Bayesian Inference With Nonlinear Generative Models: Comments on Secure Learning

Ali Bereyhi, Member, IEEE, Bruno Loureiro, Florent Krzakala, Ralf R. Müller, Fellow, IEEE, and Hermann Schulz-Baldes

Abstract—Unlike the classical linear model, nonlinear generative models have been addressed sparsely in the literature of statistical learning. This work aims to shed light on these models and their secrecy potential. To this end, we invoke the replica method to derive the asymptotic normalized cross entropy in an inverse probability problem whose generative model is described by a Gaussian random field with a generic covariance function. Our derivations further demonstrate the asymptotic statistical decoupling of the Bayesian estimator and specify the decoupled setting for a given nonlinear model. The replica solution depicts that strictly nonlinear generative models establish an all-or-nothing phase transition: there exists a critical load at which the optimal Bayesian inference changes from perfect to an uncorrelated learning. Based on this finding, we design a new secure coding scheme which achieves the secrecy capacity of the wiretap channel. This interesting result implies that strictly nonlinear generative models are perfectly secured without any secure coding. We justify this latter statement through the analysis of an illustrative model for perfectly secure and reliable inference.

Index Terms—Nonlinear generative models, Bayesian inference, Gaussian random fields, information-theoretically secure learning, replica method, decoupling principle.

I. INTRODUCTION

Inference over linear models is a widely-investigated problem in the literature of statistical learning [2], [3], [4], [5], [6], [7], [8], [9]. The popularity of linear models comes from a simple fact: they have a large scope of applications from communications and information theory [10], [11], [12], [13], [14], [15], [16] to signal processing and machine learning [17], [18], [19], [20], [21], [22], [23], [24]. Though linear models have been extensively studied, the problem of inference from observations given by a nonlinear model has been touched very sparsely [25], [26], [27]. This work bridges this gap by demonstrating the potential gains which can be achieved by nonlinear models.

Unlike in the statistical learning literature, nonlinear models are well known in physics. Indeed, various spin glass models from statistical mechanics are described by nonlinear fields [28], [29], [30], [31]. Following the close bonds between the applied concepts in information theory, such as channel and source coding, and theoretical concepts in statistical mechanics, such as Gibbs’ inequality, [32], [33], [34], [35], it is natural to ask how nonlinear models of spin glasses are connected to information-theoretic concepts. This connection has been studied only sparsely in the literature of statistical learning. The most seminal line of work in this direction was considered by Sourlas in [25], where it is shown that by designing an error-correcting code based on a nonlinear spin glass model, one can achieve an excellent performance which, under some circumstances, approaches Shannon’s limit.1 This seminal work opened the door to a key application of nonlinear models in statistical learning and led to a fruitful synergy between concepts in information theory and statistical physics; see for instance [26], [36]. Nevertheless, the development of these study lines into a general framework for learning from a nonlinear generative model has been left unaddressed in the literature of statistical learning.

In this work, we study the problem of inferring model parameters from observations given by a Gaussian nonlinear generative model. Our derivations reveal an intuitive result: while Bayesian inference from linear observations always contain information about the ground truth, the output of a nonlinear model can become completely uncorrelated to the

1Sourlas’ seminal result is discussed in more detail later in this manuscript.
true model parameters. This behavior is analytically described through asymptotic characterization of the joint statistics of the optimal Bayesian estimate and the ground truth via the replica method. Our derivations show that the overlap between the optimal Bayesian inference and the true model parameters exhibits a first-order phase transition from perfect learning to uncorrelated learning, when the generative model is strictly nonlinear. In the light of this fundamental finding, we illustrate the application of nonlinear models to the problem of secure coding. It is shown that using a strictly nonlinear generative model as an encoder, the secrecy capacity of the wiretap channel [37] is asymptotically achieved.

Our findings on the bonds between nonlinear models and the problem of secure transmission further answers the question concluded by Fyodorov in his recent study [27]. In this work, a nonlinear Gaussian field is used to encrypt data for transmission over an additive white Gaussian noise (AWGN) channel. The results of this work lead to an interesting conclusion: using strictly nonlinear fields, the average error of recovery shows a second-order phase transition with respect to the noise variance. Following this finding, Fyodorov states this question: can a nonlinear encryption strategy lead to secure transmission with a publicly known codebook? We answer to this question by linking Fyodorov’s encryption with the seminal work of Sourlas and establishing a coding scheme for secure transmission over a wiretap channel. Our derivations demonstrate that although non-zero secrecy rate is achievable by combining Fyodorov’s framework and that of Sourlas, it does not achieve the secrecy capacity. This is in contrast to our proposed secure coding scheme which achieves the secrecy capacity of the wiretap channel.

A. Contributions and Tools

We study Bayesian inference over a nonlinear generative model which is described by a Gaussian random field. We investigate the inverse probability problem for a generally mismatched Bayesian inference algorithm. For the considered generic model, we derive the normalized cross entropy as a generic performance metric, from which various information metrics can be derived as special cases, e.g., information rate. We further show the decoupling principle for this model and derive explicitly the decoupled scalar setting. This latter result enables us to evaluate a larger class of asymptotic metrics, such as a generic average input-output distortion and the average cross correlation, known as the overlap. Our derivations lead to the following key findings:

- With linear models, inference via the optimal Bayesian algorithm always contains non-zero correlation with the ground truth; however, the achievable information rate never touches Shannon’s limit.
- Strictly nonlinear models achieve the maximal information rate given by Shannon’s limit; however, they show a first-order phase transition with respect to the system load at which the learned parameters become uncorrelated with the ground truth.

The above findings lead us to an interesting result: to achieve the secure capacity of a wiretap channel, one can directly encode the information bits by passing them through a strictly nonlinear Gaussian random field. This finding not only re-disCOVERS Wyner’s result on the wiretap channel, but also suggests a new secure coding scheme which achieves the secure capacity of the wiretap channel. An intuitive conclusion from this result is that strictly nonlinear generative models are secure by nature. We give a heuristic proof to this fact by considering the problem of perfectly secure and reliable inference in a wiretap setting. For strictly nonlinear models, the maximum number of securely inferred labels per dimension meets the information-theoretic limit given by Wyner in [37]. We conclude this work by discussing some new directions for further investigations.

The key tool for large-system analysis in this work is the replica method from the theory of spin glasses and commonly utilized as an analytical tool in the context of information theory and signal processing; see for example [10], [11], [12], [38], [39], and [40]. A comprehensive introduction to the replica method and its application to the analysis of linear models can be found in [2]. The derivations in this work consider the replica symmetric ansatz. This follows from the fact that for Bayes-optimal inference the corresponding spin glass exhibits replica symmetry\(^4\) [41]. However, deviation from the indicated conditions can lead to the so-called replica symmetry breaking. More detailed discussions are given in Section III and Appendix B. The organization of this paper is presented in Table I, where the list of main sections along with their contents is given. The list of appendices is further given in Table II.

B. Notation

Scalars, vectors and matrices are represented with non-bold, bold lower-case and bold upper-case letters, respectively. The transpose of \( \mathbf{A} \) is indicated by \( \mathbf{A}^T \), \( \mathbf{I}_N \) is an \( N \times N \) identity matrix, and \( \mathbf{1}_N \) is an \( N \times N \) all-one matrix. The Euclidean norm of \( \mathbf{x} \) is denoted by \( \| \mathbf{x} \| \). For two \( K \)-dimensional vectors \( \mathbf{x} \) and \( \mathbf{y} \), we define the normalized inner-product as

\[
\langle \mathbf{x}; \mathbf{y} \rangle = \frac{\mathbf{x}^T \mathbf{y}}{K}.
\]

We extend this notation to matrix arguments with valid dimensions as

\[
\langle \mathbf{X}; \mathbf{Y} \rangle = \frac{\mathbf{X}^T \mathbf{Y}}{K},
\]

where \( K \) denotes the number of rows in \( \mathbf{X} \) and \( \mathbf{Y} \). The notation \( \log ( \cdot ) \) indicates the natural logarithm. We use \( \mathbb{E}_X \{ \cdot \} \) to refer to expectation with respect to \( X \) using the true distribution of \( X \). The notation \( \mathbb{E}_{q_X} \{ \cdot \} \) indicates expectation with respect to \( X \) when the distribution of \( X \) is replaced by \( q_X \). By using \( \mathbb{E} \{ \cdot \} \), we denote expectation with respect to

\(^2\)This question is stated by Fyodorov through an intuitive short discussion on the main results in [27, Section 1.2.1].

\(^3\)We show this connection in Section V-B.

\(^4\)This property is discussed in detail later on.
Table I
Main Sections of the Paper

| Section                          | Content                                      |
|----------------------------------|----------------------------------------------|
| I Introduction                   | I-A Contributions and Tools                  |
|                                 | I-B Notation                                 |
| II Statement of the Problem      | II-A Origin of the Higher-Order Model         |
|                                 | II-B Bayesian Inference                       |
|                                 | II-C Initial Setup                            |
|                                 | II-D Some Earlier Studies                     |
| III Corresponding Spin Glass with \(P\)-Spin Model | III-A The Variational Problem                |
|                                 | III-B Main Results                            |
|                                 | III-C Decoupling Principle                    |
| IV Linear Models versus Nonlinear Models | IV-A Classical Case: Linear Generative Model |
|                                 | IV-B Quadratic Generative Model               |
|                                 | IV-C Higher-order Nonlinear Generative Models |
| V Comments on Secure Learning    | V-A Secure Inference via Nonlinear Generative Models |
|                                 | V-B Connecting Fyodorov’s Results to Wyner’s  |
| VI Conclusions and Final Remarks |                                              |

Table II
List of Appendices

| Appendix | Content                                      |
|----------|----------------------------------------------|
| Appendix A | Fyodorov’s Nonlinear Encryption               |
| Appendix B | Derivations via the Replica Method            |
| Appendix C | Decoupling Principle                          |

All random variables involved in the argument. The entropy of a discrete random variable \(X\) and the differential entropy of a continuous random variable \(Y\) are shown by \(H(X)\) and \(h(Y)\), respectively. The notation \(\mathcal{N}(\eta, C)\) denotes the multivariate Gaussian distribution with mean vector \(\eta\) and covariance matrix \(C\). To refer to the real axis, we use the notation \(\mathbb{R}\). We further denote the unit \((D - 1)\)-sphere by \(S^{D-1}\), i.e.,

\[
S^{D-1} = \{x \in \mathbb{R}^D : ||x|| = 1\}.
\]

For brevity, \([1, \ldots, N]\) and \([N_0, \ldots, N_1]\) are abbreviated as \([N]\) and \([N_0 : N_1]\), respectively.

II. Statement of the Problem

Consider the classical regression problem: an unknown sequence of \(K\) symbols

\[
s = [s_1, \ldots, s_K]
\]

is projected onto \(\mathbb{R}^N\) via a randomly generated but known function \(\mathcal{V}(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^N\). The output of this transform, i.e.,

\[
x = \mathcal{V}(s) \in \mathbb{R}^N,
\]

is observed through an AWGN channel. This means that we are provided with

\[
y = x + w,
\]

where \(w \in \mathbb{R}^N\) is an independent and identically distributed (i.i.d.) zero-mean Gaussian vector with variance \(\sigma^2\), i.e., \(w \sim \mathcal{N}(0, \sigma^2)\). Our objective is to learn \(s\) from \(y\) considering the fact that \(\mathcal{V}(\cdot)\) is known as a side-information. The most well-known case of this setting is the linear regression model in which \(\mathcal{V}(\cdot)\) is considered to be a linear function, i.e., \(\mathcal{V}(s) = As\) for some \(A \in \mathbb{R}^{N \times K}\).

We intend to deviate from this classical assumption in this work and investigate a larger class of generative models; namely, the higher-order Gaussian random fields defined as follows:

**Definition 1 (Gaussian Random Field):** Let \(\mathcal{V}(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^N\) be a random mapping whose output entries for an input \(s \in \mathbb{R}^K\) are denoted as

\[
\mathcal{V}(s) = [\mathcal{V}_1(s), \ldots, \mathcal{V}_N(s)]^T.
\]

The mapping \(\mathcal{V}(\cdot)\) is a Gaussian random field with covariance function \(\Phi(\cdot)\), if for any \(s_1, s_2 \in \mathbb{R}^K\), the entries of \(\mathcal{V}(s_i)\) for \(i \in \{1, 2\}\) are distributed Gaussian and satisfy\(^5\)

\[
\mathbb{E}\{\mathcal{V}_m(s_1)\mathcal{V}_n(s_2)\} = 1\{m = n\} \Phi(\langle s_1; s_2 \rangle),
\]

for any \(n, m \in [N]\).

The above definition becomes clear if we consider some particular examples:

**Example 1 (Linear Model):** The most basic Gaussian field is the linear field, for which \(\Phi(x) = x\). In this case, \(\mathcal{V}(s) = As\) for an \(N \times K\) i.i.d. Gaussian matrix \(A\) whose elements are zero-mean with variance \(1/K\).

**Example 2 (Pure Quadratic Model):** As a simple extension of the linear Gaussian field, we consider the quadratic Gaussian random field with \(\Phi(x) = x^2\). In this case, a particular output entry is determined as

\[
\mathcal{V}_n(s) = s^T J_n s,
\]

where \(J_n\) for \(n \in [N]\) are independent i.i.d. \(K \times K\) random matrices whose entries are zero-mean Gaussian random variables with variance \(1/K^2\).

**Example 3 (General Quadratic Model):** A more general form of a quadratic field contains also a linear term. This

\(^5\)Note that \(\langle s_1; s_2 \rangle\) is the normalized inner-product defined in (1).
means that a particular output entry is determined as
\[ V_n(s) = s^T J_n s + a_n^T s, \] (10)
where \( J_n \in \mathbb{R}^{K \times K} \) for \( n \in [N] \) are mutually independent i.i.d. Gaussian random matrices whose entries are zero-mean with variance \( 1/K^2 \), and \( a_n \in \mathbb{R}^K \) are independent i.i.d. Gaussian vectors with mean zero and covariance \( \mathbf{I}_K/K \).

In the sequel, we consider a generic covariance function which can be represented as
\[ \Phi(u) = \sum_{\ell=1}^{\infty} c_\ell u^\ell, \] (11)
for coefficients \( c_\ell \geq 0 \).

**A. Origin of the Higher-Order Model**

The interest on higher-order Gaussian random fields comes from various sources. More relevant to the scope of this work is the so-called nonlinear Gaussian encryption method, discussed in [27]. In this method, a nonlinear Gaussian random field is used to encrypt a data sequence \( s \) into an \( x \) of larger length. As shown in [27], the non-linearity adds confusion to the system which prevents a degraded eavesdropper from overhearing the data sequence. As we show in the coming sections, this result can be observed as a non-trivial alternative derivation of Wyner’s wiretap result [37]. We discuss this aspect with more details later on in this manuscript.

Besides nonlinear Gaussian encryption, there are various other models which bring interest on higher-order Gaussian fields. An example is the multi-edge-type parity check code ensembles, in particular the low density generator matrix (LDGM) codes, whose generator can be seen as a field with a polynomial covariance function \( \Phi(x) \) [42], [43], [44]. This is the most relevant aspect connected to Sourlas’ seminal work in [25].

An alternative source of interest comes from a mathematical viewpoint. In fact, a polynomial order Gaussian field is the superposition of multiple multi-dimensional tensors with random entries. This can be seen as a natural extension of a linear field\(^6\) into higher dimensions. This viewpoint in particular is of interest in several machine learning applications [45], [46], [47], [48], [49], [50].

As mentioned in the introduction, Gaussian random fields are well-known in statistical mechanics; namely, in the theory of spin glasses, in which random fields are used to model the energy level of a spin glass. In the particular case of a linear field, the problem of analyzing the energy model in the thermodynamic limit is homomorphic to the asymptotic analysis of linear regression model.\(^7\) Considering this homomorphism, one can interpret a higher-order Gaussian field as a natural extension of linear regression to the problem of regression with a nonlinear Gaussian generative model. More precisely, the inference from the output of a linear model is mathematically equivalent to the free-energy calculation for a quadratic Hamiltonian being described via the Sherrington-Kirkpatrick model [29], [30]
\[ \mathcal{E}(s) = -\sum_{k,\ell} J_{k,\ell} s_k s_\ell. \] (12)
Here, \( \mathcal{E}(s) \) denotes the Hamiltonian of the spin glass which corresponds to the linear regression problem; see [30] for the formal definition. The entries \( J_{k,\ell} \) are further random coefficients. This model determines the energy of a spin glass in terms of pair-wise interactions. A natural extension of this model is the so-called \( P \)-spin model in which the Hamiltonian is considered to be
\[ \mathcal{E}(s) = -\sum_{\ell_1,\ldots,\ell_P} J_{\ell_1,\ldots,\ell_P} \prod_{p=1}^{P} s_{\ell_p}. \] (13)
Here, the energy is determined as the sum of \( P \)-tuple interactions among the particles. This model in the asymptotic case of \( P \uparrow \infty \) describes the random energy model (REM) in which the energy levels are assigned i.i.d. at random and do not depend on the configuration of the system particles anymore [51], [52]. The particular case of REM has been investigated extensively in the literature; see for example [53].

**B. Bayesian Inference**

Let us now get back to our main objective, i.e., reconstructing \( s \) from \( y \) given \( \mathcal{V}(\cdot) \). To this end, we estimate \( s \) via a Bayesian estimator as
\[ \hat{s} = \mathbb{E}[s|y, \mathcal{V}] = \int \text{d}Q(s|y, \mathcal{V}) s, \] (14)
where \( Q(s|y, \mathcal{V}) \) denotes the posterior cumulative distribution of \( s \) determined for some postulated i.i.d. prior belief
\[ s \sim q_S(s) = \prod_{k=1}^{K} q_S(s_k), \] (15)
and some postulated noise variance \( \hat{\sigma}^2 \). This estimator minimizes the expected error when it is calculated via the Euclidean distance for the given prior belief.

To evaluate the performance of the Bayesian estimator, we determine the normalized cross entropy between the following two distributions:
1) The true output distribution, i.e.,
\[ p_Y(y|\mathcal{V}) = (2\pi \hat{\sigma}^2)^{-N/2} \mathbb{E}_{PS} \left\{ \exp \left\{ -\frac{||y - \mathcal{V}(s)||^2}{2\hat{\sigma}^2} \right\} \right\}, \]
with the expectation being taken with respect to the true prior distribution \( PS(s) \).
2) The output distribution induced by the postulated prior and noise variance, i.e.,
\[ q_Y(y|\mathcal{V}) = (2\pi \hat{\sigma}^2)^{-N/2} \mathbb{E}_{q_S} \left\{ \exp \left\{ -\frac{||y - \mathcal{V}(s)||^2}{2\hat{\sigma}^2} \right\} \right\}. \]
We refer to this distribution in the sequel as the induced distribution.

Hence, the performance metric for fixed \( \sigma, \hat{\sigma}, PS \) and \( q_S \) is given by
\[ \Delta_N(\sigma, \hat{\sigma}, PS, q_S) = -\frac{1}{N} \mathbb{E}_{y, \mathcal{V}} \left\{ \log q_Y(y|\mathcal{V}) \right\}. \] (16)

---

\(^6\)A linear field can be seen as a two-dimensional tensor.

\(^7\)See for example [2, Chapter 3].
Here, the expectation is taken with respect to the true output distribution. It is worth noting that in the definition of \( \Delta_N \), we further determine the expectation with respect to the field. This means that the determined metric represents the expected cross entropy. The self-averaging property indicates that a particular realization of the cross entropy converges to this expectation in the large-system limit [54], [55].

In general, various metrics can be directly derived from the cross entropy. Some examples are as follows:

- The output differential entropy is directly determined by setting \( \sigma = \hat{\sigma} \) and \( p_S = q_S \), i.e.,
  \[
  \frac{1}{N} h(y|V) = \Delta_N(\sigma, \sigma, p_S, p_S). \tag{17}
  \]

- Given the fact that the conditional differential entropy \( h(y|s, V) \) in the end-to-end setting is determined via the differential entropy of the Gaussian noise process, the input-output information rate is readily written in terms of the cross entropy as
  \[
  \frac{1}{N} I(s; y|V) = \frac{1}{N} h(y|V) - \frac{1}{N} h(y|s, V) = \Delta_N(\sigma, \sigma, p_S, p_S) - \frac{1}{2} \log 2\pi e\sigma^2. \tag{18}
  \]

- The average Kullback-Leibler divergence between the true distribution and the induced distribution is given by
  \[
  \mathbb{E}\{D_{KL}(p_Y || q_Y)\} = h(y|V) - N\Delta_N(\sigma, \hat{\sigma}, p_S, q_S) = N(\Delta_N(\sigma, \sigma, p_S, p_S) - \Delta_N(\sigma, \hat{\sigma}, p_S, q_S)), \tag{19}
  \]
  where the expectation is taken with respect to the random field.

Given the performance metric, our ultimate goal is to evaluate the normalized cross entropy function in the large-system limit, i.e., \( N \uparrow \infty \). In particular, we are interested in the regime in which the data-set size, i.e., \( K \), and the number of observations, i.e., \( N \), grow significantly large with a bounded ratio. Such a regime can be studied by considering a sequence of problems indexed by \( N \), in which the data-set size is determined as function of index \( N \), i.e., \( K(N) \). We are then interested in the limit
\[
\Delta(\sigma, \hat{\sigma}, p_S, q_S) = \lim_{N \to \infty} \Delta_N(\sigma, \hat{\sigma}, p_S, q_S), \tag{20}
\]
assuming that
\[
\lim_{N \to \infty} = \frac{K(N)}{N} = R, \tag{21}
\]
for some bounded \( R \). In the remaining of this manuscript, we refer to \( R \) as the load factor of the generative model.

As shown in Section III, the asymptotic limit \( \Delta(\sigma, \hat{\sigma}, p_S, q_S) \) is derived from the free energy of a corresponding spin-glass whose Hamiltonian is described by the \( P \)-spin model.

C. An Initial Setup

To start with the derivations, we focus on a special case in which the data symbols are simply labels, i.e., \( s_k \in \{±1\} \). We further assume that the labels are uniform i.i.d. labels, and that the prior belief is perfectly matched with the true prior, i.e.,
\[
p_S(s) = q_S(s) = \prod_{k=1}^{K} B(s_k; 0.5), \tag{22}
\]
with \( B(s; f) \) describing the Bernoulli distribution at \( S = s \) for some \( S \) with
\[
Pr\{S = +1\} = 1 - Pr\{S = -1\} = f. \tag{23}
\]

The interest on this particular prior follows from the connection of this generative model with the nonlinear Gaussian encryption being studied in a related but different setting in [27]. We talk about this connection with details in the forthcoming section. It is however worth mentioning that the extension of the current derivations to more general priors is straightforward and is considered for future work.

D. Some Earlier Studies

As mentioned, the most related earlier studies to this work are Sourlas’ coding scheme explained in [25] and [36] and the nonlinear encryption technique proposed by Fyodorov in [27]. The former study by Sourlas utilizes a particular class of Gaussian random fields for channel coding and shows that this scheme sketches a cost-efficiency trade-off which is close to the information-theoretic limit given by Shannon in [56]. This seminal idea will be further illustrated throughout our investigations in Section IV.

The recent related study by Fyodorov is also related to this work in its core idea; however, the setting in this work is different in various respects. For instance, in [27], Fyodorov considers the following setting:

1. The data symbols are supposed to be drawn from a uniform distribution on a hypersphere, i.e., \( s = \sqrt{DP} s_0 \) where \( s_0 \in S^{D-1} \) with \( S^{D-1} \) denotes the \( (D-1) \)-sphere.
2. The learning task is performed via the method of least-squares, i.e.,
   \[
   \hat{s} = \arg\min_{\sqrt{2\pi} u \in S^{D-1}} \|y - V(u)\|^2. \tag{24}
   \]

In the Bayesian framework, this describes the maximum-likelihood estimation which due to the uniform prior is of the similar form as the maximum-a-posterior (MAP) estimation.

Fyodorov determines the asymptotic limit of the average normalized inner product between the ground truth \( s \) and the estimated parameter \( \hat{s} \), i.e., he finds the so-called overlap \( m^* \) which is defined as\(^8\)
\[
m^* = \lim_{N \to \infty} \frac{\langle s, \hat{s} \rangle}{P_S}, \tag{25}
\]
assuming constant ratio \( \mu = D/N \). The result of Fyodorov contains both replica symmetry (RS) and replica symmetry breaking (RSB) solutions. Despite its complicated look, the key conclusions are straightforwardly implied from the main

\(^8\)We later give more discussions on the meaning of \( m^* \).
result: for strictly nonlinear fields, \( m^* \) observes a second-order phase transition at a threshold signal-to-noise ratio (SNR), i.e., \( F_S/\sigma^2 \). Although this second-order phase transition makes a connection to the secure channel coding of Wyner [37], this connection was left unaddressed in [27], due to the fundamental differences in the inverse problem and the system model considered by Fyodorov. More details on Fyodorov’s work can be found in Appendix A.

III. CORRESPONDING SPIN GLASS WITH \( P \)-SPIN MODEL

Our key objective is to determine the limiting normalized cross entropy. We address this goal by using statistical mechanics: we define a corresponding spin glass corresponding to the generative model whose particle interactions are described via a \( P \)-spin model. We then show that the desired metrics in the inference problem are derived from the free energy of the corresponding spin glass and utilize the replica method to determine the free energy.

Let us now start with defining the corresponding spin glass:

**Definition 2 (Corresponding Spin Glass):** For a given vector of observations \( y \) and random field \( V(\cdot) \), the spin glass corresponding to the nonlinear Gaussian generative model is a thermodynamic system with \( K \) particles. The microstate of this spin glass is described by \( u \in \{\pm1\}^K \), and its energy for a given realization of the microstate \( u \) is determined via the following Hamiltonian:

\[
E(u|y,V) = \frac{||y - V(u)||^2}{2}. \tag{26}
\]

Starting from this definition, we now follow standard derivations to describe the macroscopic features of the corresponding spin glass: the conditional Boltzmann distribution of the quenched random variables at the inverse temperature \( \beta \) is given by

\[
p_{\beta}(u|y,V) = \frac{\exp\left(-\beta E(u|y,V)\right)}{Z_{\beta}(y,V)}, \tag{27}
\]

where \( Z_{\beta}(y,V) \) is the partition function, i.e.,

\[
Z_{\beta}(y,V) = \sum_{u \in \{\pm1\}^K} \exp\left(-\beta E(u|y,V)\right). \tag{28}
\]

The macroscopic parameters of this spin glass at inverse temperature \( \beta \) are fully described via the normalized average free energy which is given by

\[
\mathcal{F}_{\beta}(\beta) = -\frac{1}{N\beta} \mathbb{E} \left\{ \log Z_{\beta}(y,V) \right\}, \tag{29}
\]

where the expectation is taken over the quenched random variables, i.e., \( y \) and \( V(\cdot) \). We now show that the free energy determines our target metric, i.e., the cross entropy.

\( ^8 \)Remember that in Section II-C we restricted the labels to be uniformly generated from \( \{\pm1\} \).

\( ^9 \)By quenched random variables, we refer to \( y \) and \( V(\cdot) \) which have different orders of randomness compared to the microstate. More details can be followed in [2].

\( ^10 \)Note the subscript \( N \) refers to the fact that the free energy expression depends on the model dimensions. We however drop \( K \), as we assume \( K \) and \( N \) are related by \( K/N = R \).

A. The Variational Problem

Let us get back to our main objective, i.e., determining the cross entropy. Since in the current setting \( p_S = q_S \), we further drop the arguments \( p_S \) and \( q_S \) in the cross entropy expression and expand it as in (30), shown at the bottom of the next page.

In (30), we drop the subscript \( y,V \) for the outer expectation, as it is taken over the true distribution of the quenched random variables. Keeping in mind the definition of the partition function and the assumption on the prior distribution, i.e.,

\[
p_S(s) = 2^{-K} = 2^{-NR}, \tag{31}
\]

we can further write

\[
\mathbb{E}_s \left\{ \exp \left\{ -\frac{||y - V(s)||^2}{2\sigma^2} \right\} \right\} = 2^{-NR} \sum_{u \in \{\pm1\}^K} \exp \left\{ -\frac{||y - V(u)||^2}{2\sigma^2} \right\}, \tag{32a}
\]

(32b)

Finally, by replacing (32b) into (30c), we conclude that

\[
\Delta_N(\sigma, \hat{\sigma}) = \frac{1}{2} \log 2 \pi \sigma^2 + R \log 2 - \frac{1}{N} \mathbb{E} \left\{ \log Z_{1/\sigma^2}(y,V) \right\} \tag{33a}
\]

\[
= \frac{1}{2} \log 2 \pi \sigma^2 + R \log 2 + \frac{1}{\sigma^2} \mathcal{F}_N \left( \frac{1}{\sigma^2} \right). \tag{33b}
\]

The above identity clarifies the connection between the nonlinear generative model and the corresponding spin glass. As a direct result of this identity, we consider the variational problem of calculating the free energy of the corresponding spin glass. The variational problem is of the same level of hardness as the original problem. We hence follow the standard approach based on the replica method; see [2] and the references therein for details.

B. Main Results

We now present the final solution for the asymptotic normalized free energy, i.e.,

\[
\mathcal{F}(\beta) = \lim_{N \to \infty} \mathcal{F}_N(\beta), \tag{34}
\]

given by the replica method under the RS assumption. The derivations are skipped at this point and presented in Appendix B. The solution is referred to as the RS solution. This is known that for a generic setting, the RS assumption is not necessarily valid and hence the corresponding solution is not always reliable. However, for some particular cases the validity of RS is guaranteed. We discuss one of these cases, i.e., optimal Bayesian scenario, later on.

**Proposition I (RS solution):** Consider the generative model described by (5) and (6). Let \( \Phi(\cdot) \) denote the covariance function of \( V(\cdot) \), and \( Q \) and \( m \) be two scalars in \([0,1]\). Define \( E(Q,m) \) and \( L(Q,m) \) in terms of \( Q \) and \( m \) as

\[
L_{\beta}(Q,m) = \frac{\beta}{R} \Phi(Q) \int_{\beta g(Q)}^{\beta f(Q)} \frac{1}{g_{\beta}(Q)} d\beta, \tag{35a}
\]

\[
E_{\beta}(Q,m) = \frac{\beta}{R} \Phi(m) \frac{1}{g_{\beta}(Q)}, \tag{35b}
\]
for functions
\[ g_\beta(Q) = 1 + \beta \left[ \Phi(1) - \Phi(Q) \right], \quad (36a) \]
\[ f_\beta(Q, m) = \beta \left[ \sigma^2 + \Phi(1) + \Phi(Q) - 2\Phi(m) \right]. \quad (36b) \]

Let the function \( \Pi(Q, m) \) for a given \( Q \) and \( m \) be
\[ \Pi_\beta(Q, m) = R \left[ m E_\beta(Q, m) + \frac{1 - Q}{2} L_\beta(Q, m) \right] \]
\[ + \frac{1}{2} \left[ f_\beta(Q, m) \left/ g_\beta(Q) \right. + \log g_\beta(Q) \right]. \quad (37) \]

Assuming the RS ansatz, the normalized average free energy of the corresponding spin glass at inverse temperature \( \beta \) is given by
\[ \mathcal{F}(\beta) = \frac{F_\beta(Q^*, m^*)}{\beta}, \quad (38) \]

with \( F_\beta(Q, m) \) being defined on \([0, 1]^2\) as in (39), shown at the bottom of the next page. In (39), \( Z \) is a zero-mean Gaussian random variable with unit variance, i.e., \( Z \sim \mathcal{N}(0, 1) \), and
\[ (Q^*, m^*) = \arg\min_{(Q,m)\in\mathcal{F}} F_\beta(Q, m), \quad (40) \]

with \( \mathcal{F} \) being the union of the boundary points of the feasible set \([0, 1]^2\) and the set of all solutions to the following fixed-point equations
\[ Q = \mathbb{E}_Z \left\{ \tanh^2 \left( \sqrt{L_\beta(Q, m)Z + E_\beta(Q, m)} \right) \right\}, \quad (41a) \]
\[ m = \mathbb{E}_Z \left\{ \tanh \left( \sqrt{L_\beta(Q, m)Z + E_\beta(Q, m)} \right) \right\}. \quad (41b) \]

**Proof:** See Appendix B. \( \square \)

As mentioned, the above asymptotic result is derived under the RS ansatz, assuming that the corresponding spin glass shows the so-called *replica symmetry* at the thermal equilibrium. This is however not necessarily the case. At this point, we assume that for the settings of interest, the RS ansatz is valid. Further discussions can be followed in Appendix B. More detailed discussions in this respect can be found in [2], [3], and [57].

Considering the RS solution, the asymptotic normalized cross entropy term is readily derived from (33). This is shown in (42), shown at the bottom of the next page. In next sections, we mainly focus on the particular case of *matched* Bayesian estimation, for which it is known that the RS ansatz is valid; see for example [55], [58], [59]. Before starting with the optimal Bayesian case, let us discuss the asymptotic decoupling property of the nonlinear model.

### C. Decoupling Principle

The asymptotic result derived by the replica method leads to the *decoupled principle* for the original high-dimensional setting. This is a well-known principle developed for various inference problems with linear model [11], [39], [57], [60], [61]. The RS solution extends this principle to higher-order fields. To state the decoupling principle, consider the following scalar inference problem: the label \( S \in \{ \pm 1 \} \) distributed as \( S \sim B(s, 0.5) \) is observed through a Gaussian channel whose additive noise term is \( \rho Z \) with \( Z \sim \mathcal{N}(0, 1) \). The received symbol is hence represented by
\[ Y = S + \rho Z. \quad (43) \]

Given a particular observation \( Y = y \), the scalar Bayesian estimator determines the expectation of \( S \) via its postulated posterior distribution \( p_S(s|y) \) which assumes the noise variance to be \( \hat{\rho} \). The estimated symbol is hence given by
\[ \hat{S} = \mathbb{E}_{p_S} \{ S|Y \} = \tanh \left( \frac{Y}{\hat{\rho}^2} \right), \quad (44) \]

where the latter identity is found after few standard lines of derivations.

Considering the above scalar problem, an initial observation is that by setting
\[ \rho = \sqrt{\frac{L_{1/\hat{\rho}^2}(Q^*, m^*)}{E_{1/\hat{\rho}^2}(Q^*, m^*)}}, \quad (45a) \]
\[ \hat{\rho} = \frac{1}{\sqrt{E_{1/\hat{\rho}^2}(Q^*, m^*)}}, \quad (45b) \]

the average pair-wise distortion determined by the RS solution is given by the expected distortion between \( S \) and \( \hat{S} \) in the decoupled setting. More precisely, for a given distortion function \( D(\cdot; \cdot) : \{ \pm 1 \} \times \mathbb{R} \rightarrow \mathbb{R} \), the following identity under the RS ansatz exists:
\[ \lim_{K \uparrow \infty} \frac{1}{K} \mathbb{E} \left\{ \sum_{k=1}^{K} D \left( s_k; \hat{s}_k \right) \right\} = \mathbb{E} \left\{ D \left( S; \hat{S} \right) \right\}. \quad (46) \]

This finding is however the result of a more general principle:

**Proposition 2 (Decoupling Principle):** Assume RS is a valid ansatz. For any \( k \in [K] \), the pair \( (s_k, \hat{s}_k) \) converges in distribution to \( (S, \hat{S}) \) as \( K \uparrow \infty \), when we set \( \rho \) and \( \hat{\rho} \) according to (45) in the decoupled setting.

**Proof:** The proof follows the method of moments and is given in Appendix C. \( \square \)

A direct result of the decoupling principle is that the fixed-points \( Q^* \) and \( m^* \) in the RS solution are given in terms of the

\[ \Delta_N(\sigma, \hat{\sigma}) = -\frac{1}{N} \mathbb{E} \left\{ \log q_Y(y|V) \right\} \]
\[ = -\frac{1}{N} \mathbb{E} \left\{ \log \left( 2\pi\hat{\sigma}^2 \right)^{-N/2} \mathbb{E}_s \left\{ \exp \left\{ -\frac{||y - V(s)||^2}{2\hat{\sigma}^2} \right\} \right\} \right\} \]
\[ = \frac{1}{2} \log 2\pi\hat{\sigma}^2 - \frac{1}{N} \mathbb{E} \left\{ \log \mathbb{E}_s \left\{ \exp \left\{ -\frac{||y - V(s)||^2}{2\hat{\sigma}^2} \right\} \right\} \right\}. \quad (30c) \]
decoupled setting as follows:

\[ Q^* = \mathbb{E} \left\{ \hat{S}^2 \right\}, \quad (47a) \]
\[ m^* = \mathbb{E} \left\{ S \hat{S} \right\}. \quad (47b) \]

The implicit use of these identities is to simplify the RS solution for optimal Bayesian inference. Nevertheless, they further give operational meaning to \( m^* \) and \( Q^* \). In this respect, \( m^* \) is of particular interest: it determines the expected normalized inner product between the true symbols and the learned ones. It is hence often called the overlap.\(^{12}\) There are in principle two extreme cases for \( m^* \): 1) Perfect recovery in which \( \hat{S} = S \), and hence \( m^* = 1 \). 2) Independent inference, in which \( \hat{S} \) and \( S \) are statistically independent resulting in \( m^* = 0 \). These extreme cases limit \( m^* \in [0, 1] \) which is consistent with the constraint stated in the RS solution. They moreover provide a comprehensive interpretation for \( m^* \): larger \( m^* \) means more information.

**Remark 1:** One should note that the decoupling principle only describes the marginal distribution of each pair of true label and its estimation. Hence, it does not determine the metrics which take into account correlation among different pairs, e.g., the cross entropy.

**D. Optimal Bayesian Inference**

The mean squared error between the true labels and their recovery is minimized with respect to \( \hat{\sigma} \), when we set the postulated noise variance equal to the true noise variance, i.e., when \( \hat{\sigma} = \sigma \). Due to this property, this particular form of the Bayesian inference is referred to as the optimal Bayesian inference. This case describes the Nishimori line which lies in the RS stability region \([59]\). This means that the RS ansatz, for this particular case, is valid and no further derivations under RSB ansätze are required. In addition to RS stability, the symmetry of matched estimator leads to this basic conclusion that \( m^* = Q^* \): following the decoupling principle, \( m^* \) determines the cross correlation between the learned symbol and the true one, and \( Q^* \) calculates the second moment of the learned symbol. As the estimator is optimal Bayesian, the orthogonality principle is satisfied. This means that the estimation error is orthogonal to the estimation, i.e.,

\[ \mathbb{E} \left\{ (S - \hat{S}) \hat{S} \right\} = 0. \quad (48) \]

This identity concludes that

\[ m^* = \mathbb{E} \left\{ S \hat{S} \right\} = \mathbb{E} \left\{ \hat{S}^2 \right\} = Q^*. \quad (49) \]

Following this straightforward justification, one can see that \( m^* = Q^* \) is a fundamental property in the replica analysis of optimal Bayesian inference. This fundamental property is in fact the key observation used in \([60, \text{Appendix II}]\) to show the initial form of the decoupling principle.

From these properties of the optimal Bayesian estimator, we can write

\[ g_{1/\sigma^2} (m) = 1 + \frac{\Phi (1) - \Phi (m)}{\sigma^2}; \quad (50a) \]
\[ f_{1/\sigma^2} (m, m) = 1 + \frac{\Phi (1) - \Phi (m)}{\sigma^2} = g_{1/\sigma^2} (m). \quad (50b) \]

This leads to the conclusion that

\[ L_{1/\sigma^2} (m, m) = E_{1/\sigma^2} (m, m) = \frac{1}{R\sigma^2} \frac{\Phi' (m)}{g (m)}. \quad (51) \]

Based on these observations, we can simplify the notation by dropping the inverse temperature index,\(^{13}\) and defining \( g (m) := g_{1/\sigma^2} (m) \) and

\[ E (m) := E_{1/\sigma^2} (m, m) = \frac{1}{R\sigma^2} \frac{\Phi' (m)}{g (m)} \quad (52a) \]
\[ = - \frac{1}{R} \frac{g' (m)}{g (m)} \quad (52b) \]
\[ = - \frac{1}{R} \frac{d}{dm} \log g (m). \quad (52c) \]

We now recall that the normalized cross entropy in the matched case determines the normalized differential entropy of the output; see \((17)\). We hence replace the simplified definitions into the RS solution and conclude the following result:

**Corollary 1 (Optimal Bayesian Inference):** Let \( H (\sigma) \) denote the normalized differential entropy of the observations, i.e.,

\[ H (\sigma) = \Delta (\sigma, \sigma) = \lim_{N \to \infty} \frac{1}{N} \mathbb{E} \log g (y | \mathcal{V}). \quad (53) \]

Define function \( g (\cdot) \) as

\[ g (m) = 1 + \frac{\Phi (1) - \Phi (m)}{\sigma^2}; \quad (54) \]

and let \( E (m) \) be defined in terms of \( g (m) \) as in \((52c)\). Then, \( H (\sigma) \) is determined as

\[ H (\sigma) = H_\sigma (m^*), \quad (55) \]

\(^{12}\)Remember the performance metric defined by Fyodorov in \([27]\), i.e., \((25)\).

\(^{13}\)Since we are only operating at \( \beta = 1/\sigma^2 \).

\[ F_\beta (Q, m) := -R\mathbb{E}_Z \left\{ \log \cosh \left( \sqrt{L_\beta (Q, m)} Z + E_\beta (Q, m) \right) \right\} + \Pi_\beta (Q, m) - R \log 2 \quad (39) \]

\[ \Delta (\sigma, \sigma) = \frac{1}{2} \log 2 \pi \sigma^2 + R \log 2 + \frac{1}{\sigma^2} \mathcal{F} \left( \frac{1}{\sigma^2} \right) \quad (42a) \]
\[ = \frac{1}{2} \log 2 \pi \sigma^2 - R\mathbb{E}_Z \left\{ \log \cosh \left( \sqrt{L_{1/\sigma^2} (Q^*, m^*)} Z + E_{1/\sigma^2} (Q^*, m^*) \right) \right\} + \Pi_{1/\sigma^2} (Q^*, m^*) \quad (42b) \]
orders fair, we assume that all transformed symbol are unit identity in (18), we have
pure order optimal Bayesian case and consider a Gaussian generative model. To this end, we focus on the special cases to investigate the impacts of having a non-linear \( \ell \) in the learning stage, and thus information per dimension. This follows the fact that by rests on the RS solution. □

With optimal Bayesian estimation, we are often more interested on the information rate of the model, i.e., the mutual information per dimension. This follows the fact that by matched Bayesian inference, there is no information loss at the learning stage, and thus \( I(s; \hat{s}| \mathcal{V}) = I(s; y| \mathcal{V}) \). Using the identity in (18), we have
\[
I(\sigma) = \lim_{N \to \infty} \frac{1}{N} I(s; y| \mathcal{V}) = H(\sigma) - \frac{1}{2} \log 2\pi e \sigma^2. \tag{59}
\]
We use this metric through experimental investigations.

**IV. LINEAR MODELS VERSUS NONLINEAR MODELS**

Given the main results, we now consider some illustrative special cases to investigate the impacts of having a non-linear Gaussian generative model. To this end, we focus on the optimal Bayesian case and consider a pure order \( \lambda \) polynomial covariance function, i.e., \( \Phi(x) = c x^\lambda \) for some constant \( c \in \mathbb{R}^+ \). With this covariance function, the random field describes a \( (\lambda + 1) \)-dimensional tensor\(^{15}\) whose components are generated at random with some Gaussian prior. This means that the \( n \)-th component of the random transform is given in terms of \( s \) as
\[
\mathcal{V}_n(s) = \sum_{\ell_1, \ldots, \ell_\lambda = 1}^K J_{\ell_1, \ldots, \ell_\lambda}^n \prod_{t=1}^{\lambda} s_{t_i}, \tag{60}
\]
for some \( J_{\ell_1, \ldots, \ell_\lambda}^n \) which are i.i.d. Gaussian random variables in \( n, \ell_1, \ldots, \ell_\lambda \).

To keep the comparison among random fields of different orders fair, we assume that all transformed symbol are unit variance, i.e.,
\[
\mathbb{E}\{\mathcal{V}_n(s)^2\} = 1. \tag{61}
\]
This concludes that \( J_{\ell_1, \ldots, \ell_\lambda}^n \sim \mathcal{N}(0, 1/\sqrt{K^\lambda}) \) for \( n \in [N] \) and \( \ell_1, \ldots, \ell_\lambda \in [K] \) and
\[
\Phi(x) = x^\lambda. \tag{62}
\]

\(^{14}\)For instance what reported in [59].

\(^{15}\)Remember that a matrix is a two-dimensional tensor.

Fig. 1. Information rate against \( m \) for linear and quadratic Gaussian random fields. The overlap \( m^* \) is given by the minima of the information rate.

Noting that \( \Phi(1) = 1 \), we have \( g(m) = 1 + \xi_m(\sigma) \), where
\[
\xi_m(\sigma) = \frac{1 - m^\lambda}{\sigma^2}. \tag{63}
\]
From \( g(m) \), \( E(m) \) is determined via (52c). By substituting in Corollary 1, we conclude that the asymptotic information rate reads \( I(\sigma) = L_m(\sigma) \) with \( L_m(\sigma) \) given as follows:
\[
L_m(\sigma) = \frac{1}{2} \log (1 + \xi_m(\sigma)) + Q_m(\sigma), \tag{64}
\]
where the latter term is determined for a given \( m \) and noise variance \( \sigma \) as
\[
Q_m(\sigma) = \frac{(1 + m) R}{2 m^2(\sigma)} - \mathbb{E}_Z \left\{ \log \cosh \left( \frac{1 + \rho_m(\sigma) Z}{\rho_m(\sigma)} \right) \right\}, \tag{65}
\]
with \( \rho_m(\sigma) \) being defined as
\[
\rho_m(\sigma) = \sqrt{\frac{R \sigma^2}{\lambda m^{\lambda - 1}} (1 + \xi_m(\sigma))}. \tag{66}
\]
We refer to \( L_m(\sigma) \) as the modified free energy. This appellation becomes clear shortly.

The fixed-point \( m^* \) is further determined by minimizing \( L_m(\sigma) \) over \( m \in [0, 1] \). The minimizer is readily found by first solving
\[
m = \mathbb{E}_Z \left\{ \tanh \left( \frac{1 + \rho_m(\sigma) Z}{\rho_m(\sigma)} \right) \right\}; \tag{67}
\]
for \( m \) and then minimizing \( L_m(\sigma) \) over the union of the solution set and the set of boundary points \( \{0, 1\} \).

The fixed-point equation describes the extreme points of the free energy; see Appendix B. Following the fact that the information rate and the free energy behave the same\(^{16}\) in \( m \), it is concluded that \( m^* \) is the global minimizer\(^{17}\) of \( L_m(\sigma) \) within interval \( m \in [0, 1] \). The information rate is plotted in terms of \( m \) for \( \lambda = 1 \) and \( \lambda = 2 \) in Fig. 1, where we set \( \sigma^2 = 0.1 \) and \( R = 1.76 \). The solution \( m^* \) is further shown for both curves in the figure.

Given the above derivations, we now study the asymptotic characteristics of the optimal Bayesian learning algorithm for

\(^{16}\)Remember that the information rate is simply a shifted version of the normalized free energy; see (59).

\(^{17}\)This further clarifies the appellation.
different choices of $\lambda$. Before, we start with the investigations, let us derive a simple lower- and an upper-bound on the information rate which is useful throughout the investigations.

**Corollary 2:** Let $m^*$ be the overlap. Then, there exists $0 \leq C_m \leq 1$, such that

$$I(\sigma) = \frac{1}{2} \left[ \log (1 + \xi_m(\sigma)) + \frac{(C_m + m^*) R}{\rho_m^*(\sigma)} \right] - R \log \cosh \left( \frac{1}{\rho_m^*(\sigma)} \right).$$

\[ (68) \]

**Proof:** The proof follows Jensen’s inequality. The functions $\log(\cdot) : [0, \infty) \mapsto \mathbb{R}$ and $\log \cosh(\cdot) : \mathbb{R} \mapsto [0, \infty)$ are concave and convex, respectively. We hence can write

$$\mathbb{E} \left\{ \log \cosh \left( \frac{1 + \rho_m(\sigma) Z}{\rho_m^2(\sigma)} \right) \right\} \leq \log \mathbb{E} \left\{ \cosh \left( \frac{1 + \rho_m(\sigma) Z}{\rho_m^2(\sigma)} \right) \right\},$$

\[ (69a) \]

$$= \log \exp \left( \frac{1}{2 \rho_m^2(\sigma)} \right) \cosh \left( \frac{1}{\rho_m^2(\sigma)} \right) = \frac{1}{2 \rho_m^2(\sigma)} + \log \cosh \left( \frac{1}{\rho_m^2(\sigma)} \right),$$

\[ (69b) \]

using the concavity of $\log(\cdot)$, and

$$\mathbb{E} \left\{ \log \cosh \left( \frac{1 + \rho_m(\sigma) Z}{\rho_m^2(\sigma)} \right) \right\} \leq \frac{1}{\rho_m^2(\sigma)} + \log \cosh \left( \frac{1}{\rho_m^2(\sigma)} \right).$$

\[ (69c) \]

from the convexity of $\log \cosh(\cdot)$. By substituting into (64), we can show that $L_m(\sigma)$ is bounded from below as

$$L_m(\sigma) \geq \frac{1}{2} \left[ \log (1 + \xi_m(\sigma)) + \frac{m R}{\rho_m^2(\sigma)} \right] - R \log \cosh \left( \frac{1}{\rho_m^2(\sigma)} \right),$$

\[ (70a) \]

and from above as

$$L_m(\sigma) \leq \frac{1}{2} \left[ \log (1 + \xi_m(\sigma)) + \frac{(1 + m) R}{\rho_m^2(\sigma)} \right] - R \log \cosh \left( \frac{1}{\rho_m^2(\sigma)} \right),$$

\[ (70b) \]

We therefore conclude that there exists $0 \leq C_m \leq 1$, such that

$$L_m(\sigma) = \frac{1}{2} \left[ \log (1 + \xi_m(\sigma)) + \frac{(C_m + m) R}{\rho_m^2(\sigma)} \right] - R \log \cosh \left( \frac{1}{\rho_m^2(\sigma)} \right).$$

\[ (71) \]

By setting $I(\sigma) = L_m(\sigma)$, the proof is concluded. \hfill \Box

**A. Classical Case: Linear Generative Model**

The classical linear model is studied by setting $\lambda = 1$ in the given results. This is a widely-studied setting whose replica analysis is given in various lines of work in the literature, e.g., [3], [4], [10], [11], and [14]. We start our investigation by considering the overlap for this field. An initial finding in this respect is given in the following lemma:

**Lemma 1 (Inference Guarantee):** The overlap of linear model is always non-zero.

**Proof:** For $\lambda = 1$, the term $\rho_m(\sigma)$ reduces to

$$\rho_m(\sigma) = \sqrt{R(\sigma^2 + 1 - m^*)}.$$  \[ (74) \]

By substituting in the fixed-point equation, it is observed that $m = 0$ is not a solution for any $R$ and $\sigma$. This concludes that $m = 0$ is not an saddle point of $L_m(\sigma)$, or equivalently

$$\frac{\partial}{\partial m} L_m(\sigma)|_{m=0} \neq 0,$$  \[ (75) \]

for any choice of $R$ and $\sigma$. We now show that there exists always a point $m \in [0, 1]$ for which $L_m(\sigma)$ is smaller than $L_0(\sigma)$: let us define the Gaussian random variable $Z$ as

$$\tilde{Z} := \frac{1 + \rho_0(\sigma) Z}{\rho_0^2(\sigma)},$$  \[ (76) \]

for $Z \sim \mathcal{N}(0, 1)$. Let $\hat{L}_0$ denote the derivative of $L_m(\sigma)$ at $m = 0$, i.e.,

$$\hat{L}_0 = \frac{\partial}{\partial m} L_m(\sigma)|_{m=0}.$$  \[ (77) \]

The derivative is determined in terms of $\tilde{Z}$ as

$$\hat{L}_0 = \zeta_0 \left[ 1 - \mathbb{E} \left\{ \tanh \tilde{Z} \right\} - \rho_0^2(\sigma) \mathbb{E} \left\{ \tilde{Z} \tanh \tilde{Z} \right\} \right],$$ \[ (78) \]

where $\zeta_0$ is given by

$$\zeta_0 = \frac{R^2}{2 \rho_0^2(\sigma)}.$$ \[ (79) \]

Noting that $\tilde{Z} \sim \mathcal{N}(1/\rho_0^2(\sigma), 1/\rho_0^2(\sigma))$, one can use integration by parts to show that

$$\mathbb{E} \left\{ \tilde{Z} \tanh \tilde{Z} \right\} = \frac{1}{\rho_0^2(\sigma)} \mathbb{E} \left\{ 1 + \tanh \tilde{Z} - \tanh^2 \tilde{Z} \right\}.$$ \[ (80) \]

As the result, we have

$$\hat{L}_0 = \zeta_0 \left[ \mathbb{E} \left\{ \tanh^2 \tilde{Z} \right\} - 2 \mathbb{E} \left\{ \tanh \tilde{Z} \right\} \right] \leq 0.$$ \[ (81) \]

Considering (75) and (81), we conclude that for any $R > 0$ and $\sigma$, $L_m(\sigma)$ decreases in $m$ at $m = 0$. This leads to this conclusion that there exists a point $\bar{m}$ in the right neighborhood of $m = 0$ for which we have $L_{\bar{m}}(\sigma) \leq L_0(\sigma)$; therefore, $m = 0$ cannot be the global minimizer of $L_m(\sigma)$ on $[0, 1]$.

Non-zero overlap indicates that in the linear model the learned symbol are always correlated to the true labels. This is in fact an intuitive observation which confirms a rather known property: linear models always carry information about the model parameters.

In general, the overlap is a decreasing function of the load $R$: for smaller $R$ the problem is better determined, and hence the learned labels are more correlated to the true symbols.\footnote{This is in fact true for noisy observations. For the case of $\sigma^2 = 0$, however, the phase transition occurs with different scaling. Namely, to have an overlap smaller than one, one needs to let the load to be scaled logarithmically (or faster) with the dimension; see [62] for more details.}

This is shown in Fig. 2, where we plot the overlap $m^*$ against $R$ for $\sigma^2 = 0.1$.\footnote{This is in fact true for noisy observations. For the case of $\sigma^2 = 0$, however, the phase transition occurs with different scaling. Namely, to have an overlap smaller than one, one needs to let the load to be scaled logarithmically (or faster) with the dimension; see [62] for more details.}
We now consider the extreme cases. We start with the perfect recovery case in which \( m^* = 1 \). In this case, we can write
\[
\mathcal{I}(\sigma) = C_1 + \frac{1}{2\sigma^2} - R \log \cosh \left( \frac{1}{R\sigma^2} \right). 
\] (84)

Assuming that \( \sigma^2 \) is small enough, we can further use the approximation
\[
\log \cosh (x) \approx |x| - \log 2,
\] (85)

for large \( x \) and conclude that
\[
\mathcal{I}(\sigma) \approx \frac{C_1 - 1}{2\sigma^2} + R \log 2.
\] (86)

With small choices of \( \sigma^2 \) such that (85) is accurate, we can say that the function \( \log \cosh (\cdot) \) is approximately linear. This means that the upper-bound in Corollary 2 is tight. In other words, we can write \( C_1 \approx 1 \). One can further observe that \( C_{m^*} \) starts to decrease as \( m^* \) decreases. We hence heuristically conclude that in the perfect recovery region
\[
\mathcal{I}(\sigma) \approx R \log 2.
\] (87)

This is intuitive, since in this case we have
\[
\frac{1}{N} H(s|y,V) = 0,
\] (88)
due to perfect recovery, i.e., \( m^* = 1 \), and hence the end-to-end model reduces to a revertible noise-free transform for which the information rate is equal to the entropy rate of the labels.

Now, let us consider another extreme case: the information rate starts to deviate from the \( R \log 2 \) line, as \( m^* \) starts to decrease. Assume \( R \gg 1 \) which means that \( m^* = \epsilon \) for some small \( \epsilon > 0 \); remember Fig. 2. In this case, we can write
\[
\mathcal{I}(\sigma) = \frac{1}{2} \left[ \log \left( 1 + \frac{1 - \epsilon}{\sigma^2} \right) + \frac{C_\epsilon + \epsilon}{\sigma^2 + 1 - \epsilon} \right] - R \log \cosh \left( \frac{1}{R(\sigma^2 + 1 - \epsilon)} \right).
\] (89)

Noting that \( \log \cosh (x) \approx 0 \) for small \( x \), we conclude that
\[
\mathcal{I}(\sigma) \approx \frac{1}{2} \left[ \log \left( 1 + \frac{1}{\sigma^2} \right) + \frac{C_\epsilon + \epsilon}{\sigma^2 + 1 - \epsilon} \right].
\] (90)

By taking the limit \( R \uparrow \infty \), we can neglect \( \epsilon \) and \( C_\epsilon \) and write
\[
\lim_{R \uparrow \infty} \mathcal{I}(\sigma) = \frac{1}{2} \log \left( 1 + \frac{1}{\sigma^2} \right),
\] (91)

which is also shown in Fig. 3. This limit is in fact the information capacity of the Gaussian channel which we denote in the sequel by
\[
\mathcal{C}(\sigma) = \frac{1}{2} \log \left( 1 + \frac{1}{\sigma^2} \right).
\] (92)

This conclusion recovers similar results developed earlier in the context of spectral efficiency analysis for randomly spread code division multiple access techniques; see for example [63], [64], and [65].

---

19 Recall that the upper-bound in Corollary 2 was derived by applying Jensen’s inequality on \( \log \cosh (\cdot) \). The bound becomes tight, if the function is linear.

20 Here, \( \sigma^2 \) is small but fixed. However, by setting \( R \) small enough, we can send \( C_{m^*} \) very close to one.
This means that $L$ is always a solution of the fixed-point equation (67). This means that $L_m(\sigma)$ has a local optimum at $m = 0$. In general, $m = 0$ can be a local minimizer or maximizer. As a result, one conjectures that under some conditions, the overlap can be exactly zero for $\lambda > 1$. This is the extreme case of independent inference which is not achievable in a linear model. To validate this conjecture, we start with the special case of purely quadratic Gaussian field in this section, i.e., we consider $\lambda = 2$. The following lemma confirms the feasibility of $m^* = 0$ for the quadratic field:

**Lemma 2 (Independent Inference via the Quadratic Field):** Consider the purely quadratic field, i.e., $\lambda = 2$, and define the threshold load

$$R_{Th} = \frac{2}{\sigma^2 + 1}.$$  

(94)

For this field, the extreme point $m = 0$ is a local minimizer of the modified free energy function $L_m(\sigma)$, if $R > R_{Th}$ and is a local maximizer if $R < R_{Th}$.

**Proof:** We first note that due to the symmetry of the matched estimator, we have $m^* = Q^*$ in the RS solution.\(^{22}\) Therefore, at an extreme point $m$ we can write

$$m = \mathbb{E}_Z \left\{ \tanh \left( \frac{1 + \rho_m(\sigma) Z}{\rho^2_m(\sigma)} \right) \right\}.$$  

(95a)

$$= \mathbb{E}_Z \left\{ \tanh^2 \left( \frac{1 + \rho_m(\sigma) Z}{\rho^2_m(\sigma)} \right) \right\}.$$  

(95b)

By taking the second derivative of $L_m(\sigma)$ with respect to $m$ and using the above identity, one can show that the second derivative at $m = 0$ is positive, when $R > R_{Th}$ and is negative when $R < R_{Th}$. This concludes the proof. \qed

Considering the purely quadratic field, Lemma 2 indicates that by passing the threshold load, $m = 0$ becomes a candidate for overlap. We now examine this result by a simple experiment shown in Fig. 4. In this figure, the modified free energy $L_m(\sigma)$, has been plotted for the purely quadratic field against $m$ at two different loads $R_{low} = R_{Th} - \epsilon$ and $R_{up} = R_{Th} + \epsilon$ for $\epsilon = 0.01$ assuming $\sigma^2 = 0.1$. As the figure shows, the extreme point $m = 0$ changes its nature from a maximizer to a minimizer, as we pass the threshold load immediately. It is worth noting that variations on the vertical axis of the figure are in the order of $10^{-8}$. For instance, the lower label on the vertical axis of both sub-figures corresponds to the modified free energy at $m = 0$ which for $\sigma^2$ is given by $L_0(\sqrt{0.1}) = C(\sqrt{0.1}) = 1.989$. The upper label is only $\delta = 9.9 \times 10^{-8}$ larger than the lower one.

We now plot the modified free energy at multiple loads and take a look on the behavior of the global minimizer.

\(^{21}\)See Lemma 1.

\(^{22}\)See the discussions in Section III-D.
Interestingly, the global minimizer is jumping from $m^* = 0$ to $m^* = 1$ as the threshold load $R_{Th}$ is passed. Fig. 5 shows the global minimizer for the approximate message passing algorithm. The results; namely, the RS solution given by the global minimizer of the modified free energy. As the figure shows, below a critical load $R_{Th}$, the algorithm does not converge at $R_{Th}$. This observation reports the well-known phenomenon of critical slowing down which is in fact the channel capacity given in bits. A second-order phase transition is further observed at $R_{Th}$. To confirm the validity of this experimental conclusion, we further plot the maximum convergence time to a local extreme point against the load in Fig. 7. In this figure, we iteratively solve the fixed-point equation (67). The algorithm starts from an initial point $m_0$ and iterates till it converges to a solution. This solution is a local extreme point. For a given load, we sweep $m_0$ on a fine grid with 100 points on the interval $(0, 1)$ and find the maximum number of iterations required to converge, where we maximize over the grid. As the convergence criteria, we stop iterating at the iteration in which the rate passes the threshold load $R_{Th}$, i.e., $C(m_t) < 10^{-10}$ with $m_t$ denoting the value of $m$ in iteration $t$. The result shows a critical slowing down at load $R_{Th}$. As the figure shows, the runtime shows a small spike at $R_{Th}$.

$R^* = \frac{C(m)}{\log 2}$, (96)

which is in fact the channel capacity given in bits. A second-order phase transition is further observed at $R_{Th}$. This indicates a second-order phase transition at $R_{Th}$ and a first-order phase transition at $R^*$. The vertical axis is divided by $10^3$, i.e., label 1 denotes $10^3$ sec.

Fig. 5. Modified free energy $L_m(\sigma)$ versus $m$ for $\lambda = 2$ at load $0.8R_{Th}$ (solid line) and $1.2R_{Th}$ (dashed line). The overlap jumps from $m^* = 1$ to $m^* = 0$ as the threshold load $R_{Th}$ is passed.

Fig. 6. Overlap against load for the purely quadratic field, i.e., $\lambda = 2$. The dotted-dashed line shows the fixed-point solutions, the solid line represents the RS solution given by the global minimizer of the modified free energy, and the dashed line denotes the overlap which can be achieved via the approximate message passing algorithm.

Fig. 7. Convergence time against load $R$ for the purely quadratic field, i.e., $\lambda = 2$. The result shows a critical slow-down at $R_{Th}$. This indicates a second-order phase transition at $R_{Th}$ and a first-order phase transition at $R^*$. The vertical axis is divided by $10^3$, i.e., label 1 denotes $10^3$ sec.

- The second set of solutions to the fixed-point equations which are shown by the dotted-dashed line. This set includes both minimizers and maximizers of the modified free energy. As the figure shows, below a critical load $R_{Alg}$, the modified free energy has only two extreme points, a maxima at $m = 0$ and a minima at $m = 1$. By passing $R_{Alg}$, a local minimizer and a local maximizer start to appear at $m_1$ and $m_2$ respectively, where $0 < m_1, m_2 < 1$. As the load grows, $m_1$ moves towards zero and meets $m = 0$ as the rate passes the threshold load $R_{Th}$, at which the curvature of the curve at $m = 0$ changes and $m = 0$ becomes a local minimizer.

- The second set of results show the theoretically achievable overlap as a function of the load $R$ which corresponds to the global minimizer of the modified free energy and is shown by the solid line. Here, the figure shows a first-order phase transition at the load $R^* = \frac{C(m)}{\log 2}$, (96)

which is in fact the channel capacity given in bits. A second-order phase transition is further observed at $R_{Th}$. To confirm the validity of this experimental conclusion, we further plot the maximum convergence time to a local extreme point against the load in Fig. 7. In this figure, we iteratively solve the fixed-point equation (67). The algorithm starts from an initial point $m_0$ and iterates till it converges to a solution. This solution is a local extreme point. For a given load, we sweep $m_0$ on a fine grid with 100 points on the interval $(0, 1)$ and find the maximum number of iterations required to converge, where we maximize over the grid. As the convergence criteria, we stop iterating at the iteration in which the rate passes the threshold load $R_{Th}$, i.e., $C(m_t) < 10^{-10}$ with $m_t$ denoting the value of $m$ in iteration $t$. The result shows a critical slowing down at load $R_{Th}$. As the figure shows, the runtime shows a small spike at $R_{Alg}$ where the local extreme points in the middle of the interval $[0, 1]$ start to appear. The other spike occurs at $R_{Th}$, where the curvature at $m = 0$ changes. In the considered fine grid, the algorithm does not converge at $R_{Th}$. This observation reports the well-known phenomenon of critical slowing down which is a sign for a first-order phase transition at $R^*$ and a second-order phase transition at $R_{Th}$. In fact, the latter is due to the zero curvature of the curve which leads to an infinite number of iterations for convergence and describes a second-order phase transition. However, the former phase transition corresponds to this phenomenon that an already-existing local minimum passes the local minimum at $m = 1$ and becomes the global minimum. More details on the critical slowing down phenomenon and its connection to the phase transition can be found in [66] and [67].
The last set of the results are represented by the dashed line in the figure and predict the overlap which is achieved by the approximate message passing (AMP) algorithm [41], [68], [69]. AMP algorithms are a family of first-order algorithms designed to efficiently approximate the posterior marginals, and have been shown to be optimal among all first order methods for certain classes of problems [70]. Due to this property, AMP algorithms provide a useful benchmark to probe the computational hardness of a problem. The asymptotic performance od AMP algorithms can be exactly traced by a set of scalar state evolution equations [71], which are intimately connected with the expected cross entropy: the state evolution equations coincide with the fixed-point equations associated with the decoupled problem. Therefore, the overlap achieved by AMP from a uninformed initialization is given by the minima of the cross entropy which is closest to \( m = 0 \). As the results show, in our problem, there is a load interval in which the closest minima to \( m = 0 \) is a local minimum, showing that there exists a gap between the theoretically achievable overlap, and the one achieved by AMP from a random initialization. This gap is often referred to as the algorithmic gap and indicates the set of problem settings in which perfect inference is information-theoretically feasible however computationally hard to achieve.

Unlike the linear model, the RS solution in the purely quadratic model shows both first-order and second-order phase transitions. For this particular setting, a first-order phase transition occurs at Shannon’s limit and a second-order phase transition happens at higher loads. These phase transitions divide the behavior of system into three states:

1) The state which occurs at \( R < R^* \). In this state all labels are recovered (perfect recovery).
2) The second state is the state of some labels being inferred which occurs at \( R^* \leq R \leq R_{Th} \) (weak or partial recovery).
3) Finally, the state which happens as the load passes the threshold \( R_{Th} \). We call this new state, the state of nothing being recovered in which the Bayesian algorithm results in zero-information learning. This new state does not exist in linear models, as our conclusions illustrate that \( m^* \neq 0 \) for the linear field.

Although the first-order phase transition is generic, the second-order might not be present. This is simply observed by comparing the two critical loads, namely \( R^* \) and \( R_{Th} \). In fact, it is easily shown that there exists a critical noise variance \( \sigma_C^2 \) below which \( R_{Th} < R^* \) and hence we observe only a first-order phase transition at which the overlap jumps from perfect inference, i.e., \( m^* = 1 \), to zero-information inference, i.e., \( m^* = 0 \). This critical noise variance for the considered setting in Fig. 6 is \( \sigma_C^2 = 0.084 \). For settings below this critical variance, the inference problem shows the so-called all-or-nothing phenomenon\(^24\) [72]. As we show later on, with higher order fields, the inference problem always show the all-or-nothing behavior, regardless of the noise variance.

The transition from all-to-something behavior, i.e., \( \sigma^2 > \sigma_C^2 \), to all-or-nothing case is demonstrated in Figs. 8 and 9, where we plot the theoretic and algorithmic results for \( \sigma^2 = 0.2 > \sigma_C^2 \) and \( \sigma^2 = 0.05 < \sigma_C^2 \), respectively. Similar to Fig. 6, the fixed-point solutions, the RS solution and the algorithmically achievable overlap are shown by dotted-dashed, solid and dashed lines respectively. The algorithmic gap is also denoted by \( \Delta_{Alg} \) in the figures. By comparing Figs. 8 and 9, one can further observe that the all-or-nothing behavior\(^25\) in this case is achieved at the expense of higher algorithmic gaps. Further investigations conjecture that this behavior is generic.

Although \( R_{Th} \) and \( R^* \) are the ones which are derived analytically, we are also interested on the critical load \( R_{Alg} \) from algorithmic viewpoint. In fact, \( R_{Alg} \) represents the load

---

\(^23\)That is given by the global minimum of the modified free energy.

\(^24\)We discuss this phenomenon in details on the next subsection.

\(^25\)As we show in Section V, this behavior is favorable for secure learning.
at which we transit from the initial state of perfect learning of labels to the state of erroneous learning, when we employ the approximate message passing algorithm to implement the Bayesian estimator. More precisely, it represents the maximum load at which perfect learning is feasible computationally. It is worth noting that, unlike the theoretical results, the algorithmic results always show the three states of all, some and nothing being recovered. This is in fact the cost we pay to achieve the computationally tractability.

A summary of the findings for $\lambda = 2$ is given below:

**Conclusion 2 (Inference on a Quadratic Generative Model):** Observations from a purely quadratic generative model show a first-order and a second-order phase transition with respect to the load. Below Shannon’s limit, perfect inference is achieved. As we pass this limit, a first-order phase transition occurs, where the overlap jumps from $m^* = 1$ to some $m^* < 1$. At the threshold load $R_{Th}$, the overlap shows a second-order phase transition and becomes exactly zero. This means that by surpassing the threshold load, the learned labels contain no information about the true labels anymore. The two critical loads, i.e., Shannon’s limit and the threshold, become closer as the noise variance reduces. As the noise variance drops below a critical level, the overlap shows all-or-nothing behavior at Shannon’s limit. The algorithmic performance however demonstrates always the all-to-something-to-nothing behavior.

**C. Higher-Order Nonlinear Generative Models**

For higher-order fields, i.e., $\lambda \geq 3$, the detailed analytic investigations of the first-order phase shift is not trivial. Nevertheless, the numerical investigations and intuition behind the problem suggest the following conjecture: for larger choices of $\lambda$, the problem shows the same behavior with three major changes

1) The threshold load, at which the curvature of $L_m(\sigma)$ at $m = 0$ changes, reduces as $\lambda$ increases. The analytical derivation of the threshold load for higher-order fields is not trivial; however, numerical investigation suggests that either the threshold load drops significantly for $\lambda \geq 3$ or it is exactly zero, i.e., the curvature change happens at $R_{Th} = \epsilon$, where either $\epsilon \approx 0$ or $\epsilon = 0$ for higher-order fields. Our finite resolution searches failed to find this load for $\lambda \geq 3$. This means that for higher fields $m = 0$ is (almost) always a local minimum.

2) Since $R_{Th} \to 0$, there exists no setting at which the Bayesian algorithms gets into the state with some labels being correctly learned. This means that at higher-orders, the overlap jumps at the critical load $R^*$ directly from all recovered state, i.e., $m^* = 1$, to the state of nothing being recovered, i.e., $m^* = 0$. In other words, the Bayesian inference always shows the all-or-nothing phenomenon for $\lambda \geq 3$; the behavior which we see for $\sigma^2 \leq \sigma^2_C$ with the purely quadratic field. The all-or-nothing phenomenon is a well-known asymptotic behavior being reported in various other inference problems; see for example [72], [73], [74], [75], [76], and [77].

3) The overlap achieved by AMP from uninformed initialization tends to be zero for almost all loads. This indicates that a computationally feasible implementation of the Bayesian algorithm via the approximate message passing is unable to perform a perfect learning, even though it is theoretically achievable. A known solution to this issue is to spatially couple the random field elements [78], [79]. Investigations in this respect are, however, out of the scope of this study and are left as a potential direction for future work.

The above discussions are summarized with some heuristics as follows: for $\lambda \geq 3$, the modified free energy $L_m(\sigma)$ contains three local extreme points: $m = 0, m = m_0$ for $0 < m_0 < 1$ and $m_1 = 1$. The boundary points $m = 0$ and $m = 1$ are local minima, and $m = m_0$ is a local maximum. Starting with small loads, the local minimum at $m = 1$ is smaller than the one at $m = 0$ and hence the overlap reads $m^* = 1$. By increasing the rate, $L_1(\sigma)$ starts to increase while $L_0(\sigma) = C(\sigma)$ remains unchanged. As we surpass a critical load $L_1(\sigma) > L_0(\sigma)$ and hence $m = 0$ becomes the global minimizer which concludes $m^* = 0$. From the channel coding theorem of Shannon and our observations with the purely quadratic field, we expect the critical load to be $R^*$ defined in (96). Numerical investigations confirm this conjecture; see Figs.10 and 11. We give a heuristic proof to this conjecture later on.

The all-or-nothing phenomenon, along with a numerical validation of our conjecture, is illustrated in Fig. 10 where we plot the modified free energy $L_m(\sigma)$ against $m$ for small deviations below and above the conjectured critical load $R^* = C(\sigma)/\log 2$, i.e., we set $R = R^* - \epsilon$ (solid line) and $R = R^* + \epsilon$ (dashed line) for $\epsilon = 0.01$. As the figure shows by surpassing the critical load, $L_1(\sigma)$ passes $C(\sigma) = L_0(\sigma)$, and hence the global minimum jumps from $m^* = 1$ to $m^* = 0$. This jump is shown from another perspective in Fig. 11, in which we plot the overlap against the load for $\lambda = 3$ and 5. As the figure shows, the phase transition happens at $C(\sigma)/\log 2$ for both of them, and the curves are not numerically distinguishable. This result further explains the findings by Sourlas in [25].

**Remark 2:** The Gaussian random field can be observed as the encoder of the random coding scheme introduced by Sourlas in [25]. A related side-note in this respect is the dimensional scaling of Sourlas’ coding. Similar to Shannon’s random coding approach, Sourlas’ coding directly maps the message into a real-valued sequence. On the other hand, Fig. 11 indicates that Sourlas’ coding achieves the channel capacity with a pure cubic Gaussian random field. This observation demonstrates a significant drop in the number of random components in the coding scheme. Unlike Shannon’s random coding, in which a codebook with an exponentially large number of random entries is required, this code requires a generator function whose number of random components scales only cubically with the message length. This is an interesting finding and is worth further investigations.

We now focus on the information rate achieved by higher-order fields and try to derive heuristically the critical

$26$This is in contrast to known coding schemes whose operations are performed in a finite field.
Fig. 10. Modified free energy $L_m(\sigma)$ against $m$ for $\lambda = 3$ considering $\sigma^2 = 0.1$. The figure is plotted for $R = R^* - \epsilon$ (solid line) $R = R^* + \epsilon$ (dashed line), where we set $R^* = \mathcal{C}(\sigma)/\log 2$ and $\epsilon = 0.01$. At the critical load, the values at $m = 0$ and $m = 1$ coincide up to numerical errors.

Fig. 11. Overlap $m^\star$ against load $R$ for $\lambda = 3$ and $\lambda = 5$. The results show a first-order phase transmission at the channel capacity from $m^\star = 1$ to $m^\star = 0$. This demonstrates the all-or-nothing phenomenon in higher order models. As one can observe, the curves for $\lambda = 3$ and $\lambda = 5$ are not numerically distinguishable.

Using the approximated value for $L_1(\sigma)$, we can conclude that $m^\star = 1$, if

$$R \leq \frac{\mathcal{C}(\sigma)}{\log 2}.$$  

Consequently, $m^\star = 0$, if $R$ exceeds this critical load.

Based on this heuristic justification, as well as the numerical validations, we have the following conjecture for higher-order Gaussian fields:

**Conjecture 1:** There exists an order $\lambda_0 \geq 3$, such that for $\lambda \geq \lambda_0$, the overlap reads

$$m^\star = \begin{cases} 1 & R \leq \frac{\mathcal{C}(\sigma)}{\log 2} \\ 0 & R > \frac{\mathcal{C}(\sigma)}{\log 2} \end{cases}$$  

and the information rate is given by

$$I(\sigma) = \begin{cases} R \log 2 & R \leq \frac{\mathcal{C}(\sigma)}{\log 2} \\ \mathcal{C}(\sigma) & R > \frac{\mathcal{C}(\sigma)}{\log 2} \end{cases}.$$  

The evidence for (101) is simply what we observe in Fig 11. To examine (102), we further sketch $I(\sigma)$ against the load for $\lambda = 3$ in Fig. 12. The figure suggests that $\lambda_0 = 3$ in the conjecture, since the curve for $\lambda = 3$ meets the conjectured curve almost perfectly.

We summarize the findings for $\lambda \geq 3$ as follows:

**Conclusion 3 (Inference on Higher-Order Generative Models):** Bayesian estimation for nonlinear generative models of higher-order, i.e., $\lambda \geq 3$ experiences the all-or-nothing phenomenon: below a critical load, perfect inference, i.e., $m^\star = 1$, is achieved. By surpassing the critical load, the learned labels contain no information about the true labels anymore, i.e., the overlap jumps from $m^\star = 1$ to $m^\star = 0$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
The critical load meets Shannon’s limit. Nevertheless, for these fields, perfect inference cannot be achieved via the feasible realization of the Bayesian learning algorithm with approximate message passing.

V. COMMENTS ON SECURE LEARNING

Our findings lead us to an interesting conclusion. In settings in which we need a phase transition in the outcome of the learning algorithm, we can use nonlinear models. Instances of such settings are those with secrecy constraints. In these settings, the samples in $y$ are observed by two terminals: a legitimate terminal and a malicious one. The desired behavior in this setting is to enable the legitimate terminal to learn the labels perfectly while the malicious one receives zero information from its eavesdropping.

The secure transmission setting in its generic form is studied by Wyner in [37]. Wyner finds a maximum load for secure transmission, similar to the one given by Shannon for reliable transmission. The limit indicates that in a wiretap setting, the maximum load for which a perfectly secure transmission is achievable is given by the difference between Shannon’s limits for the legitimate and the eavesdropping channels.

Considering the conclusions in the previous section, one can conjecture that the same behavior as the one explained by Wyner is observed in our setting when the generative model is strictly nonlinear. In this case, it is enough to degrade the channel to the malicious terminal, such that its corresponding threshold load falls below the system load. This guarantees that the overlap of this terminal is $m_1^* = 1 = 0$ which corresponds to independent inference.

In this section, we try to show that this conjecture is, in fact, correct. This leads to an interesting conclusion: for secure inference, it is enough to use higher-order Gaussian random fields. This not only re-discovers Wyner’s result, but also specifies explicitly the model by which secure inference is enabled.

We start our investigations by considering a simple wiretap setting for our learning problem. A sequence of labels $s \in \{\pm 1\}^K$ is observed through a nonlinear generative Gaussian model $\mathcal{V}(\cdot)$ of order $\lambda$. The transformed sequence $x = \mathcal{V}(s) \in \mathbb{R}^N$ is then observed by two terminals which know the model, i.e., $\mathcal{V}(\cdot)$. A legitimate terminal indexed by $i = 1$ and a malicious terminal indexed by $i = 2$. Terminal $i \in \{1, 2\}$ observes

$$y_i = x + w_i,$$

where $w_i \sim \mathcal{N}(0, \sigma_i^2 I_N)$. To guarantee the feasibility of secure inference, we set $\sigma_2^2 > \sigma_1^2$. Terminal $i$ employs the optimal Bayesian algorithm to recover the labels. For this setting, we adopt the notion of perfect secrecy:

**Definition 3 (Perfectly Secure and Reliable Inference):** The inference is said to be asymptotically reliable and secure if for $N \uparrow \infty$, we have

- The overlap of the legitimate terminal is $m_1^* = 1$.
- The information rate to the malicious terminal is zero, i.e.,

$$\lim_{N \uparrow \infty} \frac{1}{N} I(s; y_2 | V) = 0.$$  

In the above definition, the first constraint restricts the legitimate terminal to recover the labels asymptotically perfectly. The second constraint moreover enforces the information leakage to the malicious terminal to be zero.

A. Secure Inference via Nonlinear Generative Models

We now consider $\mathcal{V}(\cdot)$ to be a Gaussian random field with

$$\Phi(x) = x^\lambda.$$  

Hence, the transformed sequence $x$ contains unit-variance Gaussian symbols. We further assume that $\lambda$ is chosen such that (101) and (102) are exact. To secure our learning model, we follow the strategy below:

- Generate $K_2 = \lceil NC(\sigma_2) / \log 2 \rceil$ i.i.d. uniform labels

$$r = [r_1, \ldots, r_{K_2}]^T.$$  

- Add $r$ as a prefix to $s$ and permute the extended vector via a randomly-generated permutation matrix $\Pi$, i.e., construct the new sequence of labels $\tilde{s} \in \{\pm 1\}^{K+K_2}$ as

$$\tilde{s} = \Pi [r^T, s^T]^T,$$  

for a permutation matrix $\Pi \in \{0, 1\}^{(K+K_2) \times (K+K_2)}$ which is known to all parties in the setting, and whose permuted vector $\tilde{s}$ has the following property: for $\ell \in [K_2]$, define the $\ell$-th bin $\bin_{\ell}$ of $\tilde{s}$ as

$$[\tilde{s}]_\ell = \{ \tilde{s}_k : k \in [(\ell - 1)B + 1 : \ell B] \},$$  

for the bin size

$$B = \left[ 1 + \frac{K_2}{K + K_2} \right].$$

Permutation $\Pi$ guarantees that for every choice of $\ell \in [K_2]$, there exists only one symbol of $r$ in $[\tilde{s}]_\ell$.

\footnote{Wyner’s result indicates that with $\sigma_1^2 < \sigma_2^2$, secure transmission is information-theoretically impossible.} \footnote{We use this appellation due to its connection to the random linear binning approach; see [80].}
The randomized permutation applied by $\Pi$ can be observed as the random binning approach; see [80], [81], [82], and [83] for more details on the random binning technique and its applications in information theory. Interestingly, the described property of the random permutation implies that there exist exponentially many permutations.

To secure the model, we pass $\tilde{s}$ through the model $\mathcal{V}(\cdot) : \{\pm 1\}^{K+K_2} \mapsto \mathbb{R}^N$. The effective load of this model is defined by excluding the randomly generated labels, i.e., $r$. We distinguish between the effective load and the system load by denoting the former with $R = K/N$ and the latter by $\tilde{R} = (K + K_2)/N$. Note that $R$ and $\tilde{R}$ are related as

$$\tilde{R} = R + \frac{C(\sigma_2)}{\log 2} + \epsilon_N,$$

for some $\epsilon_N = O(1/N)$ which tends to zero in the asymptotic regime.

We claim that using this simple modification, the inference in this model is perfectly secure and reliable in the large-system limit if the effective load satisfies

$$R < \frac{C(\sigma_1) - C(\sigma_2)}{\log 2}.$$

This upper-bound is in fact Wyner’s asymptotic limit for the Gaussian wiretap channel [84].

In the sequel, we give a heuristic proof to this claim. We start the analysis by investigating the security constraint. By the chain-rule, we have

$$I(\tilde{s}; y_2|\mathcal{V}) = I(r; s; y_2|\mathcal{V}) ,$$

(112a)

$$= I(s; y_2|\mathcal{V}) + I(r; y_2|s, \mathcal{V}).$$

(112b)

Hence, the information leakage to the eavesdropper is given by

$$I(s; y_2|\mathcal{V}) = I(\tilde{s}; y_2|\mathcal{V}) - I(r; y_2|s, \mathcal{V}).$$

(113)

Noting that the system load $\tilde{R} \geq C(\sigma_2)/\log 2$, we use (102) and write

$$\lim_{N \to \infty} \frac{1}{N} I(\tilde{s}; y_2|\mathcal{V}) = C(\sigma_2).$$

(114)

This limit calculates the asymptotic information rate corresponding to the first term on the right hand side of (113).

To calculate the information rate corresponding to the second term in (113), we note that $I(r; y_2|s, \mathcal{V})$ calculates the mutual information in a modified model, where the recovery algorithm tries to recover $r$ while the true labels in $s$ are revealed to it: let $s = s_0$ be known at both sides. Since $s_0$ is known at the recovery algorithm, each bin of $\tilde{s}$, which is constructed from $s_0$ and $r$ via the random permutation $\Pi$, can be seen as a uniform Bernoulli symbol.\footnote{Note that the bins are discrete random variables, and hence the information rate only depends on their distribution.} Since we have $K_2$ bins in this setting, the load in this new setting reads $K_2/N = C(\sigma_2)/\log 2$, and thus we have

$$\lim_{N \to \infty} \frac{1}{N} I(\tilde{s}; y_2|s = s_0, \mathcal{V}) = C(\sigma_2).$$

(115)

As the right hand side does not depend on $s_0$, we can conclude that

$$\lim_{N \to \infty} \frac{1}{N} I(\tilde{s}; y_2|s, \mathcal{V}) = C(\sigma_2).$$

(116)

Finally, using the fact that $I(r; y_2|s, \mathcal{V}) = I(\tilde{s}; y_2|s, \mathcal{V})$, we have

$$\lim_{N \to \infty} \frac{1}{N} I(r; y_2|s, \mathcal{V}) = \lim_{N \to \infty} \frac{1}{N} I(\tilde{s}; y_2|s, \mathcal{V})$$

$$= C(\sigma_2).$$

(117a)

(117b)

This concludes that

$$\lim_{N \to \infty} \frac{1}{N} I(s; y_2|\mathcal{V}) = 0,$$

(118)

which means that the information rate to the malicious terminal is zero for any choice of $\tilde{R}$.

For reliable inference, the legitimate terminal needs to infer perfectly all labels of $\tilde{s}$ and remove the first $K_2$ labels. From the conjecture in (101), we know that this is achievable, if

$$\tilde{R} \leq \frac{C(\sigma_1)}{\log 2}.$$

(119)

By plugging in (110) into (119) and sending $\epsilon_N$ to zero for asymptotically large $N$, we conclude that perfect inference is achieved if (111) is satisfied. Hence, (111) guarantees secure and reliable inference.

\textit{Remark 3:} Following Remark 2, one can have the same conclusion here: our proposed secure coding scheme achieves Wyner’s limit with a random code whose generator function is given by a Gaussian field. This is an alternative random coding approach whose number of randomly-generated components grows at most cubically with the message length.

\subsection{B. Connecting Fyodorov’s Results to Wyner’s}

Although the model considered in [27] is different from what we study, related behaviors are reported in both studies. Fyodorov reports a second-order phase transition with respect to $\sigma^2$ for fields which do not contain linear terms. This phase transition breaks the replica symmetry of the problem and hence is specified via the full RSB solution. This symmetry breaking comes from the fact that the performance of the MAP algorithm, used for learning by Fyodorov, is described by the free energy at zero temperature at which the symmetry of the corresponding spin glass can be broken.\footnote{For the linear field, the corresponding spin glass shows RS, even at zero temperature.}

Given the fact that our considered learning algorithm is Bayesian optimal,\footnote{And hence, it is information-theoretically optimal too.} such an asymmetry does not occur. We however observe phase transition with respect to the system setting, i.e., load. From the information-theoretic viewpoint, this is the same behavior: The input dimension, i.e., dimension of $s$, in our setting is controlled by the load,\footnote{Since we have a finite-alphabet input.} while it is controlled by the SNR in Fyodorov’s work.\footnote{Since the input in [27] is continuous.} Our final conclusion is hence similar to what Fyodorov concludes:

\begin{thebibliography}{9}

\bibitem{ref1} Title of the reference, Journal Name, Volume, Issue, Pages, Year.

\end{thebibliography}
in strictly nonlinear Gaussian models, we can achieve perfect secrecy.

Our approach sketches a direct connection between Wyner’s result and Gaussian random fields, and recovers the same limit for the Gaussian model. We now try to connect Fyodorov’s result to Wyner’s limit by investigating the same wiretap model. This investigation not only clarifies the connection between the two results, but also introduces an alternative secure coding approach. An advantage of this alternative approach is that it is more intuitive from practical viewpoints; however, as we will see, a downside is that it does not achieve Wyner’s limit. We finally discuss the key sub-optimal point of this approach and explain the solution to this point which can be considered in future work.

Let us now start again with the wiretap setting described at the beginning of this section and specified by (103). We intend to transfer securely $K$ labels to terminal 1 while keeping terminal 2 blind. Let $s \in \{±1\}^K$ denote the vector of labels. We follow these steps:

- We first apply spherical source coding. In this step, we choose $2^K$ distinct points on a scaled $(D−1)$-sphere, i.e., we find $C = \{\hat{s}_1, \ldots, \hat{s}_K\}$ on the $(D−1)$-dimensional hypersphere with radius $\sqrt{D}p_S$. We denote the minimum normalized distance between distinct points by $2d$; this means

$$2d = \min_{i \neq j \in [2^K]} \frac{||\hat{s}_i - \hat{s}_j||^2}{Dp_S}.$$  

The value of $d$ is related to the maximum normalized inner product $p \in [0, 1]$ as $d = 1 − p$, where $p$ is given by [85]

$$p = \max_{i \neq j \in [2^K]} \frac{\langle \hat{s}_i; \hat{s}_j \rangle}{p_S}.$$  

To each outcome of $s$, we allocate one distinct point $\hat{s}_k$ and refer to it as the codeword.

- In the second step, we select the codeword which corresponds to the given realization of $s$ and pass it through an order $\Lambda$ Gaussian field $V(s)$ on the hypersphere. We then add artificial noise to it. Let the selected codeword be $\hat{s}^* \in C$, the secure transformed sequence is

$$x = V(\hat{s}^*) + w_0,$$

where $w_0 \sim N(0, \theta^2I_N)$ for some noise variance $\theta^2$.

Terminal $i$ observes $y_i$ described in (103). To learn the labels, it uses the MAP algorithm:34 Terminal $i$ finds point $\hat{s}_i$ on the hypersphere whose transform by $V(\cdot)$ is closest to its observations, i.e., it finds

$$\hat{s}_i = \arg\min_{\forall \hat{s}_i \in S^{D−1}} ||y_i - V(\hat{s}_i)||^2.$$  

It then decodes the labels by finding the codeword in $C$ which is closest to $\hat{s}_i$.

To achieve perfect secrecy and reliability, we follow these arguments:

34In this particular case, the MAP algorithm reduces to least-squares.

- If the overlap between the recovery $\hat{s}_i$ and the codeword, i.e.,

$$m_i = \frac{\langle \hat{s}_i; \hat{s}^* \rangle}{p_S},$$  

is zero, the recovery and the codeword are completely uncorrelated. This leads to an error probability of 1.

- If the overlap is larger than $1−d/2$, the closest codeword to the recovery is the codeword. This leads to error probability of 0.

Hence, to guarantee secure and reliable transform, we need to design the parameters such that the asymptotic overlap between $\hat{s}_1$ and the codeword, denoted by $m_1$, is larger than $1−d/2$, and the asymptotic overlap between $\hat{s}_2$ and the codeword, denoted by $m_2$, is zero. It is worth mentioning that these conclusions are given by a slight intuition whose proof of rigor is skipped for the sake of simplicity and can be considered in more detailed studies.

From Fyodorov’s result, we can calculate the overlap explicitly. As a result, for reliability, we limit the load $R = K/N$ and $P_S$, such that the minimum distance of codewords is bounded by $2(1−m_1)$ from above, and for security we set $\theta^2$ and $P_S$, such that $m_2$ passes the phase transition and satisfies $m_2 = 0$. To this end, we consider the following results:

- Finding the maximum load $R$ for a specific minimum distance reduces to the well-known problem of spherical covering which is still an open problem. We hence approximate it with an available tight bound in the literature. We assume that Kabatiansky-Levenshtein’s bound is tight in the large-system limit, and approximate maximum load as [85]

$$R_{max} \approx \mu \left(\frac{1+q}{2q} \log_2 \frac{1+q}{2q} - \frac{1−q}{2q} \log_2 \frac{1−q}{2q}\right),$$

where $q = \sqrt{1−p^2}$ setting $p = 2m_1 − 1$, and $\mu = D/N$.35

- For the purely quadratic field with covariance function $\Phi(x) = x^2/2$, the full RSB solution for the asymptotic overlap between the recovery and codeword is given by Result 1 given in Appendix A. We hence assume $V(\cdot)$ is the purely quadratic field with the same covariance.

- Considering the codeword $\hat{s}^*$, the power of the transformed symbol is given by

$$P_T = E\{x^2\} = \Phi(\langle \hat{s}^*; \hat{s}^* \rangle) + \theta^2 = \frac{P_S^2}{2} + \theta^2.$$  

As a result, Wyner’s limit on maximum load for secure and reliable inference is

$$C_W = \frac{1}{2} \left[ \log_2 \left(1 + \frac{P_S^2}{2\sigma_1^2} + \frac{\theta^2}{\sigma_1^2}\right) - \log_2 \left(1 + \frac{P_S^2}{2\sigma_2^2} + \frac{\theta^2}{\sigma_2^2}\right) \right]$$  

Note that this bound is achievable when we use the strategy explained in Section V-A.

Fig. 13 shows the value of $R_{max}$ against $\log(1/\sigma_1^2)$ when we set $P_S = \sigma_2^2 = 1$. In this figure, for each choice of $\sigma_1^2$,
we find $\theta^2$ such that $m_2 = 0$, i.e., we set

$$\theta^2 = \frac{P_S}{\Gamma_{Th} - \sigma^2},$$

(128)

following Result 1. We then use Result 1 once again to determine $m_1$ by setting

$$\Gamma = \frac{P_S}{\sigma^2 + \theta^2}.$$  

(129)

To fulfill the reliability constraint, we set $p = 2m_1 - 1$ and determine $R_{\text{max}}$ from (125). Finally, $R_{\text{max}}$ is maximized numerically over $\mu \leq 1$. For sake of comparison, Wyner’s limit $C_W$ is further plotted in the figure by a dashed line.

As the figure shows, the achieved maximum load does not meet Wyner’s limit. This follows the fact that the considered strategy is not information-theoretically optimal in various respects. We finalize our discussions in this section by pointing out few final remarks:

- The parameters in Fig. 13 are chosen such that the approximation in (125) is tight. It is however worth mentioning that the right hand side of (125) is in general an upper bound on $R_{\text{max}}$. This means that the given approximation is slightly optimistic.
- There is a key source of sub-optimality in the strategy considered in this section: the spherical coding technique which maps the labels on the hypersphere uses spherical covering. This approach is known to be sub-optimal [86]. From the literature, we believe that the gap in Fig. 13 is mainly due to this source of sub-optimality. In fact, it is widely known in the coding theory that minimum distance is not the key player in codebook design [42], [86]. We therefore believe that the gap between the two curves shrinks by replacing the spherical covering codebook with a denser codebook.\textsuperscript{36} We leave this direction of study as a potential future work.

\textsuperscript{36}This could be simply done by random coding on hypersphere; see for example the recent work in [87].

VI. CONCLUSION AND FINAL REMARKS

Strictly nonlinear generative models show a first-order phase transition with respect to the load: at a threshold load, the overlap jumps from one to zero and the information rate changes from a linear function of load to a constant. This phase transition leads to the all-or-nothing phenomenon for strictly nonlinear fields. Numerical investigations show almost perfect match to the conjectured asymptotic behavior for orders more than two. This is a close match to the asymptotic behavior of random coding schemes described by the channel coding theorem of Shannon which cannot be seen by the linear field. This observation can be illustrated through the connection of our model to Sourlas’ coding scheme. A key finding in this respect is that using a Gaussian random field, the channel capacity can be achieved by a random code whose generator function is described by a Gaussian field. The number of random components in this code grows in terms of the message length with a polynomial order.

In strictly nonlinear models, the overlap can be set to zero. This is another behavior which is not observed in linear models. This enables us to secure the model from passive attacks in which malicious terminals try to learn the model parameters from overheard data. The heuristic derivation of maximum load, by which perfect secrecy is guaranteed, shows that achieved maximum load for large-order Gaussian fields matches the information-theoretic limit derived by Wyner. Hence, it proves Wyner’s secure coding theorem via an alternative coding scheme whose encoder is a Gaussian random field. This finding depicts the efficiency of non-linear fields for secure coding and suggests further investigations in this respect through finite dimensional analyses and simulations. To this end, one can develop AMP algorithms for Bayesian inference on nonlinear generative Gaussian models. This is an interesting perspective for future work.

The findings of this study are shown to be connected to what Fyodorov reports in [27]. This is not surprising as the models, despite their differences, are related. Using spherical covering, we can invoke this connection and design a secure coding scheme based on Fyodorov’s encryption. Nevertheless, as shown in the last section, it seems to be more challenging to achieve Wyner’s limit by this scheme. The key challenge in this respect is the codebook design on the hypersphere. Investigation of the same strategy with denser randomly generated codebooks on the hypersphere is another interesting line for future work.

All above conclusions and directions for future work require the extension of the current study to a broader set of models and target problems. The work in this direction is currently ongoing.

APPENDIX A

FYODOROV’S NONLINEAR ENCRYPTION

In Section V, we use a particular result of [27] to illustrate our contribution to secure learning. We hence state this particular result in this appendix, and represent the core finding
of Fyodorov in the shadow of this result. Let the SNR $\Gamma$ be\textsuperscript{37}

$$\Gamma = \frac{P_S}{\sigma^2}. \quad (130)$$

Fyodorov determines the overlap $m^*$ which is defined in (25). The main result of [27] represents the following findings:

1) For linear fields, the solution is RS. In this case, $m^*$ increases analytically with the SNR. Hence, $m^* \downarrow 0$ when $\Gamma \downarrow 0$, and $m^* = 0$ does not occur.

2) For higher-order fields, the solution is RSB. In such cases, $m^*$ still increases analytically with the SNR, as long as the field contains a linear term.

3) For purely quadratic fields, $m^*$ shows a second-order phase transition. This means that there exists a threshold SNR $\Gamma_{\text{Th}}$ at which $m^*$ becomes absolutely zero.

For the particular case of purely quadratic fields, Fyodorov’s result is given below:

Result 1 (Full-RSB solution for quadratic fields): Consider the quadratic Gaussian field $\mathcal{V}(\cdot)$ whose covariance function is given by

$$\Phi(x) = \frac{x^2}{2}. \quad (131)$$

Define $\Gamma^*$ and $\Gamma_{\text{Th}}$ as

$$\Gamma^* = \frac{\mu}{(\mu - 1)^2}, \quad (132a)$$

$$\Gamma_{\text{Th}} = \frac{2\mu}{2-\mu}, \quad (132b)$$

and let $z$ be the solution of

$$\mu z^3 + 3\left(\frac{1}{2} - \mu\right)z^2 + 3(\mu-1)z + \Gamma - \frac{1}{\Gamma^*} = 0, \quad (133)$$

which lies in the interval $[0, 1]$. Then, as $N, K \uparrow \infty$ with a fixed $\mu \leq 1$, we have

$$m^* = \begin{cases} \frac{1-\mu}{(1-\mu)\Gamma} & \Gamma^* \leq \Gamma \\ \mu(1-x)^3 & \Gamma_{\text{Th}} \leq \Gamma \leq \Gamma^* \\ 0 & \Gamma \leq \Gamma_{\text{Th}}. \end{cases} \quad (134)$$

Using Result 1, we plot $m^*$ against the SNR in Fig. 14. It is observed that as the SNR goes below $\Gamma_{\text{Th}}$, the overlap $m^*$ becomes zero. This key observation makes the connection between the work by Fyodorov and Wyner’s results on the wiretap setting; nevertheless, this connection was left unaddressed in [27].

**APPENDIX B**

**DERIVATIONS VIA THE REPLICA METHOD**

We first recall that $Z_\beta(y, V)$ is the partition function at inverse temperature $\beta$ determined for the Hamiltonian $\mathcal{E}(u|y, V)$ that is defined in (26). Moreover, $\mathcal{V}(\cdot): \mathbb{R}^K \rightarrow \mathbb{R}^N$ is the Gaussian random field, $y \in \mathbb{R}^N$ is the noisy observation, and $R$ is the load factor defined as the limit of $K/N$ when $K$ and $N$ grow large together. We now start the derivations by considering the scaled version of free energy, i.e.,

$$-\beta \mathcal{F}(\beta) = \frac{1}{N} \mathbb{E}\{\log Z_\beta(y, V)\}. \quad (135)$$

Using Riesz’s identity [88], we have

$$\frac{1}{N} \mathbb{E}\{\log Z_\beta(y, V)\} = \frac{1}{N} \lim_{t \uparrow 0} \frac{\partial}{\partial t} \log \mathbb{E}\{|Z_\beta(y, V)|^t\}. \quad (136)$$

Riesz’s identity reduces the logarithmic integration into the derivation of $t$-th moment, i.e.,

$$Z_\beta(t) := \mathbb{E}\{|Z_\beta(y, V)|^t\}. \quad (137)$$

Nevertheless, it is still a challenging task to determine a real-valued moment. We hence calculate the right hand side for integer $t$ as an analytic term in $t$ and continue it to the real axis by assuming the so-called replica continuity. This is a classical practice in the replica analyses; see [2], [29], [35], [38], and [55] for detailed discussions on the replica continuity assumption. The replica continuity assumption leads to

$$Z_\beta(t) = \mathbb{E}\left\{\sum_{u \in \{\pm 1\}^K} \exp\{-\beta \mathcal{E}(u|y, V)\}\right\}^t \quad (138)$$

$$= \mathbb{E}\left\{\sum_{u_1 \in \{\pm 1\}^K} \cdots \sum_{u_t \in \{\pm 1\}^K} \prod_{a=1}^t \exp\{-\beta \mathcal{E}(u_a|y, V)\}\right\}$$

$$= \mathbb{E}\left\{\sum_{\{u_a\}} \exp\{-\beta \sum_{a=1}^t \mathcal{E}(u_a|y, V)\}\right\}$$

$$= \mathbb{E}_{y, V}\left\{\sum_{\{u_a\}} \exp\{-\frac{\beta}{2} \sum_{a=1}^t \|y - \mathcal{V}(u_a)\|^2\}\right\}$$

$$= \mathbb{E}_{s, w, \mathcal{V}}\left\{\sum_{\{u_a\}} \exp\{-\frac{\beta}{2} \sum_{a=1}^t \|\mathcal{V}(s) + w - \mathcal{V}(u_a)\|^2\}\right\},$$

\textsuperscript{37}Fyodorov defines the noise-to-signal ratio in his work and denote it with $\gamma$. In this respect, $\Gamma = \gamma^{-1}$. 

---

**Fig. 14.** Overlap against the SNR for a pure quadratic Gaussian random field with covariance function $\Phi(x) = x^2/2$ and $\mu = 2/3$. The curve is sketched using the result derived by Fyodorov in [27].
where we define the notation
\[
\{u_a\} = (u_1, \ldots, u_t) \in \prod_{a=1}^{t} \{\pm 1\}^{K}. \tag{139}
\]

The sequence \(s\) is a random sequence uniformly distributed on \([2^{NR}]\). Using the independency of \(w, V\) and \(s\), the expectation is determined as given in (140), shown at the bottom of the next page. We now use the fact that outputs of the random field and Gaussian noise entries are i.i.d.. We hence can factorize the expectation as in (141), shown at the bottom of the next page. We note that the noise process and the field are both Gaussian. Using the standard Gaussian integration, \[27, \text{Appendix B}\] determines the expectation in (141) as given in (142), shown at the bottom of the next page. Here, we use the notation \(G(\cdot)\) is a \(t \times t\) matrix whose element on row \(a\) and column \(b\) is defined as shown (143), shown at the bottom of the next page, (142). Here, we use the notation \(U \in \{\pm 1\}^{K \times (t+1)}\) to shorten the sequence of replicas, i.e., \(u_1, \ldots, u_{t+1}\), as follows
\[
U = [u_1, \ldots, u_{t+1}]. \tag{144}
\]

By substituting, the moment function is finally written as
\[
Z \beta (t) = 2^{-NR} \sum_{U} (\det \{G ((U; U))\})^{-N/2}. \tag{145}
\]

We now rewrite the summation over matrices of replicas to a summation over replica correlation matrices by dividing the space of all replicas into a set of sub-shells: we define the sub-shell \(S(Q)\) for a correlation matrix \(Q \in [-1, 1]^{(t+1) \times (t+1)}\) as
\[
S(Q) = \left\{ U \in \{\pm 1\}^{K \times (t+1)} : \{U; U\} = Q \right\}. \tag{146}
\]

Note that \(Q\) is a symmetric matrix with unit diagonal, i.e.,
\[
\begin{align*}
Q_{a,a} &= 1, \quad \text{(147a)} \\
Q_{b,a} &= Q_{a,b}, \quad \text{(147b)}
\end{align*}
\]

We now can write
\[
Z \beta (t) = 2^{-NR} \int dQ \exp \{-NT(Q)\} (\det \{G(Q)\})^{-N/2}. \tag{148}
\]

where \(\exp \{-NT(Q)\}\) is the sub-shell density, defined as
\[
\exp \{-NT(Q)\} = \sum_{U} \prod_{a=1}^{t+1} \prod_{b=a+1}^{t+1} \delta \left( (u_a; u_b) - Q_{a,b} \right), \tag{149}
\]

and
\[
dQ = \prod_{a<b} dQ_{a,b}. \tag{150}
\]

The moment is hence derived by finding the sub-shell density and substituting it in (148).

### A. Deriving the Sub-Shell Density

To calculate the sub-shell density, we utilize the inverse Laplace transform of the Dirac impulse function which indicates
\[
\delta (x) = \int_{\mathcal{J}} \frac{d\mathcal{L}}{2\pi i} \exp \{Lx\}, \quad \text{(151)}
\]

with \(\mathcal{J}\) being the \(\mathbb{R} \{L\} = C\) line in the complex plane for some constant \(C\). Substituting it into the definition of the sub-shell density, we have
\[
\exp \{-NT(Q)\} = \sum_{U} \prod_{a<b} \delta \left( (u_a; u_b) - Q_{a,b} \right) \quad \text{(152a)}
\]

\[
= \sum_{U} \prod_{a<b} \int_{2\pi i} \frac{dL_{a,b}}{2\pi i} \exp \{L_{a,b} (u_a; u_b) - Q_{a,b}\} \quad \text{(152b)}
\]

\[
= \int d\mathcal{L} \sum_{U} \exp \left\{ \sum_{a<b} L_{a,b} (u_a; u_b) - Q_{a,b} \right\}. \tag{152c}
\]

Here, we define the matrix \(\mathcal{L} \in \mathbb{R}^{(t+1) \times (t+1)}\) to be a symmetric matrix with unit diagonal elements and off-diagonal elements \(L_{a,b} = L_{b,a}\) and
\[
d\mathcal{L} = \prod_{a<b} \frac{dL_{a,b}}{2\pi i}. \tag{153}
\]

We refer to \(\mathcal{L}\) as the frequency-domain correlation matrix. Following the structure of matrices \(\mathcal{L}\) and \(Q\), we have
\[
\text{tr}\{\mathcal{Q}\} = t + 1 + 2 \sum_{a<b} L_{a,b} Q_{a,b}. \quad \text{(154)}
\]

By noting that \(U^T U\) is a symmetric matrix with diagonals being all \(K\), we can further write
\[
\text{tr}\{U^T U\} = K \left[ (t+1) + 2 \sum_{a<b} L_{a,b} (u_a; u_b) \right]. \quad \text{(155)}
\]

This concludes that
\[
\sum_{a<b} L_{a,b} (u_a; u_b) - Q_{a,b} \quad \text{(156)}
\]

\[
= \frac{1}{2} \text{tr}\{U^T U\} - \frac{K}{2} \text{tr}\{\mathcal{Q}\}
\]

\[
= \frac{1}{2} \text{tr}\{U^T U\} - \frac{K}{2} \text{tr}\{\mathcal{L}\}. \quad \text{(157)}
\]

By substituting into the final expression for the sub-shell density, i.e., (152c), we have
\[
\exp \{-NT(Q)\} = \int d\mathcal{L} \sum_{U} \exp \left\{ \frac{1}{2} \text{tr}\{U^T U\} - \frac{K}{2} \text{tr}\{\mathcal{L}\} \right\}
\]

\[
= \int d\mathcal{L} \exp \left\{ -\frac{K}{2} \text{tr}\{\mathcal{L}\} \right\} \sum_{U} \exp \left\{ \frac{1}{2} \text{tr}\{U^T U\} \right\}. \quad \text{(157)}
\]

We now perform a simple variable exchange. Let \(V = U^T\) and define \(v_k \in \{\pm 1\}^{t+1}\) to be the \(k\)-th column in \(V\). We then
have
\[
\sum_{U} \exp \left\{ \frac{1}{2} \text{tr}\{ULU^T\} \right\} = \sum_{V} \exp \left\{ \frac{1}{2} \text{tr}\{V^TLV\} \right\} = \sum_{V} \exp \left\{ \frac{1}{2} \sum_{k=1}^{K} v_k^T L v_k \right\} = \prod_{k=1}^{K} \exp \left\{ \frac{1}{2} v_k^T L v_k \right\} = \left( \sum_{v \in \{\pm 1\}^{K+1}} \exp \left\{ \frac{1}{2} v^T L v \right\} \right) .
\]

Defining \( \mathcal{M}(\cdot) \) in terms of the frequency-domain correlation matrix as
\[
\mathcal{M}(L) = \log \sum_{v} \exp \left\{ \frac{1}{2} v^T L v \right\} ,
\]
we can finally write the summation over \( U \) as
\[
\sum_{U} \exp \left\{ \frac{1}{2} \text{tr}\{ULU^T\} \right\} = \exp \left\{ K \mathcal{M}(L) \right\} .
\]
This concludes that
\[
\exp \left\{ -NF(Q) \right\} = \int dL \exp \left\{ -K \left( \frac{1}{2} \text{tr}\{LQ\} - \mathcal{M}(L) \right) \right\} .
\]

### B. Moment Function in the Thermodynamic Limit

Substituting the sub-shell density into the expression for the moment function \( \mathcal{Z}_\beta(t) \), we can expand \( \mathcal{Z}_\beta(t) \) as in (162), shown at the bottom of the next page. We now send the spin glass to the thermodynamic limit, i.e., \( N \uparrow \infty \). This follows from the slightly non-rigorous assumption of limit exchange: we assume that the limits with respect to \( N \) and \( t \) exchange. Though limit exchange is not necessarily a valid assumption, it is a commonly accepted trick in the replica analysis to assume this particular exchange is valid [38]. After exchanging the limit, we deal with the following limit
\[
\lim_{N \uparrow \infty} \int dQ dL \exp \left\{ -NF(Q,L) \right\} ,
\]
where the function \( F(Q,L) \) is
\[
F(Q,L) := \frac{1}{2} \log \det \{G(Q)\} + \frac{R}{2} \text{tr}\{LQ\} - RLQ .
\]

This is a classical limit rise in large-deviations theory [89], [90], [91]. Using the saddle-point method and assuming that the exponent function has a unique minimum,\(^{38}\) we have
\[
\int dQ dL \exp \left\{ -NF(Q,L) \right\} = \exp \left\{ -NF(Q^*,L^*) \right\} ,
\]
where \( \equiv \) denotes the equivalence in logarithmic scale, and \( (Q^*,L^*) \) is the saddle-point of the exponent function.

We now find the saddle-point of \( F(Q,L) \). To this end, we first let
\[
\frac{\partial}{\partial Q} \left\{ \frac{1}{2} \log \det \{G(Q)\} + \frac{R}{2} \text{tr}\{LQ\} - RLQ \right\} = 0_{t+1},
\]

\(^{38}\) It does not lose the generality of the derivations.

---

\[
\mathcal{Z}_\beta(t) = \mathbb{E}_{w,v} \left\{ \mathbb{E}_s \left\{ \sum_{u_n} \exp \left\{ -\beta \sum_{a=1}^{t} \| \mathcal{V}(s) + w - \mathcal{V}(u_n) \|^2 \right\} \right\} \right\}
\]
\[
= \mathbb{E}_{w,v} \left\{ \sum_{u_n} \sum_{u_{i+1} \in \{\pm 1\}^{2}} 2^{-NR} \exp \left\{ -\beta \sum_{a=1}^{t} \| \mathcal{V}(u_{a+1}) + w - \mathcal{V}(u_n) \|^2 \right\} \right\}
\]
\[
= 2^{-NR} \sum_{u_{t+1},u_n} \mathbb{E}_{w,v} \left\{ \exp \left\{ -\beta \sum_{a=1}^{t} \sum_{n=1}^{N} \left( \mathcal{V}_n(u_{a+1}) + w_n - \mathcal{V}_n(u_n) \right)^2 \right\} \right\}
\]

\[
\mathcal{Z}_\beta(t) = 2^{-NR} \sum_{u_{t+1},u_n} \prod_{n=1}^{N} \mathbb{E}_{w_n,v_n} \left\{ \exp \left\{ -\beta \sum_{a=1}^{t} \left( \mathcal{V}_n(u_{a+1}) + w_n - \mathcal{V}_n(u_n) \right)^2 \right\} \right\}
\]

\[
\mathbb{E}_{w_n,v_n} \left\{ \exp \left\{ -\beta \sum_{a=1}^{t} \left( \mathcal{V}_n(u_{a+1}) + w_n - \mathcal{V}_n(u_n) \right)^2 \right\} \right\} = \frac{1}{\sqrt{\det \{G(u,U)\}}}
\]

\[
[G([U;U])]_{a,b} = \delta_{a,b} + \beta [s^2 + \Phi(1) + \Phi([u_a;u_b]) - \Phi([u_a;u_{a+1}]) - \Phi([u_b;u_{a+1}])]
\]
which concludes that

$$2RL = -\frac{\partial}{\partial Q} \log \det \{G(Q)\}.$$  \hspace{1cm} (166)

We now use the fact that for the symmetric matrix $X$, we have [92, Appendix A.4.1]

$$\frac{\partial}{\partial X} \log \det \{X\} = X^{-1}.$$  \hspace{1cm} (167)

Noting that $G(Q)$ is symmetric for a symmetric $Q$, we can use the chain rule to write that

$$\frac{\partial}{\partial Q} \log \det \{G(Q)\} = 2\beta \Phi'(Q) \odot \tilde{G}^{-1}(Q),$$  \hspace{1cm} (168)

where $\odot$ is the Hadamard product and $\tilde{G}^{-1}(Q)$ is a $(t+1) \times (t+1)$ matrix defined in terms of the inverse of $G(Q)$ as

$$\tilde{G}^{-1}(Q) = \left[ G^{-1}(Q) + dF - G^{-1}(Q)I_{t+1} \right],$$  \hspace{1cm} (169)

for some $c$ and $\tilde{c}$. This concludes that

$$L^* = -\frac{\beta}{R} \Phi'(Q^*) \odot \tilde{G}^{-1}(Q^*),$$  \hspace{1cm} (170)

where we choose $c$ and $\tilde{c}$, such that $L_{a,a} = 1$ for $a = 1, \ldots, t+1$. We further let

$$\frac{\partial}{\partial L} \left\{ \frac{1}{2} \log \det \{G(Q)\} + \frac{R}{2} \text{tr} \{LQ\} - R\mathcal{M}(L) \right\} = 0_{t+1},$$

which concludes that

$$Q^* = \sum_v vv^T \exp \left\{ \frac{1}{2} v^T L^* v \right\} \sum_v \exp \left\{ \frac{1}{2} v^T L^* v \right\}. $$  \hspace{1cm} (171)

The fixed-point equation (171) indicates that $Q^*$ is in fact the average correlation among the decoupled replicas.\footnote{This is why $Q^*$ is often called the replica correlation matrix.} To see this point clearly, let us define the **decoupled distribution of replicas** as

$$q_v (vL^*) := \frac{\exp \left\{ \frac{1}{2} v^T L^* v \right\} \sum_v \exp \left\{ \frac{1}{2} v^T L^* v \right\}}{\sum_v \exp \left\{ \frac{1}{2} v^T L^* v \right\}}. $$  \hspace{1cm} (172)

This is a Boltzmann distribution which is defined for a thermodynamic system with reduced dimension $t+1$. Using this notation, we can write

$$Q^* = E_{q_v} \{ vv^T \},$$  \hspace{1cm} (173)

which is the correlation matrix of the vector of replicas $v$.

C. Determining the Free Energy

As we exchange the limit, we can write

$$-\beta \mathcal{F}(\beta) = \lim_{t \to 0} \lim_{N \to \infty} \frac{1}{N} \log Z_\beta(t).$$  \hspace{1cm} (174)

We now replace the final expression for the moment function, i.e., the one given in (175), shown at the bottom of the next page, and use the asymptotic equivalence in the logarithmic scale to conclude (176), shown at the bottom of the next page, for $L^*$ and $Q^*$ which are given by the saddle-point equations.

D. Replica Symmetric Solution: Fixed-Point Equations

We now assume replica symmetry. This means that the saddle-point matrix $Q^*$ is a symmetric matrix of the following form

$$[Q^*]_{a,a} = 1 \quad 1 \leq a \leq t + 1,$$  \hspace{1cm} (177a)

$$[Q^*]_{a,b} = Q \quad 1 \leq a < b \leq t,$$  \hspace{1cm} (177b)

$$[Q^*]_{a,t+1} = m \quad 1 \leq a \leq t,$$  \hspace{1cm} (177c)

where $m, Q \in [0, 1]$. In this case, $G(Q^*)$ is of the following form

$$G(Q^*) = g(Q) I_t + f(Q,m) 1_t,$$  \hspace{1cm} (178)

where we define

$$g(Q) = 1 + \beta [\Phi(1) - \Phi(Q)],$$  \hspace{1cm} (179a)

$$f(Q,m) = \beta [\sigma^2 + \Phi(1) + \Phi(Q) - 2\Phi(m)].$$  \hspace{1cm} (179b)

As a result, the inverse matrix is given by

$$G^{-1}(Q) = \frac{1}{g(Q)} \left( I_t - \frac{f(Q,m)}{g(Q) + tf(Q,m)} 1_t \right).$$  \hspace{1cm} (180)

From the saddle-point equations, we conclude that the frequency-domain matrix at the saddle-point exhibits the same symmetry. Namely, its entries are given by

$$[L^*]_{a,a} = 1$$  \hspace{1cm} (181a)

for $1 \leq a \leq t + 1$,

$$[L^*]_{a,b} = \frac{1}{R} \Phi'(Q) \frac{f(Q,m)}{g(Q) + tf(Q,m)}$$  \hspace{1cm} (181b)

for $1 \leq a < b \leq t$, and

$$[L^*]_{a,t+1} = \frac{1}{R} \Phi'(m) \frac{1}{g(Q) + tf(Q,m)}.$$  \hspace{1cm} (181c)
for $1 \leq a \leq t$. By sending $t \downarrow 0$, the frequency-domain correlation matrix reduces to an RS matrix with unit diagonal and non-diagonal entries

$$L = \frac{\beta}{R} \Phi'(Q) \frac{f(Q,m)}{g^2(Q)},$$

for first $t$ columns and

$$E = \frac{\beta}{R} \Phi'(m) \frac{1}{g(Q)},$$

for the last column.

To specify $Q$ and $m$, we derive the second saddle-point equation in terms of $Q$ and $m$. We start by considering the fixed-point equation, i.e., (171). From the structure of $Q^*$, we conclude that

$$t(t-1)Q + t = \sum_v \left( \sum_{a=1}^t v_a \right)^2 \exp \left\{ \frac{1}{2} v^T L v \right\},$$

$$mt = \sum_v \left( \sum_{a=1}^t v_a \right) v_{t+1} \exp \left\{ \frac{1}{2} v^T L v \right\},$$

where the above equations are derived by summing up the $t \times t$ inner block of $Q$ and the last column. From the structure on $L$, we further have

$$v^T L v = 2E \left( \sum_{a=1}^t v_a \right) v_{t+1} + L \left( \sum_{a=1}^t v_a \right)^2$$

$$+ t (1 - L) + 1.$$  

This is then straightforward to see the identities given in (186), shown at the bottom of the next page. By substituting, we conclude the following two fixed-point equations

$$t(t-1)Q = \frac{\partial}{\partial L} \log \sum_v \exp \left\{ \frac{1}{2} v^T L v \right\},$$

$$mt = \frac{\partial}{\partial E} \log \sum_v \exp \left\{ \frac{1}{2} v^T L v \right\}.$$  

We proceed the derivations by calculating the right hand side function. Namely, we define $\mathcal{M}_c(L,E)$ as in (188), shown at the bottom of the next page. By using Gaussian integration technique, we can further write

$$\exp \left\{ \frac{L}{2} \left( \sum_{a=1}^t v_a \right)^2 \right\} = \int Dz \exp \left\{ \sqrt{L} \left( \sum_{a=1}^t v_a \right) z \right\},$$

where we define the notation

$$\int Dz = \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} \exp \left\{ -\frac{z^2}{2} \right\}. $$

We can now expand $\mathcal{M}_c(L,E)$ as given in (191), shown at the bottom of the next page.

We now define the random variable $S$ to be uniform on $\{\pm 1\}$, and the random variable $Z$ to be zero-mean unit-variance Gaussian. We then rewrite $\mathcal{M}_c(L,E)$ as in (193), shown at the bottom of the next page. Using the symmetry of $\cosh(\cdot)$, we can write

$$\mathbb{E}_{Z,S} \left\{ \cosh t \left( \sqrt{L}Z + ES \right) \right\} = \mathbb{E}_Z \left\{ \cosh t \left( \sqrt{L}Z + E \right) \right\}.$$

Defining function $W_t(\cdot)$ as

$$W_t(L,E) := \log \mathbb{E}_Z \left\{ \cosh t \left( \sqrt{L}Z + E \right) \right\},$$

we finally conclude that

$$\log \mathcal{M}_c(L,E) = \frac{(t+1) \log 2 + \frac{t(1-L)+1}{2}}{t} + W_t(L,E).$$

We now calculate the derivatives of $\log \mathcal{M}_c(L,E)$. To this end, we write the derivatives with respect to $E$ and $L$ as given in (195), shown at the bottom of page 8024. Using integration by part, one can further show that for $X \sim \mathcal{N}(\eta, \rho^2)$, we have

$$\mathbb{E} \left\{ \cosh^{t-1}(X) \sinh(X) \frac{X - \eta}{\rho^2} \right\} = \mathbb{E} \left\{ \cosh^t(X) \right\}$$

$$+ (t-1) \mathbb{E} \left\{ \cosh^{t-2}(X) \sinh^2(X) \right\}.$$  

Hence, we can determine the derivatives of $\log \mathcal{M}_c(L,E)$ with respect to $E$ and $L$ as given in (197), shown at the bottom of page 8024.

By substituting into (187), we finally have (198), shown at the bottom of page 8024, whose limits as $t \downarrow 0$ read

$$Q = \mathbb{E}_Z \left\{ \tanh^2 \left( \sqrt{L}Z + E \right) \right\},$$

$$m = \mathbb{E}_Z \left\{ \tanh \left( \sqrt{L}Z + E \right) \right\}.$$  

This concludes the fixed-point equations.

### E. Replica Symmetric Free Energy

We now calculate the free energy for the RS structure. We start with noting that with RS structured $Q^*$, matrix $G(Q^*)$ has two eigenvalues; namely,

- $g(Q)$ with multiplicity $t-1$, and
- $g(Q) + tf(Q,m)$ with multiplicity 1.

$$Z_{\beta}(t) = 2^{-NR} \exp \left\{ -N \min_{L, Q^*} \left\{ \frac{1}{2} \log \det \left\{ G(Q^*) \right\} + \frac{R}{2} \text{tr} \left\{ L^* Q^* \right\} - RM \left( L^* \right) \right\} \right\} \tag{175}$$

$$\beta \mathcal{F}(\beta) = \lim_{t \downarrow 0} \frac{\partial}{\partial t} \left( R \log 2 + \frac{1}{2} \log \det \left\{ G(Q^*) \right\} + \frac{R}{2} \text{tr} \left\{ LQ^* \right\} - RM \left( L^* \right) \right) \tag{176}$$
We therefore have
\[
\log \det G(Q) = (t - 1) \log g(Q) \\
\quad + \log (g(Q) + tf(Q, m)).
\] (200)

By replacing the RS structures, we further have
\[
\text{tr}\{LQ\} = (t + 1) + t(t - 1)QL + 2tEm,
\] (201)
and finally using the Riesz identity, we have for \(t\) in the vicinity of \(t = 0\),
\[
\mathcal{M}(L) = (t + 1) \log 2 + \frac{t(1 - L) + 1}{2} \\
\quad + tE_Z \left\{ \log \cosh \left( \sqrt{LZ + E} \right) \right\}.
\] (202)

We now apply the replica continuity assumption and determine the limit in (203), shown at the bottom of the next page, which concludes the RS solution given in Proposition 1.

**APPENDIX C**

**DECOUPLING PRINCIPLE**

This appendix aims to give a proof for Proposition 2. To this end, we use the multivariate form of Carleman’s theorem which indicates the determinacy of the moment problem with uniformly bounded moments [93].

Let us start with the decoupled setting described in Proposition 2. For this setting, the \((\kappa, \tau)\) joint moment of the label \(S\) and recovery \(\hat{S}\) is given by
\[
M_{\kappa, \tau} = \mathbb{E} \left\{ S^\kappa \hat{S}^\tau \right\}
\] (204)
\[
= \mathbb{E} \left\{ S^\kappa \tanh^{\tau} \left( S + \rho Z \right) \right\}. \quad (205)
\]

Using the odd symmetry of the \(\tanh (\cdot)\), we can conclude that
\[
M_{\kappa, \tau} = \left(1 + (-1)^{\kappa+\tau}\right) \mathbb{E} \left\{ \tanh^\tau \left( \frac{1 + \rho Z}{\rho^2} \right) \right\}. \quad (206)
\]
Since \( M_{s,\tau} \) is bounded for all non-negative integers \( \kappa \) and \( \tau \), Carleman’s theorem indicate that the mapping from corresponding distribution, i.e., the joint distribution of \( S \) and \( \hat{S} \), to the joint moments is unique. This means that any other random pairs with same joint moments have the same distributions. Therefore, it is sufficient to show that for any \( k \in [K] \), the joint \((\kappa, \tau)\) moment asymptotically converges to \( M_{s,\tau} \).

Consider index \( k \in [K] \). The joint \((\kappa, \tau)\) moment of the label \( s_k \) and its corresponding recovery is given by

\[
\mathbb{E}\{s_k^\tau \hat{s}_k^\tau\} = \mathbb{E}\left\{ \sum_u p_S(u|Y, V) u_k^\tau \right\} \tag{207a}
\]

\[
= \mathbb{E}\left\{ s_k^\tau \prod_{a=1}^{\tau} u_{a,k} p_S(u_a|Y, V) \right\} \tag{207b}
\]

where \( p_S(u|Y, V) \) is the posterior distribution of the labels given by

\[
p_s(u|Y, V) = \frac{q_y(y|u, V)}{\sum_u q_y(y|u, V)}, \tag{208}
\]

with \( q_y(y|u, V) \) denoting the likelihood of \( u \) determined for the postulated noise variance \( \hat{\sigma}^2 \).

To determine the joint moment, we define the \textit{replicated} posterior distribution

\[
p_S^{(\tau)}(\{u_a\}|y, V) := \prod_{a=1}^{t} p_S(u_a|y, V), \tag{209}
\]

which is a Boltzmann distribution with Hamiltonian

\[
\mathcal{E}^{(\tau)}(\{u_a\}|y, V) := \sum_{a=1}^{\tau} \mathcal{E}(u_a|y, V), \tag{210}
\]

at inverse temperature \( \beta = 1/\hat{\sigma}^2 \). Using this definition, the joint moment is represented as

\[
\mathbb{E}\{s_k^\tau \hat{s}_k^\tau\} = \mathbb{E}\left\{ s_k^\tau \prod_{a=1}^{\tau} u_{a,k} \right\}, \tag{211}
\]

where \((s, \{u_a\})\) are distributed with \( 2^{-K} p_S^{(\tau)}(\{u_a\}|y, V) \) conditioned by \( y \) and \( V \).

Noting the fact that the joint distribution is a Boltzmann distribution, we follow the classical averaging trick in statistical mechanics. Namely, we define the function

\[
f(s, \{u_a\}) := \frac{1}{|I_k|} \sum_{j \in I_k} s_j^\tau \prod_{a=1}^{\tau} u_{a,j}, \tag{212}
\]

for some index set \( I_k \subset [K] \) which contains \( k \) and is of width \(|I_k| = 1 + \alpha K\) for some \( \alpha \leq 1 \). We then construct a modified

\[
\frac{\partial}{\partial \mathbb{E}} W_t(L, E) = t \frac{\mathbb{E}_Z \left\{ \cosh^{-1} \left( \sqrt{LZ + E} \right) \sinh \left( \sqrt{LZ + E} \right) \right\}}{\mathbb{E}_Z \left\{ \cosh^t \left( \sqrt{LZ + E} \right) \right\}}, \tag{195a}
\]

\[
\frac{\partial}{\partial L} W_t(L, E) = \frac{t}{2\sqrt{L}} \frac{\mathbb{E}_Z \left\{ \cosh^{-1} \left( \sqrt{LZ + E} \right) \sinh \left( \sqrt{LZ + E} \right) Z \right\}}{\mathbb{E}_Z \left\{ \cosh^t \left( \sqrt{LZ + E} \right) \right\}} \tag{195b}
\]

\[
\frac{\partial}{\partial \mathbb{E}} \log \mathcal{M}_e(L, E) = \frac{t(t - 1)}{2} \frac{\mathbb{E}_Z \left\{ \cosh^{t-2} \left( \sqrt{LZ + E} \right) \sinh^2 \left( \sqrt{LZ + E} \right) \right\}}{\mathbb{E}_Z \left\{ \cosh^t \left( \sqrt{LZ + E} \right) \right\}}, \tag{197a}
\]

\[
\frac{\partial}{\partial L} \log \mathcal{M}_e(L, E) = \frac{\mathbb{E}_Z \left\{ \cosh^{t-1} \left( \sqrt{LZ + E} \right) \sinh \left( \sqrt{LZ + E} \right) \right\}}{\mathbb{E}_Z \left\{ \cosh^t \left( \sqrt{LZ + E} \right) \right\}} \tag{197b}
\]

\[
Q = \frac{\mathbb{E}_Z \left\{ \cosh^{t-2} \left( \sqrt{LZ + E} \right) \sinh^2 \left( \sqrt{LZ + E} \right) \right\}}{\mathbb{E}_Z \left\{ \cosh^t \left( \sqrt{LZ + E} \right) \right\}}, \tag{198a}
\]

\[
m = \frac{\mathbb{E}_Z \left\{ \cosh^{t-1} \left( \sqrt{LZ + E} \right) \sinh \left( \sqrt{LZ + E} \right) \right\}}{\mathbb{E}_Z \left\{ \cosh^t \left( \sqrt{LZ + E} \right) \right\}} \tag{198b}
\]

\[
\beta \mathcal{F}(\beta) = \lim_{t \uparrow 0} \frac{\partial}{\partial t} \left( R \log \frac{1}{2} \log \det \{G(Q)\} + \frac{R}{2} \text{tr} \{LQ\} - R \mathcal{M}(L) \right) \tag{203}
\]
partition function as

$$\mathcal{Z}_{\beta, h} (\tau) = \mathbb{E} \left\{ \sum_{\{ u_n \}} \exp \left\{ -\beta \mathcal{G}^{(\tau)} (u_n | y, V) + h f (s, \{ u_n \}) \right\} \right\},$$

(213)

where the expectation is taken with respect to the Boltzmann distribution in (209). It is then readily shown that at inverse temperature $\beta$, we have

$$\mathbb{E} \{ f (s, \{ u_n \}) \} = \lim_{h \downarrow 0} \frac{\partial}{\partial h} \log \mathcal{Z}_{\beta, h} (\tau).$$

(214)

We now note that the asymptotic joint moment is given at the limit $K \uparrow \infty$ and $\alpha \downarrow 0$. We hence conclude that

$$\mathbb{E} \{ s_k^\tau s_k^\tau \} = \lim_{\alpha \downarrow 0} \lim_{K \uparrow \infty} \frac{\partial}{\partial h} \log \mathcal{Z}_{\beta, h} (\tau),$$

(215)

when we set $\beta = 1/\hat{\sigma}^2$.

The modified partition function is closely similar to the replicated partition function $\mathcal{Z}_{\beta} (t)$ determined in Appendix B. We hence follow the same lines of derivations to conclude that

$$\mathcal{Z}_{\beta, h} (\tau) = 2^{-K} \sum_{\mathbf{U} \in \{ \pm 1 \}^{K \times (\tau + 1)}} \exp \left\{ h f (\mathbf{U}) \right\} \det \{ \mathbf{G} ((\mathbf{U}; \mathbf{U})) \}^{-N/2},$$

(216)

for $\mathbf{G} (\cdot)$ defined in (143), matrix $\mathbf{U} = [u_1, \ldots, u_\tau, u_{\tau+1}]$, and

$$f (\mathbf{U}) := f (u_{\tau+1}, \{ u_n \}).$$

(217)

The current form is a simple modification of the derivation in (145). We hence follow the same technique and exchange the summation variable with respect to the sub-shells defined in (146). Using the Laplace representation of the Dirac impulse, we finally conclude (218), shown at the bottom of the page, for $v \in \{ \pm 1 \}^{\tau+1}$, some integral measures $dQ$ and $dL$ and

$$\mathcal{M}_h (\mathbf{L}) = \mathbb{E}_I \left\{ \log \sum_v \exp \left\{ h I f_0 (v) + \frac{1}{2} v^T \mathbf{L} v \right\} \right\},$$

(219)

where function $f_0 (v)$ is defined as

$$f_0 (v) = v_{\tau+1}^\alpha \prod_{a=1}^\tau v_a,$$

(220)

and $I \in \{ 0, 1 \}$ is a binary random variable with

$$\Pr \{ I = 1 \} = 1 - \Pr \{ I = 0 \} = \frac{|I_k|}{K}.$$  

(221)

By standard derivation, we the identity in (222), shown at the bottom of the page, where $F (\mathbf{Q}, \mathbf{L})$ is defined as in (164), and

$$q^{(\tau)}_V (v | \mathbf{L}^*) = \frac{\exp \left\{ \frac{1}{2} v^T \mathbf{L} v \right\}}{\sum_v \exp \left\{ \frac{1}{2} v^T \mathbf{L} v \right\}}.$$  

(223)

By standard deviations arguments, we can further determine the asymptotic limit as

$$\lim_{K \uparrow \infty} \lim_{h \downarrow 0} \frac{\partial}{\partial h} \log \mathcal{Z}_{\beta, h} (\tau) = \int \mathcal{D} \mathbf{Q} \mathcal{D} \mathbf{L} \exp \left\{ -NF (\mathbf{Q}, \mathbf{L}) \right\} \sum_v f_0 (v) q^{(\tau)}_V (v | \mathbf{L}^*),$$

(224)

where $(\mathbf{Q}^*, \mathbf{L}^*)$ is the saddle point of $F (\mathbf{Q}, \mathbf{L})$. As the right-hand side does not depend on $\alpha$, we finally have

$$\mathbb{E} \{ s_k^\tau s_k^\tau \} = \sum_v f_0 (v) q^{(\tau)}_V (v | \mathbf{L}^*),$$

(225)

for $\beta = 1/\hat{\sigma}^2$.

Note that this is the same saddle point as the one derived for the replica solution in Appendix B.

---

$$\mathcal{Z}_{\beta, h} (\tau) = 2^{-K} \int dQ dL \exp \left\{ -N \left( \frac{1}{2} \log \det \{ \mathbf{G} (\mathbf{Q}) \} + \frac{R}{2} \text{tr} (LQ) - RM_h (\mathbf{L}) \right) \right\}$$

(218)

$$\lim_{h \downarrow 0} \frac{\partial}{\partial h} \log \mathcal{Z}_{\beta, h} (\tau) = \int dQ dL \frac{\exp \left\{ -NF (\mathbf{Q}, \mathbf{L}) \right\}}{\int dQ dL \exp \left\{ -NF (\mathbf{Q}, \mathbf{L}) \right\}} \sum_v f_0 (v) q^{(\tau)}_V (v)$$

(222)

$$\mathbb{E} \{ s_k^\tau s_k^\tau \} = \lim_{t \downarrow 0} \left\{ \frac{\int Dz \cosh^{t-\tau} \left( \frac{\rho z + S}{\hat{\rho}^2} \right) \sinh^{t} \left( \frac{\rho z + S}{\hat{\rho}^2} \right)}{\int Dz \cosh^{t} \left( \frac{\rho z + S}{\hat{\rho}^2} \right)} \right\}$$

(227)

$$\mathbb{E} \{ s_k^\tau s_k^\tau \} = (-1)^{\alpha + t} \lim_{t \downarrow 0} \frac{\int Dz \cosh^{t-\tau} \left( \frac{\rho z + 1}{\hat{\rho}^2} \right) \sinh^{t} \left( \frac{\rho z + 1}{\hat{\rho}^2} \right)}{\int Dz \cosh^{t} \left( \frac{\rho z + 1}{\hat{\rho}^2} \right)}$$

(228a)

$$= (-1)^{\alpha + t} \int Dz \tanh^{t} \left( \frac{\rho z + 1}{\hat{\rho}^2} \right) = M_{\kappa, t}$$

(228b)
Comparing (223) and (172), it is observed that the joint moment is given by the same replicated distribution. This is a proof for the generic form of the decoupling principle; see [2] and [57]. We however need to show that this decoupled form reduces under RS to the same form given by Proposition 2. We hence write the derived joint distribution in terms of the replicas which is continued to the real axis via the replica continuity assumption.

We however need to show that this decoupled form reduces under RS to the same form given by Proposition 2. We hence write the derived joint distribution in terms of the replicas, i.e., the one given in (172), and use the replica continuity assumption to conclude that
\[
\mathbb{E} \{ s_k^L \hat{s}_k^L \} = \lim_{t \to 0} \sum_{\mathbf{v}} f_0(\mathbf{v}) q_{\hat{V}}(\mathbf{v}, \mathbf{v} | \mathbf{L}^*),
\]
where \( \mathbf{v} = [v_{t+2}, \ldots, v_{t+1}] \). Here, \( t \) denotes the number of replicas which is continued to the real axis via the replica continuity assumption.\(^{41}\) We now set the replica correlation matrix to be RS and follow the same lines of derivations as those taken in (191) to show the identity (227), shown at the bottom of the previous page, for some uniform random variable \( S \) on \( \{ \pm 1 \} \). Using symmetry of the hyperbolic functions, we can further simplify the expression as in (228), shown at the bottom of the previous page. The proof is then concluded by applying Carleman’s theorem.

ACKNOWLEDGMENT

The author Ali Bereyhi had the honor to discuss the results of this work and receive helpful feedback from Yan Fyodorov throughout this study. He is also grateful to Yan Fyodorov for the helpful discussions and useful comments on the work and further thank Lenka Zdeborová, Nicolas Macris, Saba Asaad, and Rafael Schaefer for their comments on the work.

REFERENCES

[1] A. Bereyhi, B. Loureiro, F. Krzakala, R. R. Müller, and H. Schulz-Baldes, “Secure coding via Gaussian random fields,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2022, pp. 1193–1198.
[2] A. Bereyhi, “Statistical mechanics of regularized least squares,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Friedrich-Alexander Univ., Erlangen, Germany, Jun. 2020.
[3] A. Bereyhi, R. R. Müller, and H. Schulz-Baldes, “Statistical mechanics of MAP estimation: General replica ansatz,” IEEE Trans. Inf. Theory, vol. 65, no. 12, pp. 7896–7934, Dec. 2019.
[4] A. Bereyhi and R. R. Müller, “Maximum-a-posteriori signal recovery with prior information: Applications to compressive sensing,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Calgary, AB, Canada, Apr. 2018, pp. 4494–4498.
[5] J. Barbier, M. Dia, N. Macris, and F. Krzakala, “The mutual information in random linear estimation,” in Proc. 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton), Monticello, IL, USA, Sep. 2016, pp. 625–632.
[6] J. Barbier, N. Macris, A. Maillard, and F. Krzakala, “The mutual information in random linear estimation beyond i.i.d. matrices,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Vail, CO, USA, Jun. 2018, pp. 1390–1394.
[7] J. Barbier, N. Macris, M. Dia, and F. Krzakala, “Mutual information and optimality of approximate message-passing in random linear estimation,” IEEE Trans. Inf. Theory, vol. 66, no. 7, pp. 4270–4303, Jul. 2020.
[8] J. Barbier, W.-K. Chen, D. Panchenko, and M. Sáenz, “Performance of Bayesian linear regression in a model with mismatch,” 2021, arXiv:2107.06936.
[9] E. W. Tramel, S. Kumar, A. Giurgiu, and A. Montanari, “Statistical physics of Spin Glasses and Information Processing: An Introduction,” 2011, Oxford, U.K.: Clarendon.
[10] T. Tanaka, “A statistical-mechanics approach to large-system analysis of CDMA multiluser detectors,” IEEE Trans. Inf. Theory, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

\(^{41}\)Remember that unlike \( t, \tau \) is a fixed integer.
[39] S. Rangan, A. K. Fletcher, and V. K. Goyal, “Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing,” IEEE Trans. Inf. Theory, vol. 58, no. 3, pp. 1902–1923, Mar. 2012.

[40] B. Zaidel, R. R. Müller, A. L. Moustakas, and R. de Miguel, “Vector precoding for Gaussian MIMO broadcast channels: Impact of replica symmetry breaking,” IEEE Trans. Inf. Theory, vol. 58, no. 3, pp. 1413–1440, Mar. 2012.

[41] L. Zdeborová and F. Krzakala, “Statistical physics of inference: Thresholds and algorithms,” Adv. Probab., vol. 65, no. 5, pp. 453–552, Sep. 2016.

[42] T. Richardson and R. Urbanke, Modern Coding Theory, Cambridge, U.K.: Cambridge Univ. Press, 2008.

[43] S. Kudekar and N. Macris, “Proof of replica formulas in the high noise regime for communication using LDGM codes,” in Proc. IEEE Inf. Theory Workshop, May 2008, pp. 416–420.

[44] S. Kumar, A. J. Young, N. Macris, and H. D. Pister, “Threshold saturation for spatially coupled LDPC and LDGM codes on BMS channels,” IEEE Trans. Inf. Theory, vol. 60, no. 12, pp. 7389–7415, Dec. 2014.

[45] H.-B. Chen, J.-C. Mourrat, and J. Xia, “Statistical inference of finite-rank tensors,” 2021, arXiv:2104.05360.

[46] C. Luneau, J. Barbier, and N. Macris, “Mutual information for low-rank even-order symmetric tensor estimator,” 2019, arXiv:1904.04565.

[47] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, “Optimal errors and phase transitions in high-dimensional generalized linear models,” Proc. Nat. Acad. Sci. USA, vol. 116, no. 12, pp. 5451–5460, Mar. 2019.

[48] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, “Learning Gaussian mixtures with generalised linear models: Precise asymptotics in high-dimensions,” 2021, arXiv:2106.03791.

[49] M. Sellke, “Optimizing mean field spin glasses with external field,” 2021, arXiv:2105.03506.

[50] A. El Alaoui, A. Montanari, and M. Sellke, “Optimization of mean-field spin glasses,” Ann. Probab., vol. 49, no. 6, pp. 2922–2960, Nov. 2021.

[51] T. R. Kirkpatrick and D. Thirumalai, “Dynamics of the structural glass transition and the p-spin—Interaction spin-glass model,” Phys. Rev. Lett., vol. 58, no. 20, pp. 2091–2094, May 1987.

[52] E. Gardner, “Spin glasses with p spin interactions,” Nucl. Phys. B, vol. 257, pp. 747–765, Jan. 1985.

[53] B. Derriça, “Random-energy model: Limit of a family of disordered models,” Phys. Rev. Lett., vol. 45, no. 2, pp. 79–82, Jul. 1980.

[54] L. De Sanctis and S. Franz, “Self-averaging identities for random spin systems,” in Spin Glasses: Statics and Dynamics, Paris, France: Springer, 2009, pp. 123–142.

[55] F. Krzakala and L. Zdeborová. (2021). Statistical Physics Methods in Optimization and Machine Learning. [Online]. Available: https://spinphysteam.github.io/EPPF/DoctoralLecture2021/Notes.pdf

[56] C. E. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J., vol. 27, no. 3, pp. 379–423, Jul. 1948.

[57] A. Bereyhi, R. Müller, and H. Schulz-Baldes, “RSB decoupling property and statistical-physics-based reconstruction in compressed sensing,” Proc. IEEE Inf. Theory Workshop, vol. 61, no. 12, pp. 6635–6642, Dec. 2015.

[58] S. Verdú and S. S. Shamai, “Spectral efficiency of CDMA with random spreading,” IEEE Trans. Inf. Theory, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[59] R. R. Müller, Power and Bandwidth Efficiency of Multisystem With Random Spreading. Dürren, Germany: Shaker, 1999.

[60] M. R. Müller, “Multiuser receivers for randomly spread signals: Fundamental limits with and without decision-feedback,” IEEE Trans. Inf. Theory, vol. 47, no. 1, pp. 268–283, Jan. 2001.

[61] R. R. Müller, “On approximation, bounding & exact calculation of average block error probability for random code ensembles,” IEEE Trans. Commun., vol. 69, no. 5, pp. 2987–2996, May 2021.

[62] S. J. Lade and T. Gross, “Early-warning signals for critical transitions: A generalized modeling approach,” PloS Comput. Biol., vol. 8, no. 2, Feb. 2012, Art. no. e1002360.

[63] F. Antenucci, F. Krzakala, P. Urbani, and L. Zdeborová, “Exact asymptotics for phase retrieval and compressed sensing with random generative priors,” in Proc. Math. Soc. Mech. Learn., 2020, pp. 55–73.

[64] M. Celentano, A. Montanari, and Y. Wu, “The estimation error of general first order methods,” in Proc. Conf. Learn. Theory (COLT), vol. 125, J. D. Abernethy and S. Agarwal, Eds., Graz, Austria, Jul. 2020, pp. 1078–1141.

[65] A. Javanmard and A. Montanari, “Statue evolution for general approximate message passing algorithms, with applications to spatial coupling,” Inf. Inference, J.IMA, vol. 2, no. 2, pp. 115–144, Dec. 2013.

[66] A. Javanmard, A. Montanari, and Y. Wu, “All-or-nothing phenomenon in sparse linear regression,” in Proc. Conf. Learn. Theory (COLT), 2019, pp. 2652–2663.

[67] S. J. Lade and T. Gross, “Early warning signals for critical transitions: Statistical and computational phase transitions,” in Proc. Conf. Learn. Theory (COLT), 2020, pp. 948–953.

[68] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová, “Statistical-physics-based reconstruction in compressed sensing,” Phys. Rev. Lett., vol. 2019, no. 2, May 2019, Art. no. 023401.

[69] M. H. Yassaee, M. R. Aref, and A. Gohari, “Achievability proof via output statistics of random binning,” IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7434–7464, Nov. 2013.

[70] D. J. MacKay, “A proof of the data compression theorem of Slepian and Wolf for ergodic sources (corres.),” IEEE Trans. Inf. Theory, vol. IT-21, no. 2, pp. 226–228, Mar. 1975.

[71] D. L. Donoho, A. Javanmard, and A. Montanari, “Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing,” IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7434–7464, Nov. 2013.

[72] A. Wyner, “Recent results in the Shannon theory,” IEEE Trans. Inf. Theory, vol. IT-20, no. 1, pp. 2–10, Jan. 1974.

[73] D. J. MacKay, “A proof of the data compression theorem of Slepian and Wolf for ergodic sources (corres.),” IEEE Trans. Inf. Theory, vol. IT-21, no. 2, pp. 226–228, Mar. 1975.

[74] A. Javanmard, A. Montanari, and Y. Wu, “All-or-nothing phenomenon in sparse linear regression,” in Proc. Conf. Learn. Theory (COLT), vol. 125, J. D. Abernethy and S. Agarwal, Eds., Graz, Austria, Jul. 2020, pp. 1078–1141.

[75] F. Antenucci, F. Krzakala, P. Urbani, and L. Zdeborová, “Approximate survey propagation for statistical inference,” J. Stat. Mech., Theory Exp., vol. 2019, no. 2, Feb. 2019, Art. no. 023401.

[76] R. R. Müller and W. H. Gerster, “On the capacity loss due to separation of detection and decoding,” IEEE Trans. Inf. Theory, vol. 50, no. 8, pp. 1769–1778, Aug. 2004.

[77] M. Vehkaperä, K. Takeuchi, R. R. Müller, and T. Tanaka, “Analysis of large MIMO DS-CDMA systems with imperfect CSI and spatial correlation,” in Proc. IEEE Int. Symp. Inf. Theory, Jan. 2010, pp. 2098–2102.

[78] M. A. Sedaghat, R. R. Müller, and F. Marvasti, “On optimum asymptotic multiuser efficiency of randomly spread CDMA,” IEEE Trans. Inf. Theory, vol. 61, no. 12, pp. 6635–6642, Dec. 2015.

[79] S. Verdú and S. S. Shamai, “Spectral efficiency of CDMA with random spreading,” IEEE Trans. Inf. Theory, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[80] R. R. Müller, Power and Bandwidth Efficiency of Multisystem With Random Spreading. Dürren, Germany: Shaker, 1999.
Ali Bereyhi (Member, IEEE) received the M.Sc. degree (Hons.) in electrical engineering and communications from the Sharif University of Technology, Iran, and the Ph.D. degree (Hons.) in electrical engineering and communications from Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg. From 2020 to 2022, he was with the Institute for Digital Communications, FAU Erlangen-Nürnberg, as a Post-Doctoral Researcher and a Lecturer. In 2023, he joined the Wireless Computing Laboratory (WCL), University of Toronto. His research interests include information theory, distributed machine learning, compressive sensing, statistical mechanical tools and random matrices, and signal processing techniques for wireless systems.

Bruno Loureiro received the B.Sc. degree in mathematics and physics from King’s College London, U.K., in 2013, and the M.A.St. degree in applied mathematics and the Ph.D. degree in physics from the University of Cambridge, Cambridge, U.K., in 2014 and 2018, respectively. After two post-doctoral research positions at the Institut de Physique Théorique in Saclay and the Information, Learning and Physics Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), he joined the Department of Informatics, École Normale Supérieure. His research interests include the crossroads between machine learning, information theory, and physics.

Florent Krzakala is currently a Full Professor at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He also leads the Information, Learning and Physics Laboratory (IdePHICS), Department of Physics and the Department of Engineering, EPFL. He is the Founder and a Scientific Advisor of the startup LightOn. His research interests include statistical physics, machine learning, statistics, signal processing, computer science, and computational optics.

Hermann Schulz-Baldes received the Ph.D. degree in theoretical physics from Università-Paul-Sabatier Toulouse III. He is a Mathematical Physicist with the Department of Mathematics, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, where he has been a Professor since 2004. Many of his scientific contributions concern the analysis of models in solid-state physics, mainly using tools from index theory, probability theory, and scattering theory. He also has several contributions to random matrix theory and free probability.