# Cross validation in LASSO and its acceleration

Tomoyuki Obuchi and Yoshiyuki Kabashima

January 6, 2016

**Abstract**

We investigate leave-one-out cross validation (CV) as a determinator of the weight of the penalty term in the least absolute shrinkage and selection operator (LASSO). First, on the basis of the message passing algorithm and a perturbative discussion assuming that the number of observations is sufficiently large, we provide simple formulas for approximately assessing two types of CV errors, which enable us to significantly reduce the necessary cost of computation. These formulas also provide a simple connection of the CV errors to the residual sums of squares between the reconstructed and the given measurements. Second, on the basis of this finding, we analytically evaluate the CV errors when the design matrix is given as a simple random matrix in the large size limit by using the replica method. Finally, these results are compared with those of numerical simulations on finite-size systems and are confirmed to be correct. We also apply the simple formulas of the first type of CV error to an actual dataset of the supernovae.

## 1 Introduction

Extracting rules from data has been at the heart of modern sciences. Johannes Kepler discovered his laws of planetary motion by examining the data of planetary orbits by trial and error, which later led to classical mechanics. Max Planck proposed his law of the black body heat radiation for accurately describing experimental data, which played a key role in the discovery of quantum mechanics. As these examples imply, rule extraction has relied mainly on human thoughts.

The never-ending innovation of measurements and experimental techniques is now resulting in the ongoing creation of a large amount of high-dimensional observation data every day. This provides us with situations where rule extraction from data is desired considerably more frequently than ever. Although the entire set of DNA sequences of human beings was identified in 2003, considerable effort must still be made in the days ahead for finding out what rules are written in the dataset. Worldwide observation networks of global climate are being consolidated, but analyzing the observed data in detail is indispensable for understanding the mechanism of global warming. Unfortunately, mechanisms underlying the genome and the global climate are considerably more complicated than those of the planetary orbits and the heat radiation. This makes it difficult to discover rules only by human thoughts as has been done thus far.

Sparse modeling may be a promising framework for resolving such difficulty [1, 2, 3, 4]. This generally means methods of statistical modeling or machine learning that describe rules by using a large number of parameters and select a "sparse" model in which many of the parameters are set to zero by minimizing sparsity-inducing penalties in conjunction with imposing a good fit to the observed data. Modeling methods of this type are preferable in the sense that one can discover a simple and reasonable rule in a semi-automatic manner, with little resort to human thoughts, from a set of many rules that represent various possible relations. The least absolute shrinkage and selection operator (LASSO) is a representative method of the sparse modeling [5, 6]. In this method, many coefficients of large-dimensional linear regression are pruned by the effect of the $\ell_1$ penalty that is defined by the sum of the absolute values of the coefficients. This technique has applications in a wide variety of fields, such as image processing [7], ecology [8], genetics [9], and astronomy [10, 11]. A similar method is known for the signal recovery problem of compressed sensing [12, 13, 14, 15], which exploits the intrinsic sparsity of objective signals for enhancing the signal processing performance [16, 17, 18, 19, 20, 21, 22, 23].

LASSO is, however, required to solve another problem of determining the strength $\lambda$ of the penalty term. Cross validation (CV) is a practically useful strategy for handling this task; its basic concept is to evaluate the prediction error by examining the data under control. Smaller values of the CV error are expected to be better to express the generative model of the data. The minimum, if it exists, of the CV error when changing $\lambda$ is thus considered to obtain an optimal value of $\lambda$. Unfortunately, this reasonable strategy is not well controlled because the behavior of the CV error itself is not fully understood. In particular, there are several variants in the definition of the CV error, each of which can exhibit a different behavior and choose a different optimal value of $\lambda$. Further, conducting CV in a naive manner incurs high computational costs, which makes it difficult to systematically study the behavior of these variants. Even worse, this computational difficulty sometimes forces certain compromises such as scaling down the system size, usage of uncontrolled approximations, or even modifications in research plans.

Given the situation, in this study, we treat leave-one-out (LOO) CV and investigate two types of CV errors, to clarify their properties. Efficient formulas to calculate these two errors are proposed by using belief propagation (BP) in computer science or the cavity method in statistical mechanics [24, 25, 26], in a perturbative manner. A similar formula has also been proposed for Bayesian learning of simple perceptrons in [27]. Our derivation is analogous to that of the approximate message passing (AMP) algorithm [17, 20, 21, 28]. The resultant formulas have two advantages: The computational cost of the resultant algorithm is considerably reduced from that of the naive algorithm; this reveals a simple connection of the CV errors to the residual sums of squares (RSSs) between the reconstructed and the given measurements, in the large system limit.

In response to this second finding, we analytically assess the two CV errors and the corresponding RSSs to reveal their general properties, in the large system limit under the assumption that the measurement matrix is a random matrix, each component of which is independently identically distributed (i.i.d.) from the zero-mean normal distribution. It is commonly found that both the CV errors exhibit their unique minimums as $\lambda$ changes, but the locations of the minimums are different, and hence, the chosen "optimal" values of $\lambda$ are discriminably different between the two CV errors. We compare these two values of $\lambda$ by using the so-called receiver operating characteristic (ROC) curve and compare them to the so-called Younden's index, to find that in the weak noise case, the second CV error chooses a more preferable value of $\lambda$ than the first one. This can be attributed to the fact that the first error tends to overestimate the false positive ratio. Unfortunately, however, our analytical result also clarifies that the above simple formulas derived by the BP in a perturbative manner are not applicable to the second CV error. This is understood by an intricate discussion on the change of the chosen variables in the leave-one-out procedure. These findings are confirmed by numerical experiments on finite-size systems, and our formula is clarified to work well for moderate-size systems.

The rest of this paper is organized as follows: In sec. 2, we state LASSO in the context of compressed sensing and explain the LOO CV. In sec. 3, we explain the application of the cavity method to the evaluation of the CV errors in the LOO CV, clarifying the relation between the CV errors and the RSSs. In sec. 4, we present the analytical result in the case of a random-observation matrix. In sec. 5, we show the result of numerical experiments to support our algorithm and analytical results. An application of the proposed method to the Type Ia supernova data is also presented in this section. The last section is devoted to the conclusion.

# 2 Problem setting

Here, we state our problem setting and summarize the quantities of interest. These quantities are analyzed in the subsequent sections to clarify the behavior of CV errors in the LOO CV procedure.

## 2.1 Compressed sensing based on LASSO

In this paper, we introduce LASSO in the context of compressed sensing. Let us suppose that a vector $\boldsymbol{y} \in \mathbb{R}^M$ of measurement is generated from an unknown signal vector $\hat{\boldsymbol{x}} \in \mathbb{R}^N$, which is assumed to be sparse, through the following linear process:

$$\boldsymbol{y} = A\hat{\boldsymbol{x}} + \boldsymbol{\xi}, \tag{1}$$

where $A = \{A_{\mu i}\}_{\mu=1,\cdots,M;\, i=1,\cdots N} \in \mathbb{R}^{M\times N}$ represents a measurement (design) matrix and $\boldsymbol{\xi} \in \mathbb{R}^M$ denotes the measurement noise each component of which is drawn from the zero-mean normal distribution with variance $\sigma_\xi^2$, indicated by $\mathcal{N}(0, \sigma_\xi^2)$. The number of measurements $M$ is supposed to be smaller than the dimensions of the representation $N$. Currently, we do not specify the ensemble of $A$ but only assume the scaling of the component as $A_{\mu i} = O(1/\sqrt{N})$. On this condition, we infer the representation $\hat{\boldsymbol{x}}$ from the given measurement $\boldsymbol{y}$, by utilizing the sparseness of $\hat{\boldsymbol{x}}$. The sparsity of $\hat{\boldsymbol{x}}$ is quantified as follows:

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} |\hat{x}_i|_0 \equiv \frac{1}{N}||\hat{\boldsymbol{x}}||_0, \tag{2}$$

where $|x|_0$ results in zero if $x = 0$ and unity otherwise. The symbol $||\cdot||_0$ is called $\ell_0$-norm. We can also introduce $\ell_k$-norm as follows:

$$||\boldsymbol{x}||_k = \left(\sum_{i=1}^{N} |x_i|^k\right)^{1/k}, \tag{3}$$

Given an inferred signal $\boldsymbol{x}$, we introduce the RSS, $\mathcal{E}$, and the rate, $\epsilon$, as follows:

$$\mathcal{E}(\boldsymbol{x}) = M\epsilon(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{y} - A\boldsymbol{x}||_2^2, \tag{4}$$

The inference of $\boldsymbol{x}$ based on LASSO is expressed as follows:

$$\boldsymbol{x}^{(1)}(\lambda) = \arg\min_{\boldsymbol{x}} \left\{\mathcal{E}(\boldsymbol{x}) + \lambda||\boldsymbol{x}||_1\right\}. \tag{5}$$

Unfortunately, the result is biased; i.e., $\boldsymbol{x}^{(1)}(\lambda)$ does not agree with $\hat{\boldsymbol{x}}$ even as $M$ increases because of the presence of the penalty term $\lambda||\boldsymbol{x}||_1$ in eq. (5). A conventional alternative to $\boldsymbol{x}^{(1)}$ is obtained by minimizing the RSS on the choice of column vectors associated with $\boldsymbol{x}^{(1)}$. This can be formulated as follows:

$$\boldsymbol{x}^{(2)}(\lambda) = \arg\min_{\boldsymbol{x}} \left\{\frac{1}{2}\left|\left|\boldsymbol{y} - A\left(\left|\boldsymbol{x}^{(1)}(\lambda)\right|_0 \circ \boldsymbol{x}\right)\right|\right|_2^2\right\}, \tag{6}$$

where $\circ$ denotes the Hadamard product defined as $(\boldsymbol{v}\circ\boldsymbol{w})_i = v_i w_i$, and $|\cdot|_0$ of a vector is defined as $(|\boldsymbol{v}|_0)_i = |v_i|_0$. Corresponding to the two inferred signals $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, we define the two RSSs as follows:

$$\mathcal{E}_1(\lambda) = M\epsilon_1(\lambda) = \mathcal{E}(\boldsymbol{x}^{(1)}(\lambda)), \qquad \mathcal{E}_2(\lambda) = M\epsilon_2(\lambda) = \mathcal{E}(\boldsymbol{x}^{(2)}(\lambda)), \tag{7}$$

### 2.2 Leave-one-out cross validation

The idea of LOO CV is as follows: We select one measurement $\mu$ among $M$ measurements and leave it out while inferring the signal, which is formally written as follows:

$$\boldsymbol{x}^{(1)\backslash\mu}(\lambda) = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2} \sum_{\nu(\neq\mu)} \left( y_\nu - \sum_{i=1}^{N} A_{\nu i} x_i \right)^2 + \lambda ||\boldsymbol{x}||_1 \right\}, \tag{8}$$

where the symbol $\backslash\mu$ denotes the absence of the $\mu$th observation. Using this, we can evaluate a CV error on the $\mu$th measurement as $(1/2)\left(y_\mu - \sum_{i=1}^{N} A_{\mu i} x_i^{(1)\backslash\mu}(\lambda)\right)^2$. Summing this up for all $\mu = 1, \cdots, M$ gives the CV error of the first type:

$$\mathcal{L}_1(\lambda) = \sum_{\mu=1}^{M} \frac{1}{2M} \left( y_\mu - \sum_{i=1}^{N} A_{\mu i} x_i^{(1)\backslash\mu}(\lambda) \right)^2. \tag{9}$$

We can reduce the bias effect by the regularization as eq. (6), defining

$$\boldsymbol{x}^{(2)\backslash\mu}(\lambda) = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2} \sum_{\nu(\neq\mu)} \left( y_\nu - \sum_{i=1}^{N} A_{\nu i} |x_i^{(1)\backslash\mu}|_0 x_i \right)^2 \right\}. \tag{10}$$

The second type of the CV error can thus be expressed as follows:

$$\mathcal{L}_2(\lambda) = \sum_{\mu=1}^{M} \frac{1}{2M} \left( y_\mu - \sum_{i=1}^{N} A_{\mu i} x_i^{(2)\backslash\mu}(\lambda) \right)^2. \tag{11}$$

We particularly call these errors the LOO errors (LOOEs); they are central quantities of the analysis described in the subsequent sections.

### 2.3 Quantities to examine the quality of inference

In addition to the LOOEs, we need quantities to examine the quality of inference and compare them with the LOOEs. For this, we introduce the mean squared error (MSE) between the true and the inferred signal. Corresponding to the two inferred signals $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ in eqs. (5,6), the two MSEs can be calculated as follows:

$$\mathcal{M}_1(\lambda) = \frac{1}{N} ||\boldsymbol{x}^{(1)} - \hat{\boldsymbol{x}}||_2^2, \quad \mathcal{M}_2(\lambda) = \frac{1}{N} ||\boldsymbol{x}^{(2)} - \hat{\boldsymbol{x}}||_2^2. \tag{12}$$

Further, the ratios of the correctly or incorrectly chosen variables are important. The true positive ratio, $TP$, and the false positive ratio, $FP$, are as follows:

$$TP(\lambda) = \frac{\sum_i \delta_{1,|\hat{x}_i|_0} \delta_{1,\left|x_i^{(1)}(\lambda)\right|_0}}{\sum_i \delta_{1,|\hat{x}_i|_0}}, \quad FP(\lambda) = \frac{\sum_i \delta_{0,|\hat{x}_i|_0} \delta_{1,\left|x_i^{(1)}(\lambda)\right|_0}}{\sum_i \delta_{0,|\hat{x}_i|_0}}. \tag{13}$$

The so-called ROC curve is a plot of $TP$ against $FP$. This characterizes the quality of inference: If a ROC curve is farther from the straight line $TP = FP$, then the inference is better. Accordingly, we also refer to the so-called Youden's index and define an "optimal" value of $\lambda$ chosen according to this index, $\lambda_{\rm YI}$. The definition is as follows:

$$D(\lambda) = \min_x \left\{ (TP(\lambda) - x)^2 + (FP(\lambda) - x)^2 \right\}, \tag{14}$$

$$\lambda_{\rm YI} = \arg\max_\lambda D(\lambda). \tag{15}$$

# 3 Message passing for leave-one-out error

Suppose that the system size and the number of observations, $N$ and $M$, are sufficiently large, implying that an inferred signal does not change considerably by an addition or deletion of observations. This assumption enables us to perform a perturbative treatment, clarifying the relationship between the LOOEs and the RSSs.

## 3.1 Revisiting approximate message passing

Let us start by stating a derivation of the known AMP algorithm. This can be done by using the belief propagation (BP) or the cavity method [12, 20, 21, 28]. For this, we present a probabilistic formulation for the present problem on the basis of the prescriptions of statistical mechanics. We introduce a Hamiltonian, partition function, and Boltzmann distribution, respectively, as follows:

$$\mathcal{H}_1(\boldsymbol{x}) \equiv \mathcal{E}(\boldsymbol{x}) + \lambda||\boldsymbol{x}||_1, \tag{16}$$

$$Z_1(\beta) \equiv \int_{-\infty}^{\infty} d\boldsymbol{x}\; e^{-\beta\mathcal{H}_1}, \tag{17}$$

$$P_1(\boldsymbol{x}) = \frac{e^{-\beta\mathcal{H}_1(\boldsymbol{x})}}{Z_1} = \frac{e^{-\beta\lambda||\boldsymbol{x}||_1}\prod_\mu \Phi_\mu(\boldsymbol{x})}{Z_1}, \tag{18}$$

where $\beta$ denotes the inverse temperature and $\Phi_\mu$ represents the so-called potential function

$$\Phi_\mu(\boldsymbol{x}) = e^{-\frac{\beta}{2}\left(y_\mu - \sum_i A_{\mu i}x_i\right)^2}. \tag{19}$$

Note that $\beta$ is independent of the strength of the observation noise $\sigma_\xi$, and the limit $\beta \to \infty$ is supposed to be taken after all the calculations as we are interested in the minimum of the above Hamiltonian. We denote the average over the Boltzmann distribution by the angular brackets $\langle\cdots\rangle$. BP allows us to calculate the marginal distribution by using two types of messages $(i, j, k = 1, \cdots, N,\ \mu, \nu = 1, \cdots, M)$ as follows:

$$\hat{\phi}_{\mu\to i}(x_i) = \int \prod_{j(\neq i)} dx_j\; \Phi_\mu(\boldsymbol{x}) \prod_{j(\neq i)} \phi_{j\to\mu}(x_j), \tag{20}$$

$$\phi_{i\to\mu}(x_i) = e^{-\beta\lambda|x_i|} \prod_{\nu(\neq\mu)} \hat{\phi}_{\nu\to i}(x_i), \tag{21}$$

A crucial observation to assess eqs. (20,21) is that the exponent of the potential function has a sum of an extensive number of random variables; the central limit theorem thus justifies treating it as a Gaussian variable with the appropriate mean and variance. Hence, according to eq. (20), where $x_i$ is special, we can divide the extensive sum of $\Phi_\mu$ as follows:

$$\sum_{j=1}^{N} A_{\mu j}x_j = A_{\mu i}x_i + \sum_{j(\neq i)} A_{\mu j}x_j \approx A_{\mu i}x_i + \sum_{j(\neq i)} A_{\mu j}\bar{x}_j^{\setminus\mu} + \sqrt{V_\mu}z, \tag{22}$$

where $z$ denotes the zero-mean unit-variance Gaussian variable. The second term on the right-hand side represents the mean of $\sum_{j(\neq i)} A_{\mu j}x_j$, and thus,

$$\bar{x}_j^{\setminus\mu} = \langle x_j\rangle_{\setminus\mu}, \tag{23}$$

where the angular brackets $\langle \cdots \rangle_{\backslash \mu}$ denote the average over the Boltzmann distribution without the $\mu$th potential function. Now, let us call $\left\{ \bar{x}_j^{\backslash \mu} \right\}$ cavity magnetization. The last term is derived by calculating the variance of $\sum_{j(\neq i)} A_{\mu j} x_j$ as follows:

$$\left\langle \left( \sum_{j(\neq)i} A_{\mu j} x_j \right)^2 \right\rangle_{\backslash \mu} = \sum_{j,k(\neq i)} A_{\mu j} A_{\mu k} \left( \langle x_j x_k \rangle_{\backslash \mu} - \langle x_j \rangle_{\backslash \mu} \langle x_k \rangle_{\backslash \mu} \right)$$

$$\approx \sum_{j,k} A_{\mu j} A_{\mu k} \left( \langle x_j x_k \rangle_{\backslash \mu} - \langle x_j \rangle_{\backslash \mu} \langle x_k \rangle_{\backslash \mu} \right) = \sum_{j,k} A_{\mu j} A_{\mu k} \frac{\chi_{jk}^{\backslash \mu}}{\beta} \equiv V_\mu, \tag{24}$$

where $\chi_{jk}^{\backslash \mu}$ is called the susceptibility matrix (without the $\mu$th observation), which quantifies the correlations between the variables. The terms added at the beginning of the second line have a small contribution of the scaling $O(1/N)$; thus, they are negligible, and their addition is justified.

The application of eq. (22) in eq. (20) replaces the integration over $\boldsymbol{x}$ to that over $z$. Performing this integration yields the following:

$$\hat{\phi}_{\mu \to i}(x_i) \propto e^{\beta \left( -\frac{1}{2} \frac{A_{\mu i}^2}{1+\beta V_\mu} x_i^2 + A_{\mu i} x_i \frac{y_\mu - \sum_{j(\neq i)} A_{\mu j} \bar{x}_j^{\backslash \mu}}{1+\beta V_\mu} \right)}. \tag{25}$$

Combining this formula with eqs. (20,21), we can derive a recursion relation of $\left\{ \bar{x}_j^{\backslash \mu} \right\}$, leading to the conventional BP equation.

A more convenient recursion relationship can be obtained in terms of the full magnetization $\{ \bar{x}_j = \langle x_j \rangle \}$ instead of the cavity magnetization $\left\{ \bar{x}_j^{\backslash \mu} = \langle x_j \rangle_{\backslash \mu} \right\}$. The full marginal distribution of $x_i$ necessarily takes the following modified Gaussian form:

$$\phi_i(x_i) = e^{-\beta \lambda |x_i|} \prod_\mu \hat{\phi}_{\mu \to i}(x_i) \propto e^{\beta \left( -\frac{1}{2} \Gamma_i x_i^2 + h_i x_i - \lambda |x_i| \right)}. \tag{26}$$

where

$$\Gamma_i = \sum_{\mu=1}^M \frac{A_{\mu i}^2}{1+\beta V_\mu}, \tag{27}$$

$$h_i = \sum_{\mu=1}^M \frac{A_{\mu i}}{1+\beta V_\mu} \left( y_\mu - \sum_{j(\neq i)} A_{\mu j} \bar{x}_j^{\backslash \mu} \right) = \sum_{\mu=1}^M A_{\mu i} a_\mu + \sum_{\mu=1}^M \frac{A_{\mu i}^2}{1+\beta V_\mu} \bar{x}_i^{\backslash \mu}. \tag{28}$$

and we set

$$a_\mu \equiv \frac{1}{1+\beta V_\mu} \left( y_\mu - \sum_j A_{\mu j} \bar{x}_j^{\backslash \mu} \right). \tag{29}$$

Hereafter, we call $a_\mu$ the cavity residual. We can interpret that the cavity residual $a_\mu$ contributes the $\mu$th observation to the effective field $h_i$. Thus, the full magnetization $\bar{x}_j$ is obtained from $\bar{x}_j^{\backslash \mu}$ by adding this contribution of $a_\mu$ to the effective field in a perturbative manner. This consideration yields the following:

$$\bar{x}_j \approx \bar{x}_j^{\backslash \mu} + \sum_k \frac{\partial \bar{x}_j^{\backslash \mu}}{\partial h_k} \frac{\partial h_k}{\partial a_\mu} a_\mu = \bar{x}_j^{\backslash \mu} + \sum_k A_{\mu k} \chi_{jk}^{\backslash \mu} a_\mu. \tag{30}$$

Note that we consider only the variation of $h_k$ and do not take into account the change in $\Gamma_k$ when adding the $\mu$th observation. This is because the $\mu$th observation's contribution to $\Gamma_k$ is proportional to $A_{\mu k}^2 = O(1/N)$ and is smaller than that to $h_k$ proportional to $A_{\mu k} = O(1/\sqrt{N})$. Basically, our perturbation is connected to the smallness of $A_{\mu k} = O\left(1/\sqrt{N}\right)$, and only the linear term with respect to $A_{\mu k}$ is important. By substituting eq. (30) into eq. (29) and solving it with respect to $a_\mu$, we obtain a simple expression of $a_\mu$ in terms of the full magnetization $\{\bar{x}_i\}_i$

$$a_\mu \approx y_\mu - \sum_j A_{\mu j} \bar{x}_j. \tag{31}$$

Similarly, the substitution of eq. (30) into $m_i^{\backslash \mu}$ in eq. (28) yields the following:

$$h_i \approx \sum_{\mu=1}^{M} A_{\mu i} a_\mu + \Gamma_i \bar{x}_i. \tag{32}$$

The neglected terms are proportional to the third and higher orders of $A_{\mu i}$. The last term in eq. (32) is the well-known Onsager reaction term.

The full magnetization $\bar{x}_i$ is a function of only $\Gamma_i$ and $h_i$; thus, now, we can calculate $\bar{x}_i$ by recursion if the values of $\{\Gamma_i\}$ are determined. The functional form of $\bar{x}_i$ becomes simple in the limit $\beta \to \infty$ and is identified with $\boldsymbol{x}^{(1)}$ in eq. (5). The fixed point of the AMP is described in this limit as follows:

$$a_\mu^{(1)} = y_\mu - \sum_{j=1}^{N} A_{\mu j} x_j^{(1)}, \tag{33a}$$

$$h_i^{(1)} = \sum_{\mu=1}^{M} A_{\mu i} a_\mu^{(1)} + \Gamma_i x_i^{(1)}, \tag{33b}$$

$$x_i^{(1)} = \frac{h_i - \lambda \ \mathrm{sgn}\left(h_i^{(1)}\right)}{\Gamma_i} \Theta\left(|h_i^{(1)}| - \lambda\right), \tag{33c}$$

where $\Theta(x)$ denotes the step function giving 1 for $x \geq 0$ and 0 otherwise, and the superscript $^{(1)}$ is attached according to eq. (5).

Coefficients $\{\Gamma_i\}$ can be determined using BP in a similar manner; however, this is not an easy task. Therefore, we omit the derivation and just refer to [20, 21] for the case of weak correlations where only diagonal terms are important, $\beta V_\mu \approx \sum_{j=1}^{N} A_{\mu j}^2 \chi_{jj}^{\backslash \mu}$.

The BP algorithm can also be applied for calculating $\boldsymbol{x}^{(2)}$. The derivation is essentially the same as eq. (33), and the result is as follows:

$$a_\mu^{(2)} = y_\mu - \sum_{j=1}^{N} A_{\mu j} x_j^{(2)}, \tag{34a}$$

$$h_i^{(2)} = \sum_{\mu=1}^{M} A_{\mu i} a_\mu^{(2)} + \Gamma_i x_i^{(2)}, \tag{34b}$$

$$x_i^{(2)} = \frac{h_i^{(2)}}{\Gamma_i} \Theta\left(|h_i^{(1)}| - \lambda\right). \tag{34c}$$

A crucial difference from eq. (33) is the dependence on $h_i^{(1)}$. This implies that $\boldsymbol{x}^{(2)}$ are evaluated by solving eq. (34) conditioned by the solution of eq. (33). As explained later, this difference leads to a difficulty in evaluating $\mathcal{L}_2$, in contrast to $\mathcal{L}_1$.

### 3.2 Simple formulas of leave-one-out error

For deriving the AMP, we conducted a perturbation on $\bar{x}_j^{\backslash\mu}$. In the zero-temperature limit, $\bar{x}_j^{\backslash\mu}$ is identified with $x_j^{(1)\backslash\mu}$ in eq. (9). By inserting eq. (30) in eq. (9), we obtain the following:

$$\mathcal{L}_1(\lambda) \approx \frac{1}{2M}\sum_{\mu=1}^{M}\left(1+\sum_{i,j}A_{\mu i}A_{\mu j}\chi_{ij}^{\backslash\mu}\right)^2\left(y_\mu-\sum_i A_{\mu i}x_i^{(1)}\right)^2. \tag{35}$$

This has a considerable advantage compared to eq. (9): Eq. (9) requires us to solve the optimization problem (8) $M$ times for evaluating $\mathcal{L}_1$, but in the case of eq. (35), we need to solve the optimization of (5) only once.

The susceptibility matrix $\chi_{ij}^{\backslash\mu}$ is the origin of difficulty in the computation of $\{\Gamma_i\}$. Fortunately, once the solution of eq. (5), $\boldsymbol{x}^{(1)}$, is obtained, this can be easily calculated. LASSO separates the variables into two types: Some variables become zero as the solution of eq. (5) and are called *inactive*; the other variables take finite values and are *active*. Suppose that the active and inactive variables are known and that $\tilde{A}$ is the submatrix corresponding to the active set. The active parts of $\boldsymbol{x}$ and $\chi^{\backslash\mu}$ are also introduced as $\tilde{\boldsymbol{x}}$ and $\tilde{\chi}^{\backslash\mu}$, respectively. To evaluate $\tilde{\chi}^{\backslash\mu}$, we need to determine $\tilde{\boldsymbol{x}}$ in the case without the $\mu$th observation, and denote the solution as $\tilde{\boldsymbol{x}}^{(1)\backslash\mu}$ in accordance with eq. (8). Correspondingly, we introduce a notation $\boldsymbol{y}_{\backslash\mu}$ expressing $\boldsymbol{y}$ without the $\mu$th component and $\tilde{A}_{\backslash\mu}$ representing $\tilde{A}$ without the $\mu$th row. We assume that the active and inactive sets are stable: A small perturbation $\delta\tilde{\boldsymbol{h}}$ does not change these sets[1]. By using these notations and assumptions, we can now easily obtain $\tilde{\boldsymbol{x}}$ as follows:

$$\begin{aligned}&\min_{\tilde{\boldsymbol{x}}}\left\{\frac{1}{2}||\boldsymbol{y}_{\backslash\mu}-\tilde{A}_{\backslash\mu}\tilde{\boldsymbol{x}}||_2^2+\lambda||\tilde{\boldsymbol{x}}||_1-\delta\tilde{\boldsymbol{h}}\cdot\tilde{\boldsymbol{x}}\right\}\\ &\Rightarrow \tilde{\boldsymbol{x}}^{(1)\backslash\mu}=\left(\tilde{A}_{\backslash\mu}^{\mathrm{T}}\tilde{A}_{\backslash\mu}\right)^{-1}\left(\tilde{A}_{\backslash\mu}^{T}\boldsymbol{y}_{\backslash\mu}+\delta\tilde{\boldsymbol{h}}-\lambda\mathrm{sgn}\left(\tilde{\boldsymbol{x}}^{(1)\backslash\mu}\right)\right).\end{aligned} \tag{36}$$

Considering the variation with respect to $\delta\tilde{\boldsymbol{h}}$, we obtain the following:

$$\tilde{\chi}^{\backslash\mu}=\frac{\partial\tilde{\boldsymbol{x}}^{(1)\backslash\mu}}{\partial\boldsymbol{h}}=\left(\tilde{A}_{\backslash\mu}^{\mathrm{T}}\tilde{A}_{\backslash\mu}\right)^{-1}. \tag{37}$$

All the components of the inactive part of $\chi^{\backslash\mu}$ are zero, and thus, the susceptibility matrix is fully calculated.

Eq. (37) is seemingly inefficient in that the evaluation of $\tilde{\chi}^{\backslash\mu}$ for all $\mu$ requires $M$ inverse operations of a matrix, which is computationally expensive. Fortunately, this computational difficulty can be overcome by using the Sherman–Morrison formula. Denoting $\boldsymbol{u}_\mu^{\mathrm{T}}$ as the $\mu$th row vector of $\tilde{A}$, we obtain the following:

$$\left(\tilde{A}_{\backslash\mu}^{\mathrm{T}}\tilde{A}_{\backslash\mu}\right)^{-1}=\left(\tilde{A}^{\mathrm{T}}\tilde{A}\right)^{-1}+\frac{\left(\tilde{A}^{\mathrm{T}}\tilde{A}\right)^{-1}\boldsymbol{u}_\mu\boldsymbol{u}_\mu^{\mathrm{T}}\left(\tilde{A}^{\mathrm{T}}\tilde{A}\right)^{-1}}{1-\boldsymbol{u}_\mu^{\mathrm{T}}\left(\tilde{A}^{\mathrm{T}}\tilde{A}\right)^{-1}\boldsymbol{u}_\mu}. \tag{38}$$

Hence, $\tilde{\chi}^{\backslash\mu}$ is calculated from $\left(\tilde{A}^{\mathrm{T}}\tilde{A}\right)^{-1}$ by using a small number of simple products of matrices. The inverse operation appears in $\left(\tilde{A}^{\mathrm{T}}\tilde{A}\right)^{-1}$ just once. Eqs. (35), (37), and (38) constitute the main result presented in this paper.

[1] This assumption implies that in the evaluation of the LOOEs, we suppose that the active and inactive sets are unchanged by an addition or deletion of the measurement. Unfortunately, this assumption is not correct; however, the result on $\mathcal{L}_1$ based on this assumption is correct, while that on $\mathcal{L}_2$ is incorrect. This difference is shown in sec. 4.3 and explained in sec. 5.1.1.

A similar discussion seems to be applicable to the evaluation of eq. (10). The active and inactive sets are common with $\boldsymbol{x}^{(1)}$ by definition, and the values of active variables can be calculated as follows:

$$\min_{\tilde{\boldsymbol{x}}}\left\{\frac{1}{2}||\boldsymbol{y}_{\backslash\mu}-\tilde{A}_{\backslash\mu}\tilde{\boldsymbol{x}}||_2^2-\delta\tilde{\boldsymbol{h}}\cdot\tilde{\boldsymbol{x}}\right\}\Rightarrow\tilde{\boldsymbol{x}}^{(2)\backslash\mu}=\left(\tilde{A}^{\mathrm{T}}_{\backslash\mu}\tilde{A}_{\backslash\mu}\right)^{-1}\left(\tilde{A}^{T}_{\backslash\mu}\boldsymbol{y}_{\backslash\mu}+\delta\tilde{\boldsymbol{h}}\right). \tag{39}$$

This provides the same susceptibility matrix as eq. (37). Hence, the second type of LOOE can also be approximated as follows:

$$\mathcal{L}_2(\lambda)\approx\frac{1}{2M}\sum_{\mu=1}^{M}\left(1+\sum_{i,j}A_{\mu i}A_{\mu j}\chi_{ij}^{\backslash\mu}\right)^2\left(y_\mu-\sum_i A_{\mu i}x_i^{(2)}\right)^2. \tag{40}$$

Unfortunately, this approximation is not correct while eq. (35) is validated. The detailed reasoning is given in sec. 5.1.1.

### 3.2.1 In the large-size limit

In the case of the limit $N\to\infty$, we can obtain an analytic formula for $\tilde{\chi}_{ij}$ under certain conditions and thus, considerably simplify the computations of eqs. (35,40). We will derive this analytic formula in the next section; here, we will just refer to the result:

$$\tilde{\chi}_{ij}^{\backslash\mu}=\frac{\rho(\lambda)}{\alpha-\rho(\lambda)}\delta_{ij}. \tag{41}$$

where $\rho(\lambda)=(1/N)||\boldsymbol{x}^{(1)}(\lambda)||_0$ denotes the sparsity of the inferred signal. This result is derived under the assumption that the observation matrix is a random matrix each component of which is i.i.d. from the zero-mean normal distribution $\mathcal{N}(0,N^{-1})$. In such a case, the non-diagonal part of the susceptibility becomes irrelevant as reported in [20, 21]. The resultant formula of the LOOEs is now very simple

$$\mathcal{L}_k(\lambda)\to\left(\frac{\alpha}{\alpha-\rho(\lambda)}\right)^2\frac{1}{2M}\sum_{\mu=1}^{M}\left(y_\mu-\sum_i A_{\mu i}x_i^{(k)}\right)^2=\left(\frac{\alpha}{\alpha-\rho(\lambda)}\right)^2\epsilon_k(\lambda). \tag{42}$$

Using this formula, in the next section, we examine the behavior of the LOOEs in the limit $N\to\infty$ when $\lambda$ is changed.

Before moving to the next section, we have two comments to make on eq. (42). One is about the relationship to the Akaike information criterion (AIC). It is known that the LOOE asymptotically agrees with AIC in general. This can be directly seen by expanding eq. (42) with respect to $\rho/\alpha$ in the limit $\rho/\alpha\ll 1$:

$$2M\mathcal{L}_k(\lambda)\approx||\boldsymbol{y}-A\boldsymbol{x}^{(k)}||_2^2+2\rho\frac{||\boldsymbol{y}-A\boldsymbol{x}^{(k)}||_2^2}{\alpha}. \tag{43}$$

The second term is expected to converge to $2N\rho\sigma_\xi^2$ in the limit $\rho/\alpha\ll 1$, yielding the expression of AIC. The other comment is about the robustness of eq. (42). We have numerically examined some different ensembles of the observation matrix $A$ and found that the relation (42) between the LOOE and the RSS seems to be fairly robust, while eq. (41) is not. We have observed that the non-diagonal components of $\chi$ become important in certain ensembles in which components of $A$ are correlated. These non-diagonal components are complicated, but presumably as a result of nontrivial cancellations, the simple relation $\left(1+\sum_{ij}A_{\mu i}A_{\mu j}\chi_{ij}^{\backslash\mu}\right)^2\to(\alpha/(\alpha-\rho))^2$

seems to hold widely. This finding has an important consequence: Even for realistic situations where the observation matrix is far from the random matrix, eq. (42) can be used for accurate approximation of the LOOE. We will revisit this point later when treating the real data of the Type Ia supernovae in sec. 5. Apart from obtaining such a realistic benefit, we should examine whether the relation (42) actually holds for a wider ensemble of $A$ than the simple random matrix ensemble in a more systematic manner, which will be an important future work.

# 4 Analytic formulas on the random observation matrix

In this section, in the case of the large-system limit $N \to \infty$, we derive an analytic formula of the RSSs, $\mathcal{E}_1$ and $\mathcal{E}_2$, under the assumption that the observation matrix is a random matrix, each component of which is i.i.d. from $\mathcal{N}(0, 1/N)$. Considering this limit, we keep the ratio $\alpha = M/N(< 1)$ finite along with the sparsity of the true signal $\hat{\rho} = ||\hat{\boldsymbol{x}}||_0/N$.

As noted in sec. 2, the noise $\boldsymbol{\xi}$ is i.i.d. from the normal distribution $\mathcal{N}(0, \sigma_\xi^2)$. Moreover, the ensemble of the true signal $\hat{\boldsymbol{x}}$ is assumed to be the Bernoulli–Gaussian distribution:

$$P(\hat{\boldsymbol{x}}) = \prod_{i=1}^{N} \left\{ \frac{\rho}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{1}{2\sigma_x^2}\hat{x}_i^2} + (1 - \hat{\rho})\delta(\hat{x}_i) \right\}. \tag{44}$$

Following statistical mechanical jargon, we call the average over $A$, $\boldsymbol{\xi}$, and $\hat{\boldsymbol{x}}$ configurational average, which is represented by square brackets with appropriate subscripts. For example, the average over $\boldsymbol{\xi}$ and $\hat{\boldsymbol{x}}$ is written as $[\cdots]_{\boldsymbol{\xi},\hat{\boldsymbol{x}}}$.

In this section, we only state the outline of our analysis, give the resultant formulas of the free energies and the related quantities, and show some plots of the quantities of interest. The detailed derivations are deferred to Appendix A.

## 4.1 Outline of analysis

Following the usual prescriptions of statistical mechanics, we define the Hamiltonian $\mathcal{H}$ and the partition function $Z$. According to eqs. (5,6), we define two Hamiltonians as follows:

$$\mathcal{H}_1(\boldsymbol{x}^{(1)}|\boldsymbol{\xi}, A, \hat{\boldsymbol{x}}) = \frac{1}{2}\left|\left|\boldsymbol{y} - A\boldsymbol{x}^{(1)}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{x}^{(1)}\right|\right|_1, \tag{45}$$

$$\mathcal{H}_2(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)}, \boldsymbol{\xi}, A, \hat{\boldsymbol{x}}) = \frac{1}{2}\left|\left|\boldsymbol{y} - A\left(\left|\boldsymbol{x}^{(1)}\right|_0 \circ \boldsymbol{x}^{(2)}\right)\right|\right|_2^2. \tag{46}$$

The corresponding partition functions $Z_1$ and $Z_2$ are defined as follows:

$$Z_1(\beta|\boldsymbol{\xi}, A, \hat{\boldsymbol{x}}) = \left\{ \prod_{i=1}^{N} \int_{-\infty}^{\infty} dx_i^{(1)}\ e^{-\beta\mathcal{H}_1(\boldsymbol{x}^{(1)}|\boldsymbol{\xi},A,\hat{\boldsymbol{x}})} \right\}, \tag{47}$$

$$Z_2(\beta|\boldsymbol{x}^{(1)}, \boldsymbol{\xi}, A, \hat{\boldsymbol{x}}) = \left\{ \prod_{i=1}^{N} \int_{-\infty}^{\infty} d_{|x_i^{(1)}|_0} x_i\ e^{-\beta\mathcal{H}_2(\boldsymbol{x}|\boldsymbol{x}^{(1)},\boldsymbol{\xi},A,\hat{\boldsymbol{x}})} \right\}, \tag{48}$$

where

$$\int d_{|\xi|_0} x = \begin{cases} \int dx & (|\xi|_0 = 1) \\ 1 & (|\xi|_0 = 0) \end{cases}, \tag{49}$$

and the Boltzmann distributions can be expressed as follows:

$$p_1(\boldsymbol{x}, \beta|\boldsymbol{\xi}, A, \hat{\boldsymbol{x}}) = \frac{e^{-\beta\mathcal{H}_1(\boldsymbol{x}|\boldsymbol{\xi},A,\hat{\boldsymbol{x}})}}{Z_1(\beta|\boldsymbol{\xi}, A, \hat{\boldsymbol{x}})}, \quad p_2(\boldsymbol{x}, \beta|\boldsymbol{x}^{(1)}, \boldsymbol{\xi}, A, \hat{\boldsymbol{x}}) = \frac{e^{-\beta\mathcal{H}_2(\boldsymbol{x}|\boldsymbol{x}^{(1)},\boldsymbol{\xi},A,\hat{\boldsymbol{x}})}}{Z_2(\beta|\boldsymbol{x}^{(1)}, \boldsymbol{\xi}, A, \hat{\boldsymbol{x}})}. \tag{50}$$

Note that $\boldsymbol{x}^{(1)}$ conditioning $\mathcal{H}_2, Z_2$ and $p_2$ is drawn from $p_1$. We assume that the average over these Boltzmann distributions $p_1$ and $p_2$ is denoted by angular brackets with an appropriate subscript. We also introduce double angular brackets denoting the average over both $p_1$ and $p_2$, $\langle\langle \cdots \rangle\rangle \equiv \langle\langle \cdots \rangle_2\rangle_1$. The averaged free energies $f_1$ and $f_2$ can thus be defined as follows:

$$-\beta f_1(\beta) = \frac{1}{N}\left[\log Z_1(\beta|\boldsymbol{\xi}, A, \hat{\boldsymbol{x}})\right]_{\boldsymbol{\xi},A,\hat{\boldsymbol{x}}}\,, \tag{51}$$

$$\begin{aligned} -\beta f_2(\beta) &= \frac{1}{N}\left[\left(\prod_{i=1}^{N}\int_{-\infty}^{\infty} dx_i^{(1)}\right) p_1(\boldsymbol{x}^{(1)}, \beta|\boldsymbol{\xi}, A, \hat{\boldsymbol{x}}) \log Z_2(\beta|\boldsymbol{x}^{(1)}, \boldsymbol{\xi}, A, \hat{\boldsymbol{x}})\right]_{\boldsymbol{\xi},A,\hat{\boldsymbol{x}}} \\ &= \frac{1}{N}\left[\left\langle \log Z_2(\beta|\boldsymbol{x}^{(1)}, \boldsymbol{\xi}, A, \hat{\boldsymbol{x}})\right\rangle_1\right]_{\boldsymbol{\xi},A,\hat{\boldsymbol{x}}}\,. \end{aligned} \tag{52}$$

These are the central objects of our analysis. The rates of RSSs, $\epsilon_1$ and $\epsilon_2$, are derived in the zero-temperature limit. Other quantities of interest can also be derived from the free energies.

To take the configurational average and the average over $\boldsymbol{x}^{(1)}$ in eq. (52), we use the replica method. For the evaluation of $f_2$, we need to introduce two different replica numbers: $n$ for $1/Z_1$ in $p_1$ and $\nu$ for $\log Z_2$. Correspondingly, we introduce the following replica-generating functions:

$$\Phi_1(n, \beta) = \left[Z_1^n\right]_{\boldsymbol{\xi},A,\hat{\boldsymbol{x}}}\,, \tag{53}$$

$$\Phi_2(n, \nu, \beta) = \left[Z_1^{n-1}\left(\prod_{i=1}^{N}\int_{-\infty}^{\infty} dx_i^{(1)}\right) e^{-\beta\mathcal{H}_1(\boldsymbol{x}^{(1)},\beta|\boldsymbol{\xi},A,\hat{\boldsymbol{x}})}\left(Z_2(\beta|\boldsymbol{x}^{(1)}, \boldsymbol{\xi}, A, \hat{\boldsymbol{x}})\right)^{\nu}\right]_{\boldsymbol{\xi},A,\hat{\boldsymbol{x}}}, \tag{54}$$

We derive the free energies from $\Phi_1$ and $\Phi_2$ by using the following identities:

$$-\beta f_1 = \lim_{n\to 0}\frac{1}{nN}\log \Phi_1(n, \beta) = \lim_{n\to 0}\frac{1}{nN}\log \Phi_2(n, 0, \beta), \tag{55}$$

$$-\beta f_2 = \lim_{n\to 0}\lim_{\nu\to 0}\frac{1}{\nu N}\log \Phi_2(n, \nu, \beta). \tag{56}$$

In the actual procedure, we first assume that $n$ and $\nu$ in eqs. (53,54) are positive integers, which enables us to calculate the average over the quenched variables as well as the integrations over $\boldsymbol{x}^{(1)}$. Then, we assume the replica symmetry (RS) in the order parameters explained next, which makes it possible to analytically continue the resultant formulas of $\Phi_2$ with respect to $n$ and $\nu$. Finally, using the analytic continuation, we calculate the limits $n \to 0$ and $\nu \to 0$, yielding the free energies.

## 4.2 Free energies, order parameters, and quantities of interest

### 4.2.1 Order parameters and their significance

Let us start by summarizing the order parameters characterizing the free energies as follows:

$$m_1 = \frac{1}{N}\sum_i \left\langle \hat{x}_i x_i^{(1)}\right\rangle_1, \quad m_2 = \frac{1}{N}\sum_i \left\langle\left\langle \hat{x}_i |x_i^{(1)}|_0 x_i^{(2)}\right\rangle\right\rangle, \tag{57a}$$

$$Q_1 = \frac{1}{N}\sum_i \left\langle \left(x_i^{(1)}\right)^2\right\rangle_1, \quad Q_2 = \frac{1}{N}\sum_i \left\langle\left\langle \left(\left|x_i^{(1)}\right|_0 x_i^{(2)}\right)^2\right\rangle\right\rangle, \tag{57b}$$

$$q_1 = \frac{1}{N}\sum_i \left\langle x_i^{(1)}\right\rangle_1^2, \quad q_2 = \frac{1}{N}\sum_i \left\langle\left\langle \left|x_i^{(1)}\right|_0 x_i^{(2)}\right\rangle\right\rangle^2, \tag{57c}$$

$$Q_c = \frac{1}{N}\sum_i \left\langle\left\langle x_i^{(1)} x_i^{(2)}\right\rangle\right\rangle, \quad q_c = \frac{1}{N}\sum_i \left\langle\left\langle \left\langle x_i^{(1)}\right\rangle_1 \left|x_i^{(1)}\right|_0 x_i^{(2)}\right\rangle\right\rangle. \tag{57d}$$

Their meaning is simple: $m_{1,2}$ denote the overlaps with the true signal; $Q_{1,2}$ represent the lengths of the reconstructed signals; $q_{1,2}$ quantify the lengths of the averaged reconstructed signals; and $q_c$ and $Q_c$ indicate the overlaps between the two reconstructed signals, the former reflects the fluctuation, and the latter does not. The MSEs (12) are connected to the order parameters as follows:

$$\mathcal{M}_1 = \hat{\rho}\sigma_x^2 - 2m_1 + Q_1, \quad \mathcal{M}_2 = \hat{\rho}\sigma_x^2 - 2m_2 + Q_2. \tag{58}$$

For simplicity of notation, we also introduce the following symbols:

$$\mathcal{M}_c = \hat{\rho}\sigma_x^2 - (m_1 + m_2) + Q_c, \quad \widetilde{\mathcal{M}}_{1,2,c} = \mathcal{M}_{1,2,c} + \sigma_\xi^2. \tag{59}$$

In the calculation of the zero-temperature limit $\beta \to \infty$, the thermal fluctuation shrinks and the order parameters $Q$ and $q$ have a common value. The following order parameters become finite in the limit $\beta \to \infty$:

$$\chi_1 = \beta(Q_1 - q_1), \ \chi_2 = \beta(Q_2 - q_2), \ \chi_c = \beta(Q_c - q_c). \tag{60}$$

From the definition of the order parameters (57), we can understand that $\chi_1$ and $\chi_2$ are nothing but the average of the diagonal part of the susceptibility matrix and that an equality $\chi_1 = \chi_2$ holds as discussed above.

#### 4.2.2 $f_1$-related quantities

We are only interested in the zero-temperature limit $\beta \to \infty$, and write the explicit formula of $f_1$ in this limit as follows:

$$\begin{aligned} f_1(\beta \to \infty) = \mathop{\mathrm{Extr}}_{\Omega_1} \Bigg\{ & -\frac{1}{2}\hat{Q}_1 Q_1 + \frac{1}{2}\hat{\chi}_1\chi_1 + \hat{m}_1 m_1 \\ & -\frac{1}{\hat{Q}_1}\left(\hat{\rho}F(\theta_A) + (1-\hat{\rho})F(\theta_I)\right) + \frac{\alpha}{2}\frac{\widetilde{\mathcal{M}}_1}{1+\chi_1} \Bigg\}. \end{aligned} \tag{61}$$

where $\Omega_1 = \left\{\chi_1, Q_1, m_1, \hat{\chi}_1, \hat{Q}_1, \hat{m}_1\right\}$ and $\mathrm{Extr}_x$ represents taking an extremization condition with respect to $x$. The variable $\hat{\chi}_1$ is a conjugate order parameter of $\chi_1$, as are the other hatted variables except for $\hat{\rho}$. Further,

$$E_k(\theta) \equiv \int_\theta^\infty \frac{dz}{\sqrt{2\pi}}\, z^k = \int_\theta^\infty Dz\; z^k, \tag{62}$$

$$F(\theta) = \lambda^2 \left\{ E_0(\theta) - \frac{1}{\theta}\frac{e^{-\frac{1}{2}\theta^2}}{\sqrt{2\pi}} + \frac{1}{\theta^2}E_0(\theta) \right\}, \tag{63}$$

$$\theta_A = \frac{\lambda}{\sqrt{\hat{\chi}_1 + \hat{m}_1^2\sigma_x^2}}, \ \theta_I = \frac{\lambda}{\sqrt{\hat{\chi}_1}}. \tag{64}$$

Next, variational conditions with respect to $\Omega_1$ yield the following equations of state (EOSs) of $\Omega_1$:

$$\hat{\chi}_1 = \frac{\alpha \widetilde{\mathcal{M}}_1}{(1+\chi_1)^2}, \tag{65a}$$

$$\hat{Q}_1 = \frac{\alpha}{1+\chi_1}, \tag{65b}$$

$$\hat{m}_1 = \frac{\alpha}{1+\chi_1} = \hat{Q}_1, \tag{65c}$$

$$\chi_1 = \frac{2}{\hat{Q}_1}\left(\hat{\rho}E_0(\theta_A) + (1-\hat{\rho})E_0(\theta_I)\right), \tag{65d}$$

$$Q_1 = \frac{2}{\hat{Q}_1^2}\left(\hat{\rho}F(\theta_A) + (1-\hat{\rho})F(\theta_I)\right), \tag{65e}$$

$$m_1 = 2\frac{\hat{m}_1}{\hat{Q}_1}\hat{\rho}\sigma_x^2 E_0(\theta_A), \tag{65f}$$

Thus, the sparsity of the reconstructed signal $\rho$, the true positive ratio $TP$, and the false positive ratio $FP$ can be expressed as follows:

$$\rho = 2\left(\hat{\rho}E_0(\theta_A) + (1-\hat{\rho})E_0(\theta_I)\right), \tag{66a}$$

$$FP = 2E_0(\theta_I), \tag{66b}$$

$$TP = 2E_0(\theta_A). \tag{66c}$$

From the EOSs, we get the following simple relations:

$$\chi_1 = \frac{\rho}{\alpha-\rho}, \tag{67a}$$

$$\hat{Q}_1 = \alpha - \rho, \tag{67b}$$

The free energy $f_1$ includes the contribution of the regularization term $\lambda||\boldsymbol{x}^{(1)}||_1$. This contribution can be represented by using the relation (65) as follows:

$$\frac{\lambda}{N}\left\langle ||\boldsymbol{x}^{(1)}||_1 \right\rangle = -Q_1\hat{Q}_1 + m_1\hat{m}_1 + \hat{\chi}_1\chi_1. \tag{68}$$

Subtracting this from $f_1$ and using eq. (65) again, we obtain the following simple formula of $\epsilon_1$:

$$\epsilon_1 = \frac{1}{\alpha}\left(f_1 - \frac{1}{N}\left\langle \lambda||\boldsymbol{x}^{(1)}||_1 \right\rangle\right) = \frac{\hat{\chi}_1}{2\alpha}. \tag{69}$$

### 4.2.3 $f_2$-related quantities

Similarly, the formula of $f_2$ is as follows:

$$\begin{aligned}\alpha\epsilon_2 = f_2(\beta\to\infty) = \underset{\Omega_2}{\mathrm{Extr}}\Bigg\{ &\frac{1}{2}\frac{\alpha}{1+\chi_2}\left(\frac{\chi_c^2}{(1+\chi_1)^2}\widetilde{\mathcal{M}}_1 - 2\frac{\chi_c}{1+\chi_1}\widetilde{\mathcal{M}}_c + \widetilde{\mathcal{M}}_2\right) \\ &-\frac{1}{2}Q_2\hat{Q}_2 + \frac{1}{2}\chi_2\hat{\chi}_2 + m_2\hat{m}_2 + Q_c\hat{Q}_c + \chi_c\hat{\chi}_c \\ &-\frac{1}{2\hat{Q}_2}\Bigg(\rho\hat{\chi}_2 + \hat{m}_2^2 m_1 + 2\hat{Q}_c\left(\hat{\chi}_c\chi_1 + \hat{m}_2 m_1\right) + \hat{Q}_c^2 Q_1 \\ &+2\left\{\hat{\rho}(\hat{\chi}_c + \hat{m}_1\hat{m}_2\sigma_x^2)^2 G(\theta_A) + (1-\hat{\rho})\hat{\chi}_c^2 G(\theta_I)\right\}\Bigg)\Bigg\}\end{aligned} \tag{70}$$

where $\Omega_2 = \left\{\chi_2, Q_2, m_2, \hat{\chi}_2, \hat{Q}_2, \hat{m}_2, \chi_c, Q_c, \hat{\chi}_c, \hat{Q}_c\right\}$ and

$$G(\theta) = \frac{\theta^3}{\lambda^2}\frac{e^{-\frac{1}{2}\theta^2}}{\sqrt{2\pi}}, \tag{71}$$

The EOSs with respect to $\Omega_2$ are as follows:

$$\hat{\chi}_2 = \frac{\alpha}{(1+\chi_2)^2}\left\{\frac{\chi_c^2}{(1+\chi_1)^2}\widetilde{\mathcal{M}}_1 - 2\frac{\chi_c}{1+\chi_1}\widetilde{\mathcal{M}}_c + \widetilde{\mathcal{M}}_2\right\}, \tag{72a}$$

$$\hat{Q}_2 = \frac{\alpha}{1+\chi_2}, \tag{72b}$$

$$\hat{m}_2 = \frac{\alpha}{1+\chi_2}\left(1 - \frac{\chi_c}{1+\chi_1}\right), \tag{72c}$$

$$\hat{\chi}_c = \frac{\alpha}{(1+\chi_1)(1+\chi_2)}\left(\widetilde{\mathcal{M}}_c - \frac{\chi_c}{1+\chi_1}\widetilde{\mathcal{M}}_1\right), \tag{72d}$$

$$\hat{Q}_c = \frac{\alpha}{1+\chi_2}\frac{\chi_c}{1+\chi_1}, \tag{72e}$$

$$\chi_2 = \frac{\rho}{\hat{Q}_2}, \tag{72f}$$

$$\begin{aligned} Q_2 = \frac{1}{\hat{Q}^2}\Bigg\{&\rho\hat{\chi}_2 + \hat{m}_2^2 m_1 + 2\hat{Q}_c\left(\hat{\chi}_c\chi_1 + \hat{m}_2 m_1\right) + \hat{Q}_c^2 Q_1 \\ &+2\left\{\hat{\rho}(\hat{\chi}_c + \hat{m}_1\hat{m}_2\sigma_x^2)^2 G(\theta_A) + (1-\hat{\rho})\hat{\chi}_c^2 G(\theta_I)\right\}\Bigg\}, \end{aligned} \tag{72g}$$

$$m_2 = \frac{1}{\hat{Q}_2}\left\{m_1(\hat{m}_2 + \hat{Q}_c) + 2\hat{m}_1\sigma_y^2(\hat{\chi}_c + \hat{m}_1\hat{m}_2\sigma_x^2)G(\theta_A)\right\}, \tag{72h}$$

$$\chi_c = \frac{1}{\hat{Q}_2}\left\{\hat{Q}_c\chi_1 + 2\left\{\hat{\rho}(\hat{\chi}_c + \hat{m}_1\hat{m}_2\sigma_x^2)G(\theta_A) + (1-\hat{\rho})\hat{\chi}_c G(\theta_I)\right\}\right\}, \tag{72i}$$

$$Q_c = \frac{1}{\hat{Q}_2}\left\{\hat{\chi}_c\chi_1 + \hat{m}_2 m_1 + \hat{Q}_c Q_1\right\}. \tag{72j}$$

For the order parameters of $\Omega_1$ appearing in eq. (70), we insert the solutions of the extremization condition considered in eq. (65). From the EOSs, we obtain the following simple relations:

$$\chi_2 = \chi_1 = \frac{\rho}{\alpha - \rho}, \tag{73a}$$

$$\hat{Q}_2 = \hat{Q}_1 = \alpha - \rho, \tag{73b}$$

$$\epsilon_2 = \frac{\hat{\chi}_2}{2\alpha}. \tag{73c}$$

As expected, the two susceptibilities $\chi_1$ and $\chi_2$ coincide.

## 4.3 LOOEs, MSEs, and ROC curve

Since the non-diagonal part of $\chi_{ij}^{\backslash\mu}$ can be neglected and the diagonal part is $\chi = \rho/(\alpha-\rho)$, we have $(1 + \sum_{i,j} A_{\mu i}A_{\mu j}\chi_{ij}^{\backslash\mu})^2 = \left(\frac{\alpha}{\alpha-\rho}\right)^2$. From eqs. (42,65) and (69), the first type of LOOE becomes

$$\mathcal{L}_1 = \frac{1}{2}\widetilde{\mathcal{M}}_1. \tag{74}$$

Hence, the LOOE directly connects to the corresponding MSE and calculating its minimum is meaningful. This result is natural and can be derived from a simple consideration. Given $\boldsymbol{x}^{(1)\backslash\mu}$, let us consider the difference between $y_\mu$ and the counter part of the reconstructed data $y_\mu^{(1)} = \sum_i A_{\mu i} x_i^{(1)\backslash\mu}$. In the present situation, none of the rows of $A$ and none of the components of $\boldsymbol{\xi}$ are correlated, and hence, $\boldsymbol{x}^{(1)\backslash\mu}$ is also uncorrelated with $\{A_{\mu i}\}_i$ and $\xi_\mu$. This implies that the average of $(y_\mu - y_\mu^{(1)})^2$ over $\xi_\mu$ and $\{A_{\mu i}\}_i$ is as follows:

$$\left[(y_\mu - y_\mu^{(1)})^2\right]_{\xi_\mu,\{A_{\mu i}\}_i} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{x}_i - x_i^{(1)\backslash\mu}\right)^2 + \sigma_\xi^2 \approx \widetilde{\mathcal{M}}_1. \tag{75}$$

The last relation follows from the smallness of the difference between $\boldsymbol{x}^{(1)\backslash\mu}$ and $\boldsymbol{x}^{(1)}$. This relation immediately leads to eq. (74).

Clearly, this discussion is applied to $\mathcal{L}_2$ since $\boldsymbol{x}^{(2)\backslash\mu}$ is again uncorrelated with $\{A_{\mu i}\}_i$ and $\xi_\mu$, and $\mathcal{L}_2$ should become

$$\mathcal{L}_2 = \frac{1}{2}\widetilde{\mathcal{M}}_2. \tag{76}$$

However, our calculation based on eq. (42), combined with the replica result (72,73c), yields the following:

$$\mathcal{L}_2 = \frac{1}{2}\left\{\frac{\chi_c^2}{(1+\chi_1)^2}\widetilde{\mathcal{M}}_1 - 2\frac{\chi_c}{1+\chi_1}\widetilde{\mathcal{M}}_c + \widetilde{\mathcal{M}}_2\right\} \quad \text{(incorrect)}. \tag{77}$$

Only the last term is desired, but the other two terms appear and persist. Eq. (42) thus provides an incorrect approximation of $\mathcal{L}_2$, in contrast to $\mathcal{L}_1$.

To obtain quantitative information, we plot the LOOEs in Fig. 1. Three curves are presented in each panel: $\mathcal{L}_1 (= (1/2)\widetilde{\mathcal{M}}_1)$ (red), $\mathcal{L}_2 (= (1/2)\widetilde{\mathcal{M}}_2)$ (blue), and the incorrect evaluation of

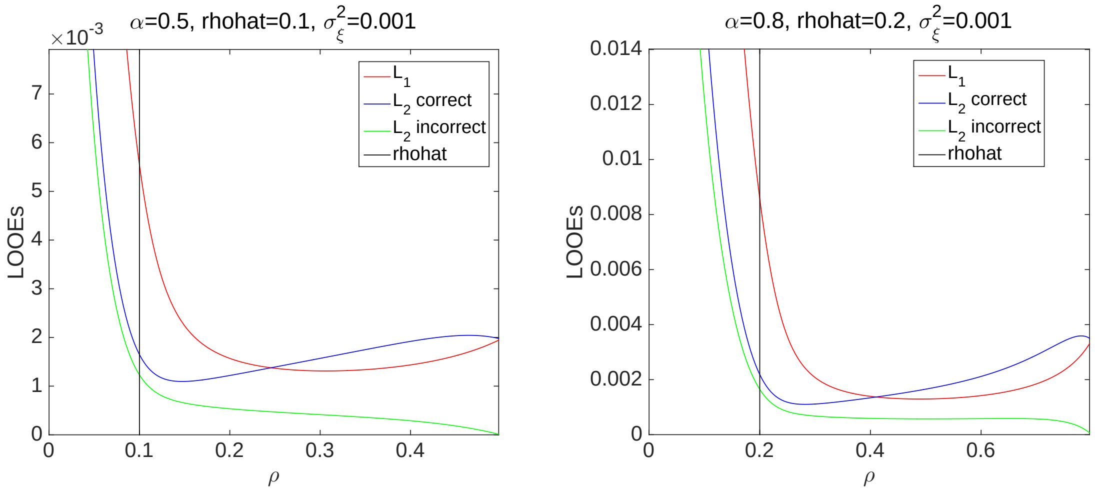

Figure 1: Plots of different estimations of LOOEs: $\mathcal{L}_1$ (red), the correct estimation of $\mathcal{L}_2$ (blue), and the incorrect estimation of $\mathcal{L}_2$ based on eq. (77). The parameters are $(\alpha, \hat{\rho}) = (0.5, 0.1)$ (left) and $(\alpha, \hat{\rho}) = (0.8, 0.2)$ (right). The noise strength is commonly set to be $\sigma_\xi^2 = 0.001$. The difference between the two estimations of $\mathcal{L}_2$ is not negligible. The minimum value of the correct $L_2$ is located at a smaller $\rho$ than that of $\mathcal{L}_1$ in both the cases.

$\mathcal{L}_2$ by eq. (77) (green). The deviation between the blue and the green curves is not negligible and is qualitatively different. The incorrect one converges to zero in the limit $\rho \to \alpha$, but the true one goes to a finite constant identical to the limiting value of $\mathcal{L}_1$. This implies that the approximation (77) is completely useless. Its reasoning will be given later in sec. 5.1.1.

We observe that $\mathcal{L}_1$ and the correct $\mathcal{L}_2$ have their unique minimums at certain values of $\rho(< \alpha)$. This implies that the minimums of the LOOEs are good determinators of the value of $\lambda$ since they are connected to the minimums of the MSEs, as shown in eqs. (74,76). To quantify the quality of inference by these two minimums of $\mathcal{L}_1$ and $\mathcal{L}_2$, we plot the ROC curves as a plot of $TP$ against $FP$ in Fig. 2. The $N \to \infty$ solution is denoted by blue points, and the scatter plots of the finite-size simulation with $N = 3600$ over 10 samples are indicated with magenta circles. This simulation is conducted using the built-in "lasso" function of MATLAB®. In the above figure, we mark the points obtained using the minimum of $\mathcal{L}_1$ as red points, those obtained using Youden's index as green points, and those obtained using the minimum of $\mathcal{L}_2$ as blue points. The best inference is achieved at the upper-most left point, $(FP, TP) = (0, 1)$, and better ROC curves are increasingly skewed to the upper-left direction. The black straight lines denote the $FP = TP$ line and are given as a reference for observing the skewness. Fig. 2 demonstrates that the inference based on LASSO has a good performance. Yet, the points obtained using the minimum values of $\mathcal{L}_1$, red points, are located a little away from the "optimal" values obtained using Youden's index. Meanwhile, the minimums of $\mathcal{L}_2$ are very close to Youden's optimal values, and in this sense, $\mathcal{L}_2$ is better than $\mathcal{L}_1$ for determining $\lambda$. However, note that for obtaining the minimum of $\mathcal{L}_2$, we have to naively conduct the LOO CV according to its definition since our approximation formulas (40,42) do not provide reliable estimates of $\mathcal{L}_2$. Unfortunately, even this naive method faces some difficulties in addition to the computational time. This will be discussed in sec. 5.2.

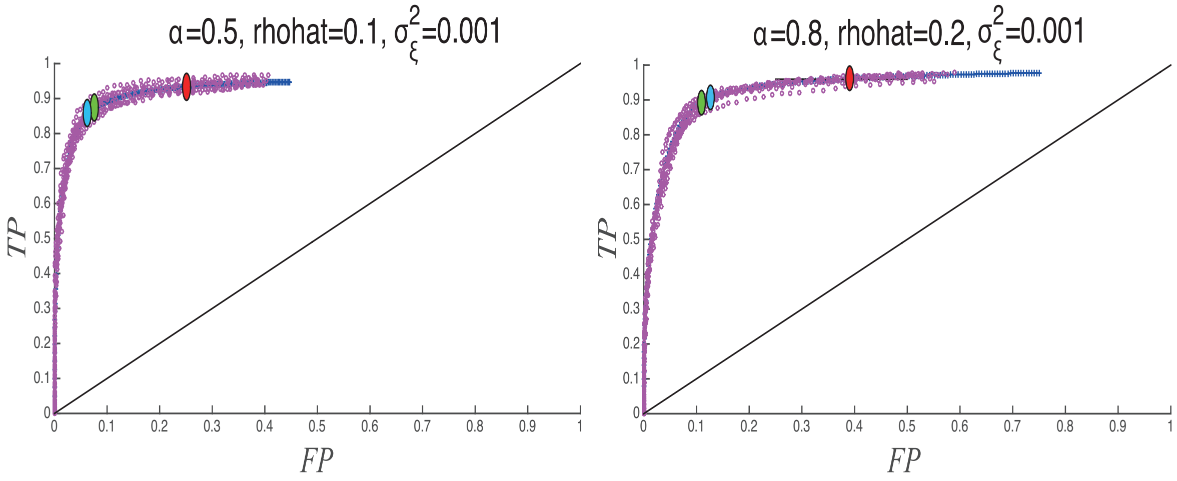

Figure 2: ROC curves for $(\alpha, \hat{\rho}) = (0.5, 0.1)$ (left) and for $(\alpha, \hat{\rho}) = (0.8, 0.2)$ (right). The minimum of $\mathcal{L}_1$ (red point), Youden's index (green point), and the minimum of $\mathcal{L}_2$ (blue point) give the coordinate values of $(FP, TP) = (0.24, 0.93)$, $(0.078, 0.087)$, and $(0.068, 86)$ in the left panel, and $(FP, TP) = (0.37, 0.96)$, $(0.11, 0.89)$, and $(0.13, 0.91)$ in the right panel, respectively. The solution for the limit $N \to \infty$ is denoted by blue points, and the scatter plots of finite-size simulations with $N = 3600$ over 10 samples are indicated with magenta circles.

Further, we have an additional remark about $\mathcal{L}_1$. The minimum of $\mathcal{L}_1$ tends to overestimate the false positive ratio as shown in Fig. 2. We have checked several different values of the parameters and confirmed that this always holds if the noise is sufficiently weak. An empirical prescription to overcome this is the so-called one-standard error rule, which chooses a larger value of $\lambda$ (corresponding to a smaller $FP$) than that by the minimum of $\mathcal{L}_1$, by using the error bar of the minimum data point. Hence, it is important to approximate not only the minimum value but also its error bar. Fortunately, our formulas eqs. (35,42) can also provide the error bar of $\mathcal{L}_1$, which will be demonstrated by a real-data application discussed in sec. 5.2.

# 5 Comparison with data on finite size systems

In this section, numerical simulations are presented to examine the validity of our analysis and to determine the finite-size effect. We also apply the proposed method to SuperNova DataBase provided by the Berkeley Supernova Ia program [29] and find that this method reproduces the obtained result [11] considerably faster than the conventional 10-fold CV.

## 5.1 Examination using artificial data

In this subsection, we numerically generate the observation matrix $A$, the true sparse signal $\hat{\boldsymbol{x}}$, and the noise $\boldsymbol{\xi}$, matching the assumptions made in our analysis. In all the simulations here, we set $\alpha = 0.8, \sigma_x^2 = 1$, and $\sigma_\xi^2 = 0.001$. Under this condition, once a set of $A, \hat{\boldsymbol{x}}$, and $\boldsymbol{\xi}$ is given, we solve eqs. (5,8) by using a versatile algorithm of convex optimization, yielding the RSSs and the LOOEs. We generate $N_s = 1000$ different samples of the set of $(A, \hat{\boldsymbol{x}}, \boldsymbol{\xi})$ and adopt the mean value in the samples as our estimate. Error bars are evaluated as standard deviations among the samples divided by $\sqrt{N_s - 1}$. In determining active sets after convex optimization, we need to introduce a certain threshold value for each signal component. Here, the threshold value is empirically chosen as $10^{-6}$.

The RSSs for $N = 16, 32, \cdots, 512$ are given in Fig. 3 and compared with the analytical curve of $N \to \infty$. We find that the finite-size effect is moderate for both $\epsilon_1$ and $\epsilon_2$, but the behavior

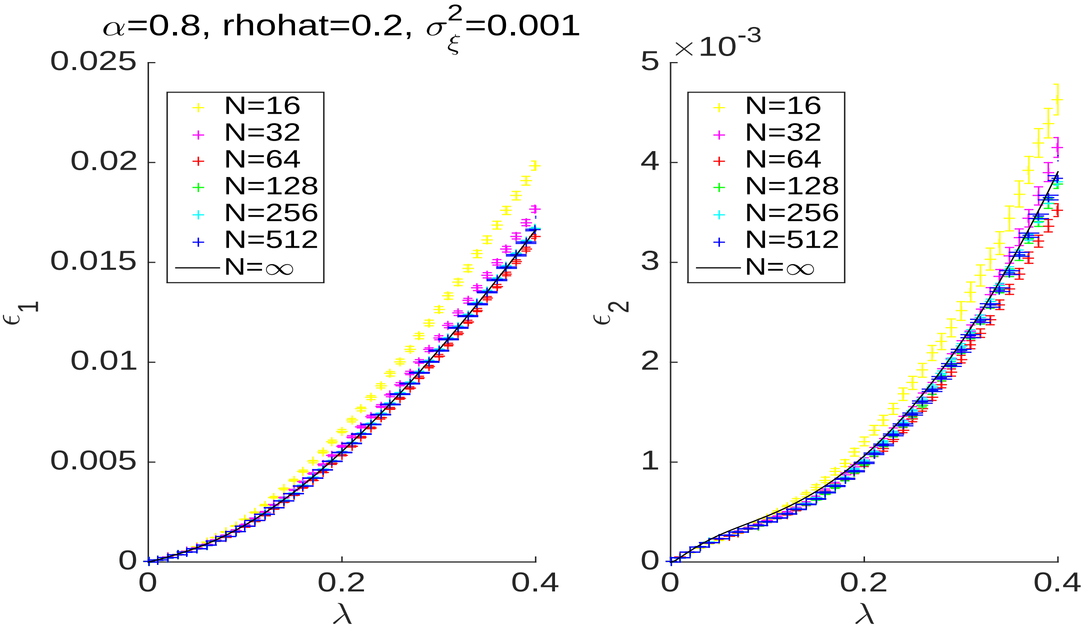

Figure 3: The RSSs $\epsilon_1$ (left) and $\epsilon_2$ (right) are plotted against $\lambda$. The behavior of $\epsilon_2$ is non-monotonic with respect to $N$ in contrast to $\epsilon_1$.

of $\epsilon_2$ is not monotonic: Up to $N = 64$, the numerical values of $\epsilon_2$ tend to become smaller as $N$ increases, but for $N > 64$, the values of $\epsilon_2$ start to increase as $N$ increases. Further, the numerical results show a fairly good agreement with the analytical curve, validating our analysis.

Next, we examine the quality of approximations of $\mathcal{L}_1$. There are two approximation methods: One is based on eq. (35) in conjunction with eqs. (37,38), which is called *Approximation 1* hereafter; the other follows eq. (42), and we call this method *Approximation 2*. They are identical in the limit $N \to \infty$, but for finite $N$, there exists a deviation. Fig. 4 provides $\mathcal{L}_1$ for $N = 16, 32, \cdots, 512$ evaluated using both Approximations 1 and 2. The finite-size effect is strong and unstable for small $N$, particularly for Approximation 1. This is reasonable because Approximation 1 requires an inversion of the matrix $\tilde{A}^{\mathrm{T}}\tilde{A}$. If the number of active variables is close to the number of observations, which is the case for a small $\lambda$, $\tilde{A}^{\mathrm{T}}\tilde{A}$ has a mode whose eigenvalue is very close to zero. This leads to the divergence of $(\tilde{A}^{\mathrm{T}}\tilde{A})^{-1}$, explaining the drastic change in the LOOEs at small values of $\lambda$ for small sizes. This effect weakens with an increase in the system size, and for $N \geq 64$, the numerical data agrees well with the analytical curve.

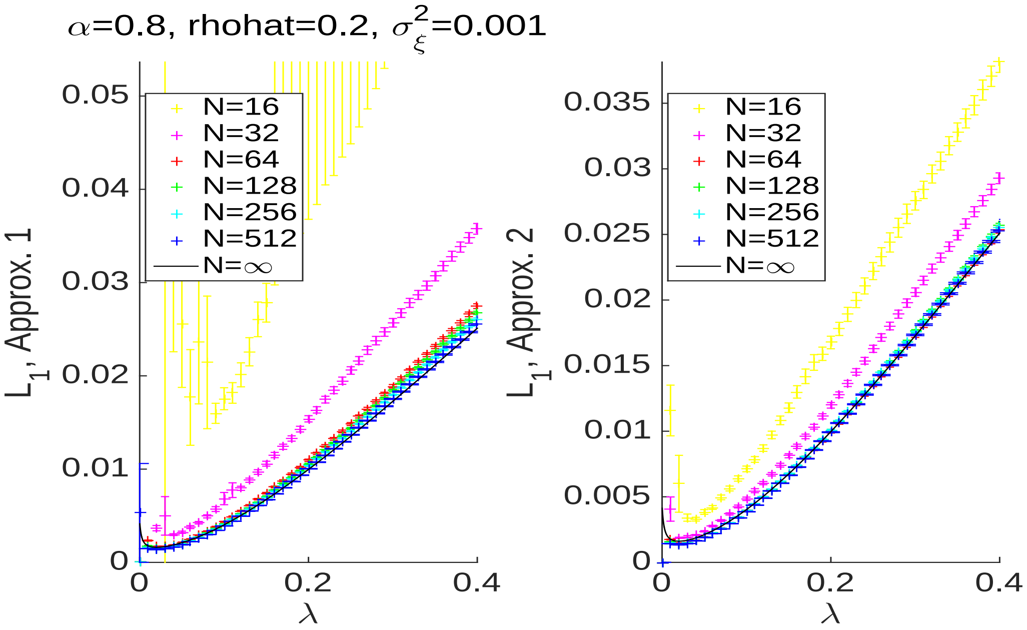

Figure 4: Approximated values of $\mathcal{L}_1$ based on eq. (35) (left) and eq. (42) (right) are plotted against $\lambda$. Numerical data (color points) show a fairly good agreement with the analytical black curve for $N \geq 64$.

Thirdly, we conduct the LOO CV directly without using any approximation and compare the result with our analytical calculations. The LOO CV is computationally expensive, and we execute it for smaller sizes up to $N = 256$. The result is given in Fig. 5. We can see that our analytical curves (black solid curves) of $\mathcal{L}_1$ (left) and $\mathcal{L}_2$ (right) show a fairly good agreement with the direct CV result for $N \geq 64$. In the right panel, we also plot the incorrect prediction based on eq. (77), and it exhibits a clear inconsistency with the numerical data.

Overall, our analytical calculations agree well with the numerical simulations for moderate system sizes, and the approximation formulas (35,42) provide reliable estimates of $\mathcal{L}_1$. The benefit of these formulas is their computational ease. For conducting the LOO CV according to its definition, the computational time required is $O(MN_{\mathrm{LASSO}})$, where $N_{\mathrm{LASSO}}$ denotes the computational time for conducting LASSO once for a desired set of $\lambda$ values, which depend on

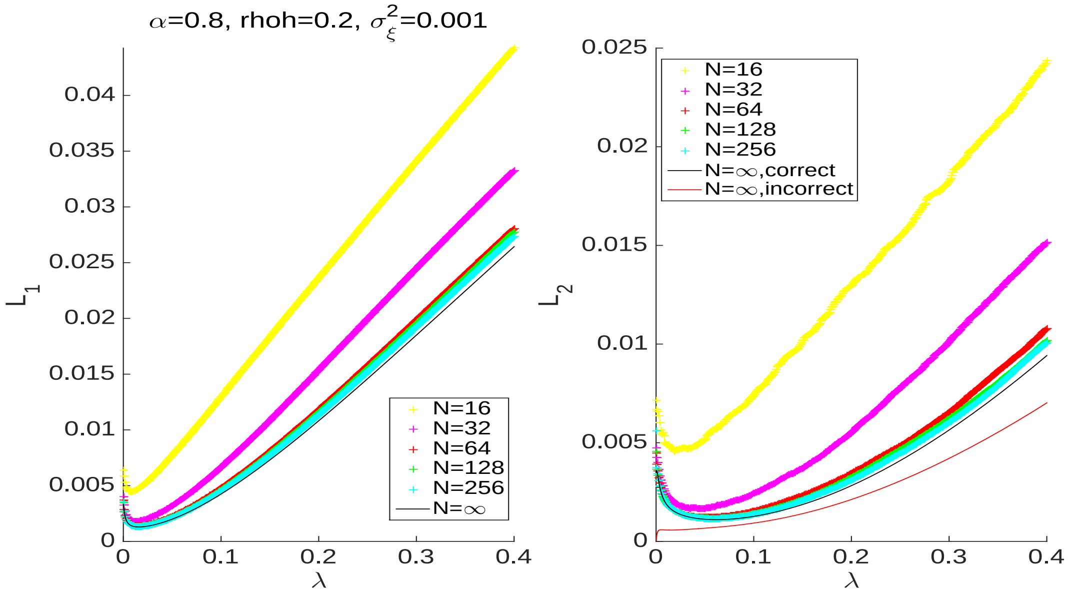

Figure 5: Numerically evaluated LOOEs, $\mathcal{L}_1$ (left) and $\mathcal{L}_2$ (right), according to their definitions, are plotted against $\lambda$. For the sake of comparison, analytical curves (black and red solid curves) are drawn. There are two analytical curves for $\mathcal{L}_2$: The black one is correct and is based on eq. (76), and the red one is incorrect and is based on eq. (77). For $N \geq 64$, the numerical data show a fairly good agreement with the analytical predictions.

the system size parameters and the algorithm used. On the other hand, on the basis of our approximation formulas, this is considerably reduced. For example, according to eq. (42), the computational time is $O(N_{\rm LASSO})$. The effect of this acceleration is discussed in sec. 5.2.

### 5.1.1 Cause of the failure on $\mathcal{L}_2$

Here, we probe the cause of the failure in approximating $\mathcal{L}_2$ based on eqs. (40,42), although the same approximation is applicable for $\mathcal{L}_1$. While deriving eqs. (35,40) and (42), we assumed that the set of active variables is stable against small perturbations and that the addition or deletion of a row of the observation matrix $A$ just leads to such small perturbations. This assumption is examined here.

In the absence of the $\mu$th row of $A$, the corresponding fixed-point equations of the AMP become

$$a_\nu^{(1)\backslash\mu} = y_\nu - \sum_i A_{\nu i} x_i^{(1)\backslash\mu}, \tag{78a}$$

$$h_i^{(1)\backslash\mu} = \sum_{\nu(\neq\mu)} A_{\nu i} a_\nu^{(1)\backslash\mu} + \Gamma_i x_i^{(1)\backslash\mu}, \tag{78b}$$

$$x_i^{(1)\backslash\mu} = \frac{h_i^{(1)\backslash\mu} - \lambda \mathrm{sgn}\left(h_i^{(1)\backslash\mu}\right)}{\Gamma_i} \Theta\left(|h_i^{(1)\backslash\mu}| - \lambda\right). \tag{78c}$$

Let us call this system the $\mu$-cavity system. The difference between eq. (78) and eq. (33) is only the term $A_{\mu i} a_\mu = O\left(\sqrt{N}^{-1}\right)$ in eq. (78b). Hence, it is expected that the difference in variables between the full and the $\mu$-cavity systems is small and can be scaled as $O\left(\sqrt{N}^{-1}\right)$. Even if this assumption is correct, it is not trivial to compute the variables of the $\mu$-cavity system. However, the discussion in sec. 3.2 implies that we estimate this difference as follows:

$$h_i^{(1)} - h_i^{(1)\backslash\mu} \approx A_{\mu i} a_\mu^{(1)}. \tag{79}$$

Further, we assume that

$$\forall i, \Theta\left(|h_i^{(1)}| - \lambda\right) = \Theta\left(|h_i^{(1)\backslash\mu}| - \lambda\right). \tag{80}$$

We have examined these two relations by numerically solving both eq. (33) and eq. (78) independently, and found that the first one is satisfied in a moderate region of $\lambda$ at a certain accuracy, while the second one is incorrect in the entire range of interest of $\lambda$. This poses another question: Why is $\mathcal{L}_1$ well approximated by eqs. (35,42)?

The violation of eq. (80) implies that the active and inactive sets are different for the full and the $\mu$-cavity systems. Some variables belong to the same sets on both the systems and are called "stable". The others change the belonging sets and are called "unstable". The behavior of the unstable variables is a crucial issue. The effective field $h^{(1)}$ of any unstable variable must satisfy the relation $|h^{(1)}| - \lambda = O\left(\sqrt{N}^{-1}\right)$, since the variation of the effective field, $\Delta h^{(1)\backslash\mu} \equiv h^{(1)} - h^{(1)\backslash\mu}$, also scales as $O\left(\sqrt{N}^{-1}\right)$ and should be comparable with $|h^{(1)}| - \lambda$. This implies that the coefficient $x^{(1)}$ of any unstable variable is zero or very small as $x^{(1)} \propto |h^{(1)}| - \lambda = O\left(\sqrt{N}^{-1}\right)$. Further, the number of unstable variables is estimated as $N_{\text{uns}} \approx |\Delta h^{(1)\backslash\mu}| \times NP(h^{(1)}) = O\left(\sqrt{N}\right)$, where $P(h^{(1)})$ denotes the distribution of the effective field of the full system and is assumed to be $O(1)$ around $|h^{(1)}| \approx \lambda$. Summarizing these scalings of $N_{\text{uns}}$ and the coefficient $x^{(1)}$, we can estimate their contribution as follows:

$$\sum_{i \in \text{UNS}} A_{\mu i} x_i^{(1)} = O\left(\sqrt{N} \times \sqrt{N}^{-1} \times \sqrt{N}^{-1}\right) = O\left(\sqrt{N}^{-1}\right) \to 0, \tag{81}$$

where UNS denotes the set of unstable variables. Note that $A_{\mu i} = O\left(\sqrt{N}^{-1}\right)$. The same is true if $x_i^{(1)}$ is replaced with $x_i^{(1)\backslash\mu}$ in the above equation. Hence, the contribution from the unstable variables vanishes in both the full and the $\mu$-cavity systems in the calculation of the cavity residuals $\boldsymbol{a}^{(1)}$ and $\boldsymbol{a}^{(1)\backslash\mu}$. As a result, our perturbative discussion assuming eq. (80) is validated to calculate macroscopic quantities such as $\mathcal{L}_1$ but cannot correctly compute microscopic information such as UNS and the associated coefficients $\{x_i^{(1)}\}_{i \in \text{UNS}}$.

The above reasoning manifests why $\mathcal{L}_2$ is not correctly evaluated by eqs. (40,42). Now, the coefficients of unstable variables, $\{x_i^{(2)}\}_{i \in \text{UNS}}$, are not proportional to $|h^{(1)}| - \lambda$ and are of $O(1)$, as seen in eq. (34). Thus, its contribution is

$$\sum_{i \in \text{UNS}} A_{\mu i} x_i^{(2)} = O\left(\sqrt{N} \times \sqrt{N}^{-1} \times 1\right) = O(1), \tag{82}$$

and is not negligible, implying that the solution of the corresponding fixed-point equations is influenced by the unstable variables. Hence, our perturbative discussion does not work even in the calculation of macroscopic quantities.

### 5.2 Application to Type Ia Supernova data

Here, we apply the proposed method for evaluating $\mathcal{L}_1$ to the data from SuperNova DataBase provided by the Berkeley Supernova Ia program. Recently, LASSO techniques have been used on these data, and a set of important variables known to be significant in explaining the Type Ia supernova data empirically has been reproduced [11]. In this study, the 10-fold CV, which is an alternative to the LOO CV when the number of variables is large and performing the LOO CV is computationally difficult, is used for determining the value of $\lambda$. We calculate $\mathcal{L}_1$ by using the proposed method on these data and compare the result with that of the 10-fold CV. The system size parameters of these data are $M = 78$ and $N = 276$.

The left panel of Fig. 6 shows the plots of $\mathcal{L}_1$ in Approximations 1 and 2, and the CV error by the 10-fold CV against $\log \lambda$ without the error bars. Clearly, the curves are very similar, and the minimum values of all the curves coincide. We also observe the quantitative similarity of not only the mean values but also the error bars in the right panel of the same figure. The error bars of the 10-fold CV are obtained using a Monte-Carlo resampling, and those of $\mathcal{L}_1$ evaluated by the proposed method are given by the standard deviation among the $M$ terms of eq. (35) or eq. (42), which is justified by a simple resampling argument using the multinomial distribution. Clearly, the largeness of the error bars is quantitatively comparable. Hence, the proposed method reproduces the 10-fold CV result at a very satisfactory level. By applying the one-standard error rule for all the three methods, we obtain df = 6 (df: Number of active variables), which agrees with an empirically validated model, as explained in [11].

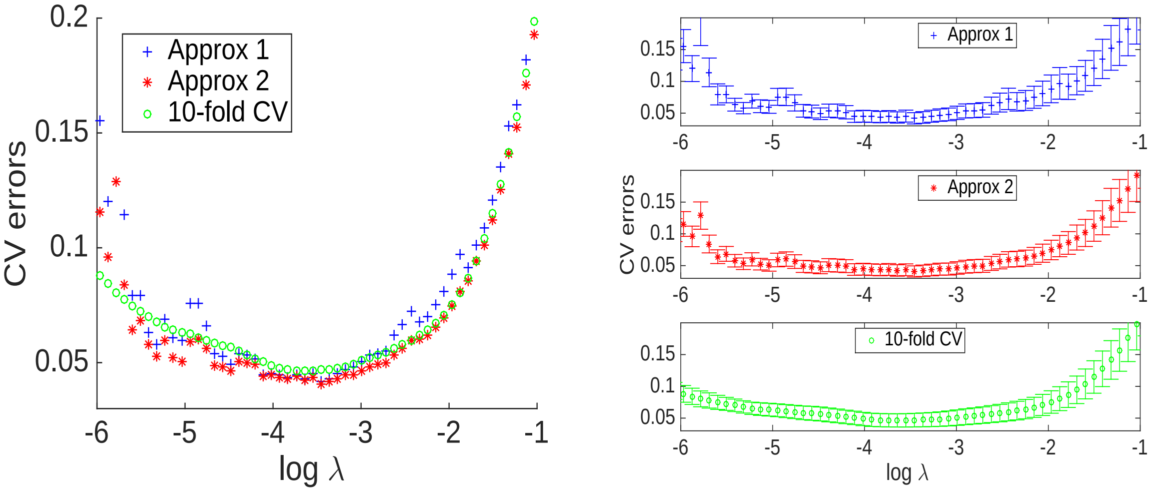

Figure 6: CV errors plotted against $\log \lambda$ for the Type Ia supernova data. Two plots (blue and red) show those obtained using the proposed method, and the other plot (green) illustrates those obtained using the 10-fold CV. The left panel is a direct comparison of the errors without the error bars, and the right panel is the same data with the error bars. All of values are in a good agreement, and the minimums coincide.

The benefit of the proposed method is apparent in the computational time. The required time for computations in an experiment is 31.6 s for the 10-fold CV, 3.20 s for Approximation 1, and 2.85 s for Approximation 2, the last two of which include the computational time of one run of LASSO. Therefore, the advantage of the proposed method is clear, and the reducing factor is about 10. If we compare with the LOO CV instead of the 10-fold CV, the factor will be considerably larger. Hence, our approximation is applicable to realistic problems and is very efficient for data with a large dimensionality. Therefore, readers are strongly encouraged to use the presented method.

The observation matrix in the Type Ia supernova data is significantly different from a simple random matrix. Hence, the success of Approximation 2 in the case of these data conversely suggests that the relation $(1+\sum_{i,j} A_{\mu i}A_{\mu j}\chi_{ij}^{\backslash \mu})^2 \approx (\alpha/(\alpha-\rho))^2$, which is used for deriving eq. (42), holds rather universally, as discussed at the end of sec. 3.2.1. As mentioned above, Approximation 2 has a relatively low computational time, and its result is stable compared to that of Approximation 1. These facts positively motivate us to find further theoretical evidence for the universality of the relation $(1+\sum_{i,j} A_{\mu i}A_{\mu j}\chi_{ij}^{\backslash \mu})^2 \approx (\alpha/(\alpha-\rho))^2$.

Now, finally, we report the behavior of $\mathcal{L}_2$ on the Type Ia supernova data. The LOO CV according to its definition is conducted, and the result is shown in Fig. 7. The minimum value of $\mathcal{L}_2$ is located at $\log\lambda \approx -1$, leading to df $= 1$. This serious reduction of df from the case of $\mathcal{L}_1$ might have a certain meaning. In fact, in [11], an appropriate preprocessing of the same data shows the same serious reduction of df, which can be attributed to some inconvenient properties of the data such as collinearity and bad statistics. The possibility that $\mathcal{L}_2$ can diminish the influence of these bad properties is suggested.

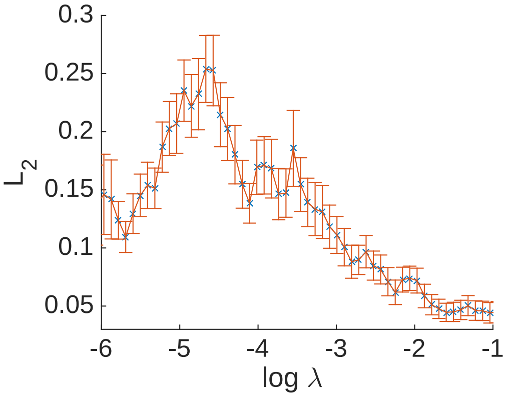

Figure 7: Second type of LOOE $\mathcal{L}_2$ plotted against $\log\lambda$ for the Type Ia supernova data.

Unfortunately, as seen in the rugged uncontrolled behavior of $\mathcal{L}_2$, the quality of the presented data is not sufficiently good to judge whether this possibility is plausible or not, although it is natural that df is reduced when using $\mathcal{L}_2$ instead of $\mathcal{L}_1$, as demonstrated in Fig. 2. Basically, $\mathcal{L}_2$ requires considerably better statistics to exhibit a meaningful behavior than $\mathcal{L}_1$. The reason for this is as follows: When changing $\lambda$, each term in $\mathcal{L}_2$ consists of several different piecewise constants, and they are not necessarily monotonic. This is because the inferred signal $\boldsymbol{x}^{(2)\backslash\mu}(\lambda)$ shows a sudden change only at certain several discrete points of $\lambda$, where the set of active variables changes, and remains unchanged in-between the neighboring pairs of these discrete points. This is in contrast to $\boldsymbol{x}^{(1)\backslash\mu}(\lambda)$, which changes continuously [6, 30]. These discrete points change among different terms in $\mathcal{L}_2$. Hence, $\mathcal{L}_2$ consists of the sum of the piecewise constants with different jumping points and heights, leading to large error bars and uncontrolled behaviors of $\mathcal{L}_2$.

Once this uncontrolled behavior of $\mathcal{L}_2$ is tamed, we expect an optimal value of $\lambda$ to be chosen by using $\mathcal{L}_2$ without employing the ad hoc one-standard error rule [30, 31, 32]. New ideas are desired for taming. An idea based on bootstrapping can be a good candidate: Increasing the statistics of the present data can diminish the abovementioned discrete behavior. Problems of rather large sizes may not pose this peculiarity in the first place, since the statistics of the

LOOE automatically improve with an increase in $M$. In less large but moderate-size problems, the $k$-fold CV for $\mathcal{L}_2$ with a moderate value of $k$ is worth trying. However, further consideration of $\mathcal{L}_2$ is desired.

## 6 Conclusion

In this study, we examined the LOO CV as the determinator of the coefficient $\lambda$ of the penalty term in LASSO. We investigated two types of CV errors by using the LOO CV, namely the LOOEs $\mathcal{L}_1$ and $\mathcal{L}_2$ corresponding to two different estimators, and for both the errors, we derived simple formulas that significantly reduce the computational cost in their evaluation. This result was derived by using the BP or cavity method and by a perturbative argument assuming that the number of observations was sufficiently large.

On the basis of this finding, we analytically evaluated the LOOEs by using the replica method when the observation matrix is a simple random matrix. This provided quantitative information about the LOOEs. Further, the locations of the minimums of the two LOOEs were found to be different, and thus, the chosen "optimal" values of $\lambda$ are different in general. Both the optimal values were examined by using ROC curves, and the one obtained using the second LOOE $\mathcal{L}_2$ was found to be preferable. However, a replica analysis clarified that our simple formulas are not useful for accurately approximating $\mathcal{L}_2$. We need further consideration to design an efficient algorithm for computing $\mathcal{L}_2$.

The above analytical calculations were compared with numerical simulations on finite-size systems. For small system sizes, there exists a deviation, but for moderate and large system sizes, the agreement between the numerical data and the analytical result is fairly good, and thus, our formulas are validated. We also applied these formulas to real Type Ia supernova data, to find that the proposed method reproduces the known result at a very satisfactory level. The benefit of our formulas is their low computational cost, and the actual reducing factor in the computational time was about 10. Further, the computation of $\mathcal{L}_2$ according to its definition was conducted on this data set, but it turned out to be difficult to obtain any meaningful result. Larger amounts of data are desired to treat $\mathcal{L}_2$.

The proposed method requires the computation of a pre-factor $\sum_{ij} A_{\mu i} A_{\mu j} \chi_{ij}^{\backslash \mu}$ and it is in fact a non-trivial task. We had recourse to an analytical formula for this quantity, which can be directly validated in the case where the observation matrix is a simple random matrix but cannot be justified in general. Some of our numerical (unreported) observations support that the analytical formula holds for a wider ensemble of the observation matrix, but deeper theoretical evidence is strongly desired.

## Acknowledgments

We would like to express our sincere gratitude to Makoto Uemura and Shiro Ikeda for their helpful discussion on the Type Ia supernova data. This work was supported by JSPS KAKENHI Grant Numbers 26870185 (TO) and 25120013 (YK).

## A Assessing the generating function $\Phi_2$

For positive integers $n$ and $\nu$, the generating function $\Phi_2(n, \nu)$ can be expressed as follows:

$$\Phi_2(n,\nu) = \widetilde{\mathop{\mathrm{Tr}}_{\{\boldsymbol{r}_a\}_{a=1}^{n}}} \mathop{\mathrm{Tr}}_{\{\boldsymbol{x}_\alpha|\boldsymbol{r}_1\}_{\alpha=1}^{\nu}} \left[ e^{-\frac{1}{2}\beta\left\{\sum_{a=1}^{n}(d_{\mu a}+\xi_\mu)^2+\sum_{\alpha=1}^{\nu}(\tilde{d}_{\mu\alpha}+\xi_\mu)^2\right\}} \right]_{\boldsymbol{\xi},A,\hat{\boldsymbol{x}}}, \qquad (83)$$

where

$$\widetilde{\mathop{\mathrm{Tr}}_{\boldsymbol{r}}} = \prod_{i=1}^{N}\left\{\int dr_i\ e^{-\beta\lambda|r_i|}\right\}, \quad \mathop{\mathrm{Tr}}_{\boldsymbol{x}|\boldsymbol{r}} = \prod_{i=1}^{N}\left\{\int d_{|r_i|_0}x_i\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_i^2}\right\}. \tag{84}$$

$$d_{\mu a} = \sum_i A_{\mu i}(\hat{x}_i - r_i^a), \quad \tilde{d}_{\mu\alpha} = \sum_i A_{\mu i}(\hat{x}_i - |r_i^1|_0 x_i^\alpha). \tag{85}$$

Hereafter, we assume that the unspecified domain of integration denotes the integration over $[-\infty:\infty]$ or $[-i\infty:i\infty]$. We also assume that indices $a, b$ run over $1,\cdots,n$, and $\alpha,\beta$ over $1,\cdots,\nu$. The quantities $d_{\mu a}$ and $\tilde{d}_{\nu\alpha}$ consist of extensive sums of random variables $\{A_{\mu i}\}_i$, implying that the central limit theorem works and that $d$ and $\tilde{d}$ can be expressed by certain Gaussian variables. The mean is clearly zero, and the covariance becomes

$$[d_{\mu a}d_{\omega b}]_A = \delta_{\mu\omega}\left(\frac{1}{N}\sum_i \hat{x}_i^2 - \frac{1}{N}\sum_i \hat{x}_i r_i^a - \frac{1}{N}\sum_i \hat{x}_i r_i^b + \frac{1}{N}\sum_i r_i^a r_i^b\right). \tag{86}$$

Accordingly, we define the following order parameters:

$$Q_{ab} = \frac{1}{N}\sum_i r_i^a r_i^b,\ m_a = \frac{1}{N}\sum_i \hat{x}_i r_i^a, \tag{87}$$

and assume the replica symmetry (RS) to be as follows:

$$Q_{ab} = Q_1\delta_{ab} + q_1(1-\delta_{ab}),\ m_a = m_1. \tag{88}$$

This RS assumption allows us to make many simplifications in dealing with $d_{\mu a}$, and $d_{\mu a}$ is represented by a sum of two independent Gaussian variables $v$ and $z$ drawn from $\mathcal{N}(0,1)$ as follows:

$$d_{\mu a} = \sqrt{Q_1 - q_1}v_{\mu a} + \sqrt{\hat{\rho}\sigma_x^2 - 2m_1 + q_1}z_\mu \equiv \sqrt{\Delta_1}v_{\mu a} + \sqrt{\widehat{M}_1}z_\mu. \tag{89}$$

Notice the relation $\hat{\rho}\sigma_x^2 = (1/N)\sum_i \hat{x}_i^2$. A similar discussion and application of the RS are possible for the covariance of $\tilde{d}$:

$$\frac{1}{N}\sum_i |r_i^1|_0 x_i^\alpha x_i^\beta = Q_2\delta_{\alpha\beta} + q_2(1-\delta_{\alpha\beta}), \tag{90a}$$

$$\frac{1}{N}\sum_i |r_i^1|_0 \hat{x}_i x_i^\alpha = m_2, \tag{90b}$$

$$\frac{1}{N}\sum_i r_i^a |r_i^1|_0 x_i^\alpha = Q_c\delta_{a1} + q_c(1-\delta_{a1}). \tag{90c}$$

Using these covariances, we have a simple representation of $\tilde{d}$ as follows:

$$\tilde{d}_{\mu\alpha} = \frac{\Delta_c}{\sqrt{\Delta_1}}v_{\mu 1} + \frac{\widehat{\mathcal{M}}_c}{\sqrt{\widehat{\mathcal{M}}_1}}z_\mu + \sqrt{\Delta_2}u_{\mu\alpha} + \sqrt{\widehat{\mathcal{M}}_2 - \frac{\widehat{\mathcal{M}}_c^2}{\widehat{\mathcal{M}}_1} - \frac{\Delta_c^2}{\Delta_1}}w_\mu. \tag{91}$$

where $u$ and $w$ denote new independent Gaussian variables from $\mathcal{N}(0,1)$, and $v$ and $z$ represent the same Gaussian variables as in eq. (89); the following abbreviations are used:

$$\widehat{\mathcal{M}}_1 = \hat{\rho}\sigma_x^2 - 2m_1 + q_1, \tag{92}$$

$$\widehat{\mathcal{M}}_2 = \hat{\rho}\sigma_x^2 - 2m_2 + q_2, \tag{93}$$

$$\widehat{\mathcal{M}}_c = \hat{\rho}\sigma_x^2 - (m_1 + m_2) + q_c, \tag{94}$$

$$\Delta_1 = Q_1 - q_1, \qquad \Delta_2 = Q_2 - q_2, \qquad \Delta_c = Q_c - q_c. \tag{95}$$

The above order parameters are nothing but those having the same symbols in eq. (57), but now, they are represented using replicas. Note that in the limit $\beta \to \infty$, all $\Delta_{1,2,c}$ vanish and $\widehat{\mathcal{M}}_{1,2,c}$ converge to $\mathcal{M}_{1,2,c}$.

On the basis of the above consideration and denoting the weight of normal distribution $\mathcal{N}(0,1)$ as $Dx = e^{-\frac{1}{2}x^2}/\sqrt{2\pi}$, we obtain the following expression:

$$\left[ e^{-\frac{1}{2}\beta\left\{\sum_{a=1}^{n}(d_{\mu a}+\xi_\mu)^2+\sum_{\alpha=1}^{m}\left(\tilde{d}_{\mu\alpha}+\xi_\mu\right)^2\right\}} \right]_{\boldsymbol{\xi},A} \equiv L^M, \tag{96}$$

where

$$L = \int D\eta Dz Dw \left(\int Dv\ e^{-\frac{1}{2}\beta h_1^2(v,z,\eta)}\right)^n \left\langle \left(\int Du\ e^{-\frac{1}{2}\beta h_2^2(v,z,u,w,\eta)}\right)^\nu \right\rangle_{v|h_1}, \tag{97}$$

where

$$h_1(v,z,\eta) = \sqrt{\Delta_1}v + \sqrt{\widehat{\mathcal{M}}_1}z + \sigma_\xi^2\eta, \tag{98}$$

$$h_2(v,z,u,w,\eta) = \frac{\Delta_c}{\sqrt{\Delta_1}}v + \frac{\widehat{\mathcal{M}}_c}{\sqrt{\widehat{\mathcal{M}}_1}}z + \sqrt{\Delta_2}u + \sqrt{\widehat{\mathcal{M}}_2 - \frac{\widehat{\mathcal{M}}_c^2}{\widehat{\mathcal{M}}_1} - \frac{\Delta_c^2}{\Delta_1}}w + \sigma_\xi^2\eta, \tag{99}$$

$$\langle \cdots \rangle_{v|h_1} = \frac{\int Dv(\cdots)e^{-\frac{1}{2}\beta h_1^2(v,z,\eta)}}{\int Dv\ e^{-\frac{1}{2}\beta h_1^2(v,z,\eta)}}. \tag{100}$$

To proceed with the calculations, we use a trick to perform a trace over $\boldsymbol{r}$ and $\boldsymbol{x}$: rewriting the order parameters as integration variables and introducing delta functions that require order parameters to take the values defined in eqs. (88,90). This yields the following:

$$\Phi_2 = \int d\Omega\ IL^M, \tag{101}$$

where $\Omega = \{Q_1, q_1, m_1, Q_2, q_2, m_2, Q_c, q_c,\}$ and

$$\begin{aligned} I = \widetilde{\operatorname{Tr}}_{\{\boldsymbol{r}_a\}_{a=1}^n} \operatorname{Tr}_{\{\boldsymbol{x}_\alpha|\boldsymbol{r}_1\}_{\alpha=1}^\nu} &\prod_{a=1}^n \delta(NQ_1 - \sum_i (r_i^a)^2) \prod_{a<b} \delta(Nq_1 - \sum_i r_i^a r_i^b) \prod_{a=1}^n \delta(Nm_1 - \sum_i \hat{x}_i r_i^a) \\ &\prod_{\alpha=1}^n \delta(NQ_2 - \sum_i |r_i^1|_0 (x_i^\alpha)^2) \prod_{\alpha<\beta} \delta(Nq_2 - \sum_i |r_i^1|_0 x_i^\alpha x_i^\beta) \prod_{a=1}^n \delta(Nm_2 - \sum_i \hat{x}_i |r_i^1|_0 x_i^\alpha) \\ &\prod_{\alpha=1}^\nu \delta(NQ_c - \sum_i r_i^1 x_i^\alpha) \prod_{a=2}^n \prod_{\alpha=1}^\nu \delta(Nq_c - \sum_i |r_i^1|_0 r_i^a x_i^\alpha). \end{aligned} \tag{102}$$

We rewrite these delta functions by using the Fourier representations. In doing so, constant factors can be applied to the Fourier integration variables, and we choose convenient factors for later calculations. For example, the delta functions of $Q_1$ and $q_1$ are expressed as follows:

$$\delta\left(NQ_1 - \sum_i (r_i^a)^2\right) = C_1 \int d\hat{Q}_1\ e^{\frac{1}{2}NQ_1\hat{Q}_1 - \frac{1}{2}\hat{Q}_1\sum_i (x_i^{a\alpha})^2} \tag{103}$$

$$\delta\left(Nq_1 - \sum_i r_i^a r_i^b\right) = c_1 \int d\hat{q}_1\ e^{-Nq_1\hat{q}_1 + \hat{q}_1\sum_i r_i^a r_i^b}, \tag{104}$$

where $C_1$ and $c_1$ denote the normalization factors but will be discarded hereafter because they do not contribute in the limit $N \to \infty$. These operations provide the following:

$$I = \int d\hat{\Omega}\; e^{N\left(S+\left[\log \widetilde{\mathrm{Tr}_{\{r_a\}_a}} \mathrm{Tr}_{\{x_\alpha|r_1\}_\alpha} e^{f(r,x|\hat{x})}\right]_{\hat{x}}\right)}, \tag{105}$$

where $\hat{\Omega} = \left\{\hat{Q}_1, \hat{q}_1, \hat{m}_1, \hat{Q}_2, \hat{q}_2, \hat{m}_2, \hat{Q}_c, \hat{q}_c,\right\}$ and

$$\begin{aligned} S = &\frac{1}{2}nQ_1\hat{Q}_1 - \frac{1}{2}n(n-1)q_1\hat{q}_1 + \frac{1}{2}\nu Q_2\hat{Q}_2 - \frac{1}{2}\nu(\nu-1)q_2\hat{q}_2 \\ &-nm_1\hat{m}_1 - \nu m_2\hat{m}_2 - \nu Q_c\hat{Q}_c - \nu(n-1)q_c\hat{q}_c, \end{aligned} \tag{106}$$

$$\begin{aligned} f(r,x|\hat{x}) = &-\frac{1}{2}\hat{Q}_1\sum_a r_a^2 + \hat{q}_1\sum_{a<b} r_a r_b + \hat{m}_1\hat{x}\sum_a r_a \\ +|r_1|_0 &\left\{-\frac{1}{2}\hat{Q}_2\sum_\alpha x_\alpha^2 + \hat{q}_1\sum_{\alpha<\beta} x_\alpha x_\beta + \hat{m}_2\hat{x}\sum_\alpha x_\alpha + \hat{\Delta}_c r_1\sum_\alpha x_\alpha + \hat{q}_c\left(\sum_{a=1}^n r_a\right)\left(\sum_\alpha x_\alpha\right)\right\}, \end{aligned} \tag{107}$$

where we set $\hat{\Delta}_c = \hat{Q}_c - \hat{q}_c$. The average over $\hat{x}$ appears as a result of the law of large numbers. As noted in the main text, we consider the Bernoulli–Gaussian distribution with respect to $\hat{x}$. Denoting its Gaussian part as $P_G(\hat{x}) = \exp(-\frac{\hat{x}^2}{2\sigma_x^2})/\sqrt{2\pi\sigma_x^2}$, we obtain the following:

$$\left[\log \widetilde{\underset{\{r_a\}_a}{\mathrm{Tr}}} \underset{\{x_\alpha|r_1\}_\alpha}{\mathrm{Tr}} e^{f(r,x|\hat{c},\hat{x})}\right]_{\hat{x}} = \hat{\rho}\int d\hat{x}\; P_G(\hat{x})\log J_A + (1-\hat{\rho})\log J_I, \tag{108}$$

where

$$J_A \equiv \widetilde{\underset{\{r_a\}_a}{\mathrm{Tr}}} \underset{\{x_\alpha|r_1\}_\alpha}{\mathrm{Tr}} e^{f(r,x|\hat{x})},\;\; J_I \equiv \widetilde{\underset{\{r_a\}_a}{\mathrm{Tr}}} \underset{\{x_\alpha|r_1\}_\alpha}{\mathrm{Tr}} e^{f(r,x|0)}. \tag{109}$$

Cross-quadratic terms in $f$ can be transformed into linear terms. First, they are transformed as follows:

$$\sum_{a<b} r_a r_b = \frac{1}{2}\left\{\left(\sum_a r_a\right)^2 - \sum_a r_a^2\right\},\;\; \sum_{\alpha<\beta}|r_1|_0 x_\alpha x_\beta = \frac{1}{2}\left\{\left(\sum_\alpha |r_1|_0 x_\alpha\right)^2 - \sum_\alpha \left(|r_1|_0 x_\alpha\right)^2\right\}. \tag{110}$$

Note that $|\cdot|_0^k = |\cdot|_0$ $(k > 0)$. Let us set $X \equiv \sum_a r_a$ and $Y \equiv \sum_\alpha |r_1|_0 x_\alpha$. The minimum number of auxiliary variables to break the quadratic terms is two, but here, we introduce three auxiliary variables to make the resultant formula interpretable. Accordingly, we have the following:

$$\begin{aligned} &\int DvDuDw\; e^{(v,u,w)\cdot(aX,bY,cX+dY)^t} = e^{\frac{1}{2}\left((a^2+c^2)X^2+(b^2+d^2)Y^2+2cdXY\right)} \\ &= e^{\frac{1}{2}\left(q_1(\sum_a r_a)^2+\hat{q}_1(\sum_\alpha |r_1|_0 x_\alpha)^2+2\hat{q}_c\left(\sum_{a=1}^n r_a\right)\left(\sum_\alpha |r_1|_0 x_\alpha\right)\right)}. \end{aligned} \tag{111}$$

A simple solution of this equation with respect to $a, b, c$, and $d$ is as follows:

$$a = \sqrt{\hat{q}_1 - \hat{q}_c},\;\; b = \sqrt{\hat{q}_2 - \hat{q}_c},\;\; c = d = \sqrt{\hat{q}_c}. \tag{112}$$

Hence, we can set $J_A = \widetilde{\mathrm{Tr}_{\{r_a\}_a}} \mathrm{Tr}_{\{x_\alpha|r_1\}_\alpha} \int DvDuDw\; e^{g_A}$ with

$$g_A = -\frac{1}{2}\left(\hat{Q}_1 + \hat{q}_1\right)\sum_a r_a^2 + A_1\sum_a r_a + |r_1|_0\left(-\frac{1}{2}\left(\hat{Q}_2 + \hat{q}_2\right)\sum_\alpha x_\alpha^2 + (A_2 + \hat{\Delta}_c r_1)\sum_\alpha x_\alpha\right), \tag{113}$$

$$A_1(\hat{m}_1) = \sqrt{\hat{q}_1 - \hat{q}_c}v + \sqrt{\hat{q}_c}w + \hat{m}_1\hat{x},\;\; A_2(\hat{m}_2) = \sqrt{\hat{q}_2 - \hat{q}_c}u + \sqrt{\hat{q}_c}w + \hat{m}_2\hat{x}. \tag{114}$$

This formula is nice because $g$ is expected to be $O(\beta)$ in the $\beta \to \infty$ limit: $\hat{q}$ and $\hat{Q}$ are $O(\beta^2)$; $(\hat{Q}+\hat{q})$ and $\hat{\Delta}_c = (\hat{Q}_c - \hat{q}_c)$ are $O(\beta)$. Now, we can easily perform the integration over $x$ as follows:

$$\mathop{\mathrm{Tr}}_{\{x_\alpha|r_1\}_\alpha} e^{g_A} = e^{-\frac{1}{2}\left(\hat{Q}_1+q_1\right)\sum_a r_a^2 + A_1(\hat{m}_1)\sum_a r_a + \frac{\nu}{2}|r_1|_0\left(\frac{\left(A_2(\hat{m}_2)+\hat{\Delta}_c r_1\right)^2}{\hat{Q}_2+\hat{q}_2} + \log\frac{2\pi}{\hat{Q}_2+\hat{q}_2}\right)}, \tag{115}$$

and thus,

$$J_A = \int DvDuDw \ \left(\mathop{\widetilde{\mathrm{Tr}}}_{r} e^{-\frac{1}{2}(\hat{Q}_1+\hat{q}_1)r^2 + A_1(\hat{m}_1)r}\right)^n \left\langle e^{\frac{\nu}{2}|r_1|_0\left(\frac{\left(A_2(\hat{m}_2)+\hat{\Delta}_c r_1\right)^2}{\hat{Q}_2+\hat{q}_2} + \log\frac{2\pi}{\hat{Q}_2+\hat{q}_2}\right)} \right\rangle_{r_1|\hat{m}_1}, \tag{116}$$

where

$$\langle \cdots \rangle_{r_1|\hat{m}_1} = \frac{\widetilde{\mathrm{Tr}}_{r_1}(\cdots) e^{-\frac{1}{2}(\hat{Q}_1+\hat{q}_1)r_1^2 + A_1(\hat{m}_1)r_1}}{\widetilde{\mathrm{Tr}}_{r_1} e^{-\frac{1}{2}(\hat{Q}_1+\hat{q}_1)r_1^2 + A_1(\hat{m}_1)r_1}}. \tag{117}$$

Setting $\hat{m}_1 = \hat{m}_2 = 0$ in $J_A$, we have the expression of $J_I$. Hence,

$$\begin{aligned}
&\phi_2(n,\nu,\beta) \equiv \frac{1}{N}\log \Phi(n,\nu,\beta) = \alpha \log L + \frac{1}{N}\log I \\
&= \alpha \log \int D\eta DzDw \left(\int Dv \ e^{-\frac{1}{2}\beta h_1^2(v,z,\eta)}\right)^n \left\langle \left(\int Du \ e^{-\frac{1}{2}\beta h_2^2(v,z,u,w,\eta)}\right)^\nu \right\rangle_{v|h_1} \\
&+\frac{1}{2}nQ_1\hat{Q}_1 - \frac{1}{2}n(n-1)q_1\hat{q}_1 + \frac{1}{2}\nu Q_2\hat{Q}_2 - \frac{1}{2}\nu(\nu-1)q_2\hat{q}_2 - nm_1\hat{m}_1 - \nu m_2\hat{m}_2 - \nu Q_c\hat{Q}_c - \nu(n-1)q_c\hat{q}_c \\
&+\hat{\rho}\int d\hat{x} \ P_G(\hat{x}) \log\left\{\int DvDuDw \ \left(\mathop{\widetilde{\mathrm{Tr}}}_{r} e^{-\frac{1}{2}(\hat{Q}_1+\hat{q}_1)r^2 + A_1(\hat{m}_1)r}\right)^n \left\langle e^{\frac{\nu}{2}|r_1|_0\left(\frac{\left(A_2(\hat{m}_2)+\hat{\Delta}_c r_1\right)^2}{\hat{Q}_2+\hat{q}_2} + \log\frac{2\pi}{\hat{Q}_2+\hat{q}_2}\right)} \right\rangle_{r_1|\hat{m}_1} \right\} \\
&+(1-\hat{\rho})\log\left\{\int DvDuDw \ \left(\mathop{\widetilde{\mathrm{Tr}}}_{r} e^{-\frac{1}{2}(\hat{Q}_1+\hat{q}_1)r^2 + A_1(0)r}\right)^n \left\langle e^{\frac{\nu}{2}|r_1|_0\left(\frac{\left(A_2(0)+\hat{\Delta}_c r_1\right)^2}{\hat{Q}_2+\hat{q}_2} + \log\frac{2\pi}{\hat{Q}_2+\hat{q}_2}\right)} \right\rangle_{r_1|0} \right\}.
\end{aligned} \tag{118}$$

Let us glance at the interdependency of the order parameters. Let us set $\tilde{\Omega}_1 = \left\{Q_1, q_1, m_1, \hat{Q}_1, \hat{q}_1, \hat{m}_1\right\}$, $\tilde{\Omega}_2 = \left\{Q_2, q_2, m_2, \hat{Q}_2, \hat{q}_2, \hat{m}_2\right\}$, and $\tilde{\Omega}_c = \left\{Q_c, q_c, \hat{Q}_c, \hat{q}_c, \right\}$. We see that

$$\phi_2(n,0,\beta) = F_1(\tilde{\Omega}_1), \ \phi_2(0,\nu,\beta) = F_2(\tilde{\Omega}_1,\tilde{\Omega}_2,\tilde{\Omega}_c), \ \phi_2(0,0,\beta) = 0. \tag{119}$$

This equation implies that $\phi_2(n,\nu)$ has multiple solutions in the saddle-point equation of the order parameters at and around $n = \nu = 0$. We should choose a solution that is analytically continued to the solution at $\nu = 0$ with respect to $\nu$. Hence, we first take the $\nu \to 0$ limit, yielding $\phi_1(n,\beta) \equiv (1/N)\log \Phi_1(n,\beta) = \phi_2(n,0,\beta)$.

## A.1 Derivation of $f_1$

The free energy $f_1$ is obtained from $\phi_1$ to $-\beta f_1 = \lim_{n\to 0}(1/n)\phi_1(n,\beta)$. Performing the variable transformation $(\sqrt{\hat{q}_1}v + \hat{m}_1\hat{x})/(\sqrt{\hat{q}_+ \hat{m}_1^2\sigma_x^2}) \to z$, we obtain the following:

$$\begin{aligned}
-\beta f_1 = \mathop{\mathrm{Extr}}_{\tilde{\Omega}_1}&\left\{\frac{1}{2}\hat{Q}_1 Q_1 + \frac{1}{2}\hat{q}_1 q_1 - \hat{m}_1 m_1 + \hat{\rho}\int Dz \log X_A + (1-\hat{\rho})\int Dz \log X_I \right. \\
&\left. -\frac{\alpha}{2}\left(\log\left(1+\beta\Delta_1\right) + \frac{\beta(\widehat{M}_1 + \sigma_\xi^2)}{1+\beta\Delta_1}\right)\right\},
\end{aligned} \tag{120}$$

where

$$X_A = \int dx\ e^{-\frac{1}{2}(\hat{Q}_1+\hat{q}_1)x^2+\sqrt{\hat{q}_1+\hat{m}_1^2\sigma_x^2}\ zx-\beta\lambda|x|}, \tag{121}$$

and the expression of $X_I$ is obtained by setting $\hat{m}_1 = 0$ in $X_A$. To take the zero-temperature limit $\beta \to \infty$, we assume the following scalings:

$$\begin{aligned}
&\beta\Delta_1 \to \chi_1, && (122a)\\
&(\hat{Q}_1 + \hat{q}_1) \to \beta\hat{\chi}_1, && (122b)\\
&\hat{q}_1 \to \beta^2\hat{q}_1, && (122c)\\
&\hat{Q}_1 \to -\beta^2\hat{q}_1, && (122d)\\
&\hat{m}_1 \to \beta\hat{m}_1. && (122e)
\end{aligned}$$

Then, the integration is dominated by the saddle point $x^*$ in the limit $\beta \to \infty$

$$X_A \to e^{\beta f_A(x^*)}, \tag{123}$$

where

$$f_A(x) = -\frac{1}{2}\hat{Q}_1 x^2 + \left\{ \begin{array}{ll} A_+x & (x \geq 0) \\ A_-x & (x < 0) \end{array} \right. , \tag{124}$$

and

$$A_\pm = \sqrt{\hat{\chi}_1 + \hat{m}_1^2\sigma_x^2}z \mp \lambda. \tag{125}$$

The saddle point $x^*$ changes the behavior depending on the value of $A_\pm$. Simple algebra yields the following:

$$\lim_{\beta\to\infty} \frac{1}{\beta} \int Dz \log X_A = \int Dz f_{\hat{p}}(x^*) = \frac{F(\theta_A)}{\hat{Q}_1}. \tag{126}$$

A similar calculation is possible for $X_I$. Summarizing the calculations, we obtain eq. (61).

## A.2 Derivation of $f_2$

A small calculation from $\phi_2$ yields the following:

$$\begin{aligned}
-\beta f_2 &= \lim_{\nu\to 0} \frac{1}{\nu}\phi_2(0,\nu,\beta)\\
&= \alpha \int D\eta Dz Dw \left\langle \log \int Du\ e^{-\frac{1}{2}\beta h_2^2(v,z,u,w,\eta)} \right\rangle_{v|h_1} + \frac{1}{2}Q_2\hat{Q}_2 + \frac{1}{2}q_2\hat{q}_2 - m_2\hat{m}_2 - Q_c\hat{Q}_c + q_c\hat{q}_c\\
&+ \frac{\hat{\rho}}{2}\int d\hat{x}\ P_G(\hat{x}) \int DvDuDw\ \left\langle |r_1|_0 \left( \frac{\left(A_2(\hat{m}_2) + \hat{\Delta}_c r_1\right)^2}{\hat{Q}_2 + \hat{q}_2} + \log\frac{2\pi}{\hat{Q}_2 + \hat{q}_2} \right) \right\rangle_{r_1|\hat{m}_1}\\
&+ \frac{1-\hat{\rho}}{2}\int DvDuDw\ \left\langle |r_1|_0 \left( \frac{\left(A_2(0) + \hat{\Delta}_c r_1\right)^2}{\hat{Q}_2 + \hat{q}_2} + \log\frac{2\pi}{\hat{Q}_2 + \hat{q}_2} \right) \right\rangle_{r_1|0}.
\end{aligned} \tag{127}$$

We assume the following scalings:

$$\beta(Q_1-q_1)\to\chi_1,\ \beta(Q_2-q_2)\to\chi_2,\ \beta(Q_c-q_c)\to\chi_c, \tag{128a}$$
$$(\hat{Q}_1+\hat{q}_1)\to\beta\hat{Q}_1,\ (\hat{Q}_2+\hat{q}_2)\to\beta\hat{Q}_2,\ (\hat{Q}_c-\hat{q}_c)\to\beta\hat{Q}_c, \tag{128b}$$
$$\hat{q}_1,-\hat{Q}_1\to\beta^2\hat{\chi}_1,\ \hat{q}_2,-\hat{Q}_2\to\beta^2\hat{\chi}_2,\ \hat{q}_c,\hat{Q}_c\to\beta^2\hat{\chi}_c, \tag{128c}$$
$$\hat{m}_1\to\beta\hat{m}_1,\ \hat{m}_2\to\beta\hat{m}_2. \tag{128d}$$

After lengthy but straightforward calculations, we obtain the following:

$$\begin{aligned}&\lim_{\beta\to\infty}\frac{1}{\beta}\int D\eta DzDw\left\langle\log\int Du\ e^{-\frac{1}{2}\beta h_2^2(v,z,u,w,\eta)}\right\rangle_{v|h_1}\\&=-\frac{1}{2}\frac{1}{1+\chi_2}\left\{\frac{\chi_c^2}{(1+\chi_1)^2}\widetilde{\mathcal{M}}_1-2\frac{\chi_c}{1+\chi_1}\widetilde{\mathcal{M}}_c+\widetilde{\mathcal{M}}_2\right\},\end{aligned} \tag{129}$$

and

$$\begin{aligned}&\lim_{\beta\to\infty}\frac{1}{\beta}\int d\hat{x}\ P_G(\hat{x})\int DvDuDw\ \left\langle|r_1|_0\left(\frac{\left(A_2(\hat{m}_2)+\hat{\Delta}_c r_1\right)^2}{\hat{Q}_2+\hat{q}_2}+\log\frac{2\pi}{\hat{Q}_2+\hat{q}_2}\right)\right\rangle_{r_1|\hat{m}_1}\\&=\frac{1}{\hat{Q}_2}\int d\hat{x}\ P_G(\hat{x})\int DvDuDw\ |r_1^*|_0\left(\sqrt{\hat{\chi}_2-\hat{\chi}_c}u+\sqrt{\hat{\chi}_c}w+\hat{m}_2\hat{x}+\hat{Q}_c r_1^*\right)^2.\end{aligned} \tag{130}$$

The saddle point $r_1^*$ depends on $v, w$, and $\hat{x}$, and we need to be careful while evaluating it. Let us set and expand

$$\begin{aligned}T&=\int d\hat{x}\ P_G(\hat{x})\int DvDuDw\ |r_1^*|_0\left(\sqrt{\hat{\chi}_2-\hat{\chi}_c}u+\sqrt{\hat{\chi}_c}w+\hat{m}_2\hat{x}+\hat{Q}_c r_1^*\right)^2\\&=(\hat{\chi}_2-\hat{\chi}_c)\int d\hat{x}\ P_G(\hat{x})\int DvDw\ |r_1^*(v,w,\hat{x})|_0+\int d\hat{x}\ P_G(\hat{x})\int DvDw\ R,\end{aligned} \tag{131}$$

where the integration of $u$ is easily performed since it is independent of $r_1^*$, and

$$R=|r_1^*(v,w,\hat{x})|_0\left(\sqrt{\hat{\chi}_c}z+\hat{m}_2\hat{x}+\hat{Q}_c r_1^*\right)^2. \tag{132}$$

The first integral is evaluated as follows:

$$\int d\hat{x}\ P_G(\hat{x})\int DvDw\ |r_1(v,w,\hat{x})|_0=2E_0(\theta_A). \tag{133}$$

This can be shown by changing the integration variable as $\sqrt{\hat{\chi}_1-\hat{\chi}_c}v+\sqrt{\hat{\chi}_c}w+\hat{m}_1\hat{x}\to\sqrt{\hat{\chi}_1+\hat{m}_1^2\sigma_x^2}z$, which appears in $\langle\cdots\rangle_{r_1|\hat{m}_1}$. Each term of $R$ is calculated in a similar manner by performing Gaussian integrations many times. Here, we summarize the result:

$$X_1=\int d\hat{x}\ P_G(\hat{x})\int DvDw\ |r_1^*|_0\ w^2=2E_0(\theta_A)+2\frac{\hat{\chi}_c}{\hat{\chi}_1+\hat{m}_1^2\sigma_x^2}\frac{\theta_A}{\sqrt{2\pi}}e^{-\frac{1}{2}\theta_A^2}, \tag{134}$$

$$X_2=\int d\hat{x}\ P_G(\hat{x})\int DvDw\ |r_1^*|_0\ w\hat{x}.=2\frac{\hat{m}_1\sigma_x^2\sqrt{\chi_c}}{\hat{\chi}_1+\hat{m}_1^2\sigma_x^2}\frac{\theta_A}{\sqrt{2\pi}}e^{-\frac{1}{2}\theta_A^2}, \tag{135}$$

$$X_3=\int d\hat{x}\ P_G(\hat{x})\int DvDw\ |r_1^*|_0\ \hat{x}^2=2\sigma_x^2E_0(\theta_A)+2\frac{\hat{m}_1^2\sigma_x^4}{\hat{\chi}_1+\hat{m}_1^2\sigma_x^2}\frac{\theta_A}{\sqrt{2\pi}}e^{-\frac{1}{2}\theta_A^2}, \tag{136}$$

$$X_4=\int d\hat{x}\ P_G(\hat{x})\int DvDw\ wr_1^*=2\frac{\sqrt{\chi_c}}{\hat{Q}_1}E_0(\theta_A), \tag{137}$$

$$X_5=\int d\hat{x}\ P_G(\hat{x})\int DvDw\ \hat{x}r_1^*=2\frac{\hat{m}_1\sigma_x^2}{\hat{Q}_1}E_0(\theta_A), \tag{138}$$

$$X_6=\int d\hat{x}\ P_G(\hat{x})\int DvDw\ (r_1^*)^2=2\frac{1}{\hat{Q}_1^2}F(\theta_A), \tag{139}$$

Collecting all the terms and rewriting them with eq. (65), we finally obtain eq. (70).

## References

[1] http://sparse-modeling.jp/index_e.html

[2] I. Rish and G. Grabarnik, *Sparse Modeling: Theory, Algorithms, and Applications*, (CRC Press, 2014)

[3] J. Mairal, F. Bach, and J. Ponce, Sparse Modeling for Image and Vision Processing, arXiv:1411.3230v2

[4] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, (CRC Press, 2015)

[5] R. Tibshirani, Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, **58**, 267–288 (1996)

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, LEAST ANGLE REGRESSION, *The Annals of Statistics*, **32**, 407–499, (2004)

[7] J. Wright, A. Y. Yang, A. Ganesh, S. . Sastry, and Y. Ma, Robust Face Recognition via Sparse Representation, *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, **31**, (2009)

[8] J. Elith, C. H. Graham, R. P. Anderson, et al. Novel methods improve prediction of species' distributions from occurrence data, *ECOGRAPHY* **29**, 129–151, (2006)

[9] J. Schafer and K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *STATISTICAL APPLICATIONS IN GENETICS AND MOLECULAR BIOLOGY*, **4**, (2005)

[10] T. Kato and M. Uemura, Period Analysis using the Least Absolute Shrinkage and Selection Operator (Lasso), *Publ. Astron. Soc. Japan*, **64**, (2012)

[11] M. Uemura, K. S. Kawabata, S. Ikeda, K. Maeda: Variable selection for modeling the absolute magnitude at maximum of Type Ia supernovae, *Publ. Astron. Soc. Japan*, **67**, 55, 1–9, (2015).

[12] D. L. Donoho, *IEEE Transactions on Information Theory*, **52**(4), 1289–1306, (2006).

[13] E. J. Candès and T. Tao, *IEEE Transactions on Information Theory*, **51**(12), 4203–4215, (2005).

[14] E. J. Candès, J. Romberg and T. Tao, *IEEE Transactions on Information Theory*, **52**(2), 489–509, (2006).

[15] E. J. Candès and T. Tao, *IEEE Transactions on Information Theory*, **52**(12), 5406–5425, (2006).

[16] D. L. Donoho and J. Tanner: *Phil. Trans. R. Soc. A*, **367**, 4273–4293, (2009)

[17] D. L. Donoho, A. Malekib, and A. Montanari: Message-passing algorithms for compressed sensing, *Proc. Natl. Acad. Sci.*, **106**, 18914–18919, (2009)

[18] Y. Kabashima, T. Wadayama, and T. Tanaka: A typical reconstruction limit for compressed sensing based on $L_p$-norm minimization, *J. Stat. Mech.*, L09003, (2009)

[19] S. Ganguli and H. Sompolinsky: Statistical Mechanics of Compressed Sensing, *Phys. Rev. Lett.*, **104**, 188701, (2010)

[20] S. Rangan, Generalized Approximate Message Passing for Estimation with Random Linear Mixing, *arXiv:1010.5141*, (2010)

[21] F. Krzakala, M. Mézard, F. Sausset, Y. Sun and L. Zdeborová, Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices, *J. Stat. Mech.*, P08009, (2012)

[22] A. Sakata and Y. Kabashima, *Europhysics Letters*, **103**, 28008–p1–28008–p6, (2013)

[23] Y. Nakanishi, T. Obuchi, M. Okada, and Y. Kabashima, arXiv:1510.02189

[24] J. S. Yedidia, W. T. Freeman, and Y Weiss, *Understanding belief propagation and its generalizations* (in Exploring artificial intelligence in the new millennium, pp. 239–269, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2003)

[25] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice*, (Neural Information Processing series, A Bradford Book, 2001)

[26] H. S. Seung, H. Sompolinsky, and N. Tishby: Statistical mechanics of learning from examples, *Phys. Rev. A*, **45**, 6056-6091, (1992)

[27] M. Opper and O. Winther, A mean field algorithm for Bayes learning in large feed-forward neural networks, *Advances in Neural Information Processing Systems 9*, NIPS, Denver, (1996)

[28] Y. Kabashima, A CDMA multiuser detection algorithm on the basis of belief propagation, *J. Phys. A* **36**, 11111–11121, (2003)

[29] http://hercules.berkeley.edu/database/index_public.html

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, (Springer Series in Statistics, 2009)

[31] L. Breiman, J. Friedman, C. J. Stone, and R.A. Olshen, *Classification and Regression Trees*, (Chapman and Hall/CRC, 1984)

[32] R. Tibshirani, G. Walther and T. Hastie, Estimating the Number of Clusters in a Data Set via the Gap Statistic, *Journal of the Royal Statistical Society*, **63**, 411–423, (2001)