Optimum Estimation via Gradients of Partition Functions and Information Measures: A Statistical–Mechanical Perspective

Neri Merhav

Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel
merhav@ee.technion.ac.il

Abstract

In continuation to a recent work on the statistical–mechanical analysis of minimum mean square error (MMSE) estimation in Gaussian noise via its relation to the mutual information (the I–MMSE relation), here we propose a simple and more direct relationship between optimum estimation and certain information measures (e.g., the information density and the Fisher information), which can be viewed as partition functions and hence are amenable to analysis using statistical–mechanical techniques. The proposed approach has several advantages, most notably, its applicability to general sources and channels, as opposed to the I–MMSE relation and its variants which hold only for certain classes of channels (e.g., additive white Gaussian noise channels). We then demonstrate the derivation of the conditional mean estimator and the MMSE in a few examples. Two of these examples turn out to be generalizable to a fairly wide class of sources and channels. For this class, the proposed approach is shown to yield an approximate conditional mean estimator and an MMSE formula that has the flavor of a single–letter expression. We also show how our approach can easily be generalized to situations of mismatched estimation.

Index Terms: Conditional mean estimation, minimum mean squared error, partition function, statistical mechanics, Fisher information.

1 Introduction

Relationships between signal estimation, signal detection, and information measures, both in discrete time and continuous time, have been known for decades [1],[3],[8] and have gained a remarkable degree of revived interest and research activity in the last several years, see, e.g., [4], [5], [6], [7], [11], [12], [13] and references therein.

In particular, in [5], Guo, Shamai and Verdú have derived a relation between the mutual information between the input and the output of an additive white Gaussian noise (AWGN) channel and the minimum mean squared error (MMSE) of non–causal estimation of the channel input based on its output. In particular, this relation, which is often called the I–MMSE relation, shows that the derivative of the mutual information with respect to (w.r.t.) the signal–to–noise (SNR) is equal to half of the MMSE, and it is intimately related to the de Bruijn identity [2, Sec. 17.7]. Later, this relation has been generalized and further developed in several directions: Guo, Shamai, and Verdú [6] and Raginsky and Coleman [12] have derived relations of the same spirit for more general additive channels. Palomar and Verdú [11] have studied relations between the covariance matrix of the MMSE estimator and arbitrary gradients of the mutual information for a general vector Gaussian channel, which allows also a linear transformation of the input signal. In [7], relations between information measures and estimation measures have been derived for Poisson channels. More recently, Verdú [13] extended the I–MMSE relation of Gaussian noise to the paradigm of mismatched conditional mean estimation, that is, to deal with an estimator that is optimally matched to a wrong probability distribution assumed on the input signal. The excess mean squared error (MSE) due to this mismatch was shown to be related to the Kullback–Leibler divergence between the channel output distributions corresponding to the true and the assumed input distributions (see also [4] for a further study in this direction). In [9], the I–MMSE relation was further investigated from a statistical physics perspective, where among other results, it was demonstrated how statistical–mechanical tools can be harnessed in order to assess the MMSE via the I–MMSE relation of [5], using the fact that in many cases, the mutual information can be viewed as the partition function of a certain physical system.

This paper is a further development in the above described direction of [9]. The main idea is that, for the purpose of evaluating the covariance matrix of the MMSE estimator, one may use a conceptually simple and more direct relationship between the MMSE covariance matrix and other information measures, that can also be presented in the form of a certain partition function and hence be analyzed using methods of statistical physics. The main advantage of the proposed approach, over those of the I–MMSE relations and its variants, is its full generality: It applies, in principle, to any joint probability function $P(\boldsymbol{x}, \boldsymbol{y})$ of the channel input signal $\boldsymbol{x} = (x_1, \ldots, x_n)$, to be estimated, and the channel output $\boldsymbol{y} = (y_1, \ldots, y_m)$ (where m and n are positive integers),

provided that certain technical regularity conditions hold. The channel $P(\boldsymbol{y}|\boldsymbol{x})$ does not even have to be additive, as opposed to the assumptions made in [6] and [12]. Moreover, the dimension m of the channel output vector $\boldsymbol{y}$ does not have to be the same as the dimension n of the input vector $\boldsymbol{x}$.

In a nutshell, the idea is to define, for a given n–vector of real–valued parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, the 'partition function'

$$Z(\boldsymbol{y}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{x}} \exp\left\{\sum_{i=1}^{n} \lambda_i x_i\right\} P(\boldsymbol{x}, \boldsymbol{y}),$$

where we have implicitly assumed that $\boldsymbol{x}$ takes on discrete values, otherwise, the sum should simply be replaced by an integral. Now, it is straightforward to show that the gradient of $\ln Z(\boldsymbol{y}, \boldsymbol{\lambda})$ w.r.t. $\boldsymbol{\lambda}$, computed at $\boldsymbol{\lambda} = 0$, gives the conditional mean estimator $\hat{\boldsymbol{x}} = \boldsymbol{E}\{\boldsymbol{X}|\boldsymbol{y}\}$, whereas the expectation of the Hessian of the same function, again, at $\boldsymbol{\lambda} = 0$, gives the error covariance matrix of the MMSE estimator. As we shall see in the sequel, $\ln Z(\boldsymbol{y}, \boldsymbol{\lambda})$ lends itself to closed form analytic evaluation (in the spirit of a single–letter formula) in a fairly wide spectrum of situations, using methods of statistical mechanics. Thus, the MMSE estimator and its performance can quite easily be derived too in these situations. Moreover, as was demonstrated extensively in [9], the statistical–mechanical perspective on estimation–theoretic problems, may offer, not only analysis techniques, but also some important insights with regard to threshold effects (whenever existent) via the inspection of possible *phase transitions* in the parallel statistical–mechanical model.

Besides the general applicability of this approach, it has several additional advantages:

1. As mentioned in the previous paragraph, it provides, not only the MMSE error covariance matrix, but also the conditional mean estimator itself.

2. As will be seen, several variants of these relations between estimation measures and information measures can be offered. In some cases, one of the relations may be more convenient to work with than the others.

3. The approach is easy to extend to the mismatched case. Furthermore, it allows mismatch in both the source and the channel (as opposed to [13], which allows mismatch in the source only).

The remaining part of this paper is organized as follows. In Section 2, we establish notation conventions. In Section 3, we first derive the basic relations between the conditional mean estimator, as well as its error covariance matrix, and the above–mentioned partition function. In the same section, we also discuss this relation and derive a few variants that involve also information measures, like the information density, the Fisher information, etc. We also outline the extension to mismatched estimation. In Section 4, we provide three examples. In Section 5, we show how two of them set the stage to the analysis of a more general class of joint distributions, $P(\boldsymbol{x}, \boldsymbol{y})$. Finally, in Section 6, we summarize and conclude the paper.

2 Notation Conventions

Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, 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 same symbols in the bold face font. Thus, for example, $\boldsymbol{X}$ will denote a random vector $(X_1, \ldots, X_n)$, and $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a specific vector value in $\mathcal{X}^n$, the n-th Cartesian power of $\mathcal{X}$. The notations y_i^j and Y_i^j, where i and j are integers and $i \leq j$, will designate segments $(y_i, \ldots, y_j)$ and $(Y_i, \ldots, Y_j)$, respectively.

Probability functions will be denoted generically by the letter P or Q. In particular, $P(\boldsymbol{x}, \boldsymbol{y})$ is the joint probability mass function (in the discrete case) or the joint density (in the continuous case) of the desired channel input vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ and the observed channel output vector $\boldsymbol{y} = (y_1, \ldots, y_m)$. Accordingly, $P(\boldsymbol{x})$ will denote the marginal of $\boldsymbol{x}$, $P(\boldsymbol{y}|\boldsymbol{x})$ will denote the conditional probability mass (or density) of $\boldsymbol{y}$ given $\boldsymbol{x}$, induced by the channel, and so on. Whenever there is room for ambiguity, these probability functions will be subscripted by the names of the random variables and the conditionings, according to standard notation conventions in probability theory and information theory. Throughout the sequel, we will assume discrete valued alphabets, mostly for the sake of simplicity and convenience. Extensions to continuous valued situations will be straightforward with summations being replaced by integrations, etc. Indeed, some of our examples will involve continuous valued random variables.

The expectation operator of a generic function $f(\boldsymbol{x}, \boldsymbol{y})$ w.r.t. the joint distribution P of $(\boldsymbol{X}, \boldsymbol{Y})$

will be denoted by $\boldsymbol{E}\{f(\boldsymbol{X},\boldsymbol{Y})\}$. The conditional expectation of the same function given that $\boldsymbol{Y}=\boldsymbol{y}$, denoted $\boldsymbol{E}\{f(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{Y}=\boldsymbol{y}\}$, and which is obviously identical to $\boldsymbol{E}\{f(\boldsymbol{X},\boldsymbol{y})|\boldsymbol{Y}=\boldsymbol{y}\}$, is, of course, a function of $\boldsymbol{y}$. On substituting $\boldsymbol{Y}$ in this function, this becomes then a random variable which will be denoted by $\boldsymbol{E}\{f(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{Y}\}$. When using vectors and matrices in a linear–algebraic format, n–dimensional vectors, like $\boldsymbol{x}$ (and $\boldsymbol{X}$), will be understood as column vectors, the operator $(\cdot)^T$ will denote vector or matrix transposition, and so, $\boldsymbol{x}^T$ would be a row vector. For two positive sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ means equivalence in the exponential order, i.e., $\lim_{n\to\infty}\frac{1}{n}\log(a_n/b_n)=0$. Finally, the indicator function of an event $\mathcal{A}$ will be denoted by $1\{\mathcal{A}\}$. I.e., $1\{\mathcal{A}\}=1$ is $\mathcal{A}$ occurs, and $1\{\mathcal{Z}\}=0$ if not.

3 MMSE Estimation Relations

This section consists of two subsections. In the first, we derive the main basic relations and in the second, we show how to extend the scope to the case of mismatched estimation.

3.1 Basic Relations

Let $\boldsymbol{X}=(X_1,\ldots,X_n)$, and $\boldsymbol{Y}=(Y_1,\ldots,Y_m)$ (n and m being positive integers), be two random vectors, jointly distributed according to a given probability function $P(\boldsymbol{x},\boldsymbol{y})$. It is further assumed that the alphabet $\mathcal{X}$, of each component of $\boldsymbol{X}$, consists of a set of real valued numbers, i.e., $\mathcal{X}\subseteq\mathbb{R}$. This assumption is obviously necessary in order to make the problem of estimating $\boldsymbol{X}$, in the MSE sense, a meaningful problem. The conditional mean estimator of $\boldsymbol{X}$ based on $\boldsymbol{Y}$, i.e., $\hat{\boldsymbol{X}}=\boldsymbol{E}\{\boldsymbol{X}|\boldsymbol{Y}\}$ is well–known to be the optimum estimator in the MSE sense, i.e., it minimizes the MSE $\boldsymbol{E}\{(X_i-\hat{X}_i)^2\}$ for all $i=1,2,\ldots,n$. The MMSE in estimating X_i is then $\boldsymbol{E}\{(X_i-\boldsymbol{E}\{X_i|\boldsymbol{Y}\})^2\}$, i.e., the expected conditional variance of X_i given $\boldsymbol{Y}$. More generally, the MMSE error covariance matrix E is an $n\times n$ matrix whose (i,j)–th element is given by $\boldsymbol{E}\{(X_i-\boldsymbol{E}\{X_i|\boldsymbol{Y}\})(X_j-\boldsymbol{E}\{X_j|\boldsymbol{Y}\})\}$. This matrix can be represented as the expectation (w.r.t. $\boldsymbol{Y}$) of the conditional covariance matrix of $\boldsymbol{X}$ given $\boldsymbol{Y}$, henceforth denoted $\mathrm{Cov}\{\boldsymbol{X}|\boldsymbol{Y}\}$. I.e.,

$$E=\boldsymbol{E}\{\mathrm{Cov}\{\boldsymbol{X}|\boldsymbol{Y}\}\}=\boldsymbol{E}\{\boldsymbol{X}\boldsymbol{X}^T\}-\boldsymbol{E}\{\boldsymbol{E}\{\boldsymbol{X}|\boldsymbol{Y}\}\cdot\boldsymbol{E}\{\boldsymbol{X}^T|\boldsymbol{Y}\}\}.$$

Defining a column vector of n real valued parameters, $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_n)^T$, consider the following

function:

$$Z(\boldsymbol{y},\boldsymbol{\lambda}) \stackrel{\Delta}{=} \sum_{\boldsymbol{x}\in\mathcal{X}^n} \exp\{\boldsymbol{\lambda}^T\boldsymbol{x}\}P(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{X}^n} \exp\{\boldsymbol{\lambda}^T\boldsymbol{x}\}P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}),$$

where it is assumed that the sum (or integral, in the continuous case) converges uniformly at least in some neighborhood of $\boldsymbol{\lambda} = 0$.[1] It is straightforward to see now that:

$$\left.\frac{\partial \ln Z(\boldsymbol{y}|\boldsymbol{\lambda})}{\partial\lambda_i}\right|_{\boldsymbol{\lambda}=0} = \frac{\sum_{\boldsymbol{x}\in\mathcal{X}^n} x_i P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{y})} = \sum_{x_i\in\mathcal{X}} x_i P(x_i|\boldsymbol{y}) = \boldsymbol{E}\{X_i|\boldsymbol{y}\}, \tag{1}$$

i.e.,

$$\boldsymbol{E}\{\boldsymbol{X}|\boldsymbol{y}\} = \nabla_{\boldsymbol{\lambda}} \ln Z(\boldsymbol{y},\boldsymbol{\lambda}), \tag{2}$$

where $\nabla_{\boldsymbol{\lambda}}$ denotes the gradient w.r.t. $\boldsymbol{\lambda}$. Similarly, upon taking second order derivatives, one obtains

$$\left.\frac{\partial^2 \ln Z(\boldsymbol{y}|\boldsymbol{\lambda})}{\partial\lambda_i\partial\lambda_i}\right|_{\boldsymbol{\lambda}=0} = E\{X_iX_j|\boldsymbol{y}\} - \boldsymbol{E}\{X_i|\boldsymbol{y}\}\cdot\boldsymbol{E}\{X_i|\boldsymbol{y}\} = \mathrm{Cov}\{X_i,X_j|\boldsymbol{y}\},$$

and so,

$$E = \boldsymbol{E}\left\{\nabla^2_{\boldsymbol{\lambda}} \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=0}\right\}, \tag{3}$$

where ∇^2_λ is the Hessian w.r.t. $\boldsymbol{\lambda}$, namely, the matrix of second order derivatives w.r.t. pairs of components of $\boldsymbol{\lambda}$. Note that here and throughout the sequel, we will always refer to gradients and Hessians of functions w.r.t. $\boldsymbol{\lambda}$, computed at the point $\boldsymbol{\lambda} = 0$. It will therefore be convenient to use, for a generic function g, the shorthand notations $\nabla_0 g(\boldsymbol{\lambda})$ and $\nabla^2_0 g(\boldsymbol{\lambda})$ to designate $\nabla_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=0}$ and $\nabla^2_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=0}$, respectively.

Another, perhaps simpler, way to look at the relations (2) and (3) is the following: Obviously, for a given $\boldsymbol{y}$, $M(\boldsymbol{y},\boldsymbol{\lambda}) = \sum_{\boldsymbol{x}} e^{\boldsymbol{\lambda}^T\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{y})$ is the moment generating function pertaining to the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ and so, its derivatives relative to $\{\lambda_i\}$, computed at $\boldsymbol{\lambda} = 0$, yield the conditional moments $\boldsymbol{E}\{X_i|\boldsymbol{y}\}$, $\boldsymbol{E}\{X_i^2|\boldsymbol{y}\}$, $\boldsymbol{E}\{X_iX_j|\boldsymbol{y}\}$, etc. Therefore, $\ln M(\boldsymbol{y},\boldsymbol{\lambda})$ is a generator of the corresponding conditional cumulants, $\boldsymbol{E}\{X_i|\boldsymbol{y}\}$, $\mathrm{Var}\{X_i|\boldsymbol{y}\}$, $\mathrm{Cov}\{X_i,X_j|\boldsymbol{y}\}$, etc. Now, observe that $\ln M(\boldsymbol{y},\boldsymbol{\lambda})$ differs from $\ln Z(\boldsymbol{y},\boldsymbol{\lambda})$ merely by the additive term $\ln P(\boldsymbol{y})$, which does not depend on $\boldsymbol{\lambda}$ anyway and hence does not affect the gradient and Hessian w.r.t. $\boldsymbol{\lambda}$. Therefore, $\ln Z(\boldsymbol{y},\boldsymbol{\lambda})$ is a generator of conditional cumulants, exactly like $\ln M(\boldsymbol{y},\boldsymbol{\lambda})$. An important point, however, is that we prefer $\ln Z(\boldsymbol{y},\boldsymbol{\lambda})$ over $\ln M(\boldsymbol{y},\boldsymbol{\lambda})$ because normally, it is more convenient to

[1] If this assumption is not met, one can instead, parametrize each component λ_i of $\boldsymbol{\lambda}$ as a purely imaginary number $\lambda_i = j\omega_i$ ($j = \sqrt{-1}$), as is done in the definition of the characteristic function.

work with the joint distribution $P(\boldsymbol{x},\boldsymbol{y})$ (or equivalently, with the source $P(\boldsymbol{x})$ and forward channel $P(\boldsymbol{y}|\boldsymbol{x})$) rather than with the backward channel (or the posterior) $P(\boldsymbol{x}|\boldsymbol{y})$.[2]

We next derive several alternative versions of this relation between the error covariance matrix of the MMSE estimator and derivatives of $\ln Z$. First, observe that $Z(\boldsymbol{y},\boldsymbol{\lambda})$ is proportional to $P_{\boldsymbol{\lambda}}(\boldsymbol{y})\cdot\Theta(\boldsymbol{\lambda})$, where

$$\Theta(\boldsymbol{\lambda}) = \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})\exp\{\boldsymbol{\lambda}^T\boldsymbol{x}\}$$

and $P_{\boldsymbol{\lambda}}(\boldsymbol{y})$ is the output marginal of $\boldsymbol{y}$ induced by the channel $P(\boldsymbol{y}|\boldsymbol{x})$ and the modified source distribution $P_{\boldsymbol{\lambda}}(\boldsymbol{x}) \stackrel{\Delta}{=} e^{\boldsymbol{\lambda}^T\boldsymbol{x}}P(\boldsymbol{x})/\Theta(\boldsymbol{\lambda})$. We therefore obtain

$$\begin{aligned} E &= \boldsymbol{E}\left\{\nabla_0^2 \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})\right\} \\ &= \boldsymbol{E}\left\{\nabla_0^2 \ln[P_{\boldsymbol{\lambda}}(\boldsymbol{Y})\cdot\Theta(\boldsymbol{\lambda})]\right\} \\ &= \nabla_0^2 \ln\Theta(\boldsymbol{\lambda}) + \boldsymbol{E}\left\{\nabla_0^2 \ln P_{\boldsymbol{\lambda}}(\boldsymbol{Y})\right\} \\ &= \mathrm{Cov}\{\boldsymbol{X}\} - J, \end{aligned} \tag{4}$$

where $\mathrm{Cov}\{\boldsymbol{X}\} = \boldsymbol{E}\{\boldsymbol{X}\boldsymbol{X}^T\} - \boldsymbol{E}\{\boldsymbol{X}\}\cdot\boldsymbol{E}\{\boldsymbol{X}^T\}$ is the covariance matrix of $\boldsymbol{X}$ and J is the Fisher information matrix of estimating $\boldsymbol{\lambda}$ based on $\boldsymbol{Y}$, computed at the point $\boldsymbol{\lambda}=0$. The Fisher information matrix J can also be expressed as

$$J = \boldsymbol{E}\left\{\nabla_0 \ln P_{\boldsymbol{\lambda}}(\boldsymbol{Y})\cdot\nabla_0^T \ln P(\boldsymbol{Y}|\boldsymbol{\lambda})\right\}.$$

Equivalently, we obtained

$$J = \mathrm{Cov}\{\boldsymbol{X}\} - E = \boldsymbol{E}\{\boldsymbol{E}\{\boldsymbol{X}|\boldsymbol{Y}\}\cdot\boldsymbol{E}\{\boldsymbol{X}^T|\boldsymbol{Y}\}\}.$$

Note that J can also be obtained as the negative expectation of the Hessian (or, equivalently, as the covariance matrix of the gradient) of the *information density* [14],

$$i_{\boldsymbol{\lambda}}(\boldsymbol{x};\boldsymbol{y}) = \ln[P(\boldsymbol{y}|\boldsymbol{x})/P_{\boldsymbol{\lambda}}(\boldsymbol{y})],$$

[2] As a side remark, we shall mention also the physical perspective: if $Z(\boldsymbol{y},\boldsymbol{\lambda})$ is thought of as the partition function of a certain statistical–mechanical model (as discussed in the Introduction), where the components of $\boldsymbol{\lambda}$ are thought of as certain generalized forces or fields that are acting on the individual particles, then the above relation between the second order derivative of $\ln Z(\boldsymbol{y},\boldsymbol{\lambda})$ w.r.t. λ_i and λ_j and the (conditional) covariances between the corresponding state variables, X_i and X_j, is known as one of the versions of the fluctuation–dissipation theorem in statistical mechanics [10, p. 32, eq. (2.44)], which relates between the linear response of the system (to an infinitesimally small perturbation in its parameters) and its fluctuations in equilibrium.

which is again, computed at $\boldsymbol{\lambda} = 0$.

Sometimes it is more convenient to square the first derivative of $\ln Z$ than to take the second derivative. In these cases, the following relationship may be useful:

$$\begin{aligned}
\Xi &\triangleq \boldsymbol{E}\left\{[\nabla_0 \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})]\cdot[\nabla_0 \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})]^T\right\} \\
&= \boldsymbol{E}\left\{[\nabla_0 \ln\{P(\boldsymbol{Y}|\boldsymbol{\lambda})\cdot\Theta(\boldsymbol{\lambda})\}]\cdot[\nabla_0 \ln\{P(\boldsymbol{Y}|\boldsymbol{\lambda})\cdot\Theta(\boldsymbol{\lambda})\}]^T\right\} \\
&= \boldsymbol{E}\left\{[\nabla_0 \ln P(\boldsymbol{Y}|\boldsymbol{\lambda})]\cdot[\nabla_0 \ln P(\boldsymbol{Y}|\boldsymbol{\lambda})]^T\right\} + [\nabla_0 \ln\Theta(\boldsymbol{\lambda})]\cdot[\nabla_0 \ln\Theta(\boldsymbol{\lambda})]^T \\
&= J + \boldsymbol{E}\{\boldsymbol{X}\}\cdot\boldsymbol{E}\{\boldsymbol{X}^T\} \\
&= \mathrm{Cov}\{\boldsymbol{X}\} + \boldsymbol{E}\{\boldsymbol{X}\}\cdot\boldsymbol{E}\{\boldsymbol{X}^T\} - E \\
&= \boldsymbol{E}\{\boldsymbol{X}\boldsymbol{X}^T\} - E
\end{aligned} \tag{5}$$

and so,

$$E = \boldsymbol{E}\{\boldsymbol{X}\boldsymbol{X}^T\} - \Xi.$$

Particularizing these results to the MMSE,

$$\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y}) \triangleq \sum_{i=1}^{n} \boldsymbol{E}\{(X_i - \boldsymbol{E}\{X_i|\boldsymbol{Y}\})^2\},$$

which is the trace of E, we have the following relations, which we formulate as a proposition.

Proposition 1. The following formulas for the MMSE hold:

$$\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = \sum_{i=1}^{n} \boldsymbol{E}\left\{\frac{\partial^2 \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i^2}\Big|_{\boldsymbol{\lambda}=0}\right\} \tag{6}$$

$$= \sum_{i=1}^{n}\left[\mathrm{Var}\{X_i\} + \boldsymbol{E}\left\{\frac{\partial^2 \ln P(\boldsymbol{Y}|\boldsymbol{\lambda})}{\partial\lambda_i^2}\Big|_{\boldsymbol{\lambda}=0}\right\}\right] \tag{7}$$

$$= \sum_{i=1}^{n}\left[\mathrm{Var}\{X_i\} - \boldsymbol{E}\left\{\left[\frac{\partial \ln P(\boldsymbol{Y}|\boldsymbol{\lambda})}{\partial\lambda_i}\right]^2\Big|_{\boldsymbol{\lambda}=0}\right\}\right] \tag{8}$$

$$= \sum_{i=1}^{n}\left[\boldsymbol{E}\{X_i^2\} - \boldsymbol{E}\left\{\left[\frac{\partial \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i}\right]^2\Big|_{\boldsymbol{\lambda}=0}\right\}\right] \tag{9}$$

In the second and the third formulas, $\ln P(\boldsymbol{Y}|\boldsymbol{\lambda})$ can be replaced by $\ln i(\boldsymbol{X};\boldsymbol{Y})$, thus relating the MMSE to the information density.

3.2 Extension to the Mismatched Case

In this short subsection, we are outlining how our approach can easily be extended to handle situations of mismatched estimation. Consider a mismatched estimator which is the conditional mean of $\boldsymbol{X}$ given $\boldsymbol{Y}$, based on an incorrect joint distribution $Q(\boldsymbol{x},\boldsymbol{y})$, whereas the true joint distribution continues to be $P(\boldsymbol{x},\boldsymbol{y})$. Denoting by $Z_P(\boldsymbol{y},\boldsymbol{\lambda})$ and $Z_Q(\boldsymbol{y},\boldsymbol{\lambda})$ the corresponding partition functions, and by $\boldsymbol{E}_P$ and $\boldsymbol{E}_Q$, the corresponding expectations, our approach can easily be generalized to handle this case as follows:

$$\begin{aligned}
E &= \boldsymbol{E}_P\Big\{(\boldsymbol{X}-\boldsymbol{E}_Q\{\boldsymbol{X}|\boldsymbol{Y}\})(\boldsymbol{X}^T-\boldsymbol{E}_Q\{\boldsymbol{X}^T|\boldsymbol{Y}\})\Big\} \\
&= \boldsymbol{E}_P\{\boldsymbol{X}\boldsymbol{X}^T\}-\boldsymbol{E}_P\{\boldsymbol{E}_P\{\boldsymbol{X}|\boldsymbol{Y}\}\boldsymbol{E}_Q\{\boldsymbol{X}^T|\boldsymbol{Y}\}\}- \\
&\quad \boldsymbol{E}_P\{\boldsymbol{E}_Q\{\boldsymbol{X}|\boldsymbol{Y}\}\boldsymbol{E}_P\{\boldsymbol{X}^T|\boldsymbol{Y}\}\}+\boldsymbol{E}_P\{\boldsymbol{E}_Q\{\boldsymbol{X}|\boldsymbol{Y}\}\boldsymbol{E}_Q\{\boldsymbol{X}^T|\boldsymbol{Y}\}\} \\
&= \boldsymbol{E}_P\{\boldsymbol{X}\boldsymbol{X}^T\}-\boldsymbol{E}_P\{[\nabla_0 \ln Z_P(\boldsymbol{Y},\boldsymbol{\lambda})]\cdot[\nabla_0 \ln Z_Q(\boldsymbol{Y},\boldsymbol{\lambda})]^T\}- \\
&\quad \boldsymbol{E}_P\{[\nabla_0 \ln Z_Q(\boldsymbol{Y},\boldsymbol{\lambda})]\cdot[\nabla_0 \ln Z_P(\boldsymbol{Y},\boldsymbol{\lambda})]^T\}+\boldsymbol{E}_P\{[\nabla_0 \ln Z_Q(\boldsymbol{Y},\boldsymbol{\lambda})]\cdot[\nabla_0 \ln Z_Q(\boldsymbol{Y},\boldsymbol{\lambda})]^T\}.
\end{aligned}$$

Thus, in particular, the MSE associated with the mismatched estimator is given by

$$\begin{aligned}
\mathrm{mse}_Q(\boldsymbol{X}|\boldsymbol{Y}) &= \sum_{i=1}^{n}\Bigg[\boldsymbol{E}_P\{X_i^2\}-2\boldsymbol{E}_P\left\{\frac{\partial \ln Z_P(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i}\Big|_{\boldsymbol{\lambda}=0}\cdot\frac{\partial \ln Z_Q(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i}\Big|_{\boldsymbol{\lambda}=0}\right\} \\
&\quad +\boldsymbol{E}_P\left\{\left[\frac{\partial \ln Z_Q(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i}\Big|_{\boldsymbol{\lambda}=0}\right]^2\right\}\Bigg]. \qquad (10)
\end{aligned}$$

4 Examples

In this section, we provide three examples, where we show how the log–partition function, $\ln Z(\boldsymbol{y},\boldsymbol{\lambda})$, can be evaluated for large n, using methods of statistical mechanics. Using the relations derived in Subsection 3.1, we then show how the conditional mean estimator and the MMSE can be approximated for large n.

4.1 Example 1 – A Codeword Transmitted Over an AWGN

Our first example is taken from [9, Subsection 5.2], but here we demonstrate how to derive the conditional mean estimator and the MMSE using Proposition 1, rather than the I–MMSE relation. For the sake of completeness and convenience, we provide here the full necessary details (with

the appropriate modifications to accommodate the method proposed herein), including those that already appear in [9]. As noted in [9], the analysis of this model is intimately related to one of the statistical mechanical techniques used in the analysis of the so called random energy model (REM) of disordered magnetic materials, a.k.a. spin glasses in the statistical physics literature (see references in [9]).

Let $\boldsymbol{X}$ be chosen uniformly at random from a codebook $\mathcal{C} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M-1}\}$ of size $M = e^{nR}$. The codebook itself is also selected at random (and then revealed to the estimator) in the following manner: Each $\boldsymbol{x}_i$ is selected independently and uniformly at random from the surface of a sphere of radius $\sqrt{nP_x}$ centered at the origin. The channel $P(\boldsymbol{y}|\boldsymbol{x})$ is an AWGN channel (hence $m = n$) whose noise variance is $1/\beta$ (keeping the same notation as in [9]). I.e.,

$$P(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{\beta}{2\pi}\right)^{n/2} \exp\left\{\frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2\right\}.$$

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

$$\begin{aligned} Z(\boldsymbol{y}, \lambda) &= \sum_{\boldsymbol{x}\in\mathcal{C}} e^{-nR} \exp\{-\beta\|\boldsymbol{y} - \boldsymbol{x}\|^2/2 + \boldsymbol{\lambda}^T\boldsymbol{x}\} \\ &= e^{-nR} \exp[-\beta\|\boldsymbol{y} - \boldsymbol{x}_0\|^2/2 + \boldsymbol{\lambda}^T\boldsymbol{x}_0] + \sum_{\boldsymbol{x}\in\mathcal{C}\setminus\{\boldsymbol{x}_0\}} e^{-nR} \exp[-\beta\|\boldsymbol{y} - \boldsymbol{x}\|^2/2 + \boldsymbol{\lambda}^T\boldsymbol{x}] \\ &\stackrel{\Delta}{=} Z_c(\boldsymbol{y}, \boldsymbol{\lambda}) + Z_e(\boldsymbol{y}, \boldsymbol{\lambda}), \end{aligned} \tag{11}$$

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

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

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

$$\overline{N(\epsilon)} = \sum_{i=1}^{M} \Pr\{\|\boldsymbol{y} - \boldsymbol{X}_i\|^2/2 - \boldsymbol{\lambda}^T\boldsymbol{X}_i/\beta \approx n\epsilon\} = e^{nR}\Pr\{\|\boldsymbol{y} - \boldsymbol{X}_1\|^2/2 - \boldsymbol{\lambda}^T\boldsymbol{X}_1/\beta \approx n\epsilon\}. \tag{12}$$

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

$$\rho(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\Delta}{=} \frac{\boldsymbol{x}^T(\boldsymbol{y} + \lambda/\beta)}{n\sqrt{P_x P_y'}} \approx \frac{\frac{1}{2}(P_x + P_y) - \epsilon}{\sqrt{P_x P_y'}} \stackrel{\Delta}{=} \frac{P_a - \epsilon}{P_g'},$$

where have defined $P_a = (P_x + P_y)/2$ and $P'_g = \sqrt{P_x P'_y}$, where $P'_y = \frac{1}{n}\sum_i (y_i + \lambda_i/\beta)^2$. The probability that a randomly chosen vector $\boldsymbol{X}$ on the sphere would have an empirical correlation coefficient ρ with a given vector $\boldsymbol{y}' = \boldsymbol{y} + \boldsymbol{\lambda}/\beta$ (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

$$\Pr\{\|\boldsymbol{y} - \boldsymbol{X}_1\|^2/2 - \boldsymbol{\lambda}^T\boldsymbol{X}_1/\beta \approx n\epsilon\} \doteq \exp\left\{n\,\Gamma\left(\frac{P_a-\epsilon}{P'_g}\right)\right\}.$$

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

$$\epsilon_1 \triangleq P_a - P'_g\sqrt{1-e^{-2R}}$$

and

$$\epsilon_2 \triangleq P_a + P'_g\sqrt{1-e^{-2R}},$$

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, it follows that for a typical realization of the randomly chosen code, we have

$$\begin{aligned} Z_e(\boldsymbol{y},\boldsymbol{\lambda}) &\doteq e^{-nR}\max_{\epsilon\in[\epsilon_1,\epsilon_2]}\exp\left\{n\left[R+\Gamma\left(\frac{P_a-\epsilon}{P'_g}\right)-\beta\epsilon\right]\right\} \\ &= \max_{\epsilon\in[\epsilon_1,\epsilon_2]}\exp\left\{n\left[\Gamma\left(\frac{P_a-\epsilon}{P'_g}\right)-\beta\epsilon\right]\right\} \\ &= \exp\left\{n\left[\max_{|\rho|\le\rho_*}\left\{\frac{1}{2}\ln(1-\rho^2)-\beta(P_a-\rho P'_g)\right\}\right]\right\}. \end{aligned} \tag{13}$$

The derivative of $\frac{1}{2}\ln(1-\rho^2)+\rho\beta P'_g$ w.r.t. ρ vanishes within $[-1,1]$ at:

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

where

$$\theta\stackrel{\Delta}{=}\frac{1}{2\beta P'_g}.$$

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 $P'_g=\sqrt{P_x(P_x+1/\beta)}$ ($\|\boldsymbol{\lambda}\|$ is small), is equivalent to $\beta<\beta_R\stackrel{\Delta}{=}(e^{2R}-1)/P_x$. Thus, for the typical code we have

$$Z_e(\beta|\boldsymbol{y})\doteq\begin{cases}\exp\left\{n\left[\frac{1}{2}\ln(1-\rho_\beta^2)-\beta(P_a-\rho_\beta P'_g)\right]\right\}, & \beta<\beta_R\\ \exp\{-n[R+\beta(P_a-\rho_* P'_g)]\}, & \beta\ge\beta_R.\end{cases}$$

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

$$Z(\boldsymbol{y},\boldsymbol{\lambda})\doteq\begin{cases}\exp\left\{n\left[\frac{1}{2}\ln(1-\rho_\beta^2)-\beta(P_a-\rho_\beta P'_g)\right]\right\}, & \beta<\beta_R\\ \exp\left\{-n\left(R+\frac{1}{2}\right)+\boldsymbol{\lambda}^T\boldsymbol{x}_0\right\}, & \beta\ge\beta_R.\end{cases}$$

A very similar analysis applies also to the derivative $\frac{\partial}{\partial\lambda_i}\ln Z(\boldsymbol{y},\boldsymbol{\lambda})$, which is essentially a weighted average of x_i with weights proportional to $\overline{N(\epsilon)}e^{-\beta\epsilon}$ for all $\epsilon\in[\epsilon_1,\epsilon_2]$. Thus, the exponentially dominant weight is due to the term that maximizes the exponent. Assuming that the correct codeword $\boldsymbol{x}_0$ is dominant ($Z_c>>Z_e$, which is the case when $R<C$), this weighted average is obviously dominated by the i–th component of $\boldsymbol{x}_0$, in which case the MMSE essentially vanishes. Otherwise, for $R>C$, Z_e dominates the partition function and the weighted average is overwhelmingly dominated by the term corresponding to the maximizing ϵ, or equivalently, the maximizing ρ, which is ρ_β. This means that the conditional mean estimator of X_i is approximately given by:

$$\begin{aligned}\boldsymbol{E}\{X_i|\boldsymbol{y}\} &\approx \frac{\partial}{\partial\lambda_i}\left[\frac{n}{2}\ln(1-\rho_\beta^2)+\beta\rho_\beta n P'_g\right]\\ &= -\frac{n\rho_\beta}{1-\rho_\beta^2}\frac{\partial\rho_\beta}{\partial\lambda_i}+\beta n P'_g\frac{\partial\rho_\beta}{\partial\lambda_i}+\beta\rho_\beta n\frac{\partial P'_g}{\partial\lambda_i}\\ &= n\frac{\partial\rho_\beta}{\partial\lambda_i}\left(-\frac{\rho_\beta}{1-\rho_\beta^2}+\beta P'_g\right)+\beta n\rho_\beta\frac{\partial P'_g}{\partial\lambda_i}\\ &= n\beta\rho_\beta\frac{\partial P'_g}{\partial\lambda_i}\\ &= \beta\rho_\beta n\cdot\frac{\sqrt{P_x}}{2\sqrt{P_y}}\cdot\frac{2y_i}{\beta n}\end{aligned}$$

$$
\begin{aligned}
&= \rho_\beta \sqrt{\frac{P_x}{P_x + 1/\beta}} \cdot y_i \\
&= \frac{P_x}{P_x + 1/\beta} \cdot y_i, \qquad (14)
\end{aligned}
$$

where in the last step we have used the identity $\rho_\beta = \sqrt{P_x/(P_x + 1/\beta)}$, which can easily be verified. This is simply the linear Wiener estimator that would have been applied had the input been zero–mean, i.i.d. Gaussian, with variance $1/\beta$ (see also [9]). According to Proposition 1, the MMSE associated with X_i is given by

$$
\boldsymbol{E}\{(X_i - \boldsymbol{E}\{X_i|\boldsymbol{Y}\})^2\} \approx P_x - \boldsymbol{E}\{\boldsymbol{E}^2(X_i|\boldsymbol{Y})\} = P_x - \left(\frac{P_x}{P_x + 1/\beta}\right)^2 \cdot (P_x + 1/\beta) = \frac{P_x}{1 + \beta P_x},
$$

as expected.

4.2 Example 2 – The Curie–Weiss Model

Consider a binary source

$$
P(\boldsymbol{x}) = C_n \exp\left\{\frac{a}{2n}\left(\sum_{i=1}^n x_i\right)^2 + b\sum_{i=1}^n x_i\right\} \qquad \boldsymbol{x} \in \{-1, +1\}^n
$$

where a and b are parameters and C_n is a normalization constant, which is immaterial for our purposes (as it is going to disappear upon taking derivatives w.r.t. $\{\lambda_i\}$, and the same comment applies to the constants C_n' and C_n'' below). Let the channel be binary and symmetric, i.e.,

$$
P(y|x) = \frac{e^{\beta xy}}{2\cosh(\beta)}, \qquad y \in \{-1, +1\}.
$$

Then, the partition function $Z(\boldsymbol{y}, \boldsymbol{\lambda})$ can be represented as a one–dimensional integral using the Hubbard–Stratonovich transform, which in turn can be assessed using saddle point methods, as is frequently done in the statistical physics literature. Specifically, we have the following:

$$
\begin{aligned}
Z(\boldsymbol{y}, \boldsymbol{\lambda}) &= C_n' \sum_{\boldsymbol{x}} \exp\left\{\frac{a}{2n}\left(\sum_{i=1}^n x_i\right)^2 + b\sum_{i=1}^n x_i + \beta\sum_{i=1}^n x_i y_i + \sum_{i=1}^n \lambda_i x_i\right\} \qquad (15) \\
&= C_n' \sum_{\boldsymbol{x}} \exp\left\{\sum_{i=1}^n x_i(\beta y_i + \lambda_i + b) + \frac{a}{2n}\left(\sum_{i=1}^n x_i\right)^2\right\} \qquad (16) \\
&= C_n'' \sum_{\boldsymbol{x}} \exp\left\{\sum_{i=1}^n x_i(\beta y_i + \lambda_i + b)\right\} \cdot \int_{-\infty}^{+\infty} \mathrm{d}\theta \exp\left\{-\frac{n\theta^2}{2a} + \theta\sum_{i=1}^n x_i\right\} \qquad (17)
\end{aligned}
$$

$$= C_n'' \int_{-\infty}^{+\infty} \mathrm{d}\theta e^{-n\theta^2/(2a)} \sum_{\boldsymbol{x}} \exp\left\{\sum_{i=1}^{n} x_i(\beta y_i + \lambda_i + b + \theta)\right\} \tag{18}$$

$$= C_n'' \int_{-\infty}^{+\infty} \mathrm{d}\theta e^{-n\theta^2/(2a)} \prod_{i=1}^{n} [2\cosh(\beta y_i + \lambda_i + b + \theta)] \tag{19}$$

$$= 2^n C_n'' \int_{-\infty}^{+\infty} \mathrm{d}\theta \exp\left\{-\frac{n\theta^2}{2a} + \sum_{i=1}^{n} \ln\cosh(\beta y_i + \lambda_i + b + \theta)\right\}. \tag{20}$$

Thus,

$$\frac{\partial \ln Z(\boldsymbol{y}, \boldsymbol{\lambda})}{\partial \lambda_i} = \frac{\int_{-\infty}^{+\infty} \mathrm{d}\theta \tanh(\beta y_i + \lambda_i + b + \theta) \exp\left\{-\frac{n\theta^2}{2a} + \sum_{i=1}^{n} \ln\cosh(\beta y_i + \lambda_i + b + \theta)\right\}}{\int_{-\infty}^{+\infty} \mathrm{d}\theta \exp\left\{-\frac{n\theta^2}{2a} + \sum_{i=1}^{n} \ln\cosh(\beta y_i + \lambda_i + b + \theta)\right\}}$$

$$\approx \tanh(\beta y_i + \lambda_i + b + \theta_*), \tag{21}$$

where θ_* is the maximizer of the expression at the exponent, i.e., it is the solution to the zero–derivative equation:

$$\theta = \frac{a}{n} \sum_{i=1}^{n} \tanh(\beta y_i + \lambda_i + b + \theta).$$

Thus, the MMSE estimator is:

$$\boldsymbol{E}\{X_i|\boldsymbol{y}\} = \frac{\int_{-\infty}^{+\infty} \mathrm{d}\theta \tanh(\beta y_i + b + \theta) \exp\left\{-\frac{n\theta^2}{2a} + \sum_{i=1}^{n} \ln\cosh(\beta y_i + b + \theta)\right\}}{\int_{-\infty}^{+\infty} \mathrm{d}\theta \exp\left\{-\frac{n\theta^2}{2a} + \sum_{i=1}^{n} \ln\cosh(\beta y_i + b + \theta)\right\}} \tag{22}$$

$$\approx \tanh(\beta y_i + b + \theta_*), \tag{23}$$

where now θ_* is understood to be taken with $\boldsymbol{\lambda} = 0$. For $b \neq 0$, the asymptotic MMSE is then given by

$$\lim_{n\to\infty} \frac{\mathrm{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n} = 1 - \boldsymbol{E}\{\tanh^2(\beta Y + b + \theta_0)\},$$

where θ_0 is the solution to the equation

$$\theta = a\boldsymbol{E}\{\tanh(\beta Y + b + \theta)\},$$

and where Y is a binary $\{\pm 1\}$ RV, with mean $m^* \tanh(\beta)$, m^* being the dominant solution to the equation $m = \tanh(am + b)$, i.e., the maximizer of $h_2((1+m)/2) + am^2/2 + bm$, where $h_2(\cdot)$ is the binary entropy function. When $b = 0$, θ_0 becomes a random variable which takes on, with equal probabilities, one of two values, each one being the solution to the above displayed equation, except that in one of them Y has mean $m^* \tanh(\beta)$ and in the other, its mean is $-m^* \tanh(\beta)$.

This calculation is intimately related to the Curie–Weiss model of magnetic spins [10, Subsection 2.5.2, pp. 40–44], where the parameter m plays the role of magnetization.

4.3 Example 3 – The Generalized Multivariate Cauchy Noise Model

Let $X_i \sim \mathcal{N}(0,\sigma^2)$ be i.i.d. RV's, and let the additive noise have a generalized multivariate Cauchy distribution, i.e.,

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{C_{n,k}}{[1+(\boldsymbol{y}-\boldsymbol{x})^T S(\boldsymbol{y}-\boldsymbol{x})]^k}$$

where $C_{n,k}$ is a normalization constant, S is a positive definite matrix, and $k>0$ is chosen large enough (as a function of n) such $\int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{z}/[1+\boldsymbol{z}^T S\boldsymbol{z}]^k < \infty$, i.e., $k > n/2$. The choice $k=(n+1)/2$ corresponds to the ordinary multivariate Cauchy distribution. Here, however, we will require moreover that k is even large enough such that the second moments exist, i.e., $\int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{z} \cdot \boldsymbol{z}^T\boldsymbol{z}/[1+\boldsymbol{z}^T S\boldsymbol{z}]^k < \infty$, which means $k > n/2+1$. For simplicity, we will take S to be the identity matrix. However, our analysis easily extends to a general positive matrix S, as well as to a general Gaussian vector $\boldsymbol{X}$, not necessarily with i.i.d. components. Using the Laplace transform identity $\int_0^\infty \mathrm{d}t \cdot t^{k-1}e^{-st} = \Gamma(k)/s^k$, we have:

$$\begin{aligned}
Z(\boldsymbol{y},\boldsymbol{\lambda}) &= \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) e^{\boldsymbol{\lambda}^T\boldsymbol{x}} \cdot \frac{C_{n,k}}{[1+\sum_{i=1}^n (y_i-x_i)^2]^k} && (24)\\
&= C_{n,k}\int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) e^{\boldsymbol{\lambda}^T\boldsymbol{x}} \int_0^\infty \mathrm{d}t \cdot \frac{t^{k-1}}{\Gamma(k)} \cdot e^{-t[1+\sum_i (y_i-x_i)^2]} && (25)\\
&= C'_{n,k}\int_0^\infty \mathrm{d}t \cdot t^{k-1}e^{-t} \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) e^{\boldsymbol{\lambda}^T\boldsymbol{x}} \cdot e^{-t\sum_i (y_i-x_i)^2} && (26)\\
&= C''_{n,k}\int_0^\infty \mathrm{d}t \cdot t^{k-1}e^{-t} \prod_{i=1}^n \int_{\mathbb{R}} \mathrm{d}x_i e^{-x_i^2/2\sigma^2} e^{\lambda_i x_i} \cdot e^{-t(y_i-x_i)^2} && (27)\\
&= C''_{n,k}\int_0^\infty \mathrm{d}t \cdot t^{k-1}e^{-t} \left(\frac{2\pi\sigma^2}{1+2t\sigma^2}\right)^{n/2} \cdot \exp\left\{-t\sum_{i=1}^n y_i^2 + \sum_{i=1}^n \frac{(ty_i+\lambda_i/2)^2}{t+1/2\sigma^2}\right\}. && (28)
\end{aligned}$$

and so,

$$\left.\frac{\partial \ln Z(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \lambda_i}\right|_{\boldsymbol{\lambda}=0} = \frac{\int_0^\infty \mathrm{d}t \cdot \frac{ty_i}{t+1/2\sigma^2} e^{-t}t^{k-1} \exp\left\{-\frac{n}{2}\ln(1+2t\sigma^2) - \frac{t}{1+2t\sigma^2}\sum_i y_i^2\right\}}{\int_0^\infty \mathrm{d}t e^{-t}t^{k-1} \exp\left\{-\frac{n}{2}\ln(1+2t\sigma^2) - \frac{t}{1+2t\sigma^2}\sum_i y_i^2\right\}}$$

which can be approximated by $\hat{t}y_i/(\hat{t}+1/2\sigma^2)$, where $\hat{t}$ is the value of t that dominates the integral, i.e.,

$$\hat{t} = \mathrm{argmax}_t \left[(k-1)\ln t - \frac{n}{2}\ln(1+2t\sigma^2) - \frac{t}{1+2t\sigma^2}\sum_i y_i^2\right].$$

The derivation of the MMSE can be done in a similar manner as in the previous examples.

5 Joint Distributions with Generalized Spherical Symmetry

Examples 2 and 3 of the previous section have one idea in common. In both of them we expressed either the source or the channel as a one–dimensional integral over a variable (t or θ, in those examples), where for each value of this variable, we have a product form measure, which enables, after applying saddle point analysis on this integral, to pass to a closed–form formula, which has the flavor of a single–letter characterization. In this section, we generalize this idea to establish a somewhat more general framework.

Suppose that $m = n$ and the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is of the form

$$P(\boldsymbol{x}, \boldsymbol{y}) = F_n(\sum_i \phi(x_i, y_i)).$$

Let $f_n(t)$ be the inverse Laplace transform of $F_n(s)$. Then, we have

$$\begin{aligned} Z(\boldsymbol{y}, \boldsymbol{\lambda}) &= \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} e^{\boldsymbol{\lambda}^T \boldsymbol{x}} P(\boldsymbol{x}, \boldsymbol{y}) \\ &= \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} e^{\boldsymbol{\lambda}^T \boldsymbol{x}} \int_0^\infty \mathrm{d}t f_n(t) \exp\left\{-t \sum_i \phi(x_i, y_i)\right\} \\ &= \int_0^\infty \mathrm{d}t f_n(t) \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} e^{\boldsymbol{\lambda}^T \boldsymbol{x}} \exp\left\{-t \sum_i \phi(x_i, y_i)\right\} \\ &= \int_0^\infty \mathrm{d}t f_n(t) \prod_i \int_{\mathbb{R}} \mathrm{d}x_i e^{\lambda_i x_i} \exp\{-t\phi(x_i, y_i)\}. \end{aligned} \tag{29}$$

Before proceeding, we should note that by using the Laplace transform, we have essentially represented the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as a mixture of product form measures, indexed by t, each being proportional to $\exp\{-t\sum_i \phi(x_i, y_i)\}$. If we normalize these measures by $Z_t^n = [\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \exp\{-t\phi(x, y)\}]^n$, and define the i.i.d. probability distribution

$$P(\boldsymbol{x}, \boldsymbol{y}|t) = \frac{\exp\{-t\sum_i \phi(x_i, y_i)\}}{Z_t^n}$$

then $P(\boldsymbol{x}, \boldsymbol{y})$ is essentially expressed here as a mixture of i.i.d. probability functions $\{P(\boldsymbol{x}, \boldsymbol{y}|t)\}$, where t can be thought of as a random parameter whose prior is given by $w_n(t) = f_n(t)Z_t^n$. However, it should be kept in mind that this integral representation goes somewhat further than being a mixture of i.i.d. distributions because $f_n(t)$, and hence also $w_n(t)$, may be negative for some ranges of t even when $F_n(s)$ is strictly positive for all s. For example, recall that the inverse Laplace transform of $s^2/(s^2 + \alpha^2)$ is $\sin(\alpha t)$ $(t \geq 0)$, and so, for $F_n(s) = \alpha^n/(s^2 + \alpha^2)^n$, $f_n(t)$ is

given by the n–fold convolution of $\sin(\alpha t)$ with itself. In such cases, $P(\boldsymbol{x}, \boldsymbol{y})$ cannot be considered a mixture of i.i.d. distributions.

Let us now denote

$$\rho(\lambda, y, t) = \ln\left[\int_{-\infty}^{\infty} \mathrm{d}x e^{\lambda x - t\phi(x,y)}\right],$$

$$\rho_0(y, t) = \rho(0, y, t) = \ln\left[\int_{-\infty}^{\infty} \mathrm{d}x e^{-t\phi(x,y)}\right],$$

and

$$\zeta(y, t) = \left.\frac{\partial \rho(\lambda, y, t)}{\partial \lambda}\right|_{\lambda=0} = \frac{\int_{\mathbb{R}} \mathrm{d}x \cdot x e^{-t\phi(x,y)}}{\int_{\mathbb{R}} \mathrm{d}x \cdot e^{-t\phi(x,y)}}.$$

Then,

$$Z(\boldsymbol{y}, \boldsymbol{\lambda}) = \int_0^{\infty} \mathrm{d}t f_n(t) e^{\sum_i \rho(\lambda_i, y_i, t)},$$

and so,

$$\boldsymbol{E}\{X_i|\boldsymbol{y}\} = \left.\frac{\partial \ln Z(\boldsymbol{y}, \boldsymbol{\lambda})}{\partial \lambda_i}\right|_{\boldsymbol{\lambda}=0} = \frac{\int_0^{\infty} \mathrm{d}t f_n(t) \zeta(y_i, t) e^{\sum_i \rho_0(y_i,t)}}{\int_0^{\infty} \mathrm{d}t f_n(t) e^{\sum_i \rho_0(y_i,t)}}$$

which is approximated by $\zeta(y_i, \hat{t})$, where $\hat{t}$ is the maximizer of the expression

$$\ln|f_n(t)| + \sum_i \rho_0(y_i, t).$$

The MMSE of estimating X_i is given by

$$\mathrm{mmse}(X_i|\boldsymbol{Y}) \approx \boldsymbol{E}\{X_i^2\} - \boldsymbol{E}\{\zeta^2(Y_i, t_0(t))\}$$

where the second term is computed as follows:

$$\boldsymbol{E}\{\zeta^2(Y_i, t_0(t))\} = \int_0^{\infty} \mathrm{d}t w_n(t) \boldsymbol{E}\{\zeta^2(Y_i, t_0(t))|t\}$$

with the inner expectation being

$$\boldsymbol{E}\{\zeta^2(Y_i, t_0(t))|t\} = \frac{\int_{\mathbb{R}} \mathrm{d}y e^{\rho_0(y,t)} \zeta^2(y, t_0(t))}{\int_{\mathbb{R}} \mathrm{d}y e^{\rho_0(y,t)}}$$

and with $t_0(t)$ being the value of t' that maximizes

$$\left[\ln|f_n(t')| + n \cdot \frac{\int_{\mathbb{R}} \mathrm{d}y \cdot e^{\rho_0(y,t)} \rho(y, t')}{\int_{\mathbb{R}} \mathrm{d}y \cdot e^{\rho_0(y,t)}}\right].$$

Thus, we have characterized both the conditional mean estimator and the MMSE in the spirit of a single–letter formula for this class of joint distributions.

The following further extensions of this formalism are conceptually straightforward:

1. The range of the variable t may not necessarily be $[0, \infty)$. Our above analysis applies to whatever range as long as the integrals exist.

2. The joint distribution $P(\boldsymbol{x}, \boldsymbol{y})$ may be a function of more than one statistic $\sum_i \phi(x_i, y_i)$, i.e.,

$$P(\boldsymbol{x}, \boldsymbol{y}) = F_n \left(\sum_{i=1}^{n} \phi_1(x_i, y_i), \ldots, \sum_{i=1}^{n} \phi_k(x_i, y_i) \right).$$

   In this case, one may apply a Laplace transform of a higher dimension

$$F(s_1, \ldots, s_k) = \int_0^\infty \cdots \int_0^\infty \mathrm{d}t_1 \cdots \mathrm{d}t_k f(t_1, \ldots, t_k) e^{-s_1 t_1 - \ldots - s_k t_k},$$

   where $s_i = \sum_i \phi_i(x_i)$, $i = 1, 2, \ldots, k$.

3. The assumption that the i–th term of $\sum_i \phi(x_i, y_i)$ depends only on the i–th coordinate of $\boldsymbol{y}$ is not really necessary. The derivation continues to hold, for example, if we allow more generally the form $\sum_i \phi(x_i, y_i, y_{i-1}, \ldots, y_{i-k})$.

4. The case where ϕ is a quadratic form can be extended to allow a quadratic form that involves all coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ collectively, using a positive definite matrix S for weighting. In other words, joint distributions with elliptic symmetry are allowed, with the form $P(\boldsymbol{x}, \boldsymbol{y}) = F_n[(\boldsymbol{x}, \boldsymbol{y})^T S(\boldsymbol{x}, \boldsymbol{y})]$, where $(\boldsymbol{x}, \boldsymbol{y})$ denotes the concatenated column vector of dimension $(n+m)$ formed by $\boldsymbol{x}$ and $\boldsymbol{y}$, and the matrix S is of dimension $(n+m) \times (n+m)$. In this case, the kernel is Gaussian and hence the estimator is linear for a given t.

6 Conclusion

In this paper, we have proposed a simple relation between MMSE estimation measures and a certain expression, which can be viewed as a partition function, and hence be analyzed using methods of statistical mechanics. This partition function is also related to several information measures, like the information density and the Fisher information. The proposed approach has several advantages over the I–MMSE relation and its variants:

1. It is conceptually simple and direct.

2. It applies in full generality, for every joint distribution of the desired random vector $\boldsymbol{X}$ and its noisy observation vector $\boldsymbol{Y}$.

3. It provides, not only the MMSE error covariance matrix, but also the conditional mean estimator itself $\hat{\boldsymbol{x}} = \boldsymbol{E}\{\boldsymbol{X}|\boldsymbol{y}\}$.

4. It offers several alternative expressions of the MMSE (see Proposition 1).

5. The approach is easy to extend to the mismatched case and it allows mismatch, not only in the marginal of $\boldsymbol{X}$, but in the entire joint density $P(\boldsymbol{x}, \boldsymbol{y})$.

Finally, considering earlier work on the I–MMSE relation and its various variants that were discussed in the Introduction, it would be natural to seek relations between MMSE estimation to the Hessian of the mutual information. One can show, using the same techniques as in Subsection 3.1, that the following relation holds:

$$E = \nabla_0^2 I_{\boldsymbol{\lambda}}(\boldsymbol{X}; \boldsymbol{Y}) + \mathrm{Cov}\{\boldsymbol{X}\} - \mathrm{Cov}\left\{(\boldsymbol{X} - \boldsymbol{E}\{\boldsymbol{X}\})(\boldsymbol{X} - \boldsymbol{E}\{\boldsymbol{X}\})^T, \ln \frac{P(\boldsymbol{Y}|\boldsymbol{X})}{Z(\boldsymbol{Y}, \boldsymbol{\lambda})}\right\},$$

where $I_\lambda(\boldsymbol{X}; \boldsymbol{Y})$ is the mutual information induced by the joint distribution

$$P_{\boldsymbol{\lambda}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{e^{\boldsymbol{\lambda}^T \boldsymbol{x}} P(\boldsymbol{x}, \boldsymbol{y})}{\Theta(\boldsymbol{\lambda})}.$$

Unfortunately, this relation seems somewhat more complicated and not as useful as the I–MMSE relation of [5] or the relations proposed in Subsection 3.1 herein.

References

[1] R. S. Bucy, "Information and filtering," *Information Sciences*, vol. 18, pp. 179–187, 1979.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Hoboken, NJ, U.S.A., 2006.

[3] T. E. Duncan, "On the calculation of mutual information," *SIAM Journal on Applied Mathematics*, vol. 19, no. 1, pp. 215–220, 1970.

[4] D. Guo, "Relative entropy and score function: new information–estimation relationships through arbitrary additive perturbations," *Proc. ISIT 2009*, Seoul, South Korea, June–July 2009.

[5] 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.

[6] D. Guo, S. Shamai, and S. Verdú, "Additive non–Gaussian noise channels: mutual information and conditional mean estimation," *Proc. 2005 IEEE Symp. on Inform. Theory (SIT 2005)*, pp. 719–723, Adelaide, Australia, September 2005.

[7] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 1187–1849, May 2008.

[8] T. Kailath, "The innovations approach to detection and estimation theory," *Proc. of the IEEE*, vol. 58, no. 5, pp. 680–695, May 1970.

[9] N. Merhav, D. Guo, and S. Shamai (Shitz), "Statistical physics of signal estimation in Gaussian noise: theory and examples of phase transitions," to appear in *IEEE Trans. Inform. Theory*, March 2010.

[10] M. Mézard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, 2009.

[11] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 141–154, January 2006.

[12] M. Raginsky and T. P. Coleman, "Mutual information and posterior estimates in channels of exponential family type," *Proc. 2009 IEEE Workshop on Inform. Theory*, pp. 399–403, Taormina, Sicily, October 2009.

[13] S. Verdú, "Mismatched estimation and relative entropy," *Proc. ISIT 2009*, Seoul, South Korea, June–July 2009.

[14] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 4, pp. 1147–1157, July 1994.