that this is essentially a constraint that the coordinates of $\mathbf{x}$ and $\mathbf{y}$ match. This is made precise by computing the large deviations rate function of such an event.

Let $\mathbb{E}_y$ denote the average with respect to the empirical measure of $\mathbf{y}$. In particular, we have $\mathbb{E}_y[y^2] = R(\mathbf{y},\mathbf{y})$. From the large deviations upper bound (see for example [37, Section 4]) it follows that for any S and M and any λ,μ

$$\lim_{\varepsilon\to0}\overline{\lim_{N\to\infty}}\frac{1}{NL}\log|B_{S,M;\varepsilon}(\mathbf{y})| \le \frac{1}{L}\Big(\mathbb{E}_y\log\sum_{x\in\Omega}e^{\lambda yx+\mu x^2} - M - \mu S\Big)\,,$$

where $\mathbb{E}_y$ is the average with respect to the empirical distribution of $\mathbf{y}$. We may take $\mu = -\frac{1}{2}\lambda$ then in the limit, we have that

$$\lim_{\varepsilon\to0}\overline{\lim_{N\to\infty}}\frac{1}{NL}\log|B_{S,M;\varepsilon}(\mathbf{y})| \le \lim_{\lambda\to\infty}\frac{1}{L}\Big(\mathbb{E}_y\log\sum_{x\in\Omega}e^{\lambda(yx-\frac{1}{2}x^2)} - M + \frac{1}{2}S\Big)\,.$$

Since $y\in\Omega$ and

$$\sup_{x\in\Omega}\Big(yx-\frac{1}{2}x^2\Big) = \frac{1}{2}y^2\,,$$

if we take $S = M = \mathbb{E}y^2$, then

$$\lim_{\lambda\to0}\Big(\lambda\Big(\frac{1}{\lambda}\mathbb{E}_y\log\sum_{x\in\Omega_K}e^{\lambda(yx-\frac{1}{2}x^2)}\Big) - \lambda\frac{1}{2}\mathbb{E}y^2\Big)$$
$$= \lim_{\lambda\to0}\Big(\lambda\Big(\frac{1}{\lambda}\mathbb{E}_y\log\sum_{x\in\Omega_K}e^{\lambda(yx-\frac{1}{2}x^2)}\Big) - \lambda\frac{1}{2}\mathbb{E}y^2\Big) = 0\,,$$

as a consequence of Laplace's method. We conclude that

$$\lim_{\varepsilon\to 0}\overline{\lim_{N\to\infty}}\frac{1}{NL}\log|B_{S,M;\varepsilon}(\mathbf{y})|\leq\frac{1}{L}\inf_{\mu,\lambda}\Big(\mathbb{E}_y\log\sum_{x\in\Omega_K}e^{\lambda yx+\mu x^2}-M-\mu S\Big)\leq 0\,.$$

as required.

We now need to control this rate function in an ε ball around $S=M=\mathbb{E}y^2$. Let $\varepsilon>0$, and notice that rate function satisfies the following bound

$$\begin{aligned}&\frac{1}{L}\inf_{\mu,\lambda}\Big(\mathbb{E}_y\log\sum_{x\in\Omega}e^{\lambda yx+\mu x^2}-M-\mu S\Big)\\&\leq\inf_{\lambda}\frac{1}{L}\Big(\mathbb{E}_y\log\sum_{x\in\Omega}e^{\lambda(yx-\frac{1}{2}x^2)}-\frac{\lambda}{2}(R_{yy}^2-\varepsilon)\Big)\\&=\inf_{\lambda}\frac{1}{L}\Big(\mathbb{E}_y\log\sum_{x\in\Omega}\frac{e^{\lambda(yx-\frac{1}{2}x^2)}}{e^{\lambda y^2/2}}+\mathbb{E}_y\log e^{\lambda y^2/2}-\frac{\lambda}{2}(R_{yy}^2-\varepsilon)\Big)\\&=\inf_{\lambda}\frac{1}{L}\Big(\mathbb{E}_y\log\sum_{x\in\Omega_K}\frac{e^{\lambda(yx-\frac{1}{2}x^2)}}{e^{\lambda y^2/2}}+\frac{\lambda}{2}\varepsilon\Big)\,.\end{aligned}$$

Since $yx-\frac{1}{2}x^2$ is maximized at $x=y\in\Omega$, and $|x-y|\geq C(\Omega)$ for all $x\neq y$ and $C(\Omega)$ depends only on the minimal distance of points in Ω, we have

$$\sum_{x\in\Omega}\frac{e^{\lambda(yx-\frac{1}{2}x^2)}}{e^{\lambda y^2/2}}=1+\sum_{x\neq y\in\Omega_K}\frac{e^{\lambda(yx-\frac{1}{2}x^2)}}{e^{\lambda y^2/2}}\leq 1+Ke^{-C(\Omega)\lambda}\,,$$

where the constant K denotes the number of points in the support of Ω. Applying the inequality $\log(1+x)\leq x$ implies that

$$\inf_{\lambda}\frac{1}{L}\Big(\mathbb{E}_y\log\sum_{x\in\Omega_K}\frac{e^{\lambda(yx-\frac{1}{2}x^2)}}{e^{\lambda y^2/2}}+\frac{\lambda}{2}\varepsilon\Big)\leq\inf_{\lambda}\frac{1}{L}(Ke^{-C\lambda}+\frac{\lambda}{2}\varepsilon)\,.$$

If we take $\varepsilon\leq 2KC$ then the function is convex and its minimum is attained at $\lambda=-\frac{1}{C}\log(\frac{KC\varepsilon}{2})$

$$\inf_{\lambda}\frac{K}{L}e^{-\lambda C}+\frac{\lambda}{2L}\varepsilon=\frac{K^2C}{2L}\varepsilon-\frac{1}{2CL}\log(\frac{KC\varepsilon}{2})\leq\frac{K^2C}{2}\varepsilon-\frac{1}{2C}\log(\frac{KC\varepsilon}{2})\,,$$

which is Hölder 1/2 with Hölder constant independent of L. Therefore,

$$\lim_{N\to\infty}\Big(\mathbb{E}F_N^L(S,M,v;\varepsilon)-\mathbb{E}F_N^L(S,M,v;\eta)\Big)\leq C(\varepsilon-\eta)^{1/2}+C(\varepsilon-\eta)\,.$$ □

The previous Lemma immediately implies the following result:

Lemma D.2 (Equicontinuity)**.** *Let $\mathbf{x}_0$ be fixed and $P>0$. The family of functions $(\frac{1}{L}\varphi_{L\bar{\beta}}(S,M))_{L\geq 1}$ restricted to $\mathcal{C}_P$ is uniformly Holder $\frac{1}{2}$ continuous in L.*

Proof. We can pick M_1,S_1,v_1 and M_2,S_2,v_2 are such that $|M_1-M_2|\leq\delta$,$|S_1-S_2|\leq\delta$, $|v_1-v_2|\leq\delta$ and $\eta<\delta<\frac{1}{2}$, then Lemma D.1 implies that

$$F_N^L(S_1,M_1,v_1;\eta)-F_N^L(S_2,M_2,v_2;\eta)\leq F_N^L(S_2,M_2,v_2;2\delta)-F_N^L(S_2,M_2,v_2;\eta)\leq|2\delta-\eta|^{\frac{1}{2}}.$$

By symmetry, we also have the reverse bound so

$$|f_N(S_1,M_1,\eta)-f_N(S_2,M_2,\eta)|\leq|2\delta-\eta|^{\frac{1}{2}}.$$

Since

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} F_N^L(S,M,\varepsilon) = \frac{1}{L}\varphi^K_{L\bar\beta}(S,M,v),$$

we can take $N\to\infty$ followed by $\varepsilon'\to 0$ to conclude that

$$\Big|\frac{1}{L}\varphi_{L\bar\beta}(S_1,M_1v_1) - \frac{1}{L}\varphi_{L\bar\beta}(S_2,M_2,v_2)\Big| \le |2\delta|^{\frac{1}{2}}.$$

which completes the proof of equicontinuity. □

From the equicontinuity, we immediately get the following characterization of the ground state of the proxy model.

Lemma D.3. *If Ω is a finite collection of poits, then for any $\bar\beta = (\beta_1,\beta_2,\beta_3,\beta_4)$,*

$$\lim_{N\to\infty}\mathbb{E}\max_{x\in\Omega^N}\frac{H_N^{\bar\beta}}{N} = \sup_{(s,m,v)}\psi_{\bar\beta}(s,m,v)$$

Proof. We now prove the main variational formula for the MLE. We fix $\bar\beta$ and define

$$\Psi_L(S,M,v) := \frac{1}{L}\varphi_{L\bar\beta}(S,Mv)$$

It suffices to show that:

$$\sup_{(S,M,v)\in\mathcal{C}}\Psi_L(S,M,v)\to\sup_{(S,M,v)\in\mathcal{C}}\Psi(S,M,v)\,,$$

We proceed as follows. For any $\eta>0$, we may find a $P\ge 1$ and a compact subset $\Omega_\eta\subset\mathcal{C}_P\subset\operatorname{int}(\mathcal{C})$ such that:

$$\sup_{(S,M,v)\in\mathcal{C}}\Psi(S,M,v)\le\sup_{(S,M,v)\in\Omega_\eta}\Psi(S,M,v)+\eta\,,$$

and by pointwise convergence of $\Psi_L(S,M,v)$ in Lemma C.1 and equicontinuity on compact subsets of the interior Lemma D.1 one has that:

$$\sup_{(S,M,v)\in\Omega_\eta}\Psi(S,M,v) = \lim_{L\to\infty}\sup_{(S,M,v)\in\Omega_\eta}\Psi^L(S,M,v)\,,$$

from which it follows that:

$$\sup_{(S,M,v)\in\mathcal{C}}\Psi(S,M,v)\le\lim_{L\to\infty}\sup_{(S,M,v)\in\mathcal{C}}\Psi^L(S,M,v)\,.$$

In establishing the lower bound we will appeal to Lemma C.1. For each $L>0$ we may find a point (S_L,M_L,v_L) such that:

$$\sup_{(S,M,v)\in\mathcal{C}}\Psi_L(S,M,v)\le\Psi_L(S_L,M_L,v_L)+\frac{1}{L}\,,$$

since $\mathcal{C}$ is compact we may extract a convergent subsequence $(S_{L'},M_{L'},v_{L'})$ converging to some point (S,M,v) in $\mathcal{C}$. By assumption on the sequence (S_L,M_L,v_L) one has that $\Psi(S,M,v)>-\infty$, and hence Lemma C.1 implies that:

$$\lim_{L\to\infty}\sup_{(S,M,v)\in\mathcal{C}}\Psi_L(S,M,v)\le\overline{\lim_{L\to\infty}}\,\Psi_L(S_L,M_L,v_L)\le\Psi(S,M,v)\le\sup_{(S,M,v)\in\mathcal{C}}\Psi_L(S,M,v)\,.$$

Concluding the proof. □

Appendix E. Proof of Limit Formulas for Continuous Parameter Spaces

Throughout this section, we assume that the parameter space Ω an interval. Since Ω is an interval, we can define $\Omega(S,M,v)=\Omega_0(S,M,v)$ to denote the set of constrained norm, overlap, and magnetization. We begin by establishing Hölder continuity of the maximum of the Hamiltonian $\frac{H_N}{N}$ on bands with fixed norm, overlap, and magnetization.

Lemma E.1. *Suppose that $\Omega=[a,b]$, and that (S,M,v) and (S',M',v') belong to $\mathcal{C}$, then:*

$$\left|\mathbb{E}\max_{x\in\Omega(S,M,v)}\frac{1}{N}H_N(x)-\mathbb{E}\max_{x\in\Omega(S',M',v')}\frac{1}{N}H_N(x)\right|\le f(S-S',M-M',v-v')+o(1)\,,$$

for f a $\frac{1}{2}$-Hölder continuous function, independent of N.

Proof. We write y for the latent vector throughout to simplify notation. We assume further that y is not a constant vector, and that the empirical distribution of y converges to a non-trivial distribution.

We proceed as follows, given the constraints on the vector x, let us define vectors $e_1,e_2\in\mathbb{R}^N$ as follows:

$$e_1=\frac{y}{\|y\|}\qquad\text{and}\qquad e_2=\frac{\mathbf{1}-e_1\langle e_1,\mathbf{1}\rangle}{\sqrt{N-N^2\frac{(\overline{y})^2}{\|y\|^2}}}\,.$$

(Note that in the case y is a vector with constant entries, one has $e_2=0$, and the calculations that follow simplify greatly.) Given $x\in\Omega(S,M,v)$ we may then write:

$$x=\alpha(x)e_1+\beta(x)e_2+w(x)\,,$$

with $w(x)$ orthogonal to e_1 and e_2. A direct computation yields:

$$\alpha(x)=\frac{NM}{\|y\|}$$
$$\beta(x)=\frac{Nv-\frac{N^2\overline{y}M}{\|y\|^2}}{\sqrt{N-\frac{N^2(\overline{y})^2}{\|y\|^2}}}$$
$$\|w(x)\|^2=NS-\frac{N^2M^2}{\|y\|^2}-\frac{(Nv-\frac{N^2\overline{y}M}{\|y\|^2})^2}{N-\frac{N^2(\overline{y})^2}{\|y\|^2}}\,.$$

Note that by definition of $\Omega(S,M,v)$ the functions $\alpha(x),\beta(x)$ are constant when fixing overlaps and sample means. Consequently, we may define $x'\in\Omega(S',M',v')$ as follows:

$$x'=\frac{NM'}{\|y\|}e_1+\frac{Nv'-\frac{N^2\overline{y}}{\|y\|^2}M'}{\sqrt{N-N^2\frac{\overline{y}^2}{\|y\|^2}}}+w(x')\,,$$

where $w(x')$ is chosen to be collinear with $w(x)$, and such that the norm squared of x' is NS'. By Collinearity of $w(x)$ and $w(x')$ one may then compute:

$$\begin{aligned}\frac{1}{N}\left\|x-x'\right\|^2&=\frac{N}{\|y\|^2}(M-M')^2+\frac{((v-v')+\overline{y}\frac{N}{\|y\|^2}(M'-M))^2}{1-\frac{N(\overline{y})^2}{\|y\|^2}}\\&+\left[\sqrt{S-\frac{NM^2}{\|y\|^2}-\frac{(v-\frac{N\overline{y}}{\|y\|^2}M)^2}{1-\frac{N\overline{y}}{\|y\|^2}}}-\sqrt{S'-\frac{N(M')^2}{\|y\|^2}-\frac{(v'-\frac{N\overline{y}}{\|y\|^2}M')^2}{1-\frac{N\overline{y}}{\|y\|^2}}}\right]^2.\end{aligned}$$

By our assumptions on the empirical distribution of y, the terms $\frac{N}{\|y\|^2}$ converge to a non-zero constant as $N \to \infty$, as does $\overline{y}$. By the non-triviality assumption of the limit, and the Cauchy-Schwarz inequality, one also has some universal constant $c_1 > 0$ so that for all N sufficiently large

$$1 - N\frac{\overline{y}^2}{\|y\|^2} > c_1 \,.$$

With this in mind let us then note that for some $\frac{1}{2}$-Hölder function f, we have a bound for all large N given by:

$$\|x - x'\| \leq \sqrt{N} f(S - S', M - M', v - v')\,,$$

which satisfies $f(0,0,0) = 0$.

To finish the proof, note for $C > 0$ sufficiently large we have $c > 0$ so that with probability $1 - e^{-cN}$ one has:

$$|H_N(x_1) - H_N(x_2)| \leq C\sqrt{N}\,\|x_1 - x_2\|\,.$$

Indeed this follows by standard bounds on the operator norm of a GOE matrix exceeding $2 + \varepsilon$ (see [6] Theorem 2.3.5), and by noting the non-random terms in H_N are $C\sqrt{N}$ Lipschitz, for some $C > 0$ depending on Ω and $\overline{\beta}$.

Now given $x \in \Omega(S, M, v)$, one may always pair it with the constructed x' above to obtain

$$\max_{x \in \Omega(S,M,v)} \frac{1}{N} H_N(x) \leq C f(S - S', M - M', v - v') + \max_{x \in \Omega(S',M',v')} \frac{1}{N} H_N(x)\,,$$

with the reverse inequality following via symmetric argument. Taking expectations and absorbing C into f, we conclude for some $c > 0$ that

$$\left|\mathbb{E} \max_{x \in \Omega(S,M,v)} \frac{1}{N} H_N(x) - \mathbb{E} \max_{x \in \Omega(S',M',v')} \frac{1}{N} H_N(x)\right| \leq f(S - S', M - M', v - v') + e^{-cN}\,.$$

Completing the proof. ☐

With Lemma E.1 we may now prove the variational characterization of the ground state for the proxy model.

Lemma E.2. *Suppose that $\Omega = [a, b]$ is an interval, then for any $\bar{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$,*

$$\lim_{N \to \infty} \mathbb{E} \max_{x \in \Omega^N} \frac{H_N^{\bar{\beta}}}{N} = \sup_{(s,m,v) \in \mathcal{C}} \psi_{\bar{\beta}}(s, m, v)$$

Proof. Define $\psi_{N,\bar{\beta}}$ by

$$\psi_{N,\bar{\beta}}(s, m, v) := \mathbb{E} \max_{\Omega(s,m,v)} \frac{H_N^{\bar{\beta}}(\mathbf{x})}{N}\,.$$

By Lemma C.1, we have that $\psi_{N,\bar{\beta}} \to \psi_{\bar{\beta}}$ pointwise as N tends to infinity. To conclude the result it will suffice to show that $\psi_{N,\bar{\beta}}$ converges uniformly to $\psi_{\bar{\beta}}$ on $\mathcal{C}$, as uniform convergence implies convergence of the maximum.

Fix $\varepsilon > 0$, then if f is as in Lemma E.1, we may pick $\delta > 0$ so that

$$f(S - S', M - M', v - v') \leq \varepsilon$$

whenever $d((s, m, v), (s', m', v')) \leq \delta$.

Let $\mathcal{C}_\delta$ be a δ-net of $\mathcal{C}$, so that for each (s, m, v) in $\mathcal{C}$ there is (s', m', v') in $\mathcal{C}_\delta$ such that $d((s, m, v), (s', m', v')) \leq \delta$. We suppose that N is large enough so that the error e^{-cN} in Lemma E.1, is at most ε. One then has:

$$|\psi_{N,\bar{\beta}}(s, m, v) - \psi_{N,\bar{\beta}}(s', m', v')| \leq 2\varepsilon\,. \tag{E.1}$$

Since $\mathcal{C}_\delta$ has finitely many points (depending on δ), there exists a N_δ such that if $N > N_\delta$, then

$$\max_{(s,m,v)\in\mathcal{C}_\delta} |\psi_{N,\bar\beta}(s,m,v) - \psi_{\bar\beta}(s,m,v)| \le \varepsilon\,. \tag{E.2}$$

Lastly, note that pointwise convergence of $\psi_{N,\bar\beta}$ to $\psi_{\bar\beta}$, and Lemma E.1 imply for every pair (s,m,v) and (s',m',v') in $\mathcal{C}$ that:

$$\Big|\psi_{\bar\beta}(s,m,v) + \psi_{\bar\beta}(s',m',v')\Big| \le f(s-s',m-m',v-v')\,. \tag{E.3}$$

Combining the estimates in (E.1), (E.2), and (E.3), via the triangle inequality, we conclude that there is N_0 such that if $N > N_0$ then:

$$\sup_{(s,m,v)\in\mathcal{C}} \Big|\psi_{N,\bar\beta}(s,m,v) - \psi_{\bar\beta}(s,m,v)\Big| \le 4\varepsilon\,,$$

and hence $\psi_{N,\bar\beta}$ converges uniformly to $\psi_{\bar\beta}$.

Recalling the concentration of the maximum of the Hamiltonian in Lemma G.1, we have that

$$\lim_{N\to\infty}\mathbb{E}\max_{x\in\Omega^N}\frac{H_N^{\bar\beta}}{N} = \max_{s,m,v}\lim_{N\to\infty}\psi_{N,\bar\beta}(s,m,v) = \max_{s,m,v}\psi_{\bar\beta}(s,m,v)\,,$$

which concludes the proof. □

Appendix F. Proofs of Variational Formulas

In this section, we will prove the variational formulas for the zero score, score biased, and score corrected models. By universality in Proposition A.2, all of these variational formulas are direct consequences of the master theorem for the proxy model which is summarized below.

Theorem F.1. *Suppose that Ω is an interval or finite collection of points. For any $\beta_1,\beta_2,\beta_3,\beta_4$ and α and constraints (S,M,v), we have*

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N} H_N^{\bar\beta,\beta^4}(x) = \sup_{S,M,v}\psi_\beta(S,M,v) \tag{F.1}$$

and

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{x\in\Omega_\varepsilon(S,M,v)} H_N^{\bar\beta,\beta^4}(x) = \psi_\beta(S,M,v). \tag{F.2}$$

If $\alpha = 0$, then for any constraints $(S,M)\in\mathcal{C}$

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N} H_N^{\bar\beta,0}(x) = \sup_{S,M}\psi_\beta(S,M) \tag{F.3}$$

and

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{x\in\Omega_\varepsilon(S,M)} H_N^{\bar\beta,0}(x) = \psi_\beta(S,M). \tag{F.4}$$

Proof. We provide the proof for the cases when $\alpha \neq 0$, because the case when $\alpha = 0$ follows from an identical argument. The limit for the unconstained maxima is given in Lemma D.3 for the discrete parameter space and Lemma E.2. The limit for the constrained model is given in Lemma G.2. □

F.1. **Proof of the Variational Formula for the Score Biased Models.** By universality, we start by showing that the maximum liklihood estimate associated with the score corrected likelihood

$$\mathcal{L}^g_{N,\alpha}(Y,x) = \sum_{i\leq j} g\Big(Y_{ij}, \frac{\lambda x_i x_j}{\sqrt{N}}\Big).$$

is equivalent to $H_N^{\bar\beta,\alpha}(x)$ where β_1,β_2,β_3 are the information parameters and $\alpha = N^{1/2}\beta_4$ defined in (B.1).

Lemma F.1. *For $g, g_0 \in \mathcal{F}_0$, we have*

$$\lim_{N\to\infty}\Big|\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N}\mathcal{L}^g_{N,\alpha}(Y,x) - \frac{1}{N}\mathbb{E}\max_{x\in\Omega_N} H_N^{\bar\beta,N^{\frac12}\beta^4}(x)\Big| = 0.$$

Proof. This is a restatement of Lemma A.1, which shows that the pseudo maximum likelihood is equal to the maximizer of (A.1). □

In particular, it suffices to study the maximizers of the function

$$H_N^{\bar\beta,N^{\frac12}\beta^4}(x) = \frac{\beta_1}{\sqrt{N}}\sum_{ij} g_{ij}x_ix_j + \frac{N\beta_2}{2}R_{10}^2 - \frac{N\beta_3}{4}R_{11}^2 + \frac{N^{3/2}\beta_4}{2}\bar x^2 + o_N(1). \tag{F.5}$$

Notice that the last term $\frac{N^{3/2}\beta_4}{2}\bar x^2$ is the leading order term. This leading order term does not depend on the unknown variable, but dictates the performance of the MLE. If $\beta_4 > 0$, then the estimator must maximize this term, which is the statement of Theorem 2.3.

Proof of Theorem 2.3. If $\beta_4 > 0$, notice that $H_N^{\bar\beta,N^{\frac12}\beta^4}(x)$ is maximized when $(\bar x)^2$ is maximized. In particular, we have that $x = x_+\mathbf{1}$, where $\mathbf{1}$ is the all 1's vector and x_+ was the largest point in Ω. □

If $\beta_4 < 0$, then the estimator must minimize the leading order term in (F.5), which is the conclusion of Theorem 2.4

Proof of Theorem 2.4. If $\beta_4 < 0$, notice that $H_N^{\bar\beta,N^{\frac12}\beta^4}(x)$ is maximized when $(\bar x)^2$ is minimized. In particular, we have that $x = x_-$, where x_- is the smallest point in the convex hull of Ω. Note however that x_- may not lie within Ω^N, but up to introducing a term of order $\frac{C}{N}$ for some $C > 0$, we may assume it does. Taking $\varepsilon_N = \frac{C}{N}$ for a large constant C, we then have with probability at least $1 - e^{-cN}$ that:

$$|\max_{x\in\Omega_N} H_N^{\bar\beta,N^{\frac12}\beta^4}(x) - \max_{x\in\Omega:\bar x\approx_{\varepsilon_N} x_-} H_N^{\bar\beta,N^{\frac12}\beta^4}(x)| \leq \frac{C'}{\sqrt{N}},$$

where the bound above comes from tail bounds on the operator norm of a GOE matrix, see [6].

We now maximize the constrained maximization problems

$$\max_{x\in\Omega:\bar x\approx x_-} H_N^{\bar\beta,N^{\frac12}\beta^4}(x) = \sup_{S,M\in\mathcal{C}}\ \max_{,R_{11}\approx S, R_{10}\approx M,\bar x\approx x_-} H_N^{\bar\beta,N^{\frac12}\beta^4}(x)\,.$$

Subtracting the leading order term and taking limits implies that

$$\begin{aligned}
&\lim_{\varepsilon\to0}\lim_{N\to\infty}\max_{x\in\Omega_N}\frac{1}{N}\Big(H_N^{\bar\beta,N^{\frac12}\beta^4}(x) - \frac{N^{3/2}\beta_4}{2}\bar x^2\Big)\\
&= \lim_{\varepsilon\to0}\lim_{N\to\infty}\sup_{S,M\in\mathcal{C}}\ \max_{,R_{11}\approx S,R_{10}\approx M\bar x = x_-}\frac{1}{N}\big(H_N^{\bar\beta,N^{\frac12}\beta^4}(x) - \frac{N^{3/2}\beta_4}{2}\bar x^2\big)\\
&= \lim_{\varepsilon\to0}\lim_{N\to\infty}\sup_{S,M\in\mathcal{C}}\ \max_{R_{11}\approx S,R_{10}\approx M,\bar x = x_-}\frac{1}{N}H_N^{\bar\beta,0}(x)
\end{aligned}$$

$$= \sup_{S,M\in\mathcal{C}} \psi_-(S,M,x_-)\,.$$

where the last equality follows from (F.2) and concentration of the ground state Lemma G.1. $\square$

F.2. **Proof of the Variational Formula for the Score Corrected Model.** We prove Theorem 2.5. We start by showing that the maximum liklihood estimate associated with the score corrected likelihood

$$\mathcal{L}^g_{N,\alpha}(Y,x) = \sum_{i\le j} g\Big(Y_{ij}, \frac{\lambda x_i x_j}{\sqrt{N}}\Big) - N^{\frac{3}{2}}\hat{\beta}_4\bar{x}^2 + N\alpha\bar{x}^2.$$

is equivalent to $H_N^{\bar{\beta},\alpha}(x)$ where β_1,β_2,β_3 are the information parameters and $\alpha = \beta_2[\mathbb{E}_{\mathbb{Q}}x_0]^2 + \alpha$ defined in (B.1).

Lemma F.2. *For $g, g_0 \in \mathcal{F}_0$, we have*

$$\lim_{N\to\infty}\left|\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N}\mathcal{L}^g_{N,\alpha}(Y,x) - \frac{1}{N}\mathbb{E}\max_{x\in\Omega_N} H_N^{\bar{\beta},\alpha}(x)\right| = 0\,.$$

Proof. We define

$$\hat{\beta}_4 = \frac{1}{N^2}\sum_{ij}\partial_w g(Y_{ij},0)\,,$$

which approximates

$$\mathbb{E}\hat{\beta}_4 = \frac{1}{N^2}\sum_{ij}\mathbb{E}\partial_w g(Y_{ij},0) = \mathbb{E}[\partial_w g(Y,0)] = \beta_4 + \frac{\beta_2}{\sqrt{N}}\Big(\frac{1}{N}\sum_i x_i^0\Big)^2 + O(N^{-1})\,.$$

This is not an immediate consequence of universality (Proposition A.2) because $\hat{\beta}_4$ depends on all entries of Y. We will show that we can replace $\hat{\beta}_4$ with its expected value. We define the likelihood

$$\bar{\mathcal{L}}^g_{N,\alpha}(Y,x) = \sum_{i\le j} g\Big(Y_{ij}, \frac{\lambda x_i x_j}{\sqrt{N}}\Big) - N^{\frac{3}{2}}\mathbb{E}[\hat{\beta}_4]\bar{x}^2 + N\alpha\bar{x}^2\,,$$

which replaces $\hat{\beta}_4$ in $\bar{\mathcal{L}}^g_{N,\alpha}(Y,x)$ with its expected value. We will prove that

$$\left|\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N}\mathcal{L}^g_{N,\alpha}(Y,x;) - \frac{1}{N}\mathbb{E}\max_{x\in\Omega_N}\bar{\mathcal{L}}^g_{N,\alpha}(Y,x)\right| \le O(N^{-\frac{1}{2}})\,.$$

By Jensen's inequality, it suffices to show that

$$\mathbb{E}\left|\frac{1}{N}\max_{x\in\Omega_N}\mathcal{L}^g_{N,\alpha}(Y,x) - \frac{1}{N}\max_{x\in\Omega_N}\bar{\mathcal{L}}^g_{N,\alpha}(Y,x)\right| \le O(N^{-\frac{1}{2}})\,.$$

We make use of the obvious inequality that if $f(x)\ge g(x)$ or $g(x)\ge f(x)$, then

$$|\max f(x) - \max g(x)| = \max(\max g(x) - \max f(x), \max f(x) - \max g(x)) \le \max|f(x) - g(x)|\,,$$

to conclude that

$$\left|\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N}\mathcal{L}^g_{N,\alpha}(Y,x) - \frac{1}{N}\mathbb{E}\max_{x\in\Omega_N}\bar{\mathcal{L}}^g_{N,\alpha}(Y,x)\right| \le \sqrt{N}\mathbb{E}\left|\max_{x\in\Omega_N}|\hat{\beta}_4 - \mathbb{E}\hat{\beta}_4|(\bar{x})^2\right|.$$

Since Ω is bounded by C, we have

$$\mathbb{E}\left|\frac{1}{N}\max_{x\in\Omega_N}\mathcal{L}^g_{N,\alpha}(Y,x) - \frac{1}{N}\max_{x\in\Omega_N}\bar{\mathcal{L}}^g_{N,\alpha}(Y,x)\right| \le C\sqrt{N}\mathbb{E}|\hat{\beta}_4 - \mathbb{E}\hat{\beta}_4| \le C\sqrt{N}(\mathbb{E}(\hat{\beta}_4 - \mathbb{E}\hat{\beta}_4)^2)^{1/2}\,.$$

We have

$$\hat{\beta}_4 - \mathbb{E}\hat{\beta}_4 = \frac{1}{N^2}\sum_{i,j=1}^N[\partial_w g(Y_{ij},0) - \mathbb{E}\partial_w g(Y_{ij},0)]\,,$$

which has variance

$$\mathrm{Var}(\hat{\beta}_4 - \mathbb{E}\hat{\beta}_4) = \frac{1}{N^4}\sum_{i,j=1}^{N} \mathbb{E}[\partial_w g(Y_{ij},0) - \mathbb{E}\partial_w g(Y_{ij},0)]^2 = O(N^{-2})\,,$$

since $g \in \mathcal{F}_0$ we have that $\mathrm{Var}([\partial_w g(Y_{ij},0) - \mathbb{E}\partial_w g(Y_{ij},0)])$ is uniformly bounded for all i,j independent of N. This bound implies that

$$\mathbb{E}\left|\frac{1}{N}\max_{x\in\Omega_N} \mathcal{L}^g_{N,\alpha}(Y,x) - \frac{1}{N}\max_{x\in\Omega_N} \bar{\mathcal{L}}^g_{N,\alpha}(Y,x)\right| \le \sqrt{N}(\mathbb{E}(\hat{\beta}_4 - \mathbb{E}\hat{\beta}_4)^2)^{1/2} = O(N^{-1/2})\,.$$

By Proposition A.2 applied to $\bar{\mathcal{L}}^g_{N,\alpha}(Y,x)$ we conclude that

$$\lim_{N\to\infty}\left|\mathbb{E}\frac{1}{N}\max_{x\in\Omega_N} \bar{\mathcal{L}}^g_{N,\alpha}(Y,x) - \frac{1}{N}\mathbb{E}\max_{x\in\Omega_N} H_N^{\bar{\beta},\alpha}(x)\right| = 0\,.$$

□

Theorem 2.5 now follows by applying the variational formulas in Theorem F.1 to $H_N^{\bar{\beta},\alpha}(x)$.

Proof of Theorem 2.5. By Lemma F.2, it suffices to compute the limit of

$$\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N} H_N^{\bar{\beta},\alpha}(x)\,.$$

The maximum of such functions are precisely the one computed in (F.1), so

$$\mathbb{E}\frac{1}{N}\max_{x\in\Omega_N} \bar{\mathcal{L}}^g_{N,\alpha}(Y,x) \to \sup_{(S,M,v)} \psi_{\bar{\beta},\alpha}(S,M,v)\,.$$

□

F.3. **Proof of the Variational Formula for the Zero Score Model.** For completeness, we also provide the proof for zero score models, which follows from a simple modification of the previous arguments for score biased score models.

Proof of Theorem 2.1. By Lemma A.2, it suffices to compute the limit of

$$\frac{1}{N}\mathbb{E}\max_{x\in\Omega_N} H_N^{\bar{\beta},\alpha}(x)\,.$$

The maximum of such functions are precisely the one computed in (F.3), so

$$\mathbb{E}\frac{1}{N}\max_{x\in\Omega_N} \bar{\mathcal{L}}^g_{N,\alpha}(Y,x) \to \sup_{(S,M,v)} \psi_{\bar{\beta},\alpha}(S,M,v)\,.$$

□

Appendix G. Characterizations of the Overlaps for MLE

In this section, we will prove the second part of Theorem 2.1, Theorem 2.4, and Theorem 2.5. Throughout this section, for any two vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$, we define

$$R(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i\,,$$

to denote its normalized inner product. Similarly, we define

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} x_i\,,$$

to denote the sample mean of $\mathbf{x}$. For fixed S and M, to simplify notation we define

$$\psi_{\bar\beta}(S,M) = \begin{cases} \sup_v \psi_{\bar\beta,\alpha}(S,M,v) & \text{if } \alpha \neq 0 \\ \psi_{\bar\beta,0}(S,M) & \text{if } \alpha = 0. \end{cases}$$

We show that the limiting overlaps of the ground state variational formula are given by the corresponding maximizers of the ground state free energy provided the maximizers are unique. In the case the maximizers are not unique, the limits of the overlaps converge to one of the maximizers of ψ, and will be dealt with separately at the end of this section.

This hypothesis ensures that a notion of the limiting overlap $R(\hat{\mathbf{x}}_{\text{MLE}}, \mathbf{x}_0)$ is well defined, since in general $\hat{\mathbf{x}}_{\text{MLE}}$ may not unique, so its normalized inner product with $\mathbf{x}_0$ may depend on the choice of maximizer. We show that under some uniqueness assumptions, the normalized inner products of an equivalent model encoded by the information parameters $\bar\beta$ can only take one value.

Assumption 2. Given $\bar\beta$, suppose that $\psi_{\bar\beta}(s,m)$ has unique maximizers s_*, m_* up to a sign. More precisely, this implies that $\psi_{\bar\beta}(s,m)$ can have at most 2 maximizers, $(s_*, \pm m_*)$. In particular, we have that (s_*, m_*^2) is unique.

This assumption will imply that the maximum likelihood estimator is well defined in the following sense. Recall the restricted parameter space in (G.1), we define

$$\Omega_\varepsilon(S,M) = \{R_{10} \in A_M, R_{11} \in B_S\}. \tag{G.1}$$

Suppose that there exists a unique (s_*, m_*) such that all maximizers of $H_N^{\bar\beta}(\mathbf{x})$ are attained on the set $\Omega_\varepsilon(s_*, m_*)$

$$\max_{\mathbf{x}} H_N^{\bar\beta}(\mathbf{x}) = \max_{s,m} \max_{\Omega_\varepsilon(s,m)} H_N^{\bar\beta}(\mathbf{x}) = \max_{\Omega_\varepsilon(s^*,m^*)} H_N^{\bar\beta}(\mathbf{x}).$$

If this holds, then

$$\hat{\mathbf{x}}_{\text{mis}} = \arg\max_{\mathbf{x}} H_N^{\beta}(\mathbf{x})$$

satisfies $R(\hat{\mathbf{x}}_{\text{MLE}}, \mathbf{x}_0) \approx_\varepsilon m_*$, $R(\hat{\mathbf{x}}_{\text{MLE}}, \hat{\mathbf{x}}_{\text{MLE}}) \approx_\varepsilon s_*$. In particular, any maximizer of $H_N^{\beta}(\mathbf{x})$ maximizes the overlap with the underlying signal. The rest of this section will be devoted to show that the maximizing m_* is given by the largest maximizer of $\psi_{\bar\beta}$, for models such that the Fisher score parameters $\bar\beta$ satisfy (2).

We begin by showing a concentration result that implies that we can consider the average overlaps.

Lemma G.1. *Let $\overline{\beta}$ be fixed. There exists a universal constant C that depends only on $\bar\beta$ such that*

$$\mathbb{P}\Big(\Big|\frac{1}{N}\max_{\Omega_\varepsilon(S,M)} H_N^{\bar\beta}(x) - \mathbb{E}\frac{1}{N}\max_{\Omega_\varepsilon(S,M)} H_N^{\bar\beta}(x)\Big| \geq t\Big) \leq e^{-Ct^2N},$$

for any $(s,m) \in \mathcal{C}$. Furthermore,

$$\mathbb{E}\Big|\frac{1}{N}\max_{\Omega_\varepsilon(s,m)} H_N^{\bar\beta}(\mathbf{x}) - \frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s,m)} H_N^{\bar\beta}(\mathbf{x})\Big| \leq \frac{C}{\sqrt{N}}.$$

Proof. Notice that $\sup_{\Omega_\varepsilon(S,M)} |H_N(x)|$ is almost surely finite, and

$$\sup_{\Omega_\varepsilon(S,M)} \operatorname{Var}(H_N(x)) \leq NS \leq NC^2$$

where C is the maximal point in Ω. By the Borrell–TIS inequality [3, Section 2.1], it follows that

$$\mathbb{P}\Big(\Big|\frac{1}{N}\max_{\Omega_\varepsilon(S,M)} H_N(x) - \mathbb{E}\frac{1}{N}\max_{\Omega_\varepsilon(S,M)} H_N(x)\Big| \geq t\Big) \leq e^{-Ct^2N}.$$

This immediately implies the L^1 bound by integrating the tail see [65, Theorem 1.2]. ∎

We now show that the limit of the constrained proxy maximization problem is given by $\psi_{\bar\beta}(s,m)$.

Lemma G.2. *For any $(s,m)\in\mathcal{C}$*

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})=\psi_{\bar\beta}(s,m).$$

Proof. Let $\varepsilon>0$. We consider the constrained free energy

$$\frac{1}{L}F_N^{L\bar\beta}(s,m)=\frac{1}{NL}\mathbb{E}\log\int\mathbb{1}(R_{10}\approx m,R_{11}\approx s)e^{LH_N^{\bar\beta}(\mathbf{x})}\,d\,\mathbb{P}_X^{\otimes N}(\mathbf{x}).$$

Without loss of generality, we may assume that $\mathbb{P}_X$ is uniform over Ω_N.

Using the bounds of the ground state with the free energy (A.3), we have

$$\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})+o_N(L)\le\frac{1}{L}F_N^{L\bar\beta}(s,m)\le\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})\,.$$

Therefore, it suffices to compute a limit of $\frac{1}{L}F_N^{L\bar\beta}(s,m)$ for large N and fixed L.

For every L, we have as $N\to\infty$ by the finite temperature case that

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{L}F_N^{L\bar\beta}(s,m)=\varphi_{L\bar\beta}(s,m)\,.$$

By applying Lemma C.1, we conclude that

$$\lim_{L\to\infty}\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{L}F_N^{L\bar\beta}(s,m)=\lim_{L\to\infty}\varphi_{L\bar\beta}(s,m)=\psi_{\bar\beta}(s,m).$$

Therefore, using the ground state bounds (A.3) we have

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})=\psi_{\bar\beta}(s,m)\,,$$

which is what we needed to show. □

It now remains to show that the proxy model characterizes the maximum likelihood estimator of the original inference problem. To this end, given a model g and the corresponding Fisher score parameters $\bar\beta$, we define

$$\hat{\mathbf{x}}^g_{\mathrm{PMLE}}=\arg\max_{x\in\Omega^N}\sum_{i\le j}g\Big(Y_{ij},\frac{x_ix_j}{\sqrt{N}}\Big)\,,.$$

as was defined in (2.2). The following Lemma is a universality statement for the overlaps of the ground state.

Lemma G.3. *If the Fisher score parameters $\bar\beta$ corresponding to $g\in\mathcal{F}_0$ satisfies Hypothesis 2,then for any choice of maximizer $\hat{\boldsymbol{x}}^g_{\mathrm{PMLE}}$ one has almost surely*

$$\lim_{N\to\infty}R(\hat{\mathbf{x}}^g_{\mathrm{PMLE}},\mathbf{x}_0)^2=\lim_{N\to\infty}R(\hat{\mathbf{x}}^{\bar\beta}_{\mathrm{PMLE}},\mathbf{x}_0)^2=(m_*)^2\,,$$

and

$$\lim_{N\to\infty}R(\hat{\mathbf{x}}^g_{\mathrm{PMLE}},\hat{\mathbf{x}}^g_{\mathrm{PMLE}})=\lim_{N\to\infty}R(\hat{\mathbf{x}}^{\bar\beta}_{\mathrm{PMLE}},\hat{\mathbf{x}}^{\bar\beta}_{\mathrm{PMLE}})=s_*\,,$$

where $(s_*,(m_*)^2)$ is the maximizing pair of $\psi_{\bar\beta}$ given in Hypothesis 2.

Proof. From the universality of the restricted free energies Proposition A.2. We know that uniformly for s and m,

$$\left|\frac{1}{L}F_N^{L\bar\beta}(s,m)-\frac{1}{L}F_N^{Lg}(s,m)\right|\le o_N(L).$$

Furthermore, we have using the finite temperature formulas in Theorem B.3 that

$$\lim_{N\to\infty}\frac{1}{L}F_N^{Lg}(s,m)=\lim_{N\to\infty}\frac{1}{L}F_N^{L\bar\beta}(s,m)=\lim_{N\to\infty}\frac{1}{L}\varphi_{\bar\beta}(s,m)\,,\tag{G.2}$$

and so by (F.2) we obtain

$$\lim_{L\to\infty}\lim_{\varepsilon\to 0}\frac{1}{L}F_N^{Lg}(s,m)=\lim_{L\to\infty}\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{L}F_N^{L\bar\beta}(s,m)=\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})=\psi_{\bar\beta}(s,m). \tag{G.3}$$

An identical argument using the unconstrained limit of Theorem B.3 and (F.1) implies that

$$\lim_{L\to\infty}\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{L}F_N^{Lg}=\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\mathbf{x}}H_N^{\bar\beta}(\mathbf{x})=\sup_{s,m}\psi_{\bar\beta}(s,m).$$

The following is where Hypothesis 2 plays its most crucial role. Since the square of the maximizers of $\psi_{\bar\beta}(s,m)$ are unique, we have for all $m^2\neq m_*^2$,

$$\psi_{\bar\beta}(s,m)<\psi_{\bar\beta}(s^*,m^*).$$

Furthermore, since $\psi_{\bar\beta}(s,m)$ only depends on m through m^2, we have for the case that $m^2=m_*^2$ that

$$\psi_{\bar\beta}(s,m)=\psi_{\bar\beta}(s^*,m^*).$$

Using the characterization in (F.2), this implies that for all s,m such that $m^2\neq m_*^2$

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})<\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s_*,m_*)}H_N^{\bar\beta}(\mathbf{x})\,,$$

where equality is attained when $m^2=m_*^2$. By partitioning our state space, we have that

$$\frac{1}{N}\mathbb{E}\max_{\mathbf{x}}H_N^{\bar\beta}(\mathbf{x})=\frac{1}{N}\mathbb{E}\max_{s,m}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})\,,$$

and concentration in Lemma G.1 implies that for every $\varepsilon>0$,

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{s,m}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})=\lim_{N\to\infty}\frac{1}{N}\max_{s,m}\mathbb{E}\max_{\Omega_\varepsilon(s,m)}H_N^{\bar\beta}(\mathbf{x})\,.$$

Taking limits and applying (F.2) and (G.2), (G.3) implies

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\mathbf{x}}H_N^{g}(\mathbf{x})=\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\mathbf{x}}H_N^{\bar\beta}(\mathbf{x})=\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\max_{\Omega_\varepsilon(s_*,m_*)}H_N^{\bar\beta}(\mathbf{x})=\psi_{\bar\beta}(s_*,m_*).$$

The latter implies that the maximum is attained on the set $\Omega_\varepsilon(s,m)$, which implies that

$$R(\hat{\mathbf{x}}^{g}_{\mathrm{PMLE}},\mathbf{x}_0)^2=R(\hat{\mathbf{x}}^{\bar\beta}_{\mathrm{PMLE}},\mathbf{x}_0)^2=(m_*)^2,$$

where m_* is the largest maximizer of $\psi_{\bar\beta}(s,m)$ and

$$R(\hat{\mathbf{x}}^{g}_{\mathrm{PMLE}},\hat{\mathbf{x}}^{g}_{\mathrm{MLE}})=R(\hat{\mathbf{x}}^{\bar\beta}_{\mathrm{PMLE}},\hat{\mathbf{x}}^{\bar\beta}_{\mathrm{PMLE}})=s_*\,.$$

□

Combining all of the above implies the following lemma which characterizes the cosine similarity and mean squared error in the high dimensional limit.

Lemma G.4. *If our model and associated Fisher parameters $\bar\beta_g$ satisfy Hypothesis 2, we have*

$$|\mathbf{CS}(\hat{\mathbf{x}}_{\mathrm{MLE}},\mathbf{x}_0)|\to\frac{|m_{\beta_1,\beta_2,\beta_3}|}{\sqrt{s_{\beta_1,\beta_2,\beta_3}}\sqrt{\mathbb{E}_{\mathbb{Q}}(x^0)^2}}\qquad a.s.$$

where $m_{\beta_1,\beta_2,\beta_3}$ is the largest maximizer of $\sup_{(s,m)\in\mathcal{C}}\psi_{\bar\beta}(s,m)$.

Proof. This follows immediately from the characterization of the normalized inner products in Lemma G.3 and the fact that the mean squared error and cosine similarity are determined by the normalized inner products.

Indeed, notice that

$$\mathbf{CS}(\hat{\mathbf{x}}_{\mathrm{MLE}}, \mathbf{x}^0) = \frac{R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \mathbf{x}^0)}{\sqrt{R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \hat{\mathbf{x}}^g_{\mathrm{MLE}}) R(\mathbf{x}^0, \mathbf{x}^0)}} ,$$

and furthermore, Lemma G.3 implies that

$$|R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \mathbf{x}_0)| \to |m_*| \text{ and } R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \hat{\mathbf{x}}^g_{\mathrm{MLE}}) \to s_* ,$$

which finishes the proof. □

We close this section by arguing that the technical assumption Hypothesis 2 is equivalent to a regularity condition on $\psi_{\bar{\beta}}$ with respect to its information parameters. In particular, we will show that the overlaps of the proxy problem are uniquely characterized by the maximizers of $\psi_{\bar{\beta}}$ on the sets where $\psi_{\bar{\beta}}$ is differentiable. Consider the function

$$f(\beta_1, \beta_2, \beta_3) = \sup_{s,m} \psi_{\bar{\beta}}(S, M) ,$$

and define the set

$$\mathcal{D}_{\beta_2} = \{(\beta_1, \beta_2, \beta_3) \mid \partial_{\beta_2} f \text{ exists}\} \text{ and } \mathcal{D}_{\beta_3} = \{(\beta_1, \beta_2, \beta_3) \mid \partial_{\beta_3} f \text{ exists}\} .$$

We show that the characterization of the overlap for the proxy model is valid at points where $f(\beta_1, \beta_2, \beta_3)$ is differentiable.

Lemma G.5. *We have for all $\beta \in \mathcal{D}_{\beta_2}$ that*

$$R(\hat{\mathbf{x}}^{\bar{\beta}}_{\mathrm{MLE}}, \mathbf{x}_0)^2 \to m^2_{\beta_1,\beta_2,\beta_3} ,$$

and for all $\beta \in \mathcal{D}_{\beta_3}$ that

$$R(\hat{\mathbf{x}}^{\bar{\beta}}_{\mathrm{MLE}}, \hat{\mathbf{x}}^{\bar{\beta}}_{\mathrm{MLE}}) \to s_{\beta_1,\beta_2,\beta_3} ,$$

where $(s_{\beta_1,\beta_2,\beta_3}, m_{\beta_1,\beta_2,\beta_3})$ are maximizers $\psi_{\bar{\beta}}$. Furthermore, the optimizers of $\psi_{\bar{\beta}}(s, m)$ are unique up to a sign for $\beta \in \mathcal{D}_{\beta_2} \cap \mathcal{D}_{\beta_3}$.

Proof. This proof follows from an application of the envelope theorem and the fact that the ground state variational formula is the limit of the finite dimensional ground state. We start by characterizing the overlaps $R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \mathbf{x}_0)$. By universality, we know that the square of $R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \mathbf{x}_0)$ and $R(\hat{\mathbf{x}}^{\bar{\beta}}{}_{\mathrm{MLE}}, \mathbf{x}_0)$ converge to the same value.

Step 1: We fix the parameters β_1 and β_3 and consider

$$f_N(\beta_2) = \frac{1}{N}\mathbb{E} \max H^{\bar{\beta}}_N(x) = \frac{1}{N}\mathbb{E} \max \frac{\beta_1}{\sqrt{N}} \sum_{i<j} g_{ij} x_i x_j + \frac{\beta_2}{N} \sum_{i<j} x^0_i x^0_j x_i x_j - \frac{\beta_3}{2N} \sum_{i<j} x^2_i x^2_j$$

as a function of β_2 only. By the envelope theorem we have that

$$f'_N(\beta_2) = \frac{1}{2}\mathbb{E}R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \mathbf{x}_0)^2 ,$$

where $\hat{x}_{\mathrm{MLE}}$ denotes a maximizer of $H^{\bar{\beta}}_N(x)$.

Step 2: We now need to relate f'_N with the derivative of ψ with respect to β. Notice that $f_N(\beta_2)$ is convex in β_2 because it is the pointwise limit in L of convex functions in β_2 by (A.3),

$$\lim_{L\to\infty} \frac{1}{L} F_N(L\beta_2) := \lim_{L\to\infty} \frac{1}{L} F_N(L\beta_1, L\beta_2, L\beta_3) = f_N(\beta_2).$$

Since the derivatives of convex functions converge to the derivative of the limit on all points were the limit is differentiable we have

$$f_N'(\beta) \to \frac{d}{d\beta_2} \sup_{s,m} \psi_{\bar{\beta}}(s,m)\,.$$

Step 3: It remains to see that the limiting object is characterized by a maximizer of $\psi_{\bar{\beta}}$. By another application of the envelope theorem we see that

$$\frac{d}{d\beta_2} \sup_{s,m} \psi_{\bar{\beta}}(s,m) = \frac{m_*^2}{2}\,,$$

where m_* is a maximizer of $\psi(S,M)$. In particular, there can only be at most two maximizers of $\psi(S,M)$ and they are unique up to a sign. We have

$$\mathbb{E}R(\hat{\mathbf{x}}_{\mathrm{PMLE}}, \mathbf{x}_0)^2 \to m_*^2\,,$$

where m_* is a maximizer of $\psi(S,M)$. By Lemma G.1, we conclude

$$R(\hat{\mathbf{x}}_{\mathrm{PMLE}}, \mathbf{x}_0)^2 \to m_*^2\,,$$

almost surely.

The proof for the characterization of the limit of $R(\hat{\mathbf{x}}_{\mathrm{MLE}}, \hat{\mathbf{x}}_{\mathrm{MLE}})$ is identical, differentiating in β^3 instead. It uses again that the only dependence of ψ in β_3 is in the linear last term of the formula. □

The next result shows that the differentiability condition is necessary and sufficient for $\psi_{\bar{\beta}}$ to have a unique maximizer.

Lemma G.6. *We have*

$$\{(\beta_1,\beta_2,\beta_3) \mid \partial_{\beta_2} f \text{ exists}, \partial_{\beta_3} f \text{ exists}\} = \{(\beta_1,\beta_2,\beta_3) \mid \psi_{\bar{\beta}} \text{ satisfies Assumption 2}\}\,.$$

Proof. Suppose that $\psi_{\bar{\beta}}$ satisfies Assumption 2. More formally suppose for fixed β there exists an open interval I such that $\beta \in I$, and for every $\beta' \in I$ there exists a unique maximizing pair $(s_{\beta'}, \pm m_{\beta'})$ for $\psi_{\beta'}(S,M)$, then the map $\beta \mapsto \psi_{\beta_1,\beta,\beta_3}$ is differentiable on I. (The analogous statement holds for uniqueness in S.)

We proceed using the convexity argument as in the classical proof of the differentiability of the Parisi formula [65]. To simplify notation, we fix β_1 and β_3 and treat the functions f as only a function of $\beta_2 = \beta$. The case with β_3 will be handled in a similar manner.

Indeed suppose the hypothesis above holds, to show differentiability at β it suffices to show there is a unique subgradient, as the map $\beta \mapsto \psi_{\beta_1,\beta,\beta_3}$ is convex. Letting a denote a subgradient we have for every $y > 0$ that:

$$a \le \frac{f(\beta+y) - f(\beta)}{y}\,,$$

set $y_n = n^{-1/2}$, then immediately it follows that

$$a \le \frac{1}{y_n}\left(\psi_{\beta+y_n}(S_n, M_n) - \psi_\beta(S_n, M_n)\right),$$

where (S_n, M_n) is the minimizing pair at $\beta + y_n$, and (S,M) the minimizing pair at β. We thus have the corresponding Parisi functionals, and by definition of the infimum we may take a sequence $\zeta_n, \mu_n, \lambda_n$ so that:

$$\inf_{\zeta,\lambda,\mu} \mathcal{P}_\beta + \frac{1}{n} \ge \mathcal{P}_\beta(\zeta_n, \lambda_n, \mu_n)\,,$$

plugging in this inequality into both terms we then get:

$$a \le \frac{M_n^2}{2} + \frac{1}{\sqrt{n}}\,.$$

Proceeding as above taking we obtain a matching lower bound:

$$\frac{(M_n')^2}{2} - \frac{1}{\sqrt{n}} \leq a \leq \frac{M_n^2}{2} + \frac{1}{\sqrt{n}} \,.$$

By the uniqueness hypothesis and continuity of the map $\beta \to \psi_\beta$ one then has that $M_n^2, (M_n')^2 \to M^2$, and hence the is a unique subgradient if there is a unique maximizer. (This argument follows verbatim when differentiating in β_3 and gives differentiability in S.

The converse is immediate, because if ψ is differentiable in β_2 and β_3, then one immediately obtains uniqueness for maximizing pairs by Danskin's envelope theorem as was shown in part 3 of the proof of Lemma G.5. □

We end this section by showing that although in some cases the maximizers of $\psi_{\overline{\beta}}$ may not be unique, the performance of the MLE is still characterized by the maximizers of ψ.

Lemma G.7. *Let $\overline{\beta}$ be fixed, and suppose that (S, M) are such that $-\infty < \psi(S, M) < \sup \psi$. Let GS_N denote the (random) collection of maximizers of $H_N^{\overline{\beta}}$ in Ω^N, then for $\varepsilon > 0$ sufficiently small, one has:*

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \mathbb{P}(GS_N \cap \{R_{1,1} \approx_\varepsilon S, R_{1,0} \approx_\varepsilon M\} \neq \emptyset) < 0\,.$$

Furthermore, one has that the collection of all limit points, taken over all sequences of near maximizers $\mathbf{x}_N$, for the sequence $(S_N(\mathbf{x}), M_N(\mathbf{x}))$ is equal to $\mathcal{C}_{\bar{\beta}}$.

Proof. We proceed as follows, since $\psi(S, M) < \sup \psi$, there are constant $\varepsilon, \delta > 0$ such that if $A_\delta = [S - \delta, S + \delta] \times [M - \delta, M + \delta] \cap \mathcal{C}$, then

$$\sup_{(s,m)\in A_\delta} \psi + \varepsilon < \sup_{(s,m)\in\mathcal{C}} \psi\,.$$

By Lemma G.1 one has that:

$$\mathbb{P}\left(\frac{1}{N}\left|\mathbb{E} \sup_{\Omega_\delta(s,m)} H_n(x) - \sup_{\Omega_\delta(s,m)} H_n(x)\right| \geq \frac{\varepsilon}{2}\right) \leq e^{-\frac{CN\varepsilon^2}{4}}\,,$$

and so by Theorem 2.1, we have with with probability at least $1 - 2e^{-\frac{1}{4}CN\varepsilon^2}$ for any $x \in GS_N$ that:

$$\frac{1}{N}\left|H_N(x) - \max_{\Omega_\delta(S,M)} H_N\right| \geq |\sup_{(s,m)\in\mathcal{C}} \psi - \sup_{(s,m)\in A_\delta} \psi| - \frac{\varepsilon}{2} + o(1)\,.$$

Hence, for N sufficiently large one has that $\mathrm{GS}_N \cap \{R_{1,1} \approx_\delta S, R_{1,0} \approx_\delta M\} \neq \emptyset$, with probability at most $2e^{-\frac{1}{4}C\varepsilon^2 N}$. Consequently:

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \mathbb{P}(\mathrm{GS}_N \cap \{R_{1,1} \approx_\varepsilon S, R_{1,0} \approx_\varepsilon M\} \neq \emptyset) \leq -\frac{C\varepsilon^2}{2}\,,$$

which proves the first claim. The second claim follows from Portmanteau's theorem. Suppose that $(S_N(\mathbf{x}_N), M_N(\mathbf{x}_N))$ converges weakly to $(\mathscr{S}, \mathscr{M})$, then for any sufficiently small δ neighbourhood U of (S, M), such that $\mathcal{C}_\beta \cap \overline{U}$ is empty, the large deviation upper bound from before implies that:

$$\begin{aligned}\mathbb{P}((\mathscr{S}, \mathscr{M}) \in U) &\leq \lim_{N\to\infty} \mathbb{P}((S_N(\mathbf{x}_N), M_N(\mathbf{x}_N) \in U) \\ &\leq \overline{\lim_{N\to\infty}} \mathbb{P}(\mathrm{GS}_N \cap \{R_{1,1} \approx_\delta S, R_{1,0} \approx_\delta M\} \neq \emptyset) = 0\,,\end{aligned}$$

and hence the support of $(\mathscr{S}, \mathscr{M})$ is contained in $\mathcal{C}_{\bar{\beta}}$.

Now let us fix a point (S, M) in $\mathcal{C}_{\bar{\beta}}$, then by definition of a near maximizer and concentration of H_N/N, one has for any fixed $\varepsilon > 0$, and for N sufficiently large that:

$$\max_{x\in\Omega_\varepsilon(S,M)} \frac{H_N}{N} > \psi(S, M) - \varepsilon\,,$$

with probability $1 - o(1)$. Consequently there is $\hat{\boldsymbol{x}}_N \in \Omega_\varepsilon(S, M)$ achieving this bound. Let us consider the sequence of overlaps $(R(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{x}}_N), R(\hat{\boldsymbol{x}}_N\hat{\boldsymbol{x}}^{0,N}))$, we note that this sequence is tight, and so by passing to a subsequence we may assume that $(R(\hat{\boldsymbol{x}}_{N,\varepsilon}, \hat{\boldsymbol{x}}_{N,\varepsilon}), R(\hat{\boldsymbol{x}}_N\hat{\boldsymbol{x}}^{0,N}))$ converges almost surely as $N \to \infty$ to some random variables $(\mathscr{S}_\varepsilon, \mathscr{M}_\varepsilon)$ By definition of Ω_ε we have that

$$|\mathscr{S}_\varepsilon - S| \le \varepsilon \qquad \text{and} \qquad |\mathscr{M}_\varepsilon - M| \le \varepsilon\,,$$

hence taking $\varepsilon \to 0$ implies that $\mathcal{M}_\varepsilon \to M$ and $\mathcal{S}_\varepsilon \to S$ almost surely, finishing the proof.

□

We now state an impossibility result for models when $\beta_2 = 0$ and $\mathbf{x}^0$ is balanced, i.e. its sample mean converges to 0. This follows from the following fact about exchangeable independent sums.

Lemma G.8. *Let $\mathbf{y^N}$ be a sequence such that $\frac{1}{N}\sum_{i=1}^N y_i^N \to 0$ and let $\boldsymbol{x}_N$ a triangular array of uniformly bounded exchangable vectors independent of y. Then*

$$\frac{1}{N}\sum_{i=1}^N x_i y_i \to 0\,,$$

in probability.

Proof. By Markov's inequality, it suffices to show that

$$\mathbb{E}\Big(\frac{1}{N}\sum_{i=1}^N x_i y_i\Big)^2 \to 0\,.$$

We have

$$\begin{aligned}
\mathbb{E}\Big(\frac{1}{N}\sum_{i=1}^N x_i y_i\Big)^2 &= \frac{1}{N^2}\Big[\sum_{i,j}\mathbb{E}[x_i x_j] y_i y_j\Big] \\
&= \frac{1}{N^2}\Big[a_N\sum_{i=1}^N y_i^2 + b_N \sum_{1\le i\neq j\le N} y_i y_j\Big] \\
&= b_N\Big(\frac{1}{N}\sum_{i=1}^N y_i\Big)^2 + O(N^{-1})\,,
\end{aligned}$$

where $a_N = \mathbb{E}[x_1]$ and $b_N = \mathbb{E}[x_1 x_2]$. These numbers a_N and b_N are bounded, so

$$\lim_{N\to\infty}\mathbb{E}\Big(\frac{1}{N}\sum_{i=1}^N x_i y_i\Big)^2 = 0\,.$$

□

We now have the following characterization of performance when $\beta_2 = 0$.

Lemma G.9. *If $\beta_2 = 0$ and $\mathbb{E}_\mathbb{Q} x_0 = 0$, then* $\mathbf{CS}(\mathbf{x}^0, \hat{\mathbf{x}}_{\mathrm{PMLE}}) = 0$.

Proof. By Theorem G.7 we have that (R_{11}, R_{10}) have limit points in the set $\mathcal{C}_{\bar{\beta}}$. It suffices to show that $\mathcal{C}_{\bar{\beta}}$ only contains points of the form $(s, 0)$.

When $\beta_2 = 0$, we have $H_N^{\bar\beta}(\mathbf{x})$ defined in (A.1) does not depend on $\mathbf{x}_0$ nor $\mathbb{Q}$. By symmetry, any near maximizer $\hat{\mathbf{x}}$ of $H_N^{\bar\beta}(\mathbf{x})$ has exchangeable bounded entries and is independent of $\mathbf{x}_0$. By Lemma G.8 it follows that any near maximizer satisfies

$$R_{10} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_i x_i^0 \to 0\,,$$

in probability. Since the possible limit points of the overlaps of near maximizers determine $\mathcal{C}_{\bar\beta}$ by Theorem G.7, the set $\mathcal{C}_{\bar\beta}$ only contains points of the form $(s, 0)$. □

Appendix H. Coarse Equivalence of Psuedo Estimators

In this section, we prove Theorem 3.2, using results proved in Sections A and G.

Proof of Theorem 3.2. We first consider the case of well-scored models. Given two well-scored loglikelihood functions g_1 and g_2, we let $\bar\beta(g_1)$ and $\bar\beta(g_2)$ to be the Fisher score parameter vectors corresponding to g_1 and g_2. Note that if the ratios satisfy:

$$\frac{\sqrt{\beta_1(g^1)}}{\sqrt{\beta_1(g^2)}} = \frac{\beta_2(g^1)}{\beta_2(g^2)} = \frac{\beta_3(g^1)}{\beta_3(g^2)}\,,$$

then $\bar\beta(g_1) = C\bar\beta(g_2)$ for some constant C. By Lemma A.1 this implies that

$$H_N^{\bar\beta(g^2)}(x) = C H_N^{\bar\beta(g^1)}(x)\,,$$

and hence both functions have the same maximizers. Therefore,

$$R(\hat{\mathbf{x}}^{\bar\beta(g_1)}_{\text{MLE}}, \mathbf{x}_0) = R(\hat{\mathbf{x}}^{\bar\beta(g_2)}_{\text{MLE}}, \mathbf{x}_0) \text{ and } R(\hat{\mathbf{x}}^{\bar\beta(g_1)}_{\text{MLE}}, \hat{\mathbf{x}}^{\bar\beta(g_1)}_{\text{MLE}}) = R(\hat{\mathbf{x}}^{\bar\beta(g_2)}_{\text{MLE}}, \hat{\mathbf{x}}^{\bar\beta(g_2)}_{\text{MLE}})\,.$$

On the other hand, by Theorem 2.1 and Lemma G.7, the maximizers satisfy

$$R(\hat{\mathbf{x}}^{g_1}_{\text{MLE}}, \mathbf{x}_0) = R(\hat{\mathbf{x}}^{\bar\beta(g_1)}_{\text{MLE}}, \mathbf{x}_0) \to m(g_1) \quad \text{and} \quad R(\hat{\mathbf{x}}^{g_1}_{\text{MLE}}, \hat{\mathbf{x}}^{g_1}_{\text{MLE}}) = R(\hat{\mathbf{x}}^{\bar\beta(g_1)}_{\text{MLE}}, \hat{\mathbf{x}}^{\bar\beta(g_1)}_{\text{MLE}}) \to s(g_1)$$
$$R(\hat{\mathbf{x}}^{g_2}_{\text{MLE}}, \mathbf{x}_0) = R(\hat{\mathbf{x}}^{\bar\beta(g_2)}_{\text{MLE}}, \mathbf{x}_0) \to m(g_2) \quad \text{and} \quad R(\hat{\mathbf{x}}^{g_2}_{\text{MLE}}, \hat{\mathbf{x}}^{g_2}_{\text{MLE}}) = R(\hat{\mathbf{x}}^{\bar\beta(g_2)}_{\text{MLE}}, \hat{\mathbf{x}}^{\bar\beta(g_2)}_{\text{MLE}}) \to s(g_2)\,,$$

where $s(g_i), m(g_i)$ are maximizers of $\psi_{\bar\beta(g_1)}$ and $\psi_{\bar\beta(g_2)}$ respectively. We conclude that the set of all limit points of the overlaps coincide so appealing again to Lemma G.7, we can conclude that $\mathcal{C}_{\bar\beta(g_1)} = \mathcal{C}_{\bar\beta(g_2)}$, which is the definition of coarsely equivalent PMLEs.

To prove part (b) of Theorem 3.2, note that if Ω satisfies $|\boldsymbol{x}| = C$ for every $\boldsymbol{x} \in \Omega$, then the third term in $H_N^{\bar\beta(g^1)}(x)$ and $H_N^{\bar\beta(g^2)}(x)$ are constant. Consequently one has for some constants C, D that:

$$C H_N^{\bar\beta(g^1)}(x) + D = H_N^{\bar\beta(g^2)}(x)\,,$$

and hence $H_N^{\bar\beta(g^1)}(x)$ and $H_N^{\bar\beta(g^2)}(x)$ have the same maximizers. The result then follows from the same argument as above.

□

The proof for illscored models is similar.

Proof of Theorem 3.3. For illscored models, notice that if $\beta_4(g^1)$ and $\beta_4(g^2)$ are non-zero and satisfy

$$\frac{\sqrt{\beta_1(g^1)}}{\sqrt{\beta_1(g^2)}} = \frac{\beta_2(g^1)}{\beta_2(g^2)} = \frac{\beta_3(g^1)}{\beta_3(g^2)} = \frac{\beta_4(g^1)}{\beta_4(g^2)} = C, \tag{H.1}$$

then Lemma A.1 implies that

$$H_N^{\bar\beta(g^2),\beta_4(g^2)}(x) = C H_N^{\bar\beta(g^1),\beta_4(g^1)}(x)\,.$$

From here the remainder of the proof is identical to the above, appealing to Theorem 2.5 instead of Theorem 2.1. □

We conclude with the proof of Theorem 3.1.

Proof of Theorem 3.1. For a given inference problem (g_0, g_1) with information parameters $\bar{\beta}$, the coarsely equivalent model given in (A.1) is equal to:

$$H_N^{\bar{\beta}}(x) = \frac{\sqrt{\beta_1}}{\sqrt{N}} \sum_{1 \le i \le j \le N} g_{ij} x_i x_j + \frac{\beta_2}{N} \sum_{1 \le i \le j \le N} x_i^0 x_j^0 x_i x_j - \frac{\beta_3}{2N} \sum_{1 \le i \le j \le N} x_i^2 x_j^2 + \beta_4 \sum_{1 \le i \le j \le N} \frac{x_i x_j}{\sqrt{N}}\,.$$

Let Y denote the matrix given by:

$$Y = \sqrt{\beta_1} G + \frac{\beta_2}{\sqrt{N}} x^0 (x^0)^T + \beta_4 \mathbf{1}\,,$$

with G a symmetric matrix whose entries are i.i.d $\mathcal{N}(0,1)$, and $\mathbf{1}$ is the matrix with all entries equal to 1. For any $x \in \Omega^N$, one then has the following equality:

$$\begin{aligned} H_N(x) &= -\frac{1}{2} \left\| Y - \frac{1}{\sqrt{N}} x x^T \right\|_F^2 - \frac{\beta_3 - 1}{N} \left\| x x^T \right\|_F^2 \\ &= \sum_{i,j=1}^N g_{U,1}^{\bar{\beta}}\Big(Y_{ij}, \frac{x_i x_j}{\sqrt{N}}\Big) - g_{U,1}^{\bar{\beta}}(Y_{ij}, 0)\,, \end{aligned}$$

and hence maximizing $H_N(x)$ corresponds exactly to the pseudo-likelihood inference task corresponding to the pair $(g_{U,0}^{\bar{\beta}}, g_{U,1}^{\bar{\beta}})$. This gives the interpretation of least squares with a correction term.

An alternative approach to proving theorem 3.1 is to compute the information parameters for the pair $(g_{U,0}^{\bar{\beta}}, g_{U,1}^{\bar{\beta}})$ directly to see they are equal to $\bar{\beta}$. We omit the calculations as they are straightforward. □

Appendix I. Properties of the Ruelle Probability Cascades

For a textbook introduction to the Ruelle probability cascades, we refer to [65, Chapter 2]. In this section, we recall only the essentials to understand the notation in the Appendices, and remind readers of its connection with the Parisi PDE. The Ruelle probability cascades are a measure on a Hilbert space indexed by $\mathbb{N}^r$ parameterized by sequences

$$\zeta_{-1} = 0 < \zeta_0 < \cdots < \zeta_{r-1} < 1 \tag{I.1}$$

and

$$0 = Q_0 \le Q_1 \le \cdots \le Q_{r-1} \le Q_r = S. \tag{I.2}$$

The weights of the Ruelle probability cascades is indexed by $\mathbb{N}^r$, the leaves of the infinite rooted tree with depth r encoded by the sequence of parameters ζ. Every leaf of the tree $\alpha = (n_1, \dots, n_r) \in \mathbb{N}^r$ can be encoded by a path along the vertices,

$$\alpha_{|1} = (n_1),\ \alpha_{|2} = (n_1, n_2),\ \dots,\ \alpha_{|r-1} = (n_1, n_2, \dots, n_{r-1}),\ \alpha = \alpha_{|r} = (n_1, \dots, n_r)$$

with the convention that $\alpha_{|0} = \emptyset$ is the root of the tree, and $k \le r$ denotes the distance from the vertex $\alpha_{|k} \in \mathbb{N}^k$ to the root. Each vertex $\beta_{|k} = (n_1, \dots, n_{k-1}, n_k)$ of the tree will be associated with a random variable $u_{\beta_{|k}}$ defined as follows: Let $\beta_{|k-1} = (n_1, \dots, n_{k-1})$ denote the parent of $\beta_{|k}$ and let

$$u_{(\beta_{|k-1},1)} > u_{(\beta_{|k-1},2)} > \cdots > u_{(\beta_{|k-1},n_k)} > \dots.$$

be the points from a Poisson process with mean measure $\zeta_{k-1}x^{-1-\zeta_{k-1}}$ arranged in decreasing order, and define

$$u_{\beta_{|k}} = u_{(n_1,\dots,n_{k-1},n_k)} = u_{(\beta_{|k-1},n_k)}.$$

We further assume that these points are generated independently for different parent vertices. For each leaf $\alpha \in \mathbb{N}^r$, the weights of the Ruelle probability cascades v_α is the product of these points along the path from the root to the leaf:

$$v_\alpha = \frac{u_{\alpha_{|1}} \cdots u_{\alpha_{|r}}}{\sum_{\beta \in \mathbb{N}^r} u_{\beta_{|1}} \cdots u_{\beta_{|r}}}.$$

We consider the Gaussian processes $Z(\alpha)$ and $Y(\alpha)$ indexed by points on the infinite tree $\mathbb{N}^r$ with covariances

$$\mathbb{E}Z(\alpha^1)Z(\alpha^2) = Q_{\alpha^1\wedge\alpha^2} \quad \mathbb{E}Y(\alpha^1)Y(\alpha^2) = \frac{1}{2}Q^2_{\alpha^1\wedge\alpha^2}.$$

The notation $\alpha^1 \wedge \alpha^2$ denotes the least common ancestor of the paths leaves α^1 and α^2 of the infinite tree indexed by $\mathbb{N}^r$,

$$\alpha^1 \wedge \alpha^2 = \min\Big\{0 \le j \le r \mid \alpha^1_{|1} = \alpha^2_{|1}, \dots, \alpha^1_{|j} = \alpha^2_{|j}, \alpha^1_{|j+1} \ne \alpha^2_{|j+1}\Big\}$$

These averages with respect to the Ruelle probability cascades variable α can be computed using the following recursive formulation from [65, Theorem 2.9].

Lemma I.1 (Averages with Respect to the Ruelle Probability Cascades)**.** *Let $C : \mathbb{R} \to \mathbb{R}$ be an increasing non-negative function. Suppose that there exists a Gaussian process $g(\alpha)$ by $\alpha \in \mathbb{N}^r$ with covariance*

$$\mathbb{E}g(\alpha^1)g(\alpha^2) = C(Q_{\alpha^1\wedge\alpha^2})$$

independent of v_α. For a function $f : \mathbb{R} \to \mathbb{R}$ we define

$$X_r = f\Big(\sum_{k=1}^r (C(Q_k) - C(Q_{k-1}))^{1/2} z_k\Big) \qquad X_p = \frac{1}{\zeta_p}\log \mathbb{E}_{z_{k+1}} e^{\zeta_p X_{p+1}} \quad \text{for } 0 \le p \le r-1$$

where z_k are iid standard Gaussians. If $\mathbb{E}e^{\zeta_{r-1}X_r} < \infty$ then

$$\mathbb{E}\log\sum_\alpha v_\alpha e^{f(g(\alpha))} = X_0.$$

The average on the outside is over the randomness in the Gaussian processes and the random measure v_α.

This is applied in Section B in the following way. We start by defining recursively the random variables $X_r, X_{r-1}, \dots, X_0$ that depend on x^0, the sequences (I.1) and (I.2), and real parameters λ, μ. Let X_r be the random variable

$$X_r = \log\int e^{\beta\sum_{j=1}^r z_i x + \lambda x^2 + \mu x x^0}\, d\,\mathbb{P}_X(x)$$

where z_j are Gaussian random variables with covariance

$$\mathrm{Var}(z_j) = Q_j - Q_{j-1}$$

and x^0 is an independent random variable with distribution $\mathbb{P}_0$. We define recursively for $0 \le p \le r-1$ the random variables

$$X_j = \frac{1}{\zeta_j}\log\mathbb{E}_{z_{j+1}} e^{\zeta_j X_{j+1}}. \tag{I.3}$$

Then Lemma I.1 implies that

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\log\sum_\alpha v_\alpha\int e^{\beta\sum_{i=1}^N Z_i(\alpha)x_i + \lambda x_i^2 + \mu x_i x_i^0}\, d\,\mathbb{P}_X(x) = \mathbb{E}_{\mathbb{Q}}X_0(x^0).$$

By continuity, one can represent the averages with respect to the Ruelle probability cascades as the solution to the Parisi PDE, which is the form we are using in this work. The details of this reduction can be found in [65, Section 4.1]. Consider a distribution function $\zeta(t)$ such that

$$\zeta(t) = \zeta_p \qquad Q_p \leq t < Q_{p+1}$$

for $p = 0, \dots, r$. The following Lemma shows that we can approximate the discrete distributions

Lemma I.2. *For every discrete distribution function $\zeta(t)$ encoded by the parameters (I.1) and (I.2), then*

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E} \log \sum_\alpha v_\alpha \int e^{\beta \sum_{i=1}^N Z_i(\alpha) x_i + \lambda x_i^2 + \mu x_i x_i^0} \, d\,\mathbb{P}_X(x) = \mathbb{E}_{\mathbb{Q}} X_0(x^0) = \Phi_\zeta(0,0).$$

Furthermore, by continuity [65, Lemma 4.1] we can extend this result to any distribution function $\zeta(t)$, which gives us the representation in (B.24).

We end this section by stating a an upper bound of the Ruelle probability cascades of a partition with respect to its maximum value on each partition. The proof can be found in [67, Lemma 6].

Lemma I.3 (Upper Bound of the Ruelle Probability Cascades)**.** *Let $g(\alpha)$ be a Gaussian process indexed by $\alpha \in \mathbb{N}^r$ with covariance*

$$\mathbb{E} g(\alpha^1) g(\alpha^2) = C(Q_{\alpha^1 \wedge \alpha^2})$$

independent of v_α. If $A_j : \mathbb{R} \to \mathbb{R}$ are positive functions of the same Gaussian process $g(\alpha)$ for $1 \leq j \leq n$ then

$$\mathbb{E} \log \sum_{\alpha \in \mathbb{N}^r} v_\alpha \sum_{j \leq n} A_j(g(\alpha)) \leq \frac{\log n}{\zeta_0} + \max_{j \leq n} \mathbb{E} \log \sum_{\alpha \in \mathbb{N}^r} v_\alpha A_j(g(\alpha)),$$

where $\zeta_0 > 0$ is the smallest point in the sequence (I.1).

References

1. Emmanuel Abbe, *Community detection and stochastic block models: Recent developments*, Journal of Machine Learning Research **18** (2018), no. 177, 1–86.
2. Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall, *Exact recovery in the stochastic block model*, IEEE Transactions on Information Theory **62** (2016), no. 1, 471–487.
3. Robert J. Adler and Jonathan E. Taylor, *Random fields and geometry*, Springer Monographs in Mathematics, Springer, New York, 2007. MR 2319516
4. Michael Aizenman, Robert Sims, and Shannon L. Starr, *Extended variational principle for the sherrington-kirkpatrick spin-glass model*, Phys. Rev. B **68** (2003), 214403.
5. Ahmed El Alaoui, Florent Krzakala, and Michael Jordan, *Fundamental limits of detection in the spiked Wigner model*, The Annals of Statistics **48** (2020), no. 2, 863 – 885.
6. Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, Cambridge Studies in Advanced Mathematics, vol. 118, Cambridge University Press, Cambridge, 2010. MR 2760897
7. Antonio Auffinger and Wei-Kuo Chen, *Parisi formula for the ground state energy in the mixed p-spin model*, The Annals of Probability **45** (2017), no. 6B, 4617 – 4631.
8. Jean Barbier, Wei-Kuo Chen, Dmitry Panchenko, and Manuel Sáenz, *Performance of bayesian linear regression in a model with mismatch*, 2021.
9. Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová, *Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula*, Proceedings of the 30th International Conference on Neural Information Processing Systems (Red Hook, NY, USA), NIPS'16, Curran Associates Inc., 2016, p. 424–432.
10. Jean Barbier, TianQi Hou, Marco Mondelli, and Manuel Sáenz, *The price of ignorance: how much does it cost to forget noise structure in low-rank matrix estimation?*, 2022.
11. Jean Barbier and Galen Reeves, *Information-theoretic limits of a multiview low-rank symmetric spiked matrix model*, 2020 IEEE International Symposium on Information Theory (ISIT), IEEE, 2020, pp. 2771–2776.
12. Gérard Ben Arous and Aukosh Jagannath, *Spectral gap estimates in mean field spin glasses*, Comm. Math. Phys. **361** (2018), no. 1, 1–52. MR 3825934

13. Shankar Bhamidi, Partha Dey, and Andrew Nobel, *Energy landscape for large average submatrix detection problems in gaussian random matrices*, Probability Theory and Related Fields **168** (2017).
14. Cristina Butucea and Yuri I. Ingster, *Detection of a sparse submatrix of a high-dimensional noisy matrix*, Bernoulli **19** (2013), no. 5B, 2652 – 2688.
15. Francesco Camilli, Pierluigi Contucci, and Emanuele Mingione, *An inference problem in a mismatched setting: a spin-glass model with Mattis interaction*, SciPost Phys. **12** (2022), no. 4, Paper No. 125, 27. MR 4409513
16. Philippe Carmona and Yueyun Hu, *Universality in Sherrington-Kirkpatrick's spin glass model*, Ann. Inst. H. Poincaré Probab. Statist. **42** (2006), no. 2, 215–222. MR 2199799
17. Michael Celentano, Zhou Fan, and Song Mei, *Local convexity of the TAP free energy and AMP convergence for -synchronization*, The Annals of Statistics **51** (2023), no. 2, 519 – 546.
18. Michael Celentano, Zhou Fan, and Song Mei, *Local convexity of the tap free energy and amp convergence for z 2-synchronization*, The Annals of Statistics **51** (2023), no. 2, 519–546.
19. Hong-Bin Chen, Jean-Christophe Mourrat, and Jiaming Xia, *Statistical inference of finite-rank tensors*, 2021.
20. Hong-Bin Chen and Jiaming Xia, *Hamilton-Jacobi equations for inference of matrix tensor products*, Ann. Inst. Henri Poincaré Probab. Stat. **58** (2022), no. 2, 755–793. MR 4421607
21. Cosme Louart, Zhenyu Liao, and Romain Couillet, *A random matrix approach to neural networks*, The Annals of Applied Probability **28** (2018), no. 2, 1190–1248.
22. J R L de Almeida and D J Thouless, *Stability of the sherrington-kirkpatrick solution of a spin glass model*, Journal of Physics A: Mathematical and General **11** (1978), no. 5, 983.
23. A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, second ed., Applications of Mathematics (New York), vol. 38, Springer-Verlag, New York, 1998. MR 1619036
24. Yash Deshpande and Andrea Montanari, *Information-theoretically optimal sparse pca*, 2014 IEEE International Symposium on Information Theory, IEEE, 2014, pp. 2197–2201.
25. Tomas Dominguez and Jean-Christophe Mourrat, *Mutual information for the sparse stochastic block model*, 2022.
26. Tomas Dominguez and Jean-Christophe Mourrat, *Mutual information for the sparse stochastic block model*, The Annals of Probability **52** (2024), no. 2, 434–501.
27. David L. Donoho, Arian Maleki, and Andrea Montanari, *Message-passing algorithms for compressed sensing*, Proceedings of the National Academy of Sciences **106** (2009), no. 45, 18914–18919.
28. Carl Eckart and G. Marion Young, *The approximation of one matrix by another of lower rank*, Psychometrika **1** (1936), 211–218.
29. Zhou Fan, *Approximate message passing algorithms for rotationally invariant matrices*, The Annals of Statistics **50** (2022).
30. Zhou Fan, Song Mei, and Andrea Montanari, *Tap free energy, spin glasses and variational inference*, The Annals of Probability **49** (2021), no. 1, 1–45.
31. Michael J Feldman, *Spectral properties of elementwise-transformed spiked matrices*, arXiv preprint arXiv:2311.02040 (2023).
32. R. A. Fisher, *On the mathematical foundations of theoretical statistics*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **222** (1922), 309–368.
33. Teng Fu, YuHao Liu, Jean Barbier, Marco Mondelli, ShanSuo Liang, and TianQi Hou, *Mismatched estimation of non-symmetric rank-one matrices corrupted by structured noise*, 2023 IEEE International Symposium on Information Theory (ISIT), 2023, pp. 1178–1183.
34. C. Gourieroux, A. Monfort, and A. Trognon, *Pseudo maximum likelihood methods: Theory*, Econometrica **52** (1984), no. 3, 681–700.
35. Francesco Guerra, *Broken replica symmetry bounds in the mean field spin glass model*, Communications in mathematical physics **233** (2003), no. 1, 1–12.
36. Alice Guionnet, Justin Ko, Florent Krzakala, Pierre Mergny, and Lenka Zdeborová, *Spectral phase transitions in non-linear Wigner spiked models*, arXiv preprint arXiv:2310.14055 (2023).
37. Alice Guionnet, Justin Ko, Florent Krzakala, and Lenka Zdeborová, *Estimating rank-one matrices with mismatched prior and noise: universality and large deviations*, 2023.
38. Bruce Hajek, Yihong Wu, and Jiaming Xu, *Information limits for recovering a hidden community*, IEEE Transactions on Information Theory **63** (2017), no. 8, 4729–4745.
39. ———, *Recovering a hidden community beyond the kesten–stigum threshold in o (| e| log*| v|) time*, Journal of Applied Probability **55** (2018), no. 2, 325–352.
40. Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer, *The power of sum-of-squares for detecting hidden structures*, 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2017, pp. 720–731.

41. Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer, *Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors*, Proceedings of the forty-eighth annual ACM symposium on Theory of Computing, 2016, pp. 178–191.
42. Samuel B Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, Conference on Learning Theory, PMLR, 2015, pp. 956–1006.
43. Yukito Iba, *The nishimori line and bayesian statistics*, Journal of Physics A: Mathematical and General **32** (1999), no. 21, 3875.
44. Aukosh Jagannath, Patrick Lopatto, and Leo Miolane, *Statistical thresholds for tensor pca*, The Annals of Applied Probability **30** (2020), no. 4, 1910–1933.
45. Aukosh Jagannath and Subhabrata Sen, *On the unbalanced cut problem and the generalized sherrington-kirkpatrick model*, Ann. Inst. Henri Poincaré **8** (2020).
46. Aukosh Jagannath and Ian Tobasco, *A dynamic programming approach to the parisi functional*, Proceedings of the American Mathematical Society **144** (2016), no. 7, 3135–3150.
47. Aukosh Jagannath and Ian Tobasco, *Low temperature asymptotics of spherical mean field spin glasses*, Communications in Mathematical Physics **352** (2016), 979–1017.
48. Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi, *Phase transitions in semidefinite relaxations*, Proceedings of the National Academy of Sciences **113** (2016), no. 16, E2218–E2223.
49. Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi, *Phase transitions in semidefinite relaxations*, Proceedings of the National Academy of Sciences **113** (2016), no. 16, E2218–E2223.
50. Iain M Johnstone and Arthur Yu Lu, *On consistency and sparsity for principal components analysis in high dimensions*, Journal of the American Statistical Association **104** (2009), no. 486, 682–693.
51. Noureddine El Karoui, *On information plus noise kernel random matrices*, The Annals of Statistics **38** (2010), no. 5, 3191 – 3216.
52. ———, *The spectrum of kernel random matrices*, The Annals of Statistics **38** (2010), no. 1, 1 – 50.
53. Brian Karrer and M. E. J. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E **83** (2011), 016107.
54. Dmitriy Kunisky, *Low coordinate degree algorithms i: Universality of computational thresholds for hypothesis testing*, arXiv preprint arXiv:2403.07862 (2024).
55. Marc Lelarge and Léo Miolane, *Fundamental limits of symmetric low-rank matrix estimation*, Conference on Learning Theory, PMLR, 2017, pp. 1297–1301.
56. Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová, *Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel*, 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2015, pp. 680–687.
57. Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová, *Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications*, Journal of Statistical Mechanics: Theory and Experiment **2017** (2017), no. 7, 073403 (en).
58. Clément Luneau, Jean Barbier, and Nicolas Macris, *Mutual information for low-rank even-order symmetric tensor estimation*, Information and Inference: A Journal of the IMA **10** (2020), no. 4, 1167–1207.
59. S.C. Madeira and A.L. Oliveira, *Biclustering algorithms for biological data analysis: a survey*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **1** (2004), no. 1, 24–45.
60. Vaishakhi Mayya and Galen Reeves, *Mutual information in community detection with covariate information and correlated networks*, 2019 57th annual allerton conference on communication, control, and computing (allerton), IEEE, 2019, pp. 602–607.
61. Pierre Mergny, Justin Ko, Florent Krzakala, and Lenka Zdeborová, *Fundamental limits of non-linear low-rank matrix estimation*, Proceedings of Thirty Seventh Conference on Learning Theory (Shipra Agrawal and Aaron Roth, eds.), Proceedings of Machine Learning Research, vol. 247, PMLR, 30 Jun–03 Jul 2024, pp. 3873–3873.
62. Léo Miolane, *Phase transitions in spiked matrix estimation: information-theoretic analysis*, arXiv preprint arXiv:1806.04343 (2018).
63. Andrea Montanari and Alexander S Wein, *Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation*, arXiv preprint arXiv:2212.06996 (2022).
64. Dmitry Panchenko, *The Parisi ultrametricity conjecture*, Ann. of Math. (2) **177** (2013), no. 1, 383–393. MR 2999044
65. ———, *The Sherrington-Kirkpatrick model*, Springer Monographs in Mathematics, Springer, New York, 2013. MR 3052333
66. ———, *Free energy in the mixed p-spin models with vector spins*, Ann. Probab. **46** (2018), no. 2, 865–896. MR 3773376
67. ———, *Free energy in the Potts spin glass*, Ann. Probab. **46** (2018), no. 2, 829–864. MR 3773375

68. Giorgio Parisi, *Infinite number of order parameters for spin-glasses*, Physical Review Letters **43** (1979), no. 23, 1754.
69. Jeffrey Pennington and Pratik Worah, *Nonlinear random matrix theory for deep learning*, Advances in Neural Information Processing Systems (I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
70. Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra, *Optimality and sub-optimality of pca i: Spiked random matrix models*, The Annals of Statistics **46** (2018), no. 5, 2416–2451.
71. Amelia Perry, Alexander S. Wein, Afonso S. Bandeira, and Ankur Moitra, *Optimality and sub-optimality of PCA i: Spiked random matrix models*, The Annals of Statistics **46** (2018), no. 5.
72. Vanessa Piccolo and Dominik Schröder, *Analysis of one-hidden-layer neural networks via the resolvent method*, Advances in Neural Information Processing Systems (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, eds.), vol. 34, Curran Associates, Inc., 2021, pp. 5225–5235.
73. Farzad Pourkamali and Nicolas Macris, *Mismatched estimation of non-symmetric rank-one matrices under gaussian noise*, 2022 IEEE International Symposium on Information Theory (ISIT), 2022, pp. 1288–1293.
74. Sundeep Rangan and Alyson K Fletcher, *Iterative estimation of constrained rank-one matrices in noise*, 2012 IEEE International Symposium on Information Theory Proceedings, IEEE, 2012, pp. 1246–1250.
75. Galen Reeves, *Information-theoretic limits for the matrix tensor product*, IEEE Journal on Selected Areas in Information Theory **1** (2020), 777–798.
76. Romain Couillet and Florent Benaych-Georges, *Kernel spectral clustering of large dimensional data*, Electronic Journal of Statistics **10** (2016), no. 1, 1393–1454.
77. M.J. Schervish, *Theory of statistics*, Springer Series in Statistics, Springer New York, 2012.
78. David Sherrington and Scott Kirkpatrick, *Solvable model of a spin-glass*, Phys. Rev. Lett. **35** (1975), 1792–1796.
79. A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Applied and Computational Harmonic Analysis **30** (2011), no. 1, 20–36.
80. A. Singer and Y. Shkolnisky, *Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming*, SIAM Journal on Imaging Sciences **4** (2011), no. 2, 543–572.
81. Michel Talagrand, *The Parisi formula*, Ann. of Math. (2) **163** (2006), no. 1, 221–263. MR 2195134
82. F. L. Toninelli, *About the almeida-thouless transition line in the sherrington-kirkpatrick mean-field spin glass model*, Europhysics Letters **60** (2002), no. 5, 764.
83. Hui Zou, Trevor Hastie, and Robert Tibshirani, *Sparse principal component analysis*, Journal of computational and graphical statistics **15** (2006), no. 2, 265–286.

(Curtis Grant) Northwestern University
Email address: curtisgrant2026@u.northwestern.edu

(Aukosh Jagannath) University of Waterloo
Email address: a.jagannath@uwaterloo.ca

(Justin Ko) University of Waterloo
Email address: justin.ko@uwaterloo.ca