Statistical Physics of Signal Estimation in Gaussian Noise: Theory and Examples of Phase Transitions

Neri Merhav, Dongning Guo, and Shlomo Shamai (Shitz)

Abstract

We consider the problem of signal estimation (denoising) from a statistical mechanical perspective, using a relationship between the minimum mean square error (MMSE), of estimating a signal, and the mutual information between this signal and its noisy version. The paper consists of essentially two parts. In the first, we derive several statistical–mechanical relationships between a few important quantities in this problem area, such as the MMSE, the differential entropy, the Fisher information, the free energy, and a generalized notion of temperature. We also draw analogies and differences between certain relations pertaining to the estimation problem and the parallel relations in thermodynamics and statistical physics. In the second part of the paper, we provide several application examples, where we demonstrate how certain analysis tools that are customary in statistical physics, prove useful in the analysis of the MMSE. In most of these examples, the corresponding statistical–mechanical systems turn out to consist of strong interactions that cause phase transitions, which in turn are reflected as irregularities and discontinuities (similar to threshold effects) in the behavior of the MMSE.

Index Terms

Gaussian channel, denoising, de Bruijn's identity, MMSE estimation, phase transitions, random energy model, spin glasses, statistical mechanics.

The work of D. Guo is supported by the NSF under grant CCF-0644344 and DARPA under grant W911NF-07-1-0028 . The work of S. Shamai is supported in part by the Israel Science Foundation.

N. Merhav is with the Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel. E–mail: merhav@ee.technion.ac.il

D. Guo is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, U.S.A. E–mail: dguo@northwestern.edu

S. Shamai is with the Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel. E–mail: sshlomo@ee.technion.ac.il

I. INTRODUCTION

The relationships and the interplay between Information Theory and Statistical Physics have been recognized and exploited for several decades by now. The roots of these relationships date back to the celebrated papers by Jaynes from the late fifties of the previous century [15], [16], but their aspects and scope have been vastly expanded and deepened ever since. Much of the research activity in this interdisciplinary problem area revolves around the identification of 'mappings' between problems in Information Theory and certain many–particle systems in Statistical Physics, which are analogous at least as far as their mathematical formalisms go. One important example is the paralellism and analogy between random code ensembles in Information Theory and certain models of disordered magnetic materials, known as *spin glasses*. This analogy was first identified by Sourlas (see, e.g., [27], [28]) and has been further studied in the last two decades to a great extent. Beyond the fact that these paralellisms and analogies are academically interesting in their own right, they also prove useful and beneficial. Their utility stems from the fact that physical insights, as well as statistical mechanical tools and analysis techniques can be harnessed in order to advance the knowledge and the understanding with regard to the information–theoretic problem under discussion.

In this context, our work takes place at the meeting point of Information Theory, Statistical Physics, and yet another area – Estimation Theory, where the bridge between information–theoretic and the estimation–theoretic ingredients of the topic under discussion is established by an identity [12, Theorem 2], equivalent to the de Bruijn identity (cf. e.g., [3, Theorem 17.7.2]), which relates the minimum mean square error (MMSE), of estimating a signal in additive white Gaussian noise (AWGN), to the mutual information between this signal and its noisy version. We henceforth refer to this relation as the *I–MMSE relation*. It should be pointed out that the present work is not the first to deal with the interplay between the I–MMSE relation and statistical mechanics. In an earlier paper by Shental and Kanter [26], the main theme was an attempt to provide an alternative proof of the I–MMSE relation, which is rooted in thermodynamics and statistical physics. However, to this end, the authors of [26] had to generalize the theory of thermodynamics.

Our study is greatly triggered by [26] (in its earlier versions), but it takes a substantially different route. Rather than proving the I–MMSE relation, we simply use it in conjunction with analysis techniques used in statistical physics. The basic idea that is underlying our work is that when the channel input signal is rather complicated (but yet, not too complicated), which is the case in certain applications, the mutual information with its noisy version can be evaluated using statistical–mechanical analysis techniques, and

then related to the MMSE using the I–MMSE relation. This combination proves rather powerful, because it enables one to distinguish between situations where irregular (i.e., non–smooth or even discontinuous) behavior of the mean square error (as a function of the signal–to–noise ratio) is due to artifacts of a certain ad–hoc signal estimator, and situations where these irregularities are inherent in the model, in the sense that they are apparent even in optimum estimation. In the latter situations, these irregularities (or threshold effects) are intimately related to *phase transitions* in the parallel statistical–mechanical systems.

These motivations set the stage for our study of the relationships between the MMSE and statistical mechanics, first of all, in the general level, and then in certain concrete applications. Accordingly, the paper consists of two main parts. In the first, which is a general theoretical study, we derive several statistical–mechanical relationships between a few important quantities such as the MMSE, the differential entropy, the Fisher information, the free energy, and a generalized notion of temperature. We also draw analogies and differences between certain relations pertaining to the estimation problem and the parallel relations in thermodynamics and statistical physics. In the second part of the paper, we provide several application examples, where we demonstrate how certain analysis tools that are customary in statistical physics (in conjunction with large deviations theory) prove useful in the analysis of the MMSE. In light of the motivations described in the previous paragraph, in most of these examples, the corresponding statistical–mechanical systems turn out to consist of strong interactions that cause phase transitions, which in turn are reflected as irregularities and discontinuities in the behavior of the MMSE.

The remaining part of this paper is organized as follows: In Section II, we establish a few notation conventions and we formalize the setting under discussion. In Section III, we provide the basic background in statistical physics that will be used in the sequel. Section IV is devoted to the general theoretical study, and finally, Section V includes application examples, where the MMSE will be analyzed using statistical–mechanical tools.

II. Notation Conventions, Formalization and Preliminaries

A. Notation Conventions

Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, like X and Y, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with the same symbols in the boldface font. Thus, for example, $\boldsymbol{X}$ will denote a random n-vector $(X_1, \ldots, X_n)$, and $\boldsymbol{x} = (x_1, ..., x_n)$ is a specific vector value in $\mathcal{X}^n$, the n-th Cartesian power of $\mathcal{X}$.

Sources and channels will be denoted generically by the letters P and Q. The expectation operator will be denoted by $\boldsymbol{E}\{\cdot\}$. When the underlying probability measure is indexed by a parameter, say, β, then it will used as a subscript of P, p and $\boldsymbol{E}$, unless there is no ambiguity.

For two positive sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n \dot{=} b_n$ means that a_n and b_n are asymptotically of the same exponential order, that is, $\lim_{n\to\infty} \frac{1}{n} \ln \frac{a_n}{b_n} = 0$. Similarly, $a_n \dot{\leq} b_n$ means that $\limsup_{n\to\infty} \frac{1}{n} \ln \frac{a_n}{b_n} \leq 0$, etc. Information theoretic quantities like entropies and mutual informations will be denoted following the usual conventions of the Information Theory literature.

B. Formalization and Preliminaries

We consider the simplest variant of the signal estimation problem setting studied in [12], with a few slight modifications in notation. Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a pair of random vectors in $\mathbb{R}^n$, related by the Gaussian channel

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{N}$ is a random vector (noise), whose components are i.i.d., zero–mean, Gaussian random variables (RV's) whose variance is $1/\beta$, where β is a given positive constant designating the signal–to–noise ratio (SNR), or the inverse temperature in statistical–mechanical point of view (cf. Section III). It is assumed that $\boldsymbol{X}$ and $\boldsymbol{N}$ are independent. Upon receiving $\boldsymbol{Y}$, one is interested in inferring about the (desired) random vector $\boldsymbol{X}$. As is well known, the best estimator of $\boldsymbol{X}$ given the observation vector $\boldsymbol{Y}$, in the mean square error (MSE) sense, i.e., the MMSE estimator, is the conditional mean $\hat{\boldsymbol{X}} = \boldsymbol{E}(\boldsymbol{X}|\boldsymbol{Y})$ and the corresponding MMSE, $\boldsymbol{E}\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|^2$ will denoted by $\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})$. Theorem 2 in [12], which provides the I–MMSE relation, relates the MMSE to the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ (defined using the natural base logarithm) according to

$$\frac{\mathrm{d}I(\boldsymbol{X};\boldsymbol{Y})}{\mathrm{d}\beta} = \frac{\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{2}. \tag{2}$$

For example, if $n = 1$ and $X \sim \mathcal{N}(0,1)$, then $I(X;Y) = \frac{1}{2}\ln(1+\beta)$, which leads to $\mathrm{mmse}(X|Y) = 1/(1+\beta)$, in agreement with elementary results. The relationship has been used in [24] to compute the mutual information achieved by low-density parity-check (LDPC) codes over Gaussian channels through evaluation of the marginal estimation error.

A very important function, which will be pivotal to our derivation of both $\boldsymbol{E}(\boldsymbol{X}|\boldsymbol{Y})$ and $\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})$, as well as to the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$, is the posterior distribution. Denoting the probability mass

function of $\boldsymbol{x}$ by $Q(\boldsymbol{x})$ and the channel induced by (1) by $P(\boldsymbol{y}|\boldsymbol{x})$, then

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{Q(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})}{\sum_{\boldsymbol{x}'} Q(\boldsymbol{x}')P(\boldsymbol{y}|\boldsymbol{x}')} = \frac{Q(\boldsymbol{x})\exp[-\beta\cdot\|\boldsymbol{y}-\boldsymbol{x}\|^2/2]}{Z(\beta|\boldsymbol{y})}, \tag{3}$$

where we defined

$$Z(\beta|\boldsymbol{y}) \stackrel{\triangle}{=} \sum_{\boldsymbol{x}} Q(\boldsymbol{x})\exp[-\beta\cdot\|\boldsymbol{y}-\boldsymbol{x}\|^2/2] = (2\pi/\beta)^{n/2}P_\beta(\boldsymbol{y}) \tag{4}$$

where $P_\beta(\boldsymbol{y})$ is the channel output density. Here we have assumed that $\boldsymbol{x}$ is discrete, as otherwise Q should be replaced by the probability density function (pdf) and the summation over $\{\boldsymbol{x}'\}$ should be replaced by an integral. The function $Z(\beta|\boldsymbol{y})$ is very similar to the so-called *partition function*, which is well known to play a very central role in statistical mechanics, and will also play a central role in our analysis. In the next section, we then give some necessary background in statistical mechanics that will be essential to our study.

III. Physics Background

Consider a physical system with n particles, which can be in a variety of microscopic states ('microstates'), defined by combinations of physical quantities associated with these particles, e.g., positions, momenta, angular momenta, spins, etc., of all n particles. For each such microstate of the system, which we shall designate by a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, there is an associated energy, given by a *Hamiltonian* (energy function), $\mathcal{E}(\boldsymbol{x})$. For example, if $x_i = (\boldsymbol{p}_i, \boldsymbol{r}_i)$, where $\boldsymbol{p}_i$ is the momentum vector of particle number i and $\boldsymbol{r}_i$ is its position vector, then classically, $\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^N \left[\frac{\|\boldsymbol{p}_i\|^2}{2m} + mgz_i\right]$, where m is the mass of each particle, z_i is its height – one of the coordinates of $\boldsymbol{r}_i$, and g is the gravitation constant.

One of the most fundamental results in statistical physics (based on the law of energy conservation and the basic postulate that all microstates of the same energy level are equiprobable) is that when the system is in thermal equilibrium with its environment, the probability of finding the system in a microstate $\boldsymbol{x}$ is given by the *Boltzmann–Gibbs* distribution

$$P(\boldsymbol{x}) = \frac{e^{-\beta\mathcal{E}(\boldsymbol{x})}}{Z(\beta)} \tag{5}$$

where $\beta = 1/(kT)$, k being Boltmann's constant and T being temperature, and $Z(\beta)$ is the normalization constant, called the *partition function*, which is given by

$$Z(\beta) = \sum_{\boldsymbol{x}} e^{-\beta\mathcal{E}(\boldsymbol{x})},$$

assuming discrete states. In case of continuous state space, the partition function is defined as

$$Z(\beta) = \int \mathrm{d}\boldsymbol{x}\, e^{-\beta\mathcal{E}(\boldsymbol{x})},$$

and $P(\boldsymbol{x})$ is understood as a pdf. The role of the partition function is by far deeper than just being a normalization factor, as it is actually the key quantity from which many macroscopic physical quantities can be derived, for example, the free energy[1] is $F(\beta) = -\frac{1}{\beta}\ln Z(\beta)$, the average internal energy is given by $\bar{E} \stackrel{\triangle}{=} \boldsymbol{E}\{\mathcal{E}(\boldsymbol{X})\} = -(\mathrm{d}/\mathrm{d}\beta)\ln Z(\beta)$ with $\boldsymbol{X} \sim P(\boldsymbol{x})$, the heat capacity is obtained from the second derivative, etc. One of the ways to obtain eq. (5), is as the maximum entropy distribution under an average energy constraint (owing to the second law of thermodynamics), where β plays the role of a Lagrange multiplier that controls the average energy.

An important special case, which is very relevant both in physics and in the study of AWGN channel considered here, is the case where the Hamiltonian $\mathcal{E}(\boldsymbol{x})$ is additive and *quadratic* (or "harmonic" in the physics terminology), i.e., $\mathcal{E}(x) = \sum_{i=1}^{n}\frac{1}{2}\kappa x_i^2$, for some constant $\kappa > 0$, or even more generally, $\mathcal{E}(x) = \sum_{i=1}^{n}\frac{1}{2}\kappa_i x_i^2$, which means that the components $\{x_i\}$ are Gaussian and independent. A classical result in this case, known as the *equipartition theorem of energy*, which is very easy to show, asserts that each particle (or, more precisely, each degree of freedom) contributes an average energy of $\boldsymbol{E}\{\frac{1}{2}\kappa_i X_i^2\} = 1/(2\beta) = kT/2$ independently of κ (or κ_i).

Returning to the case of a general Hamiltonian, it is instructive to relate the Shannon entropy, pertaining to the Boltzmann–Gibbs distribution, to the quantities we have seen thus far. Specifically, the Shannon entropy $S(\beta) = -\boldsymbol{E}\{\ln P(\boldsymbol{X})\}$ associated with $P(\boldsymbol{x}) = e^{-\beta\mathcal{E}(\boldsymbol{x})}/Z(\beta)$, is given by

$$S(\beta) = \boldsymbol{E}\ln\left[\frac{Z(\beta)}{e^{-\beta\mathcal{E}(\boldsymbol{x})}}\right] = \ln Z(\beta) + \beta\cdot\bar{E},$$

where, as mentioned above,

$$\bar{E} = -\frac{\mathrm{d}\ln Z(\beta)}{\mathrm{d}\beta} \tag{6}$$

is the average internal energy. This suggests the differential equation

$$\dot{\psi}(\beta) - \frac{\psi(\beta)}{\beta} = \frac{S(\beta)}{\beta}, \tag{7}$$

where $\psi(\beta) = -\ln Z(\beta)$ and $\dot{\psi}$ means the derivative of ψ. Equivalently, eq. (7) can be rewritten as:

$$\beta\frac{\mathrm{d}}{\mathrm{d}\beta}\left[\frac{\psi(\beta)}{\beta}\right] = \frac{S(\beta)}{\beta}, \tag{8}$$

[1]The free energy means the maximum work that the system can carry out in any process of fixed temperature. The maximum is obtained when the process is reversible (slow, quasi–static changes in the system).

whose solution is easily found to be

$$\psi(\beta) = \beta E_0 - \beta \int_\beta^\infty \frac{d\hat{\beta} S(\hat{\beta})}{\hat{\beta}^2}, \tag{9}$$

where $E_0 = \min_{\boldsymbol{x}} \mathcal{E}(\boldsymbol{x})$ is the *ground–state energy*, here obtained as a constant of integration by examining the limit of $\beta \to \infty$. Thus, we see that the log–partition function at a given temperature can be expressed as a *heat integral* of the entropy, namely, as an integral of a function that consists of the entropy at all lower temperatures. This is different from the other relations we mentioned thus far, which were all 'pointwise' in the temperature domain, in the sense that all quantities were pertaining to the same temperature. Taking the derivative of $\psi(\beta)$ according to eq. (9), we obtain the average internal energy:

$$\bar{E} = \dot{\psi}(\beta) = E_0 - \int_\beta^\infty \frac{d\hat{\beta} S(\hat{\beta})}{\hat{\beta}^2} + \frac{S(\beta)}{\beta}, \tag{10}$$

where the first two terms form the free energy.[2]

As a final remark, we should note that although the expression $Z(\beta|\boldsymbol{y})$ of eq. (4) is similar to that of $Z(\beta)$ defined in this section (for a quadratic Hamiltonian), there is nevertheless a small difference: The exponentials in (4) are weighted by probabilities $\{Q(\boldsymbol{x})\}$, which are independent of β. However, as explained in [17, p. 3713], this is not an essential difference because these weights can be interpreted as degeneracy of states, that is, as multiple states (whose number is proportional to $Q(\boldsymbol{x})$) of the same energy.

IV. Theoretical Derivations

Consider the Gaussian channel (1) and the corresponding posterior (3). Denoting by $\boldsymbol{E}_\beta$ the expectation operator w.r.t. joint pdf of $(\boldsymbol{X}, \boldsymbol{Y})$ induced by β, we have:

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= \boldsymbol{E}_\beta \left\{ \ln \frac{\exp[-\beta \cdot \|\boldsymbol{Y}-\boldsymbol{X}\|^2/2]}{Z(\beta|\boldsymbol{Y})} \right\} \\ &= -\frac{\beta}{2} \boldsymbol{E}_\beta \left\{ \|\boldsymbol{Y}-\boldsymbol{X}\|^2 \right\} - \boldsymbol{E}_\beta \left\{ \ln Z(\beta|\boldsymbol{Y}) \right\} \\ &= -\frac{n}{2} - \boldsymbol{E}_\beta \left\{ \ln Z(\beta|\boldsymbol{Y}) \right\} \end{aligned} \tag{11}$$

[2]By changing the integration variable from β to T, this is identified with the relation $F = E_0 - \int_0^T S \mathrm{d}T'$, which together with $F = \bar{E} - ST$, complies with the relation $\bar{E} = E_0 + \int_0^S T \mathrm{d}S' = E_0 + \int_0^Q \mathrm{d}Q'$, accounting for the simple fact that in the absence of any external work applied to the system, the internal energy is simply the heat accumulated as temperature is raised from 0 to T.

where we use the fact that $\boldsymbol{E}_\beta\left\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2\right\} = \boldsymbol{E}_\beta\left\{\|\boldsymbol{N}\|^2\right\} = n/\beta$. Taking derivatives w.r.t. β, and using the I–MMSE relation, we then have:

$$\frac{\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{2} = \frac{\partial I(\boldsymbol{X};\boldsymbol{Y})}{\partial\beta} = -\frac{\partial}{\partial\beta}\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\}. \tag{12}$$

and so, we obtain a very simple relation between the MMSE and the partition function of the posterior:

$$\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = -2\frac{\partial}{\partial\beta}\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\} \tag{13}$$

By calculating the derivative of the right-hand side (r.h.s.) more explicitly, one further obtains the following:

$$\begin{aligned}-\frac{\partial}{\partial\beta}\boldsymbol{E}_\beta \ln Z(\beta|\boldsymbol{Y}) &= -\frac{\partial}{\partial\beta}\int_{\mathbb{R}^n} d\boldsymbol{y}\cdot P_\beta(\boldsymbol{y})\ln Z(\beta|\boldsymbol{y}) \\ &= -\int_{\mathbb{R}^n} d\boldsymbol{y}\cdot P_\beta(\boldsymbol{y})\frac{\partial \ln Z(\beta|\boldsymbol{y})}{\partial\beta} - \int_{\mathbb{R}^n} d\boldsymbol{y}\cdot\frac{\partial P_\beta(\boldsymbol{y})}{\partial\beta}\cdot\ln Z(\beta|\boldsymbol{y}).\end{aligned} \tag{14}$$

Now, the first term at the right–most side of (14) can easily be computed by using the fact that $\ln Z(\beta|\boldsymbol{y})$ is a log–moment generating function of the energy (as is customarily done in statistical mechanics, cf. eq. (6)), which implies that it is given by $\boldsymbol{E}_\beta\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2\} = n/(2\beta) = nkT/2$, just like in the energy equipartition theorem for quadratic Hamiltonians. As for the second term, we have

$$\begin{aligned}&\int_{\mathbb{R}^n} d\boldsymbol{y}\cdot\frac{\partial P_\beta(\boldsymbol{y})}{\partial\beta}\cdot\ln Z(\beta|\boldsymbol{y}) \\ &= \int_{\mathbb{R}^n} d\boldsymbol{y}\cdot P_\beta(\boldsymbol{y})\cdot\frac{\partial\ln P_\beta(\boldsymbol{y})}{\partial\beta}\cdot\ln Z(\beta|\boldsymbol{y}) \\ &= \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{y}\cdot\left(\frac{2\pi}{\beta}\right)^{-n/2}\sum_{\boldsymbol{x}} Q(\boldsymbol{x})\left[\frac{n}{2\beta}-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\right]\cdot\exp\{-\beta\|\boldsymbol{y}-\boldsymbol{x}\|^2/2\}\ln Z(\beta|\boldsymbol{y}) \\ &= -\frac{1}{2}\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2,\ln Z(\beta|\boldsymbol{Y})\}.\end{aligned} \tag{15}$$

The MMSE is then given by

$$\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = -2\frac{\partial}{\partial\beta}\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\} = \frac{n}{\beta} + \mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2,\ln Z(\beta|\boldsymbol{Y})\}, \tag{16}$$

which can then be viewed as a variant of the energy equipartition theorem with a correction term that stems from the fact the pdf of $\boldsymbol{Y}$ depends on β.

Another look, from an estimation–theoretic point of view, at this expression reveals the following: The first term, $n/\beta = \boldsymbol{E}\|\boldsymbol{Y}-\boldsymbol{X}\|^2$, is the amount of noise in the raw data $\boldsymbol{Y}$, without any processing. The second term, which is always negative, designates then the *noise suppression level* due to MMSE estimation relative to the raw data. The intuition behind the covariance term is that when the 'correct' $\boldsymbol{x}$ (the one that actually feeds the Gaussian channel) dominates the partition function then $\ln Z(\beta|\boldsymbol{Y}) \approx$

$-\beta\|\boldsymbol{Y}-\boldsymbol{X}\|^2/2$, and so, there is a very strong negative correlation between $\|\boldsymbol{Y}-\boldsymbol{X}\|^2$ and $\ln Z(\beta|\boldsymbol{Y})$. In particular,

$$\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, -\beta\|\boldsymbol{Y}-\boldsymbol{X}\|^2/2\} = -\frac{n}{\beta}, \tag{17}$$

which exactly cancels the above–mentioned first term, n/β, and so, the overall MMSE essentially vanishes. When the correct $\boldsymbol{x}$ is not dominant, this correlation is weaker. Also, note that since

$$\boldsymbol{E}\|\boldsymbol{Y}-\boldsymbol{X}\|^2 = \mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y}) + \boldsymbol{E}\|\boldsymbol{Y}-\boldsymbol{E}(\boldsymbol{X}|\boldsymbol{Y})\|^2, \tag{18}$$

then this implies that

$$\boldsymbol{E}\|\boldsymbol{Y}-\boldsymbol{E}(\boldsymbol{X}|\boldsymbol{Y})\|^2 = -\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \ln Z(\beta|\boldsymbol{Y})\}. \tag{19}$$

It is now interesting to relate the noise suppression level

$$\Delta \stackrel{\triangle}{=} \boldsymbol{E}\|\boldsymbol{Y}-\boldsymbol{E}(\boldsymbol{X}|\boldsymbol{Y})\|^2 = -\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \ln Z(\beta|\boldsymbol{Y})\}$$

to the Fisher information matrix and then to a new generalized notion of temperature due to Narayanan and Srinivasa [21] via the de Bruijn identity. According to de Bruijn's identity, if $\boldsymbol{W}$ is a vector of i.i.d. standard normal components, independent of $\boldsymbol{X}$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}h(\boldsymbol{X}+\sqrt{t}\boldsymbol{W}) = \frac{1}{2}\mathrm{tr}\{J(\boldsymbol{X}+\sqrt{t}\boldsymbol{W})\}$$

where $h(\boldsymbol{Y})$ is differential entropy and $J(\boldsymbol{Y})$ is the Fisher information matrix associated with $\boldsymbol{Y}$ w.r.t. a translation parameter, namely,

$$\mathrm{tr}\{J(\boldsymbol{Y})\} = \sum_{i=1}^{n}\boldsymbol{E}\left\{\left[\frac{\partial \ln P_\beta(\boldsymbol{y})}{\partial y_i}\bigg|_{\boldsymbol{y}=\boldsymbol{Y}}\right]^2\right\} = \sum_{i=1}^{n}\int_{\mathbb{R}^n}\frac{\mathrm{d}\boldsymbol{y}}{P_\beta(\boldsymbol{y})}\left[\frac{\partial P_\beta(\boldsymbol{y})}{\partial y_i}\right]^2.$$

Note that since $P_\beta(\boldsymbol{y})$ and $Z(\beta|\boldsymbol{y})$ differ only by a multiplicative factor of $(\beta/2\pi)^{n/2}$, it is obvious that $\partial \ln P_\beta(\boldsymbol{y})/\partial y_i = \partial \ln Z(\beta|\boldsymbol{y})/\partial y_i$ and so, the Fisher information can also be related directly to the free energy by

$$\begin{aligned}\mathrm{tr}\{J(\boldsymbol{Y})\} &= \sum_{i=1}^{n}\boldsymbol{E}\left\{\left[\frac{\partial \ln Z(\beta|\boldsymbol{y})}{\partial y_i}\bigg|_{\boldsymbol{y}=\boldsymbol{Y}}\right]^2\right\}\\ &= \sum_{i=1}^{n}\boldsymbol{E}\{[\boldsymbol{E}\{-\beta(Y_i-X_i)|\boldsymbol{Y}\}]^2\}\\ &= \beta^2\sum_{i=1}^{n}\boldsymbol{E}\{\boldsymbol{E}^2(N_i|\boldsymbol{Y})\},\end{aligned} \tag{20}$$

where $N_i = Y_i - X_i$ and where we have used the fact that the derivative of $\exp\{-\beta\|\boldsymbol{y}-\boldsymbol{x}\|^2\}$ w.r.t. y_i is given by $-\beta(y_i - x_i)\cdot\exp\{-\beta\|\boldsymbol{y}-\boldsymbol{x}\|^2\}$. Now, as is also shown in [12]:

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{X}+\boldsymbol{N}) &= I(\boldsymbol{X};\boldsymbol{X}+\boldsymbol{W}/\sqrt{\beta}) \\ &= h(\boldsymbol{X}+\boldsymbol{W}/\sqrt{\beta}) - h(\boldsymbol{W}/\sqrt{\beta}) \\ &= h(\boldsymbol{X}+\boldsymbol{W}/\sqrt{\beta}) - \frac{n}{2}\ln\left(2\pi e/\beta\right). \end{aligned} \tag{21}$$

Thus,

$$\begin{aligned} \mathrm{mmse}(\boldsymbol{X}|\boldsymbol{X}+\boldsymbol{N}) &= 2\cdot\frac{\partial I(\boldsymbol{X};\boldsymbol{X}+\boldsymbol{N})}{\partial\beta} \\ &= 2\cdot\frac{\partial h(\boldsymbol{X};\boldsymbol{X}+\boldsymbol{W}/\sqrt{\beta})}{\partial\beta} + \frac{n}{\beta} \\ &= -\frac{1}{\beta^2}\mathrm{tr}\{J(\boldsymbol{Y})\} + \frac{n}{\beta}, \end{aligned} \tag{22}$$

where the factor $-1/\beta^2$ in front of the Fisher information term accounts for the passage from the variable t to the variable $\beta = 1/t$, as $\mathrm{d}t/\,\mathrm{d}\beta = -1/\beta^2$. Combining this with the previously obtained relations, we see that the noise suppression level due to MMSE estimation is given by

$$\Delta = \frac{\mathrm{tr}\{J(\boldsymbol{Y})\}}{\beta^2}.$$

In [21, Theorem 3.1], a generalized definition of the inverse temperature is proposed, as the response of the entropy to small energy perturbations, using de Bruijn's identity. As a consequence of that definition, the generalized inverse temperature in [21] turns out to be proportional to the Fisher information of $\boldsymbol{Y}$, and thus, in our setting, it is also proportional to $\beta^2\Delta$.[3] It should be pointed out that whenever the system undergoes a phase transition (as is the case with most of our forthcoming examples), then Δ, and hence also the effective temperature, may exhibit a non–smooth behavior, or even a discontinuity.

Additional relationships can be obtained in analogy to certain relations in statistical thermodynamics that were mentioned in Section III: Consider again the chain of equalities (11), but this time, instead using the relation $\boldsymbol{E}_\beta\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2\} = n/\beta$, in the passage from the second to the third line, we use the

[3]As is shown in [21], the generalized inverse temperature coincides with the ordinary inverse temperature when $\boldsymbol{Y}$ is purely Gaussian with variance proportional to $1/\beta$, i.e., the ordinary Boltzmann distribution with a quadratic Hamiltonian. In our setting, on the other hand, $\boldsymbol{Y}$ is given by a mixture of Gaussians whose weights are independent of β. To avoid confusion, it is important to emphasize that the original parameter β, in our setting, pertains to the Boltzmann form of the distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ according to the posterior $P(\boldsymbol{x}|\boldsymbol{y})$, whereas the current discussion concerns the temperature associated with the (unconditional) ensemble of $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N}$.

relation $\boldsymbol{E}_\beta\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2\} = -\boldsymbol{E}_\beta\{\frac{\mathrm{d}}{\mathrm{d}\beta}\ln Z(\beta|\boldsymbol{Y})\}$ in conjunction with the identity (cf. eq. (14)):

$$\begin{aligned}\boldsymbol{E}_\beta\left\{\frac{\mathrm{d}\ln Z(\beta|\boldsymbol{Y})}{\mathrm{d}\beta}\right\} &= \frac{\mathrm{d}\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\}}{\mathrm{d}\beta} - \int_{\mathbb{R}^n}\mathrm{d}\boldsymbol{y}\frac{\mathrm{d}P_\beta(\boldsymbol{y})}{\mathrm{d}\beta}\cdot\ln Z(\beta|\boldsymbol{y}) \\ &= \frac{\mathrm{d}\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\}}{\mathrm{d}\beta} + \frac{1}{2}\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \ln Z(\beta|\boldsymbol{Y})\}, \end{aligned} \tag{23}$$

to obtain

$$\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\} - \beta\cdot\frac{\mathrm{d}}{\mathrm{d}\beta}\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y}) = \frac{\beta}{2}\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \ln Z(\beta|\boldsymbol{Y})\} - I(\boldsymbol{X};\boldsymbol{Y}). \tag{24}$$

Thus, redefining the function $\psi(\beta)$ as

$$\psi(\beta) = -\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\}, \tag{25}$$

we obtain the following differential equation which is very similar to (7):

$$\dot{\psi}(\beta) - \frac{\psi(\beta)}{\beta} = \frac{\Sigma(\beta)}{\beta} \tag{26}$$

where

$$\Sigma(\beta) = \frac{\beta}{2}\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \ln Z(\beta|\boldsymbol{Y})\} - I(\boldsymbol{X};\boldsymbol{Y}). \tag{27}$$

Thus, the solution to this equation is precisely the same as (9), except that $S(\beta)$ is replaced by $\Sigma(\beta)$ and the ground–state energy E_0 is redefined as

$$E_0 = \boldsymbol{E}_\beta\{\min_{\boldsymbol{x}}\|\boldsymbol{Y}-\boldsymbol{x}\|^2\}.$$

Consequently, $\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = 2\dot{\psi}(\beta)$, where

$$\dot{\psi}(\beta) = E_0 - \int_\beta^\infty \frac{\mathrm{d}\hat{\beta}\Sigma(\hat{\beta})}{\hat{\beta}^2} + \frac{\Sigma(\beta)}{\beta}$$

and one can easily identify the contributions of the free energy and the internal energy (heat), as was done in Section III.

To summarize, we see that the I-MMSE relation gives rise essentially similar relations as in statistical thermodynamics except that the "effective entropy" $\Sigma(\beta)$ includes correction terms that account for the fact that our ensemble corresponds to a posterior distribution $P(\boldsymbol{x}|\boldsymbol{y})$ and the fact that the distribution of $\boldsymbol{Y}$ depends on β.

V. EXAMPLES

In this section, we provide a few examples where we show how the asymptotic MMSE can be calculated by using the I–MMSE relation in conjunction with statistical–mechanical techniques for evaluating the mutual information, or the partition function pertaining to the posterior distribution.

After the first example, of a Gaussian i.i.d. channel input, which is elementary, we turn to explore three examples where the channel input is a randomly selected codebook vector from a certain ensemble of codebooks that comply with a power constraint $\frac{1}{n}\boldsymbol{E}\{\|\boldsymbol{X}\|^2\} \leq P_x$. There could be various motivations for MMSE estimation when the desired signal is a codeword: One example is that of a user that, in addition to its desired signal, receives also a relatively strong interfering signal, which carries digital information (a codeword) intended to other users, and which comes from a codebook whose rate exceeds the capacity of this crosstalk channel between the interferer and our user, so that the user cannot fully decode this interference. Nonetheless, our user would like to estimate it as accurately as possible in order to subtract it and thereby perform interference cancellation.

In the first example of a code ensemble (Subsection V-B), we deal with a simple ensemble of block codes, and we demonstrate that the MMSE exhibits a phase transition at the value of β for which the channel capacity $C(\beta) = \frac{1}{2}\ln(1+\beta P_x)$ agrees with the coding rate R. The second ensemble (Subsection V-C) consists of an hierarchical structure which is suitable for the Gaussian broadcast channel. Here, we will observe two phase transitions, one corresponding to the weak user and one – to the strong user. The third ensemble (Subsection V-D) is also hierarchical, but in a different way: here the hierarchy corresponds to that of a tree structured code that works in two (or more) segments. In this case, there could be either one phase transition or two, depending on the coding rates at the two segments (see also [19]). Our last example is not related to coding applications, and it is based on a very simple model of sparse signals which is motivated by compressed sensing applications. Here we show that phase transitions can be present when the signal components are strongly correlated.

The statistical–mechanical considerations in this section provide unique insight into the coding and estimation problems, in particular by examining the typical behavior of the geometry of the free energy. This is in fact related to the notion of joint typicality for proving coding theorems, but more concrete geometry is seen due to the special structures of the code ensembles. In some of the ensuing examples, the mutual information can also be obtained through existing channel capacity results from information theory. In the last example pertaining to sparse signals (Subsection V-E), however, we are not aware of any alternative to the calculation using statistical mechanical techniques.

A. Gaussian I.I.D. Input

Our first example is very simple: Here, the components of $\boldsymbol{X}$ are zero–mean, i.i.d., Gaussian RV's with variance P_x. In this case, we readily obtain

$$Z(\beta|\boldsymbol{y}) = \frac{\exp\{-\|\boldsymbol{y}\|^2/[2(P_x+1/\beta)]\}}{(1+\beta P_x)^{n/2}},$$

thus

$$\ln Z(\beta|\boldsymbol{y}) = -\frac{n}{2}\ln(1+\beta P_x) - \frac{\|\boldsymbol{y}\|^2}{2(P_x+1/\beta)}.$$

Clearly,

$$\boldsymbol{E}_\beta \ln Z(\beta|\boldsymbol{Y}) = -\frac{n}{2}\ln(1+\beta P_x) - \frac{n}{2}$$

and its negative derivative is $nP_x/[2(1+\beta P_x)]$, which is indeed half of the MMSE. Here, we have:

$$\Delta = \frac{n}{\beta} - \frac{nP_x}{1+\beta P_x} = \frac{n}{\beta(1+\beta P_x)}$$

and

$$\mathrm{tr}\{J(\boldsymbol{Y})\} = n\boldsymbol{E}\left[\frac{Y}{P_x+1/\beta}\right]^2 = \frac{n\beta}{1+\beta P_x}$$

and so, the relation $\mathrm{tr}\{J(\boldsymbol{Y})\} = \beta^2\Delta$ is easily verified. Thus, the generalized temperature here is $\beta/(1+\beta P_x)$, which is the reciprocal of the variance of the Gaussian output.

B. Random Codebook on a Sphere Surface

Let $\boldsymbol{X}$ assume a uniform distribution over a codebook $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, $M = e^{nR}$, where each codeword $\boldsymbol{x}_i$ is drawn independently under the uniform distribution over the surface of the n–dimensional sphere, which is centered at the origin, and whose radius is $\sqrt{nP_x}$. The code is capacity achieving (the input becomes essentially i.i.d. Gaussian as $n \to \infty$). In the following we show that the MMSE vanishes if the code rate R is below channel capacity, but is no different than that of i.i.d. Gaussian input (without code structure) if R exceeds the capacity. We note that such a phase transition has been shown for good binary codes in general in [25] using the I-MMSE relationship.

Here, for a given $\boldsymbol{y}$, we have:

$$\begin{aligned} Z(\beta|\boldsymbol{y}) &= \sum_{\boldsymbol{x}\in\mathcal{C}} e^{-nR}\exp[-\beta\|\boldsymbol{y}-\boldsymbol{x}\|^2/2] \\ &= e^{-nR}\exp[-\beta\|\boldsymbol{y}-\boldsymbol{x}_0\|^2/2] + \sum_{\boldsymbol{x}\in\mathcal{C}\setminus\{\boldsymbol{x}_0\}} e^{-nR}\exp[-\beta\|\boldsymbol{y}-\boldsymbol{x}\|^2/2] \\ &\stackrel{\triangle}{=} Z_c(\beta|\boldsymbol{y}) + Z_e(\beta|\boldsymbol{y}) \end{aligned} \tag{28}$$

where, without loss of generality, we assume $\boldsymbol{x}_0$ to be the transmitted codeword. Now, since $\|\boldsymbol{y}-\boldsymbol{x}_0\|^2$ is typically around n/β, $Z_c(\beta|\boldsymbol{y})$ would typically be about $e^{-nR}e^{-\beta\cdot n/(2\beta)} = e^{-n(R+1/2)}$. As for $Z_e(\beta|\boldsymbol{y})$, we have:

$$Z_e(\beta|\boldsymbol{y}) \doteq e^{-nR}\int_{\mathbb{R}} \mathrm{d}\epsilon N(\epsilon)e^{-\beta n\epsilon},$$

where $N(\epsilon)$ is the number of codewords $\{\boldsymbol{x}\}$ in $\mathcal{C}-\{\boldsymbol{x}_0\}$ for which $\|\boldsymbol{y}-\boldsymbol{x}\|^2/2 \approx n\epsilon$, namely, between $n\epsilon$ and $n(\epsilon+\mathrm{d}\epsilon)$. Now, given $\boldsymbol{y}$, $N(\epsilon) = \sum_{i=1}^{M} 1\{\boldsymbol{x}_i : \ \|\boldsymbol{y}-\boldsymbol{x}_i\|^2/2 \approx n\epsilon\}$ is the sum of M i.i.d. Bernoulli RV's and so, its expectation is

$$\overline{N(\epsilon)} = \sum_{i=1}^{M} \Pr\{\|\boldsymbol{y}-\boldsymbol{X}_i\|^2/2 \approx n\epsilon\} = e^{nR}\Pr\{\|\boldsymbol{y}-\boldsymbol{X}_1\|^2/2 \approx n\epsilon\}. \tag{29}$$

Denoting $P_y = \frac{1}{n}\sum_{i=1}^{n} y_i^2$ (typically, P_y is about $P_x + 1/\beta$), the event $\|\boldsymbol{y}-\boldsymbol{x}\|^2/2 \approx n\epsilon$ is equivalent to the event $\langle \boldsymbol{x},\boldsymbol{y}\rangle \approx [(P_x+P_y)/2-\epsilon]n$ or equivalently,

$$\rho(\boldsymbol{x},\boldsymbol{y}) \stackrel{\triangle}{=} \frac{\langle \boldsymbol{x},\boldsymbol{y}\rangle}{n\sqrt{P_xP_y}} \approx \frac{\frac{1}{2}(P_x+P_y)-\epsilon}{\sqrt{P_xP_y}} \stackrel{\triangle}{=} \frac{P_a-\epsilon}{P_g}$$

where have defined $P_a = (P_x+P_y)/2$ and $P_g = \sqrt{P_xP_y}$ (the arithmetic and the geometric means between P_x and P_y, respectively). The probability that a randomly chosen vector $\boldsymbol{X}$ on the sphere would have an empirical correlation coefficient ρ with a given vector $\boldsymbol{y}$ (that is, $\boldsymbol{X}$ falls within a cone of half angle $\arccos(\rho)$ around $\boldsymbol{y}$) is exponentially $\exp[\frac{n}{2}\ln(1-\rho^2)]$. For convenience, let us define

$$\Gamma(\rho) = \frac{1}{2}\ln\left(1-\rho^2\right)$$

so that we can write

$$\Pr\{\|\boldsymbol{y}-\boldsymbol{X}_1\|^2/2 \approx n\epsilon\} \doteq \exp\left\{n\,\Gamma\left(\frac{P_a-\epsilon}{P_g}\right)\right\}.$$

From this point and onward, our considerations are very similar to those that have been used in the random energy model (REM) of spin glasses in statistical mechanics [5]–[7], a model of disordered magnetic materials where the energy levels pertaining to the various configurations of the system $\{\mathcal{E}(\boldsymbol{x})\}$ are i.i.d. RV's. These considerations have already been applied in the analogous analysis of random code ensemble performance, where the randomly chosen codewords give rise to random scores that play the same role as the random energies of the REM. The reader is referred to [27], [28], [20, Chapters 5,6], and [18] for a more detailed account of these ideas.

Applied to the random code ensemble considered here, the line of thought is as follows: If ϵ is such that

$$\Gamma\left(\frac{P_a-\epsilon}{P_g}\right) > -R,$$

then the energy level ϵ will be typically populated with an exponential number of codewords, concentrated very strongly around its mean

$$\overline{N(\epsilon)} \doteq \exp\left\{n\left[R+\Gamma\left(\frac{P_a-\epsilon}{P_g}\right)\right]\right\},$$

otherwise (which means that $\overline{N(\epsilon)}$ is exponentially small), the energy level ϵ will not be populated by any codewords typically. This means that the populated energy levels range between

$$\epsilon_1 \stackrel{\triangle}{=} P_a - P_g\sqrt{1-e^{-2R}}$$

and

$$\epsilon_2 \stackrel{\triangle}{=} P_a + P_g\sqrt{1-e^{-2R}},$$

or equivalently, the populated values of ρ range between $-\rho_*$ and $+\rho_*$ where $\rho_* = \sqrt{1-e^{-2R}}$. By large deviations and saddle–point methods [4], [11], it follows that for a typical realization of the randomly chosen code, we have

$$\begin{aligned} Z_e(\beta|\boldsymbol{y}) &\doteq e^{-nR}\max_{\epsilon\in[\epsilon_1,\epsilon_2]}\exp\left\{n\left[R+\Gamma\left(\frac{P_a-\epsilon}{P_g}\right)-\beta\epsilon\right]\right\} \\ &= \max_{\epsilon\in[\epsilon_1,\epsilon_2]}\exp\left\{n\left[\Gamma\left(\frac{P_a-\epsilon}{P_g}\right)-\beta\epsilon\right]\right\} \\ &= \exp\left\{n\left[\max_{|\rho|\le\rho_*}\left\{\frac{1}{2}\ln(1-\rho^2)-\beta(P_a-\rho P_g)\right\}\right]\right\}\ . \end{aligned}$$

The derivative of $\frac{1}{2}\ln(1-\rho^2)+\rho\beta P_g$ w.r.t. ρ vanishes within $[-1,1]$ at:

$$\rho = \rho_\beta \stackrel{\triangle}{=} \sqrt{1+\theta^2}-\theta$$

where

$$\theta \stackrel{\triangle}{=} \frac{1}{2\beta P_g}.$$

This is the maximizer as long as $\sqrt{1+\theta^2}-\theta \le \rho_*$, namely, $\theta > e^{-2R}/2\rho_*$, or equivalently, $\beta < \rho_* e^{2R}/P_g$, which for $P_g = \sqrt{P_x(P_x+1/\beta)}$, is equivalent to $\beta < \beta_R \stackrel{\triangle}{=} (e^{2R}-1)/P_x$. Thus, for the typical code we have

$$\phi_e(\beta,R) \stackrel{\triangle}{=} \lim_{n\to\infty}\frac{\ln Z_e(\beta|\boldsymbol{y})}{n} = \begin{cases} \frac{1}{2}\ln(1-\rho_\beta^2)-\beta(P_a-\rho_\beta P_g), & \beta<\beta_R \\ -R-\beta(P_a-\rho_* P_g), & \beta\ge\beta_R\ . \end{cases}$$

Taking now into account $Z_c(\beta|\boldsymbol{y})$, it is easy to see that for $\beta \ge \beta_R$ (which means $R < C$), $Z_c(\beta|\boldsymbol{y})$ dominates $Z_e(\beta|\boldsymbol{y})$, whereas for $\beta < \beta_R$ it is the other way around. It follows then that

$$\phi(\beta,R) \stackrel{\triangle}{=} \lim_{n\to\infty}\frac{\ln Z(\beta|\boldsymbol{y})}{n} = \begin{cases} \frac{1}{2}\ln(1-\rho_\beta^2)-\beta(P_a-\rho_\beta P_g), & \beta<\beta_R \\ -R-\frac{1}{2}, & \beta\ge\beta_R\ . \end{cases}$$

On substituting $P_a = P_x + 1/(2\beta)$, $P_g = \sqrt{P_x(P_x + 1/\beta)}$ and

$$\rho_\beta = \sqrt{1+\theta^2} - \theta = \sqrt{\frac{\beta P_x}{1+\beta P_x}},$$

we then get:

$$\psi(\beta) = -\lim_{n\to\infty} \frac{\ln Z(\beta|\boldsymbol{y})}{n} = \begin{cases} \frac{1}{2}\ln(1+\beta P_x) + \frac{1}{2}, & \beta < \beta_R \\ R + \frac{1}{2} & \beta \geq \beta_R \end{cases}.$$

Note that $\psi(\beta)$ is a continuous function but it is not smooth at $\beta = \beta_R$. Now,

$$\lim_{n\to\infty} \frac{\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n} = 2\frac{\mathrm{d}\psi(\beta)}{\mathrm{d}\beta} = \begin{cases} \frac{P_x}{1+\beta P_x}, & \beta < \beta_R \\ 0, & \beta \geq \beta_R \end{cases}. \tag{30}$$

which means that there is a first order phase transition[4] in the MMSE: As long as $\beta \geq \beta_R$, which means $R < C$, the MMSE essentially vanishes since the correct codeword can be reliably decoded, whereas for $R > C$, the MMSE behaves as if the inputs were i.i.d. Gaussian with variance P_x (cf. Subsection V-A).

C. Hierarchical Code Ensemble for the Degraded Broadcast Channel

Consider the following hierarchical code ensemble: First, randomly draw $M_1 = e^{nR_1}$ cloud–center vectors $\{\boldsymbol{u}_i\}$ on the $\sqrt{n}$–sphere. Then, for each $\boldsymbol{u}_i$, randomly draw $M_2 = e^{nR_2}$ codewords $\{\boldsymbol{x}_{i,j}\}$ according to $\boldsymbol{x}_{i,j} = \alpha\boldsymbol{u}_i + \sqrt{1-\alpha^2}\,\boldsymbol{v}_{i,j}$, where $\{\boldsymbol{v}_{i,j}\}$ are randomly drawn uniformly and independently on the $\sqrt{n}$–sphere. This means that $\|\boldsymbol{x}_{i,j} - \alpha\boldsymbol{u}_i\|^2 = n(1-\alpha^2) \stackrel{\triangle}{=} nb$. Without essential loss of generality, here and in Subsection V-D, we take the channel input power to be $P_x = 1$.

Let $\boldsymbol{x}_{0,0}$, belonging to cloud center $\boldsymbol{u}_0$, be the input to the Gaussian channel (1). It is easy to see that if the SNR of the Gaussian channel is high enough, the codeword $\boldsymbol{x}_{i,j}$ can be decoded; while at certain lower SNR only the cloud center $\boldsymbol{u}_i$ can be decoded but not $\boldsymbol{v}_{i,j}$. In the following we show the phase transitions of the MMSE as a function of the SNR.

[4]By "first–order phase transition", we mean, in this context, that the MMSE is a discontinuous function of β.

We will decompose the partition function as follows:

$$\begin{aligned} Z(\beta|\boldsymbol{y}) &= e^{-nR}\sum_{i,j}\exp(-\beta\|\boldsymbol{y}-\boldsymbol{x}_{i,j}\|^2/2) \\ &= e^{-nR}\exp(-\beta\|\boldsymbol{y}-\boldsymbol{x}_{0,0}\|^2/2) + e^{-nR}\sum_{j\geq 1}\exp(-\beta\|\boldsymbol{y}-\boldsymbol{x}_{0,j}\|^2/2) \\ &\qquad + e^{-nR}\sum_{i\geq 1}\sum_{j}\exp(-\beta\|\boldsymbol{y}-\boldsymbol{x}_{i,j}\|^2/2) \\ &\stackrel{\triangle}{=} Z_c(\beta|\boldsymbol{y}) + Z_{e1}(\beta|\boldsymbol{y}) + Z_{e2}(\beta|\boldsymbol{y}) \end{aligned} \tag{31}$$

where once again, $Z_c(\beta|\boldsymbol{y})$ – the contribution of the correct codeword, is typically about $e^{-n(R+1/2)}$. The other two terms $Z_{e1}(\beta|\boldsymbol{y})$ and $Z_{e2}(\beta|\boldsymbol{y})$ correspond to contributions of incorrect codewords from the same cloud and from other clouds, respectively.

Let us consider $Z_{e1}(\beta|\boldsymbol{y})$ first. The distance $\|\boldsymbol{y}-\boldsymbol{x}_{0,j}\|^2$ is decomposed as follows:

$$\begin{aligned} \|\boldsymbol{y}-\boldsymbol{x}_{0,j}\|^2 &= \|(\boldsymbol{y}-\alpha\boldsymbol{u}_0)+(\alpha\boldsymbol{u}_0-\boldsymbol{x}_{0,j})\|^2 \\ &= \|\boldsymbol{y}-\alpha\boldsymbol{u}_0\|^2 + \|\alpha\boldsymbol{u}_0-\boldsymbol{x}_{0,j}\|^2 + 2\langle \boldsymbol{y}-\alpha\boldsymbol{u}_0, \alpha\boldsymbol{u}_0-\boldsymbol{x}_{0,j}\rangle \ . \end{aligned} \tag{32}$$

Now, $\|\boldsymbol{y}-\alpha\boldsymbol{u}_0\|^2$ is typically about $n/\beta+nb \stackrel{\triangle}{=} na$ and $\|\alpha\boldsymbol{u}_0-\boldsymbol{x}_{0,j}\|^2 = nb$. Thus, for $\|\boldsymbol{y}-\boldsymbol{x}_{0,j}\|^2/2$ to be around $n\epsilon$, $\langle \boldsymbol{y}-\alpha\boldsymbol{u}_0, \alpha\boldsymbol{u}_0-\boldsymbol{x}_{0,j}\rangle$ must be around $n[\epsilon-(a+b)/2] \stackrel{\triangle}{=} n[\epsilon-P_a]$. Now, the question is this: Given $\boldsymbol{y}-\alpha\boldsymbol{u}_0$, what is the typical number of codewords in cloud 0 for which $\langle \boldsymbol{y}-\alpha\boldsymbol{u}_0, \alpha\boldsymbol{u}_0-\boldsymbol{x}_{0,j}\rangle = n[\epsilon-P_a]$. Similarly as before, the answer is the following:

$$N(\epsilon) \doteq \begin{cases} \exp\left\{n\left[R_2+\Gamma\left(\frac{\epsilon-P_a}{P_g}\right)\right]\right\}, & \epsilon\in[P_a-\rho_2P_g, P_a+\rho_2P_g] \\ 0, & \text{elsewhere} \end{cases} \tag{33}$$

where $P_g \stackrel{\triangle}{=} \sqrt{ab}$ and $\rho_2 = \sqrt{1-e^{-2R_2}}$. Thus,

$$\begin{aligned} Z_{e1}(\beta|\boldsymbol{y}) &\doteq e^{-nR}\exp\left\{n\left[\max_{|\rho|\leq\rho_2}\{R_2+\Gamma(\rho)-\beta(P_a-\rho P_g)\}\right]\right\} \\ &= e^{-nR_1}\exp\left\{n\left[\max_{|\rho|\leq\rho_2}\left\{\frac{1}{2}\ln(1-\rho^2)+\beta\rho P_g\right\}-\beta P_a\right]\right\}. \end{aligned} \tag{34}$$

As before, the derivative of $[\frac{1}{2}\ln(1-\rho^2)+\rho\beta P_g]$ w.r.t. ρ vanishes within $[-1,1]$ at:

$$\rho = \rho_\beta \stackrel{\triangle}{=} \sqrt{1+\theta^2}-\theta$$

where

$$\theta \stackrel{\triangle}{=} \frac{1}{2\beta P_g}.$$

This is the maximizer as long as $\sqrt{1+\theta^2}-\theta \leq \rho_2$, namely, $\theta > e^{-2R_2}/2\rho_2$, or equivalently, $\beta < \rho_2 e^{2R_2}/P_g$, which for $P_g = \sqrt{b(b+1/\beta)}$, is equivalent to $\beta < \beta(R_2) \stackrel{\triangle}{=} (e^{2R_2}-1)/b$. Thus, for the typical code we have

$$\psi_{e1}(\beta) \stackrel{\triangle}{=} -\lim_{n\to\infty}\frac{\ln Z_{e1}(\beta|\boldsymbol{y})}{n} = \begin{cases} R_1 - \frac{1}{2}\ln(1-\rho_\beta^2) + \beta(P_a - \rho_\beta P_g), & \beta < \beta(R_2) \\ R + \beta(P_a - \rho_2 P_g), & \beta \geq \beta(R_2)\ . \end{cases}$$

Similarly as before, it is easy to see that

$$Z_c + Z_{e1} \doteq \exp\left\{-n\left[R_1 + \min\left\{R_2, \frac{1}{2}\ln(1+b\beta)\right\} + \frac{1}{2}\right]\right\}.$$

Turning now to $Z_{e2}(\beta|\boldsymbol{y})$, we have the following consideration. Given $\boldsymbol{u}_i$, $i \geq 1$, let $\boldsymbol{y}' = \boldsymbol{y} - \alpha\boldsymbol{u}_i$ and $\boldsymbol{v}_{i,j} = \boldsymbol{x}_{i,j} - \alpha\boldsymbol{u}_i$. We would like to estimate how many codewords in cloud i, $N_i(\epsilon)$, contribute $\|\boldsymbol{y}-\boldsymbol{x}_{i,j}\|^2/2 = \|\boldsymbol{y}'-\boldsymbol{v}_{i,j}\|^2/2 = n\epsilon$. Similarly as before, $N_i(\epsilon)$ is given by exactly the same formula as (33) where this time, $P_a = (1-\alpha^2+\|\boldsymbol{y}-\alpha\boldsymbol{u}_i\|^2/n)/2$ and $P_g = \sqrt{(1-\alpha^2)\|\boldsymbol{y}-\alpha\boldsymbol{u}_i\|^2/n}$. Thus, we have expressed the typical number of codewords that cloud i contributes with energy ϵ as $N_i(\epsilon) = \exp\{nF(\|\boldsymbol{y}-\alpha\boldsymbol{u}_i\|^2/n, \epsilon)\}$, and the total number is $N(\epsilon) = \sum_i N_i(\epsilon)$. Now let $M(\delta)$ be the number of $\{\boldsymbol{u}_i\}$ for which $\|\boldsymbol{y}-\alpha\boldsymbol{u}_i\|^2/n = \delta$. Then,

$$N(\epsilon) \doteq \sum_\delta M(\delta) e^{nF(\delta,\epsilon)}.$$

Now,

$$M(\delta) = \begin{cases} \exp\left\{n\left[R_1 + \Gamma\left(\frac{\delta/2-P_a'}{P_g'}\right)\right]\right\}, & \delta \in [\delta_1, \delta_2]\ , \\ 0, & \text{elsewhere} \end{cases}$$

where $P_a' = (1+1/\beta+\alpha^2)/2$, $P_g' = \alpha\sqrt{1+1/\beta}$, $\delta_1 = 2(P_a' - P_g'\sqrt{1-e^{-2R_1}}) \stackrel{\triangle}{=} 2(P_a' - \rho_1 P_g')$ and $\delta_2 = 2(P_a' + P_g'\rho_1)$. Thus,

$$N(\epsilon) \doteq \exp\left\{n \max_{\delta_1\leq\delta\leq\delta_2}\left[R_1 + \Gamma\left(\frac{P_a'-\delta}{P_g'}\right) + F(\delta,\epsilon)\right]\right\}.$$

Putting it all together, we get:

$$\begin{aligned}\psi_{e2}(\beta) \stackrel{\triangle}{=} -\lim_{n\to\infty}\frac{\ln Z_{e2}(\beta|\boldsymbol{y})}{n} = -\max_{|r_1|\leq\rho_1}\ \max_{|r_2|\leq\rho_2(r_1)}\Bigg\{&\frac{1}{2}\ln(1-r_1^2) + \frac{1}{2}\ln(1-r_2^2) - \\ &\beta\left[\frac{1-\alpha^2}{2} + P_a' - r_1 P_g' - r_2\sqrt{2(1-\alpha^2)(P_a' - r_1 P_g')}\right]\Bigg\},\end{aligned} \tag{35}$$

where $\rho_1 = \sqrt{1-e^{-2R_1}}$, $\rho_2(r_1) = \sqrt{1-e^{-2R}/(1-r_1^2)}$, $P_a' = (1+1/\beta+\alpha^2)/2$, and $P_g' = \alpha\sqrt{1+1/\beta}$. The above expression does not seem to lend itself to closed form analysis in an easy manner. Numerical results (cf. Fig. 1) show a reasonable match (within the order of magnitude of 1×10^{-5}) between values

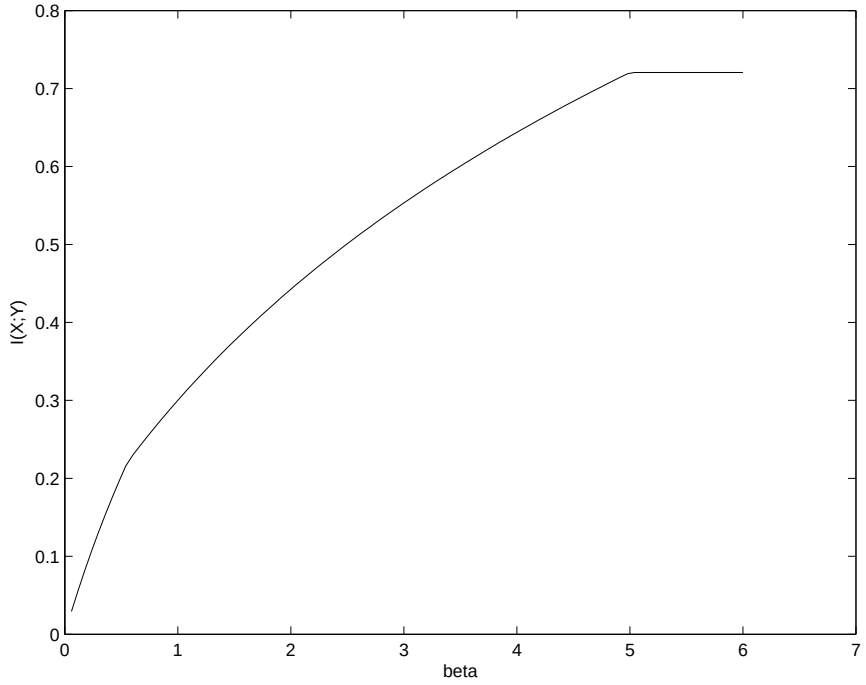

Fig. 1. Graph of $\lim_{n\to\infty} I(\boldsymbol{X};\boldsymbol{Y})/n = -\boldsymbol{E}_\beta\{\ln Z(\beta|\boldsymbol{Y})\}/n - 1/2$ as a function of β for $R_1 = 0.1$, $R_2 = 0.6206$, and $\alpha = 0.7129$, which result in $\beta_1 = 0.5545$ and $\beta_2 = 5.001$. As can be seen quite clearly, there are phase transitions at these values of β.

of $\lim_{n\to\infty} I(\boldsymbol{X};\boldsymbol{Y})/n$ obtained numerically from the asymptotic exponent of $\boldsymbol{E}_\beta \ln Z(\beta|\boldsymbol{Y})$ and those that are obtained from the expected behavior in this case:

$$\lim_{n\to\infty} \frac{I(\boldsymbol{X};\boldsymbol{Y})}{n} = \begin{cases} \frac{1}{2}\ln(1+\beta), & \beta < \beta_1 \\ R_1 + \frac{1}{2}\ln(1+\beta b), & \beta_1 \le \beta < \beta_2 \\ R = R_1 + R_2, & \beta \ge \beta_2 \end{cases}$$

where

$$\beta_1 \stackrel{\triangle}{=} \frac{e^{2R_1}-1}{1-be^{2R_1}}\ , \quad \beta_2 \stackrel{\triangle}{=} \frac{e^{2R_2}-1}{1-b}\ ,$$

and it is assumed that the parameters of the model (R_1, R_2 and α) are chosen such that $\beta_1 < \beta_2$. Accordingly, the MMSE undergoes two phase transitions, where it behaves as if the input was: (i) Gaussian i.i.d. with unit variance for $\beta < \beta_1$ (where no information can be decoded), (ii) Gaussian input of a smaller variance (corresponding to the cloud), in the intermediate range (where the cloud center is decodable, but the refined message is not), and (iii) the MMSE altogether vanishes for $\beta > \beta_2$, where

both messages are reliably decodable.

The hierarchical code ensemble takes the superposition code structure which achieves the capacity region of the Gaussian broadcast channel. Consider two receivers, referred to as receiver 1 and receiver 2, with β_1 and β_2 respectively. Receiver 1 can decode the cloud center, whereas receiver 2 can decode the entire codeword. In other words, suppose the hierarchical code ensemble with rate pair (R_1, R_2) and parameter α is sent to two receivers with fixed SNR of γ_1 and γ_2 respectively. Then the minimum decoding error probability vanishes as long as (R_1, R_2, α) are such that

$$R_1 < \frac{1}{2}\log\left(1 + \frac{\alpha^2\gamma_1}{1+(1-\alpha^2)\gamma_1}\right), \tag{36}$$

$$R_2 < \frac{1}{2}\log\left(1+\alpha^2\gamma_2\right). \tag{37}$$

In particular, all boundary points of the capacity region can be achieved by varying the power distribution coefficient α. This capacity region result also leads to the fact that if only the cloud center is decodable, then the MMSE for the codeword $\boldsymbol{v}_{i,j}$ is no different to that if the elements of $\boldsymbol{v}_{i,j}$ were i.i.d. standard Gaussian. Knowledge of the codebook structure of $\{\boldsymbol{v}_{i,j}\}$ does not reduce the MMSE because otherwise the code cannot achieve the capacity region of the Gaussian broadcast channel.

D. Hierarchical Tree–Structured Code

Consider next an hierarchical code with the following structure: The block of length n is partitioned into two segments, the first is of length $n_1 = \lambda_1 n$ ($\lambda_1 \in (0,1)$) and the second is of length $n_2 = \lambda_2 n$ ($\lambda_2 = 1-\lambda_1$). We randomly draw $M_1 = e^{n_1 R_1}$ first–segment codewords $\{\boldsymbol{x}_i\}$ on the surface of the $\sqrt{n_1}$–sphere, and then, for each $\boldsymbol{x}_i$, we randomly draw $M_2 = e^{n_2 R_2}$ second–segment codewords $\{\boldsymbol{x}'_{i,j}\}$ on the surface of the $\sqrt{n_2}$–sphere. The total message of length $nR = n_1R_1 + n_2R_2$ (thus $R = \lambda_1 R_1 + \lambda_2 R_2$) is encoded in two parts: The first–segment codeword depends only on the first n_1R_1 bits of the message whereas the second–segment codeword depends on the entire message.

Let $(\boldsymbol{x}_0, \boldsymbol{x}_{0,0})$ be the transmitted codeword, and let $\boldsymbol{y}$ and $\boldsymbol{y}'$ be the corresponding segments of the channel output vector $(\boldsymbol{y}, \boldsymbol{y}')$. The partition function is as follows:

$$\begin{aligned} Z(\beta|\boldsymbol{y}) &= e^{-nR}\exp\{-\beta[\|\boldsymbol{y}-\boldsymbol{x}_0\|^2 + \|\boldsymbol{y}'-\boldsymbol{x}_{0,0}\|^2]/2\} \\ &\quad + e^{-nR}\exp\{-\beta[\|\boldsymbol{y}-\boldsymbol{x}_0\|^2/2\}\sum_j \exp\{-\beta\|\boldsymbol{y}'-\boldsymbol{x}_{0,j}\|^2]/2\} \\ &\quad + e^{-nR}\sum_{i\geq 1}\sum_j \exp\{-\beta[\|\boldsymbol{y}-\boldsymbol{x}_i\|^2/2\}\exp\{-\beta\|\boldsymbol{y}'-\boldsymbol{x}_{i,j}\|^2]/2\} \\ &\stackrel{\triangle}{=} Z_c + Z_{e1} + Z_{e2}. \end{aligned} \tag{38}$$

Now, as before, $Z_c \doteq e^{-n(R+1/2)}$. As for Z_{e1}, it can also be treated as in Subsection V-B: The first factor contributes $e^{-nR} \cdot e^{-n\lambda_1/2}$. The second factor is $e^{-n\lambda_2[\min\{R_2,C(\beta)\}+1/2]}$, where $C(\beta) = \frac{1}{2}\ln(1+\beta)$. Thus,

$$Z_{e1}(\beta|\boldsymbol{y}) + Z_c \doteq \exp\left\{-n\left[\lambda_1 R_1 + \ \lambda_2 \min\{R_2, C(\beta)\} + \frac{1}{2}\right]\right\}.$$

Consider next the term Z_{e2}. Let $r_1 = \langle \boldsymbol{x}, \boldsymbol{y}\rangle/(n_1 P_g)$ and $r_2 = \langle \boldsymbol{x}', \boldsymbol{y}'\rangle/(n_2 P_g)$ where P_g is as in Subsection V-B. Of course, $\langle (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{y}, \boldsymbol{y}')\rangle/(nP_g) = \lambda_1 r_1 + \lambda_2 r_2$. What is the typical number of codewords $(\boldsymbol{x}_i, \boldsymbol{x}'_{i,j})$ of Z_{e2} whose correlation with $(\boldsymbol{y}, \boldsymbol{y}')$ is exactly r? The answer is

$$\lim_{n\to\infty} \frac{\ln N(r)}{n} = \max_{|r_1|\le\rho(R_1)} \left\{\lambda_1 R_1 + \lambda_1 \Gamma(r_1) + \lambda_2 R_2 + \lambda_2 \Gamma\left(\frac{r - \lambda_1 r_1}{\lambda_2}\right)\right\},$$

where $\rho(x) = \sqrt{1 - e^{-2x}}$. This expression behaves differently depending on whether $R_1 > R_2$ or $R_1 < R_2$. In the first case, it behaves exactly as in the ordinary ensemble, that is:

$$\lim_{n\to\infty} \frac{\ln N(r)}{n} = \begin{cases} R + \frac{1}{2}\ln(1 - r^2), & |r| \le \rho(R) \\ 0, & |r| > \rho(R)\ . \end{cases}$$

and then, of course, Z_{e2} is as before:

$$Z_{e2} + Z_c \doteq \exp\{-n[\min\{R, C(\beta)\} + 1/2]\}.$$

When $R_1 < R_2$, however, we have two phase transitions:

$$\lim_{n\to\infty} \frac{\ln N(r)}{n} = \begin{cases} R + \Gamma(r), & |r| \le \rho(R_1) \\ \lambda_2\left[R_2 + \Gamma\left(\frac{r - \lambda_1\rho(R_1)}{\lambda_2}\right)\right], & \rho(R_1) \le |r| \le \lambda_1\rho(R_1) + \lambda_2\rho(R_2) \\ 0, & |r| > \lambda_1\rho(R_1) + \lambda_2\rho(R_2)\ . \end{cases}$$

In this case, we get:

$$\lim_{n\to\infty} \frac{\ln(Z_{e2} + Z_c)}{n} = \begin{cases} -C(\beta) - \frac{1}{2}, & \beta \le \beta(R_1) \\ -\lambda_1 R_1 - \lambda_2 C(\beta) - \frac{1}{2}, & \beta(R_1) < \beta \le \beta(R_2) \\ -R - \frac{1}{2}, & \beta > \beta(R_2) \end{cases}$$

where $\beta(R)$ is the solution β to the equation $C(\beta) \equiv \frac{1}{2}\ln(1+\beta) = R$. To summarize, we have the following: $Z_c \doteq e^{-n(R+1/2)}$, $Z_{e1} + Z_c \doteq \exp\{-n[\lambda_1 R_1 + \lambda_2 \min\{R_2, C(\beta)\} + 1/2]\}$ and

$$Z_{e2} + Z_c \doteq \begin{cases} \exp\{-n[\min\{R, C(\beta)\} + 1/2]\}, & R_1 > R_2 \\ \exp\{-n[\lambda_1 \min\{R_1, C(\beta)\} + \lambda_2 \min\{R_2, C(\beta)\} + 1/2]\}, & R_1 \le R_2\ . \end{cases}$$

Clearly, if $R_1 \leq R_2$ then $Z_{e2} + Z_c$ dominates $Z_{e1} + Z_c$. If $R_1 > R_2$, we note that

$$\min\{\lambda_1 R_1 + \lambda_2 \min\{R_2, C(\beta)\}, \min\{R, C(\beta)\}\} \equiv \min\{R, C(\beta)\}.$$

Thus,

$$Z \doteq \begin{cases} \exp\{-n[\min\{R, C(\beta)\} + 1/2]\}, & R_1 > R_2 \\ \exp\{-n[\lambda_1 \min\{R_1, C(\beta)\} + \lambda_2 \min\{R_2, C(\beta)\} + 1/2]\}, & R_1 \leq R_2 \ . \end{cases}$$

The MMSE then is as in (30) in Subsection V-B when $R_1 > R_2$, and given by

$$\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = \begin{cases} \frac{1}{1+\beta}, & \beta \leq \beta(R_1) \\ \frac{\lambda_2}{1+\beta}, & \beta(R_1) < \beta \leq \beta(R_2) \\ 0, & \beta > \beta(R_2) \end{cases} \tag{39}$$

when $R_1 < R_2$. This dichotomy between these two types of behavior have their roots in the behavior of the GREM, a generalized version of the random energy model, where the random energy levels of the various system configurations are correlated (rather than being i.i.d.) in an hierarchical structure [8]–[10]. The GREM turns out to have an intimate analogy with the tree–structured code ensemble considered here. The reader is referred to [19] for a more elaborate discussion on this topic.

The preceding result on the MMSE is consistent with the analysis based solely on information theoretic considerations. In case $R_1 < R_2$, the first segment code is decodable as long as $R_1 < (1/2)\log(1+\beta)$, whereas the second segment code is decodable if also $R_2 < (1/2)\log(1+\beta)$. Hence the MMSE is given by (39). In case $R_1 > R_2$, the second-segment code is decodable if and only if the first-segment is also decodable, i.e., the two codes can be decoded jointly. This requires $R_2 < (1/2)\log(1+\beta)$, $\lambda_1 R_1 < \lambda_1 \log(1+\beta) + \lambda_2 \log(1+\beta)$ and $R = \lambda_1 R_1 + \lambda_2 R_2 < \log(1+\beta)$. The last inequality dominates, hence the MMSE is given by (30).

E. Estimation of Sparse Signals

Let the components of $\boldsymbol{X}$ be given by $X_i = S_i U_i$, $i = 1, 2, \ldots, n$, where $S_i \in \{0, 1\}$ and $\{U_i\}$ are $\mathcal{N}(0, \sigma^2)$ i.i.d. and independent of $\{X_i\}$. As before $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N}$, where the components of $\boldsymbol{N}$ are i.i.d. Gaussian $\mathcal{N}(0, 1/\beta)$. One motivation of this simple model is in compressed sensing applications, where the signal $\boldsymbol{X}$ (possibly, in some transform domain) is assumed to possess a limited fraction of non–zero components, here designated by the non–zero components of $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$. The signal $\boldsymbol{X}$ is considered sparse if the relative fraction of 1's in $\boldsymbol{S}$ is small. We will assume that $\boldsymbol{S}$, whose realization is not revealed to the estimator, is governed by a given probability distribution $P(\boldsymbol{s})$. We first derive an

expression of the partition function for a general $P(\boldsymbol{s})$ and then particularize our study to a certain form of $P(\boldsymbol{s})$. First, we have the following:

$$\begin{aligned} P(\boldsymbol{x}) &= \sum_{\boldsymbol{s}} P(\boldsymbol{s}) P(\boldsymbol{x}|\boldsymbol{s}) \\ &= \sum_{\boldsymbol{s}} P(\boldsymbol{s}) \prod_{i:\ s_i=0} \delta(x_i) \prod_{i:\ s_i=1} \Big[(2\pi\sigma^2)^{-1/2} \exp\{-x_i^2/(2\sigma^2)\} \Big] \\ &= \sum_{\boldsymbol{s}} P(\boldsymbol{s}) \prod_{i=1}^{n} \Big[(2\pi s_i\sigma^2)^{-1/2} \exp\{-x_i^2/(2s_i\sigma^2)\} \Big] \end{aligned} \tag{40}$$

where a zero–variance Gaussian distribution is understood to be equivalent to the Dirac delta–function. Thus,

$$\begin{aligned} Z(\beta|\boldsymbol{y}) &= \int_{\mathbb{R}^n} d\boldsymbol{x} P(\boldsymbol{x}) \exp\{-\beta\|\boldsymbol{y}-\boldsymbol{x}\|^2/2\} \\ &= \sum_{\boldsymbol{s}} P(\boldsymbol{s}) \prod_{i=1}^{n} \left[\int_{-\infty}^{\infty} \mathrm{d}x_i (2\pi s_i\sigma^2)^{-1/2} \exp\{-x_i^2/(2s_i\sigma^2)\} \cdot \exp\{-\beta(y_i-x_i)^2/2\} \right] \\ &= \sum_{\boldsymbol{s}} P(\boldsymbol{s}) \prod_{i=1}^{n} \left[(1+qs_i)^{-1/2} \exp\left\{ -\frac{\beta y_i^2}{2(1+qs_i)} \right\} \right] \\ &= \sum_{\boldsymbol{s}} P(\boldsymbol{s}) \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2} \left[\frac{\beta y_i^2}{1+qs_i} + \ln(1+qs_i) \right] \right\} \end{aligned} \tag{41}$$

where we have used the notation[5] $q = \beta\sigma^2$. Transforming $\boldsymbol{s}$ to "spins" $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ by the relation $\mu_i = 1 - 2s_i \in \{-1, +1\}$, we get:

$$\frac{\beta y_i^2}{1+qs_i} + \ln(1+qs_i) = \frac{(1+q/2)\beta y_i^2}{1+q} + \frac{1}{2}\ln(1+q) - 2\mu_i h_i$$

where

$$h_i = -\frac{\beta^2\sigma^2 y_i^2}{4\big(1+\beta\sigma^2\big)} + \frac{1}{4}\ln\big(1+\beta\sigma^2\big). \tag{42}$$

On substituting back into the partition function we get:

$$Z(\beta|\boldsymbol{y}) = (1+q)^{-n/4} \cdot \exp\left\{ -\frac{\beta(1+q/2)}{2(1+q)} \|\boldsymbol{y}\|^2 \right\} \cdot \sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) \exp\left\{ \sum_{i=1}^{n} \mu_i h_i \right\}. \tag{43}$$

Thus h_i is given the statistical–mechanical interpretation of the random 'local' magnetic field felt by the i–th spin.

Eq. (43) holds for a general distribution $P(\boldsymbol{s})$ or equivalently, $P(\boldsymbol{\mu})$. To further develop this expression, we must make some assumptions on one of these distributions. At this point, we have the freedom to

[5]The quantity q is proportional to the SNR.

examine certain models of $P(\boldsymbol{\mu})$, and by viewing the expression $\sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) \exp\{\sum_i \mu_i h_i\}$ as the partition function of a certain spin system with a non–uniform, random field $\{H_i\}$ (whose realization is $\{h_i\}$), we can borrow techniques from statistical physics to analyze its behavior. Evidently, for every spin glass model that exhibits phase transitions, it is conceivable that there will be analogous phase transitions in the corresponding signal estimation problem.

Assuming certain symmetry properties among the various components of $\boldsymbol{s}$, it would be plausible to postulate that all $\{\boldsymbol{s}\}$ with the same number of 1's are equally likely, or equivalently, all spin configurations $\{\boldsymbol{\mu}\}$ with the same magnetization

$$m(\boldsymbol{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \mu_i$$

have the same probability. This means that $P(\boldsymbol{\mu})$ depends on $\boldsymbol{\mu}$ only via $m(\boldsymbol{\mu})$. Consider then the form

$$P(\boldsymbol{\mu}) = C_n \exp\{n f(m(\boldsymbol{\mu}))\},$$

where $f(m)$ is an arbitrary function and C_n is a normalization constant. Further, let us assume that f is twice differentiable with finite first derivative on $[-1, 1]$. Clearly,

$$\begin{aligned} C_n &= \left(\sum_{\boldsymbol{\mu}} \exp\{n\, f(m(\boldsymbol{\mu}))\}\right)^{-1} \\ &\dot{=} \exp\left\{-n \max_m \{\mathcal{H}_2((1+m)/2) + f(m)\}\right\} \\ &= \exp\left\{-n\left(\mathcal{H}_2((1+m_a)/2) + f(m_a)\right)\right\} \end{aligned} \tag{44}$$

where $\mathcal{H}_2(\cdot)$ denotes the binary entropy function and m_a is the maximizer of $\mathcal{H}_2((1+m)/2) + f(m)$. In other words, m_a is the *a–priori* magnetization, namely the magnetization that dominates $P(\boldsymbol{\mu})$. Of course, when $f(m)$ is linear in m, the components of $\boldsymbol{\mu}$ are i.i.d. Note that if f is monotonically increasing in m, then $P(\boldsymbol{\mu})$ has a sharp peak at $m = 1$, which corresponds to a vanishing fraction of sites with $s_i = 1$, i.e., a sparse signal. Our derivation, however, will take place for general f.

1) General Solution: On substituting the above expression of $P(\boldsymbol{\mu})$ into that of $Z(\beta|\boldsymbol{y})$, our main concern is then how to deal with the expression

$$\hat{Z}(\beta|\boldsymbol{h}) \stackrel{\triangle}{=} \sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) e^{\sum_i \mu_i h_i} = C_n \sum_{\boldsymbol{\mu}} \exp\left\{n\left[f(m(\boldsymbol{\mu})) + \frac{1}{n}\sum_i \mu_i h_i\right]\right\}. \tag{45}$$

We investigate the typical behavior of the partition function, or more precisely, calculate the following quantity:

$$\frac{1}{n}\log \boldsymbol{E}\left\{\hat{Z}(\beta|\boldsymbol{H})\right\} = \frac{1}{n}\log\left[C_n \boldsymbol{E}\left\{\sum_{\boldsymbol{\mu}} \exp\left\{n\left[f(m(\boldsymbol{\mu})) + \frac{1}{n}\sum_i \mu_i H_i\right]\right\}\right\}\right] \tag{46}$$

where $\boldsymbol{H}$ consists of i.i.d. random variables with arbitrary distribution $p(H)$.

Using large deviations theory, as $n \to \infty$, the dominant value of m in (46), henceforth denoted as m^* is shown to satisfy

$$m^* = \boldsymbol{E}\{\tanh(f'(m^*) + H)\} \tag{47}$$

and

$$\boldsymbol{E}\{\tanh^2(f'(m^*) + H)\} > 1 - \frac{1}{f''(m^*)}. \tag{48}$$

The detailed analysis is relegated to Appendix A. Clearly, m^* is the dominant magnetization *a–posteriori*, i.e., the one that dominates the posterior of $m(\boldsymbol{\mu})$ given (a typical) $\boldsymbol{y}$. It is also shown in Appendix A that

$$\lim_{n\to\infty} \frac{1}{n} \log \boldsymbol{E}\left\{\hat{Z}(\beta|\boldsymbol{H})\right\} = \lim_{n\to\infty} \frac{1}{n} \log C_n - \psi(m^*) \tag{49}$$

where

$$\psi(m^*) \stackrel{\triangle}{=} f'(m^*)\, m^* - f(m^*) - \boldsymbol{E}\left\{\log\left[2\cosh(f'(m^*) + H)\right]\right\} \tag{50}$$

and the normalized exponent of C_n is given by (44). Thus the asymptotic normalized mutual information is expressed as

$$\lim_{n\to\infty} \frac{I(\boldsymbol{X};\boldsymbol{Y})}{n} = -\frac{1}{2} + \frac{1}{4}\ln(1+q) + \frac{\beta(1+q/2)\boldsymbol{E}\{Y^2\}}{2(1+q)} - \lim_{n\to\infty} \frac{\ln C_n}{n} + \psi(m^*). \tag{51}$$

For the sparse signal model described by (40), H is defined by (42) with y_i replaced by Y and the expectation over Y is w.r.t. a mixture of two Gaussians: $\mathcal{N}(0, 1/\beta)$ with weight $(1+m_a)/2$, and $\mathcal{N}(0, \sigma^2 + 1/\beta)$ with weight $(1 - m_a)/2$.

The solution to

$$\boldsymbol{E}\{\tanh^2(f'(m) + H)\} = 1 - \frac{1}{f''(m)} \tag{52}$$

is known as a *critical point*, beyond which the solution to (47) ceases to be a local maximum and it becomes a local minimum. The dominant m^* must jump elsewhere. Also, as we vary one of the other parameters of the model, it might happen that the global maximum jumps from one local maximum to another.

2) Special Case with Quadratic Exponent: In the case where f is quadratic[6] in m, i.e.,

$$f(m) = am + bm^2/2. \tag{53}$$

[6]A quadratic model can be thought of as consisting of the first few terms of the Taylor expansion of a smooth function f.

This is similar though not identical to the *random–field Curie-Weiss model* (RFCW model) of spin systems[7] (cf. e.g., [2] and references therein). Eq. (47) becomes

$$m = \boldsymbol{E}\{\tanh(bm + a + H)\},$$

similarly as in the mean field model with a random field [2]. Eq. (52) for the critical point satisfies

$$\boldsymbol{E}\{\tanh^2(bm + a + H)\} = 1 - (1/b). \tag{54}$$

To demonstrate that the global maximum might jump from one local maximum to another, consider the quadratic case and assume that β and σ^2 are so small that the fluctuations in H can be neglected. Equation (47) can then be approximated by

$$m = \tanh(bm + a),$$

which is actually the same the equation of the magnetization as in the Curie–Weiss model (a.k.a. the mean field model or the infinite–range model) of spin arrays (cf. e.g., [22, Sect. 4.2], [1, Chap. 3], [14, Sect. 4.5.1]), which is actually a special case of the above with $H_i \equiv 0$ for all i. For $a = 0$ and $b > 1$, this equation has two symmetric non–zero solutions $\pm m_0$, which both dominate the partition function. If $a \neq 0$ but small, then the symmetry is broken, and there is only one dominant solution which is about $m_0\,\mathrm{sgn}(a)$. To approximate m_0 for the case where $|a|$ is small and b is only slightly larger than 1, one can use the Taylor expansion of the function $\tanh(\cdot)$ (as is customarily done in the theory of the infinite range Ising model; see e.g., [22, p. 188, eqs. (4.21a), (4.21b)]) and get

$$m \approx bm + a - \frac{(bm + a)^3}{3}.$$

Neglecting the contribution of a, we get a simple quadratic equation whose solutions are $\pm m_0$ with $m_0 = \frac{1}{b}\sqrt{3(1 - 1/b)}$. Thus, for small values of $|a|$ and $b - 1$,

$$m^* \approx m_0 \cdot \mathrm{sgn}(a),$$

and so, m^* jumps between $+m_0$ and $-m_0$ as a crosses the origin. Similarly, for $a = 0$, m^* jumps from zero to $+m_0$ or $-m_0$ as b passes the value $b = 1$ while increasing.

[7]There is a certain difference in the sense that in the RFCW $\{H_i\}$ are i.i.d., whereas here each H_i depends on the corresponding μ_i because the variance of y_i depends on whether $\mu_i = -1$ or $\mu_i = +1$. Also as a result, $\{H_i\}$ here are not i.i.d. because they depend on each other via the dependence between $\{\mu_i\}$. These differences are not crucial, however.

By (51), the asymptotic normalized mutual information of this model is given by

$$\begin{aligned}\lim_{n\to\infty}\frac{I(\boldsymbol{X};\boldsymbol{Y})}{n} &= -\frac{1}{2}+\frac{1}{4}\ln(1+q)+\frac{\beta(1+q/2)}{2(1+q)}\left[\frac{1+m_a}{2}\cdot\frac{1}{\beta}+\frac{1-m_a}{2}\left(\sigma^2+\frac{1}{\beta}\right)\right]\\ &\quad+\mathcal{H}_2\left(\frac{1+m_a}{2}\right)+f(m_a)+\psi(m^*)\\ &= -\frac{1}{2}+\frac{1}{4}\ln(1+q)+\frac{1+q/2}{2(1+q)}\left(1+\frac{1-m_a}{2}\cdot q\right)+\mathcal{H}_2\left(\frac{1+m_a}{2}\right)\\ &\quad+am_a+\frac{bm_a^2}{2}-\boldsymbol{E}\{\ln[2\cosh(bm^*+a+H)]\}+\frac{b(m^*)^2}{2}.\end{aligned}\tag{55}$$

In this special case of quadratic exponent, the Hubbard-Stratonovich transformation can be used to obtain an alternative, more straightforward derivation of the mutual information result (55). The details are provided in Appendix B.

The MMSE is equal to twice the derivative of (55) w.r.t. β. Note that the dominant value m^* is dependent on β. In Appendix C, we carry out the calculation and obtain

$$\begin{aligned}&\lim_{n\to\infty}\frac{\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n}\\ &=\frac{\sigma^2 q}{2(1+q)^2}+\frac{(1-m_a)\sigma^2}{2}\left[1-\frac{q(1+q/2)}{(1+q)^2}\right]\\ &\quad+\frac{1+m_a}{2}\left[\mathrm{Cov}_0\{Y^2,\ln[2\cosh(bm^*+a+H)]\}+\boldsymbol{E}_0\{H'\tanh(bm^*+a+H)\}\right]\\ &\quad+\frac{1-m_a}{2}\left[\frac{1}{(1+q)^2}\cdot\mathrm{Cov}_1\{Y^2,\ln[2\cosh(bm^*+a+H)]\}+\boldsymbol{E}_1\{H'\tanh(bm^*+a+H)\}\right]\end{aligned}\tag{56}$$

where H' is defined by

$$H'=-\frac{\sigma^2}{2(1+q)}+\frac{q(q+2)}{2(1+q)^2}\cdot Y^2\tag{57}$$

which is in fact the derivative of (42) w.r.t. β. To ease understanding of the MMSE, we evaluate its value in two extreme cases in Appendix D.

3) Discussion: Returning now to the general expression of the MMSE, it is reasonable to expect that at the critical points, where m^* jumps from one solution of eq. (47) to another as the parameters of the model vary, the MMSE may also undergo an abrupt change, and so the MMSE may be discontinuous (w.r.t. these parameters) at these points. A related abrupt change takes place also in the *response* of the MMSE estimator itself at the critical points: Note that m^* is the dominant magnetization *a–posteriori*. Thus, as m^* jumps, say, from $m^*=m_1$ to $m^*=m_2$, the conditional mean estimator, which is a weighted average of $\{\boldsymbol{x}\}$, transfers most of the weight from a set of $\boldsymbol{x}$–vectors whose binary support vectors $\{\boldsymbol{s}\}$ correspond to magnetization m_1, into another set of $\boldsymbol{x}$–vectors supported by $\{\boldsymbol{s}\}$ with magnetization m_2.

It is not surprising then that this abrupt change in the response of the estimator is accompanied by a corresponding sudden drop in the MMSE.

It is instructive to compare the type of the phase transition in our example to those of the ordinary Curie–Weiss model. In the Curie-Weiss model, we have:

- A first order phase transition w.r.t. the magnetic field (below the critical temperature), i.e., the first derivative of the free energy w.r.t. the magnetic field (which is exactly the magnetization) is discontinuous (at the point of zero field).
- A second order phase transition w.r.t. temperature, i.e., the first derivative of the free energy w.r.t. temperature (which is related to the internal energy) is continuous, but the second derivative (which is related to the specific heat) is not.

Here, on the other hand, in physics terms, what we observe is a first order phase transition w.r.t. temperature. The reason for this discrepancy is that in our model, the dependency of the free energy on temperature is introduced via the variables $\{h_i\}$ that play the role of magnetic fields.

In case of quadratic exponent (53), $b = 0$ corresponds to the special case of i.i.d. $\{S_i\}$. In this case, our problem is analogous to a system of non-interacting particles, where of course, no phase transitions can exist. Therefore, what we learn from statistical physics here is that phase transitions in the MMSE estimator cannot be a property of the sparsity alone (because sparsity may be present also for the i.i.d. case with $P\{S_i = 1\}$ small), but rather a property of strong dependency between $\{S_i\}$, whether it comes with sparsity or not.

ACKNOWLEDGEMENT

N. Merhav would like to thank Yonina Eldar for a few interesting discussions concerning the example of estimating sparse signals (in Subsection V-E) during the early stages of this work.

APPENDIX A
ESTIMATION OF SPARSE SIGNALS: THE DOMINANT MAGNETIZATION

For the time being let us assume that H_i, $i = 1, \ldots, n$ take on values from a discrete set $\{h_1, \ldots, h_K\}$, where of the n variables, $q_k n$ of them taking the value of h_k. The sum in (46) can be rewritten as

$$\sum_{\boldsymbol{\mu}} \exp\left\{ nf(m(\boldsymbol{\mu})) + \sum_{k=1}^{K} h_k \sum_{i=1}^{q_k n} \mu_{ki} \right\} \tag{58}$$

where we relabel μ_i as μ_{ki} with $i = 1, \ldots, q_k n$ for each k. The expectation on the r.h.s. of (46) can be viewed as an integral

$$2^n \int_{-1}^{1} \cdots \int_{-1}^{1} \exp\left\{ nf(m) + \sum_{k=1}^{K} h_k(q_k n) m_k \right\} N(\,\mathrm{d}m_1, \cdots, \,\mathrm{d}m_K) \tag{59}$$

where N is a probability measure proportional to the number of sequences $\boldsymbol{\mu}$ with $\frac{1}{q_k n} \sum_{i=1}^{q_k n} \mu_{ki} \approx m_k$. Here $m = \sum_{k=1}^{K} q_k m_k$. For $\boldsymbol{\mu}$ uniformly randomly chosen from ± 1 sequences, the probability measure satisfies large deviations property, the rate function (or entropy) of which is obtained as (using the Legendre-Fenchel transform)[8]

$$I(m_1, \ldots, m_K) = \sum_{k=1}^{K} q_k \left(\log 2 - \mathcal{H}_2\left(\frac{1+m_k}{2} \right) \right). \tag{61}$$

Not surprisingly, the rate function achieves its maximum at $m_k = 0$, $k = 1, \ldots, K$, where the number of ± 1's in each subsequence μ_{ki}, $i = 1, \ldots, q_k n$ is balanced. Due to large deviations property, the integral (59) is dominated by unique values of m_k, $k = 1, \ldots, K$. Specifically, we use Varadhan's Theorem [4], [11] to obtain[9]

$$\begin{aligned} &\frac{1}{n} \log \int \cdots \int \exp\left\{ nf(m) + \sum_{k=1}^{K} h_k(q_k n) m_k \right\} N(\,\mathrm{d}m_1, \ldots, \,\mathrm{d}m_k) \\ &\quad \to \sup_{m_1, \ldots, m_K \in [-1,1]} \left\{ f(m) + \sum_{k=1}^{K} h_k q_k m_k - I(m_1, \ldots, m_K) \right\} \\ &\quad = 2^{-n} \cdot \sup_{m_1, \ldots, m_K \in [-1,1]} \psi(m_1, \ldots, m_K) \end{aligned} \tag{63}$$

where we use (61) and define

$$\psi(m_1, \ldots, m_K) \stackrel{\triangle}{=} f\left(\sum_{k=1}^{K} q_k m_k \right) + \sum_{k=1}^{K} h_k q_k m_k + \sum_{k=1}^{K} q_k \mathcal{H}_2\left(\frac{1+m_k}{2} \right). \tag{64}$$

[8]By Cramér's theorem [11, Theorem II.4.1], the probability measure of the empirical mean $\frac{1}{n} X_i$ of i.i.d. random variables X_i satisfy, as $n \to \infty$, the large deviations property with some rate function $I(m)$. The rate of the probability measure is given by the Legendre-Fenchel transform of the cumulant generating function (logarithm of the moment generating function) [4], [11]:

$$I(m) = \sup_{\eta} \left[\eta\, m - \log \boldsymbol{E}\left\{ e^{\eta X} \right\} \right]. \tag{60}$$

It is straightforward to generalize to the product measure of the means of subgroups of i.i.d. random variables.

[9]The Varadhan's Theorem basically states that, if the sequence of probability measures N_n on $\mathbb{R}$ satisfies large deviations property with rate function $I(m)$, and that F is continuous and upper bounded on $\mathbb{R}$, then

$$\lim_{n \to \infty} \frac{1}{n} \log \int_{\mathbb{R}} \exp\{F(m)\} N_n(\,\mathrm{d}m) = \sup_{m} \{F(m) - I(m)\}\ . \tag{62}$$

The result can also be generalized to multiple dimensions.

The maximum of ψ is achieved by an internal point in $(-1,1)^K$. This is because $\mathcal{H}_2$ is concave with infinite derivative at the boundary $m_k = \pm 1$, whereas the derivative of f is finite by assumption. Because the function ψ is twice differentiable, at its maximum, the gradient of ψ w.r.t. every m_k should be equal to 0, whereas the Hessian of ψ should be negative definite. It can be shown by taking derivative of ψ w.r.t. m_k that zero gradient is achieved by setting

$$m_k = \tanh\left(f'\left(\sum_{l=1}^{K} q_l m_l\right) + h_k\right) \tag{65}$$

for all k, so that

$$m = \sum_{k=1}^{K} q_k \tanh\left(f'(m) + h_k\right). \tag{66}$$

The Hessian of ψ is determined by noting that

$$\frac{\partial^2 \psi}{\partial m_k \partial m_l} = q_k q_l f''(m) - q_k \frac{\delta_{k,l}}{1 - m_k^2} \tag{67}$$

where $\delta_{k,l}$ is equal to 1 if $k = l$ and equal to 0 otherwise. The Hessian is negative definite if and only if

$$\left(\sum_{k=1}^{K} q_k x_k\right)^2 f''(m) \le \sum_{k=1}^{K} q_k \frac{x_k^2}{1 - m_k^2} \tag{68}$$

for all $x_k \in \mathbb{R}$, $k = 1, \ldots, K$, which is equivalent to

$$f''(m) \le \min_{x_1,\ldots,x_K} \frac{\sum_{k=1}^{K} q_k x_k^2/(1-m_k^2)}{\left(\sum_{k=1}^{K} q_k x_k\right)^2}. \tag{69}$$

Using Lagrange multiplier, the minimum on the r.h.s. of (69) is obtained as $1 - \sum_{k=1}^{K} q_k m_k^2$. Further, by (65), the condition (69) reduces to

$$f''(m) \le \frac{1}{1 - \sum_{k=1}^{K} q_k \tanh^2(f'(m) + h_k)}. \tag{70}$$

In other words, a solution of (65) is a local maximum of ψ if and only if it also satisfies (70). In multiple such solutions exist, the global supremum is identified by comparing the corresponding values of ψ.

In the limit $n \to \infty$, the requirement that H_i take discrete values is not necessary (the continuous distribution can be regarded as the limit of a degenerate discrete one). Using (66) and (70), the dominant magnetization m^* satisfy (47) and (48) for general distribution of H. This can be made precise by formulating a variational problem.

We also note an alternative technique for evaluating the free energy (46) using Fourier transform and saddle point method, which is standard in statistical mechanics (often without rigorous justification). Usage of this technique in information theory can be found in e.g., [23].

APPENDIX B
ESTIMATION OF SPARSE SIGNALS: AN ALTERNATIVE DERIVATION OF (55)

In case of quadratic exponent (53), the partition function (45) can be written using the Hubbard–Stratonovich transformation as

$$\begin{aligned}\sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) e^{\sum_i \mu_i h_i} &= C_n \sum_{\boldsymbol{\mu}} \exp\left\{ a \sum_i \mu_i + \sum_i \mu_i h_i + \frac{b}{2n} \Big(\sum_i \mu_i \Big)^2 \right\} \\ &= C_n \sqrt{\frac{nb}{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}m \exp\left\{ -\frac{nbm^2}{2} \right\} \sum_{\boldsymbol{\mu}} \exp\left\{ a \sum_i \mu_i + \sum_i \mu_i h_i + bm \sum_i \mu_i \right\} \\ &= C_n \sqrt{\frac{nb}{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}m \exp\left\{ -\frac{nbm^2}{2} \right\} \prod_{i=1}^{n} \left[2\cosh\left(a + bm + h_i\right) \right] \\ &= C_n \sqrt{\frac{nb}{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}m \exp\left\{ n \Big[-\frac{bm^2}{2} + \frac{1}{n} \sum_{i=1}^{n} \ln[2\cosh(a + bm + h_i)] \Big] \right\}. \end{aligned} \tag{71}$$

Thus, we have $-\ln \hat{Z} \approx n \min_m \psi(m) - \ln C_n$, where ψ is defined by (50), whose minimum is attained at $m^* = m^*(\beta)$, one of the solutions to the equation $m = \boldsymbol{E}\{\tanh(bm + a + H\}$, as before.[10] The mutual information is then obtained as (55).

APPENDIX C
ESTIMATION OF SPARSE SIGNALS: THE MMSE

The MMSE is equal to twice the derivative of (55) w.r.t. β. We will denote hereafter H_i as given by (42) with y_i replaced by Y_i and $\boldsymbol{H} = (H_1, \ldots, H_n)$. Let us present the asymptotic MMSE per sample, $\lim_{n\to\infty} \mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})/n$, as $A + B$, where A is the double derivative of the first three terms, and B is the contribution of the other terms. The easy part is the former:

$$A = \frac{\sigma^2 q}{2(1+q)^2} + \frac{(1-m_a)\sigma^2}{2} \left[1 - \frac{q(1+q/2)}{(1+q)^2} \right].$$

As for B, we have the following consideration: The first three terms depend only on m_a, which in turn is independent of β, therefore their derivatives w.r.t. β all vanish. For the last two terms, pertaining to

[10]The function $\psi(m)$ is (within a factor of the inverse temperature) identified with the Landau free energy function for this problem [22, p. 186, eq. (4.15a)], [14, Sect. 4.6].

$\psi(m^*)$, it proves useful to return to the original expression of the Gaussian integral (71), i.e.,

$$\begin{aligned}
B &= -\frac{2}{n}\frac{\partial}{\partial\beta}\boldsymbol{E}\{\ln\hat{Z}(\beta|\boldsymbol{H})\} \\
&= -\frac{2}{n}\frac{\partial}{\partial\beta}\boldsymbol{E}\left\{\ln\int_{-\infty}^{\infty}\frac{d\nu}{\sqrt{2\pi}}\exp\left\{n\left[-\frac{(\nu-a)^2}{2b}+\frac{1}{n}\sum_{i=1}^{n}\ln[2\cosh(\nu+h_i)]\right]\right\}\right\} \\
&= -\frac{2}{n}\frac{\partial}{\partial\beta}\int_{\mathbb{R}^n}d\boldsymbol{y}P_\beta(\boldsymbol{y})\ln\int_{-\infty}^{\infty}\mathrm{d}m\exp\left\{n\left[-\frac{bm^2}{2}+\frac{1}{n}\sum_{i=1}^{n}\ln[2\cosh(bm+a+h_i)]\right]\right\} \\
&= -\frac{2}{n}\int_{\mathbb{R}^n}d\boldsymbol{y}\frac{\partial P_\beta(\boldsymbol{y})}{\partial\beta}\ln\int_{-\infty}^{\infty}\mathrm{d}m\exp\left\{n\left[-\frac{bm^2}{2}+\frac{1}{n}\sum_{i=1}^{n}\ln[2\cosh(bm+a+h_i)]\right]\right\} \\
&\quad -\frac{2}{n}\int_{\mathbb{R}^n}d\boldsymbol{y}P_\beta(\boldsymbol{y})\frac{\partial}{\partial\beta}\ln\int_{-\infty}^{\infty}\mathrm{d}m\exp\left\{n\left[-\frac{bm^2}{2}+\frac{1}{n}\sum_{i=1}^{n}\ln[2\cosh(bm+a+h_i)]\right]\right\} \\
&\stackrel{\triangle}{=} B_1+B_2. \qquad (72)
\end{aligned}$$

Now, $P_\beta(\boldsymbol{y})$ is the mixture of Gaussians weighted by $\{P(\boldsymbol{\mu})\}$, where the dominant $\boldsymbol{\mu}$–configurations are those with $(1+m_a)/2$ $(+1)$'s and $(1-m_a)/2$ (-1)'s. Each such configuration contributes the same quantity to B_1 and B_2, because for every given such $\boldsymbol{\mu}$, the random variables $\{Y_i\}$ (and hence also $\{H_i\}$) are all independent, a fraction $(1+m_a)/2$ of them are $\mathcal{N}(0,1/\beta)$ and the remaining fraction of $(1-m_a)/2$ are $\mathcal{N}(0,\sigma^2+1/\beta)$. Thus, it is sufficient to confine attention to one such sequence, call it $\boldsymbol{\mu}^*$, whose first $n_1\stackrel{\triangle}{=}n(1-m_a)/2$ components are all -1 and last $n-n_1=n(1+m_a)/2$ components are all $+1$. Thus,

$$\begin{aligned}
B_1 &\approx -\frac{2}{n}\int_{\mathbb{R}^n}d\boldsymbol{y}\frac{\partial P_\beta(\boldsymbol{y}|\boldsymbol{\mu}^*)}{\partial\beta}\ln\int_{-\infty}^{\infty}\mathrm{d}m\exp\left\{n\left[-\frac{bm^2}{2}+\frac{1}{n}\sum_{i=1}^{n}\ln[2\cosh(bm+a+h_i)]\right]\right\} \\
&\approx \frac{1}{n}\mathrm{Cov}\left\{\sum_{i=1}^{n_1}Y_i^2+\frac{1}{(1+q)^2}\sum_{i=n_1+1}^{n}Y_i^2,\sum_{i=1}^{n}\ln[2\cosh(bm^*+a+H_i)]\right\} \\
&= \frac{1+m_a}{2}\cdot\mathrm{Cov}_0\{Y^2,\ln[2\cosh(bm^*+a+H)]\} \\
&\quad +\frac{1-m_a}{2}\cdot\frac{1}{(1+q)^2}\cdot\mathrm{Cov}_1\{Y^2,\ln[2\cosh(bm^*+a+H)]\}. \qquad (73)
\end{aligned}$$

where $\mathrm{Cov}_s\{\cdot,\cdot\}$ denotes covariance with respect to $\mathcal{N}(0,\sigma^2 s+1/\beta)$, $s=0,1$. Finally, for B_2, we have:

$$\begin{aligned}
B_2 &= -\frac{2}{n}\int_{\mathbb{R}^n} d\boldsymbol{y} P_\beta(\boldsymbol{y})\frac{\partial}{\partial\beta}\ln\int_{-\infty}^{\infty}\mathrm{d}m\exp\left\{n\left[-\frac{bm^2}{2}+\frac{1}{n}\sum_{i=1}^{n}\ln[2\cosh(bm+a+h_i)]\right]\right\}\\
&= \frac{1}{n}\int_{\mathbb{R}^n} d\boldsymbol{y} P_\beta(\boldsymbol{y})\cdot\frac{\int_{-\infty}^{\infty}\mathrm{d}m\,[\sum_i h_i'\tanh(bm+a+h_i)]\,e^{-n\psi(m)}}{\int_{-\infty}^{\infty}\mathrm{d}m e^{-n\psi(m)}}\\
&\approx \boldsymbol{E}\left\{\frac{1}{n}\sum_{i=1}^{n}H_i'\tanh(bm^*+a+H_i)\right\}\\
&\approx \frac{1+m_a}{2}\cdot\boldsymbol{E}_0\{H'\tanh(bm^*+a+H)\}+\frac{1-m_a}{2}\cdot\boldsymbol{E}_1\{H'\tanh(bm^*+a+H)\},
\end{aligned}\tag{74}$$

where $\boldsymbol{E}_s$ denotes expectation w.r.t. $\mathcal{N}(0,\sigma^2 s+1/\beta)$, $s=0,1$, and H' is given by (57), and correspondingly, h_i' and H_i' are given by the same formula with Y replaced by y_i and Y_i' respectively. Collecting all terms, A, B_1, and B_2, we have (56).

APPENDIX D
ESTIMATION OF SPARSE SIGNALS: TWO EXTREME CASES

Two extreme cases, where it is relatively easy to examine the resulting expression are as follows:

- When $b\gg 1$ and $a\ll -1$, we have $m_a\approx -1$ and $m^*\approx -1$ (which means that most $s_i=1$), and so we can approximate

  $$\ln[2\cosh(bm^*+a+H)]\approx\ln[2\cosh(-b+a+H)]\approx b-a-H$$

  and $\tanh(bm^*+a+H)\approx -1$, and we get

  $$\lim_{n\to\infty}\frac{\mathrm{MMSE}(\boldsymbol{X}|\boldsymbol{Y})}{n}\approx\frac{\sigma^2}{1+q},$$

  the classical Wiener expression, as expected.[11]

- When $b\gg 1$ and $a\gg 1$, we have $m_a\approx 1$ and $m^*\approx 1$ (which means that most $s_i=0$), and then $\ln[2\cosh(bm^*+a+H)]\approx b+a+H$ and $\tanh(bm^*+a+H)\approx 1$, so we get

  $$\lim_{n\to\infty}\frac{\mathrm{MMSE}(\boldsymbol{X}|\boldsymbol{Y})}{n}\approx\frac{1-m_a}{2}\cdot\sigma^2,$$

  which means the conditional–mean estimator simply outputs essentially the all–zero sequence without attempting to detect (explicitly or implicitly) which of the few signal components are active. The intuition behind this behavior is that when there are so few active components of the clean signal,

[11]Here, by $\lim_{n\to\infty}\mathrm{MMSE}(\boldsymbol{X}|\boldsymbol{Y})/n \approx F(a,b,\beta,\sigma^2)$, for a generic function F, we mean that $\lim_{a\to-\infty}\lim_{b\to\infty}\lim_{n\to\infty} nF(a,b,\beta,\sigma^2)/\mathrm{MMSE}(\boldsymbol{X}|\boldsymbol{Y})=1$. A similar comment applies to item number 2 below.

then even if there are nevertheless a few observations $\{y_i\}$ with large absolute values (and hence could have been suspected to stem from places where $s_i = 1$), it is still more plausible for the estimator to "assume" that they simply belong to the tail of $\mathcal{N}(0, 1/\beta)$ (with $s_i = 0$) rather than to $\mathcal{N}(0, \sigma^2 + 1/\beta)$ with $s_i = 1$. This because the prior for $s_i = 1$ is so small that it becomes comparable to the tail probability of $\mathcal{N}(0, 1/\beta)$.[12]

REFERENCES

[1] R. J. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, 1982.

[2] A. Bianchi, A. Bovier, and D. Ioffe, "Sharp asymptotics for metastability in the random field Curie–Weiss model,"arXiv:0806.4478v1 [math.PR] 27 Jun 2008.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, second edition, Hoboken, New Jersey, U.S.A., 2006.

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer, 1998.

[5] B. Derrida, "Random–energy model: limit of a family of disordered models," *Phys. Rev. Lett.*, vol. 45, no. 2, pp. 79–82, July 1980.

[6] B. Derrida, "The random energy model," *Physics Reports* (Review Section of Physics Letters), vol. 67, no. 1, pp. 29–35, 1980.

[7] B. Derrida, "Random–energy model: an exactly solvable model for disordered systems," *Phys. Rev. B*, vol. 24, no. 5, pp. 2613–2626, September 1981.

[8] B. Derrida, "A generalization of the random energy model which includes correlations between energies," *J. de Physique – Lettres*, vol. 46, L–401-107, May 1985.

[9] B. Derrida and E. Gardner, "Solution of the generalised random energy model," *J. Phys. C: Solid State Phys.*, vol. 19, pp. 2253–2274, 1986.

[10] B. Derrida and E. Gardner, "Magnetic properties and the function $q(x)$ of the generalised random–energy model," *J. Phys. C: Solid State Phys.*, no. 19, pp. 5783–5798, 1986.

[11] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*, ser. A series of comprehensive studies in mathematics. Springer-Verlag, 1985, vol. 271.

[12] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean–square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.

[13] D. Guo and T. Tanaka, "Generic multiuser detection and statistical physics," in *Advances in Multiuser Detection* (M. Honig, ed.), Wiley, 2009. to be published.

[14] J. Honerkamp, *Statistical physics: an advanced approach with applications*, Springer–Verlag, 2002.

[15] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev. A*, vol. 106, pp. 620–630, May 1957.

[16] E. T. Jaynes, "Information theory and statistical mechanics - II," *Phys. Rev. A*, vol. 108, pp. 171–190, October 1957.

[17] N. Merhav, "An identity of Chernoff bounds with an interpretation in statistical physics and applications in information theory," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3710–3721, August 2008.

[12]To see this, it is instructive to think of a simple binary hypothesis testing problem where an observer is required to decide whether an observation comes from $\mathcal{N}(0, 1/\beta)$ or $\mathcal{N}(0, \sigma^2 + 1/\beta)$ and the priors are very much in favor of the former.

[18] N. Merhav, "Relations between random coding exponents and the statistical physics of random codes," to appear in *IEEE Trans. Inform. Theory*, January 2009.

[19] N. Merhav, "The generalized random energy model and its application to the statistical physics of ensembles of hierarchical codes," to appear in *IEEE Trans. Inform. Theory*.

[20] M. Mézard and A. Montanari, *Information, Physics and Computation*, draft, November 9, 2007. Available on–line at: http://www.stanford.edu/∼montanar/BOOK/book.html.

[21] K. R. Narayanan and A. R. Srinivasa, "On the thermodynamic temperature of a general distribution," arXiv:0711.1460v2 [cond-mat.stat-mech], Nov. 10, 2007.

[22] J. W. Negele and H. Orland, *Quantum many–particles systems*, Frontier in Physics Lecture Notes, Addison–Wesley, 1988.

[23] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, ser. Number 111 in International Series of Monographs on Physics. Oxford University Press, 2001.

[24] D. P. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Trans. Inform. Theory*, vol. 53, no. 2, pp. 453–470, February 2008.

[25] M. Peleg, A. Sanderovich and S. Shamai (Shitz), "On extrinsic information of good codes operating over Gaussian channels," European Transactions on Telecommunications, Vol. 18, No. 2, pp. 133-139, 2007.

[26] O. Shental and I. Kanter, "Shannon meets Carnot: generalized second thermodynamic law," http://arxiv.org/PS_cache/arxiv/pdf/0806/0806.3763v1.pdf

[27] N. Sourlas, "Spin–glass models as error–correcting codes," *Nature*, pp. 693–695, vol. 339, June 1989.

[28] N. Sourlas, "Spin glasses, error–correcting codes and finite–temperature decoding," *Europhysics Letters*, vol. 25, pp. 159–164, 1994.