Signal recovery using expectation consistent approximation for linear observations

Yoshiyuki Kabashima
Dept. of Comput. Intell. & Syst. Sci.
Tokyo Institute of Technology
Yokohama 226–8502, Japan
Email: kaba@dis.titech.ac.jp

Mikko Vehkaperä
Department of Signal Processing and Acoustics
Aalto University
Otakaari 5 A FI-02150 Espoo, Finland
Email: mikko.vehkapera@aalto.fi

***Abstract*—A signal recovery scheme is developed for linear observation systems based on expectation consistent (EC) mean field approximation. Approximate message passing (AMP) is known to be consistent with the results obtained using the replica theory, which is supposed to be exact in the large system limit, when each entry of the observation matrix is independently generated from an identical distribution. However, this is not necessarily the case for general matrices. We show that EC recovery exhibits consistency with the replica theory for a wider class of random observation matrices. This is numerically confirmed by experiments for the Bayesian optimal signal recovery of compressed sensing using random row-orthogonal matrices.**

I. Introduction

Let us suppose that an original N-dimensional vector $\mathbf{x} = (x_i) \in \mathbb{R}^N$ is transformed into an M-dimensional vector $\mathbf{y} = (y_\mu) \in \mathbb{R}^M$ by a matrix $A = (A_{\mu i}) \in \mathbb{R}^{M\times N}$ as

$$\mathbf{y} = A\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{n} = (n_\mu) \in \mathbb{R}^M$ is provided randomly. Many problems of signal processing are formulated using this form. Equation (1) describes the basic signal sampling scheme if we identify $\mathbf{y}$, $\mathbf{x}$, and A as sampled signal values, Fourier coefficients, and a Fourier matrix, respectively. In code division multiple access (CDMA) systems, $\mathbf{x}$ and A correspond to transmitted signals of N users and a set of signature sequences, whereas $\mathbf{y}$ is the signal observed at a base station. In multi-input and multi-output communication (MIMO) systems, $\mathbf{x}$, $\mathbf{y}$, and A represent the transmitted and received signals by N and M antennas, and the signal transmission efficiency between the input and output antennas. In compressive sensing (CS), $\mathbf{x}$, $\mathbf{y}$, and A represent a sparse signal, its measurement, and a measurement matrix.

For simplicity, we hereafter assume that A is known exactly. Then, a major problem is to design a computationally efficient scheme for recovering $\mathbf{x}$ from $\mathbf{y}$ accurately. A standard approach for this is to follow the least-square principle; minimizing $||\mathbf{y} - A\mathbf{x}||^2$ in conjunction with appropriate l_2-regularization terms with respect to $\mathbf{x}$ yields a signal recovery scheme that performs with a low computational cost using operations of linear algebra. Unfortunately, the optimality of inference accuracy is not guaranteed for the resulting scheme unless $\mathbf{x}$ follows a distribution of a specific class. In recent years, significant attention has been paid to the usage of the l_1-norm regularization when $\mathbf{x}$ is supposed to be a sparse signal. The l_1-based recovery is capable of recovering sparse signals with a computational cost of the polynomial order of N. However, this still does not achieve the optimal accuracy in general [1], whereas perfect recovery is possible for the noiseless case if the observation ratio $\alpha = M/N$ is sufficiently large [2], [3], [4].

When the prior distribution of $\mathbf{x}$ and the distribution of the observation noise $\mathbf{n}$ are known, the Bayesian framework offers an optimal recovery scheme in minimum mean square error (MMSE) sense although its exact execution is computationally difficult in most cases. The purpose of this paper is to develop a computationally feasible approximate scheme for the Bayesian signal recovery for a class of random observation matrix A. For this, we employ an advanced mean field method known as *expectation consistent* (EC) approximation developed in statistical mechanics [5], [6] and machine learning [7]. The developed scheme exhibits consistency with the replica theory, which is supposed to provide exact predictions in the large system limit.

II. Related work

Reference [8] used the replica method to find a decoupled formulation for the input-output statistics of a CS system whose measurement matrix is composed of independently and identically distributed (i.i.d.) entries. As a corollary, this leads to a computationally feasible characterization of the MMSE as well. The MMSE of a similar i.i.d. setup was later evaluated directly in [9] by using mathematically rigorous methods. Numerical results therein verified the accuracy of the earlier replica analysis. Finally, non-i.i.d. sensing matrices where considered in [10], where the replica method was used to find the support recovery performance of a class of CS systems.

To the best of our knowledge, computationally feasible algorithms approximately performing the Bayesian recovery were initially developed for a simple perceptron (linear classifier) [11] and later for CDMA [12], [13]. Recently, a similar idea was applied for CS [4], [14], [15] as *approximate message passing* (AMP), and was summarized as a general formulation termed *generalized approximate message passing* (GAMP) [16]. However, these studies rely on the assumption that each entry of A, A_{ij}, is i.i.d., and the appropriateness for the employment to other ensembles is not guaranteed. In fact, the necessity for considering a certain characteristic feature of A in constructing the approximation was pointed out in [5], and its significance was tested for the simple perceptron [17], CDMA [18], and MIMO [19]. Here, we show how this approach is employed for the signal recovery of linear observations and examine its significance for an example of CS.

III. Problem setup

A. Model specification

In the following, we suppose that each entry of $\mathbf{x}$, x_i ($i = 1, 2, \ldots, N$), is generated from a distribution $P(x)$ independently of one another. For simplicity, we focus on the case where the observation noise $\mathbf{n}$ obeys a memoryless zero mean Gaussian distribution, so that the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is provided as

$$P(\mathbf{y}|\mathbf{x}, A) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{||\mathbf{y} - A\mathbf{x}||^2}{2\sigma^2}\right). \quad (2)$$

General treatment that includes the case of non-Gaussian noise can be found in [20]. Further, we assume that for eigenvalue decomposition $A^{\mathrm{T}}A = ODO^{\mathrm{T}}$, where O is the right eigenbasis of A and $D = (d_i\delta_{ij})$ is the diagonal matrix composed of eigenvalues d_i of $A^{\mathrm{T}}A$, O can be regarded as a random sample from the uniform distribution of $N \times N$ orthogonal matrices and $\rho_{A^{\mathrm{T}}A}(\lambda) = N^{-1}\sum_{i=1}^{N}\delta(\lambda - d_i)$ asymptotically converges to a certain distribution $\rho(\lambda)$ with a probability of unity as $N \to \infty$. This assumption holds when A_{ij}'s are generated independently of one another from a zero mean Gaussian distribution. Further, this is also the case when A is constructed by randomly selecting M rows from a randomly generated $N \times N$ orthogonal matrix.

B. Bayesian recovery and expected performance

Let $\hat{\mathbf{x}}(\mathbf{y})$ be an arbitrary recovery scheme given $\mathbf{y}$. Under the above assumption, the mean square error $\mathsf{mse} = N^{-1}\int d\mathbf{x}d\mathbf{y}P(\mathbf{x},\mathbf{y}|A)||\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})||^2$ is minimized by the Bayesian recovery

$$\hat{\mathbf{x}}^{\mathrm{Bayes}}(\mathbf{y}) \equiv \int d\mathbf{x}\mathbf{x}P(\mathbf{x}|\mathbf{y}, A), \quad (3)$$

which achieves the minimum value of mse (MMSE) as

$$\mathsf{mmse} = N^{-1}\left(\left\langle ||\mathbf{x}||^2 \right\rangle - \int d\mathbf{y}P(\mathbf{y}|A)\left|\left|\langle \mathbf{x}\rangle_{|\mathbf{y}}\right|\right|^2\right), \quad (4)$$

where $P(\mathbf{x},\mathbf{y}|A) = P(\mathbf{y}|\mathbf{x}, A)\prod_{i=1}^{N}P(x_i)$, $P(\mathbf{y}|A) = \int d\mathbf{x}P(\mathbf{x},\mathbf{y}|A)$. $\langle\cdots\rangle$ and $\langle\cdots\rangle_{|\mathbf{y}}$ denote averages with respect to the prior and posterior distributions $P(\mathbf{x}) = \prod_{i=1}^{N}P(x_i)$ and $P(\mathbf{x}|\mathbf{y}, A) = P(\mathbf{x},\mathbf{y}|A)/P(\mathbf{y}|A)$, respectively.

Although optimality of (3) is guaranteed, evaluating the MMSE is generally difficult. The replica method from statistical mechanics enables the evaluation for the large system limit $N, M \to \infty$ keeping $\alpha = N/M \sim O(1)$ although its mathematical validity is still open. For generality, let us suppose that the true prior and the variance of Gaussian noise, $P_0(x)$ and σ_0^2, may be different from $P(x)$ and σ^2, respectively. The replica symmetric (RS) computation [21], [22] evaluates the performance of the Bayesian recovery as follows.

Theorem 1: The RS evaluation offers the typical value of mse as

$$\mathsf{mse} = q - 2m + Q_0. \quad (5)$$

Here, $Q_0 = \int dx x^2 P_0(x)$, and q and m are determined by extremizing the variational free energy density

$$\begin{aligned}\phi(q, m, Q, \hat{q}, \hat{m}, \hat{Q}) = &-\frac{\hat{Q}Q}{2} - \frac{\hat{q}q}{2} + \hat{m}m - G\left(-\frac{Q-q}{\sigma^2}\right) \\ &+ \left(\frac{q - 2m + Q_0}{\sigma^2} - \frac{\sigma_0^2(Q-q)}{\sigma^4}\right)G'\left(-\frac{Q-q}{\sigma^2}\right) \\ &- \int dx^0 P_0(x^0)\mathrm{D}z \ln\left[\int dx P(x)\mathrm{e}^{-\frac{\hat{Q}+\hat{q}}{2}x^2 + (\sqrt{\hat{q}}z + \hat{m}x^0)x}\right], \end{aligned} \quad (6)$$

where $\mathrm{D}z = \frac{dz}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{2}}$ stands for the Gaussian measure. $G(x) = \mathrm{extr}_{\Lambda}\left\{-\frac{1}{2}\int d\lambda\rho(\lambda)\ln|\Lambda - \lambda| + \frac{1}{2}\Lambda x\right\} - \frac{1}{2}\ln|x| - \frac{1}{2}$, where $\mathrm{extr}_X\{\cdots\}$ denotes the extremization with respect to X, means the asymptotic form of the single rank Harish-Chandra-Itykson-Zuber integral of $A^{\mathrm{T}}A$ [23], which is linked to the R-transform [24] as $R_{A^{\mathrm{T}}A}(x) = G'(x)$.

Proof: Use techniques employed in [18], [25][1]. ■

When the correct prior and variance, $P(x) = P_0(x)$ and $\sigma^2 = \sigma_0^2$, are used, the replica symmetry ensures that the dominant solution extremizing (6) satisfies $Q = Q_0$, $q = m$, $\hat{Q} = 0$, and $\hat{q} = \hat{m}$, which yields $\mathsf{mmse} = Q_0 - q$. It is strongly conjectured that solutions of this type are always thermodynamically dominant offering exact (but not rigorous) predictions in the large system limit [26], [15]. Therefore, our goal is to develop a computationally feasible scheme that approximately evaluates (3) and becomes consistent with the results predicted by (6) as the system size tends to infinity.

IV. Expectation consistent signal recovery

A. Gibbs free energy formalism

The following theorem constitutes the basis of our approximation.

Theorem 2: Let us define *Gibbs free energy* as

$$\Phi(\mathbf{m}) = \mathop{\mathrm{extr}}_{\mathbf{h}}\left\{\mathbf{h}\cdot\mathbf{m} - \ln\left[\int d\mathbf{x}P(\mathbf{x}|\mathbf{y}, A)\mathrm{e}^{\mathbf{h}\cdot\mathbf{x}}\right]\right\}. \quad (7)$$

The global minimizer of $\Phi(\mathbf{m})$ is $\mathbf{m} = \langle\mathbf{x}\rangle_{|\mathbf{y}}$.

Proof: The extremization of (7) offers

$$m_i = \frac{\int d\mathbf{x}x_i P(\mathbf{x}|\mathbf{y}, A)\mathrm{e}^{\mathbf{h}\cdot\mathbf{x}}}{\int d\mathbf{x}P(\mathbf{x}|\mathbf{y}, A)\mathrm{e}^{\mathbf{h}\cdot\mathbf{x}}}. \quad (8)$$

This means that for a given value of $\mathbf{m}$, $\mathbf{h}$ is determined so that the average of $\mathbf{x}$ for a modified distribution $P(\mathbf{x}|\mathbf{y}, A, \mathbf{h}) = P(\mathbf{x}|\mathbf{y}, A)\mathrm{e}^{\mathbf{h}\cdot\mathbf{x}}/\int d\mathbf{x}P(\mathbf{x}|\mathbf{y}, A)\mathrm{e}^{\mathbf{h}\cdot\mathbf{x}}$ coincides with $\mathbf{m}$. In particular, $\mathbf{h} = \mathbf{0}$ offers $\mathbf{m} = \langle\mathbf{x}\rangle_{|\mathbf{y}}$ and corresponds to an extremum point of $\Phi(\mathbf{m})$ since $\partial\Phi(\mathbf{m})/\partial m_i = h_i = 0$ holds. Furthermore, $\mathbf{m} = \langle\mathbf{x}\rangle_{|\mathbf{y}}$ is characterized as the globally minimum point, which is shown as follows. For any value of $\mathbf{m}$, the Hessian of $\Phi(\mathbf{m})$ is evaluated as $\left(\frac{\partial^2\Phi(\mathbf{m})}{\partial m_i \partial m_j}\right) = \left(\frac{\partial h_j}{\partial m_i}\right) = \left(\frac{\partial m_j}{\partial h_i}\right)^{-1}$. However, (8) indicates that $\frac{\partial m_j}{\partial h_i}$ coincides with the covariance of x_i and x_j evaluated by $P(\mathbf{x}|\mathbf{y}, A, \mathbf{h})$. Therefore, both matrices $\left(\frac{\partial m_j}{\partial h_i}\right)$ and $\left(\frac{\partial^2\Phi(\mathbf{m})}{\partial m_i \partial m_j}\right) = \left(\frac{\partial m_j}{\partial h_i}\right)^{-1}$ are positive

[1] In [25], the free energy is expressed using the Stieltjes transform. The two expressions are, however, mathematically equivalent, and always transformable to each other

definite. This means that $\Phi(\mathbf{m})$ is a convex downward function and has a unique minimum point. ■

B. Expectation consistent approximation

Theorem 2 indicates that Bayesian recovery can be performed using the techniques of convex optimization if $\Phi(\mathbf{m})$ is correctly evaluated. Unfortunately, this is also practically unfeasible in most cases as the assessment of $\Phi(\mathbf{m})$ is computationally difficult in general. One could exceptionally evaluate $\phi(\mathbf{m})$ with a low computational cost if $P(\mathbf{x}|\mathbf{y}, A)$ were a factorized distribution as $P(\mathbf{x}) = \prod_{i=1}^N P(x_i)$. Reference [27] developed an approximation scheme based on Taylor's expansion around the factorized distribution by introducing an expansion parameter β in the interaction terms that result in computational difficulty. In the current case, this implies that the evaluation of $\Phi(\mathbf{m}) = \tilde{\Phi}(\mathbf{m}; \beta = 1)$ is performed such that $\Phi(\mathbf{m}) = \tilde{\Phi}(\mathbf{m}; 0) + \frac{\partial}{\partial\beta}\tilde{\Phi}(\mathbf{m}; 0) + \frac{\partial^2}{2!\partial\beta^2}\tilde{\Phi}(\mathbf{m}; 0) + \ldots$ by introducing generalized Gibbs free energy

$$\tilde{\Phi}(\mathbf{m}; \beta) = const + \underset{\mathbf{h}}{\mathrm{extr}}\,\Big\{\mathbf{h}\cdot\mathbf{m} - \ln\left[\int d\mathbf{x}\mathrm{e}^{-\frac{\beta}{2\sigma^2}||\mathbf{y}-A\mathbf{x}||^2}\prod_{i=1}^N\left(P(x_i)\mathrm{e}^{h_i x_i}\right)\right]\Big\}. \quad (9)$$

This treatment leads to asymptotically exact results for some systems as statistical properties of the interaction matrix allow us to truncate the expansion up to the second order [27] or enable us to sum up all relevant terms in Taylor series analytically [28]. In fact, when A_{ij}'s are independently generated from the zero mean variance M^{-1} Gaussian distribution (i.i.d. Gaussian ensemble), the expansion yields an expression

$$\Phi(\mathbf{m}) \simeq \underset{Q,E,\mathbf{h}}{\mathrm{extr}}\left\{\frac{1}{2\sigma^2}||\mathbf{y}-A\mathbf{m}||^2 + \frac{M}{2}\ln\left(1+\frac{Q-q}{\alpha\sigma^2}\right) - \frac{NEQ}{2} + \mathbf{h}\cdot\mathbf{m} - \sum_{i=1}^N \ln\left[\int dx_i P(x_i)\mathrm{e}^{-\frac{E}{2}x_i^2 + h_i x_i}\right]\right\} + const, \quad (10)$$

for large N and M owing to the latter property, where $q = N^{-1}||\mathbf{m}||^2$. Notation of "$\simeq$" means that the equation holds approximately. Under appropriate conditions, its minimum is guaranteed to converge to the fixed point of AMP for large systems [15], and the treatment becomes asymptotically exact.

Unfortunately, summing up all the relevant terms in Taylor series for generic matrices is technically difficult. For avoiding this difficulty, we employ an alternative approach based on an identity $\tilde{\Phi}(\mathbf{m}; 1) - \tilde{\Phi}(\mathbf{m}; 0) = \int_0^1 d\beta \frac{\partial}{\partial\beta}\tilde{\Phi}(\mathbf{m}; \beta) = \frac{1}{2\sigma^2}\int_0^1 d\beta \left\langle ||\mathbf{y}-A\mathbf{x}||^2\right\rangle_\beta$, following [5], [6]. Here, $\langle\cdots\rangle_\beta$ denotes the average with respect to the modified distribution $P_\beta(\mathbf{x}|\mathbf{y}, A) \propto \mathrm{e}^{-\frac{\beta}{2\sigma^2}||\mathbf{y}-A\mathbf{x}||^2}\prod_{i=1}^N\left(P(x_i)\mathrm{e}^{h_i x_i}\right)$ in which $\mathbf{h}$ is determined so that $\langle\mathbf{x}\rangle_\beta = \mathbf{m}$ holds for each β. As a decomposition $\left\langle ||\mathbf{y}-A\mathbf{x}||^2\right\rangle_\beta = ||\mathbf{y}-A\mathbf{m}||^2 + \mathrm{Tr}\left(A^{\mathrm{T}}AC_\beta\right)$ is allowed, where $C_\beta = (\langle x_i x_j\rangle_\beta - m_i m_j)$, evaluating the second moment $\langle x_i x_j\rangle_\beta$ is necessary to perform the integral of the last expression. Here, we approximately perform this by replacing $P_\beta(\mathbf{x}|\mathbf{y}, A)$ with a Gaussian distribution $P_\beta^{\mathrm{G}}(\mathbf{x}|\mathbf{y}, A) \propto \mathrm{e}^{-\frac{\beta}{2\sigma^2}||\mathbf{y}-A\mathbf{x}||^2 - \frac{\Lambda}{2}||\mathbf{x}||^2 + \mathbf{h}^G\cdot\mathbf{x}}$, where $\mathbf{h}^{\mathrm{G}}$ and Λ are determined so that the first moment $\mathbf{m} = \langle\mathbf{x}\rangle_\beta$ and the macroscopic second moment $Q = N^{-1}\left\langle ||\mathbf{x}||^2\right\rangle_\beta$ are consistent between $P_\beta(\mathbf{x}|\mathbf{y}, A)$ and $P_\beta^{\mathrm{G}}(\mathbf{x}|\mathbf{y}, A)$. Such an approximation scheme is often termed the *expectation consistent (EC)* approximation. This yields the following theorem.

Theorem 3: EC approximation offers

$$\Phi(\mathbf{m}) \simeq \underset{Q,E,\mathbf{h}}{\mathrm{extr}}\left\{\frac{1}{2\sigma^2}||\mathbf{y}-A\mathbf{m}||^2 - NG\left(-\frac{Q-q}{\sigma^2}\right) - \frac{NEQ}{2} + \mathbf{h}\cdot\mathbf{m} - \sum_{i=1}^N \ln\left[\int dx_i P(x_i)\mathrm{e}^{-\frac{E}{2}x_i^2 + h_i x_i}\right]\right\} + const, \quad (11)$$

for large systems.

Proof: For considering the consistency of $\mathbf{m}$ and Q, we define the generalized Gibbs free energy as $\tilde{\Phi}(\mathbf{m}, Q; \beta) = \mathrm{extr}_{\mathbf{h},E}\Big\{-\frac{NEQ}{2} + \mathbf{h}\cdot\mathbf{m} - \ln\Big[\int d\mathbf{x}P(\mathbf{x})\mathrm{e}^{-\frac{\beta}{2\sigma^2}||\mathbf{y}-A\mathbf{x}||^2 - \frac{E}{2}||\mathbf{x}||^2 + \mathbf{h}\cdot\mathbf{x}}\Big]\Big\}$, and denote its Gaussian approximation as $\tilde{\Phi}^{\mathrm{G}}(\mathbf{m}, Q; \beta)$. EC approximation offers an expression $\tilde{\Phi}(\mathbf{m}, Q; 1) \simeq \tilde{\Phi}(\mathbf{m}, Q; 0) + \tilde{\Phi}^{\mathrm{G}}(\mathbf{m}, Q; 1) - \tilde{\Phi}^{\mathrm{G}}(\mathbf{m}, Q; 0)$. Each part on the right-hand side is evaluated as follows: $\tilde{\Phi}(\mathbf{m}, Q; 0) = \mathrm{extr}_{\mathbf{h},E}\Big\{-\frac{NEQ}{2} + \mathbf{h}\cdot\mathbf{m} - \sum_{i=1}^N \ln\Big[\int dx_i P(x_i)\mathrm{e}^{-\frac{E}{2}x_i^2 + h_i x_i}\Big]\Big\}$. $\tilde{\Phi}^{\mathrm{G}}(\mathbf{m}, Q; 1) = \frac{1}{2\sigma^2}||\mathbf{y}-A\mathbf{m}||^2 + \mathrm{extr}_\Lambda\left\{\frac{1}{2}\ln\left|\det\left(\Lambda - A^{\mathrm{T}}A\right)\right| + \frac{N\Lambda(Q-q)}{2\sigma^2}\right\} + const$. $\tilde{\Phi}^{\mathrm{G}}(\mathbf{m}, Q; 0) = -\frac{N}{2}\ln\left(\frac{Q-q}{\sigma^2}\right) - \frac{N}{2} + const$. For $N, M \gg 1$, one can replace $\ln\left|\det\left(\Lambda - A^{\mathrm{T}}A\right)\right|$ with $N\int d\lambda\rho(\lambda)\ln|\Lambda-\lambda|$. Substituting the three expressions in conjunction with this replacement into the identity $\Phi(\mathbf{m}) = \mathrm{extr}_Q\left\{\tilde{\Phi}(\mathbf{m}, Q; 1)\right\} \simeq \mathrm{extr}_Q\Big\{\tilde{\Phi}(\mathbf{m}, Q; 0) + \tilde{\Phi}^{\mathrm{G}}(\mathbf{m}, Q; 1) - \tilde{\Phi}^{\mathrm{G}}(\mathbf{m}, Q; 0)\Big\}$ yields (11). ■

Here, two points are worth noting. First, for the current characterization of A based on the eigenvalue decomposition $A^{\mathrm{T}}A = ODO^{\mathrm{T}}$, all statistical features of A are summarized in $G(x)$, which is defined for the asymptotic eigenvalue distribution $\rho(\lambda)$, in (11). This means that the functional form to be optimized for computing the Bayesian recovery varies depending on the employed matrix ensemble. For instance, $G(x) = -\frac{\alpha}{2}\ln\left(1-\alpha^{-1}x\right)$ should be used for the i.i.d. Gaussian ensemble, which reduces (11) to (10). However, when A is constructed by randomly selecting M rows from a randomly generated $N\times N$ orthogonal matrix (row-orthogonal ensemble), the proper function to be employed is given by $G(x) = \mathrm{extr}_\Lambda\left\{-\frac{1-\alpha}{2}\ln\Lambda - \frac{\alpha}{2}\ln\left|\Lambda - \alpha^{-1}\right| + \frac{1}{2}\Lambda x\right\} - \frac{1}{2}\ln|x| - \frac{1}{2}$. This implies that the employment of AMP (in general, GAMP), the fixed point of which asymptotically extremizes (10), for generic matrix ensembles may not be a theoretically appropriate treatment even if it leads to a satisfiable approximation accuracy [29]. Second, although we imposed the consistency of the second moment in a macroscopic manner, one can construct a more accurate approximation by achieving the consistency in a component wise manner as for $Q_i = \left\langle x_i^2\right\rangle_\beta$. Such an approximation was once tested for CDMA demodulation [30]; however, it incurs $O(N^3)$ computational costs and is difficult to use for large systems.

C. Consistency with the replica theory

Following the argument of [5], one can show that EC approximation becomes asymptotically consistent with the replica theory for matrix ensembles of the current characterization. For this, we denote the function to be extremized in (11) as $\Phi(\mathbf{m}, Q, \mathbf{h}, E)$, and introduce the auxiliary partition function $Y(Q, E; \beta) = \int d\mathbf{h} d\mathbf{m} e^{-\beta\Phi(\mathbf{m},Q,\mathbf{h},E)}$. In the limit $\beta \to \infty$, $Y(Q, E; \beta)$ is dominated by the values of $\mathbf{m}$ and $\mathbf{h}$ for which $\Phi(\mathbf{m}, Q, \mathbf{h}, E)$ is stationary, provided the paths of integration are chosen such that the integral exists. Further, assuming the stationarity with respect to Q and E, we have an expression of free energy density as $f = N^{-1} \min_{\mathbf{m}} \{\Phi(\mathbf{m})\} = N^{-1} \mathrm{extr}_{Q,E} \left\{-\lim_{\beta\to\infty} \beta^{-1} \ln Y(Q, E; \beta)\right\}$. Variation with respect to Q offers $E = \frac{2}{\sigma^2} G' \left(-(Q - q)/\sigma^2\right)$.

For assessing the average of f with respect to A, $\mathbf{x}^0$, and $\mathbf{n}$, we employ the replica method using the average under the replica symmetric ansatz, which offers

$$\frac{1}{N} \ln \left[\exp\left[-\frac{\beta}{2\sigma^2} \sum_{a=1}^{n} \left|\left|A(\mathbf{x}^0 - \mathbf{m}^a) + \mathbf{n}\right|\right|^2 \right] \right]_{A,\mathbf{n}}$$
$$= n\left(-\left(\frac{\beta(\overline{q} - 2m + Q_0)}{\sigma^2} - \frac{\sigma_0^2 \beta^2 (q - \overline{q})}{\sigma^4} \right) G'\left(-\frac{\beta(q - \overline{q})}{\sigma^2} \right) + G\left(-\frac{\beta(q - \overline{q})}{\sigma^2} \right) \right) + O(n^2), \quad (12)$$

where we set $q = N^{-1} ||\mathbf{m}^a||^2$, $\overline{q} = N^{-1} \mathbf{m}^a \cdot \mathbf{m}^b$ $(a \neq b)$, and $m = N^{-1} \mathbf{x}^0 \cdot \mathbf{m}^a$. It is worth noting that $\overline{q} \to q$ holds for $\beta \to \infty$ and we can identify $\lim_{\beta\to\infty} \beta(q - \overline{q}) = Q - q \equiv \chi$ by a linear response argument. For $\beta \to \infty$, the integrations over m_i^a and h_i^a can be performed by using the saddle-point method. This yields $m_i^a = 0$ and $h_i^a = \sqrt{\hat{q}} z_i + \hat{m} x_i^0$ as the saddle point, where

$$\hat{q} = \frac{2}{\sigma^2} \left(\frac{q - 2m + Q_0}{\sigma^2} - \frac{\sigma_0^2 \chi}{\sigma^4} \right) G''\left(-\frac{\chi}{\sigma^2} \right) + \frac{2\sigma_0^2}{\sigma^4} G'\left(-\frac{\chi}{\sigma^2} \right), \quad (13)$$

$$\hat{m} = \frac{2}{\sigma^2} G'\left(-\frac{\chi}{\sigma^2} \right) = E, \quad (14)$$

and z_i is a standard Gaussian random variable. Combining all these, we find the consistency between EC approximation and the replica theory as

$$[f]_{A,\mathbf{x}^0,\mathbf{n}} = \underset{q,m,Q,\hat{q},\hat{m},\hat{Q}}{\mathrm{extr}} \left\{ \phi(q, m, Q, \hat{q}, \hat{m}, \hat{Q}) \right\} \quad (15)$$

by identifying $E = \hat{Q} + \hat{q}$ in (6).

V. Experimental validation

We performed numerical experiments for the signal recovery of compressed sensing using the Bernoulli-Gaussian prior

$$P(x) = (1 - \rho)\delta(x) + \rho \frac{\exp\left(-\frac{x^2}{2\sigma_X^2}\right)}{\sqrt{2\pi\sigma_X^2}}, \quad (16)$$

for examining the accuracy of the developed scheme. In the experiments, we set $\rho = 0.1$, $\sigma_X^2 = 1$, and $\sigma^2 = 0.01$, and the correct prior and noise value were used for simulating the Bayesian optimal recovery. The performance was examined for i) row-orthogonal and i.i.d. Gaussian ensembles. In addition to these, iii) random M row selection from discrete cosine transform matrix (random DCT), which does not follow a rotationally invariant distribution but shares the same eigenvalue distribution with the row-orthogonal ensemble, was tested for investigating the significance of rotational invariance.

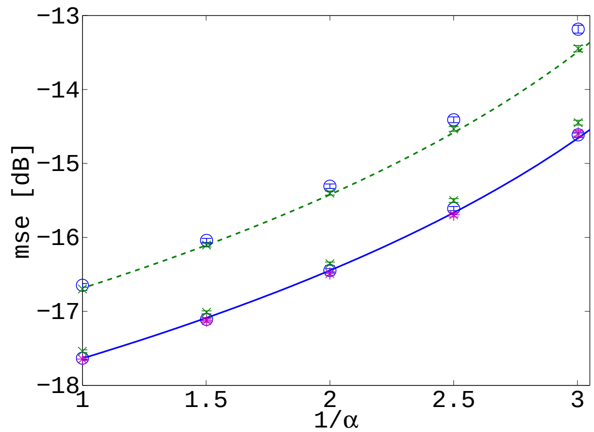

Fig. 1. The normalized mean square error $\mathbb{E}\left\{||\mathbf{m} - \mathbf{x}^0||^2/||\mathbf{x}^0||^2\right\}$ versus $1/\alpha = N/M$. Full (blue) and broken (green) curves represent the theoretical prediction assessed by the replica method for the row-othogonal and i.i.d. Gaussian ensembles, respectively. Circle (blue) indicates experimental results obtained by the EC recovery designed for the row-orthogonal ensemble, whereas cross (green) stands for those by AMP, which is suitable for the i.i.d. Gaussian ensemble. Asterisk (magenta) shows the result for random DCT obtained by the EC recovery designed for the row-orthogonal ensemble.

The equation to be solved for the recovery can be read as

$$\mathbf{h} = \frac{1}{\sigma^2} A^{\mathrm{T}} (\mathbf{y} - A\mathbf{m}) + E\mathbf{m}, \quad (17)$$

$$m_i = \frac{\rho Z(h_i, E)}{1 - \rho + \rho Z(h_i, E)} \frac{h_i}{E + \sigma_X^{-2}}, \quad (18)$$

$$Q_i = \frac{\rho Z(h_i, E)}{1 - \rho + \rho Z(h_i, E)} \left(\frac{1}{E + \sigma_X^{-2}} + \left(\frac{h_i}{E + \sigma_X^{-2}} \right)^2 \right), \quad (19)$$

where $Z(h_i, E) = (1 + \sigma_X^2 E)^{-1/2} \exp\left(\frac{h_i^2}{2(E + \sigma_X^{-2})} \right)$, $\chi = N^{-1} \sum_{i=1}^{N} (Q_i - m_i^2)$, and $E = \frac{2}{\sigma^2} G'\left(-\chi/\sigma^2\right)$. The naive iterative substitution scheme did not exhibit a good convergence property. Therefore, we introduced a dumping factor γ and updated m_i and χ as $(1 - \gamma)m_i + \gamma m_i^{\mathrm{new}} \to m_i$ and $(1 - \gamma)\chi + \gamma\chi^{\mathrm{new}} \to \chi$, where m_i^{new} and χ^{new} are the values evaluated from the right-hand sides of (18) and (19). For all experiments, we truncated the updates up to 3×10^3 iterations setting $\gamma = 0.05$, which led to no divergent behavior but exhibited slower convergence as α decreases.

We constructed the EC approximation assuming the row-orthogonal ensemble. For comparison, we also tested the performance of AMP designed for (16), which is suitable for the i.i.d. Gaussian ensemble. Symbols in Fig. 1 show the signal recovery performance evaluated from 10^3 experiments of $N = 2^{10}$ systems while curves stand for the theoretical prediction assessed by the replica method. These indicate the superiority of the row-orthonal to the i.i.d. Gaussian ensembles in the noisy setting, which was also reported for l_p-recovery in [25]. Excellent agreement between the circles/crosses and the full/broken curves experimentally validates the consistency between the ensemble-dependent proper approximations and the replica theory. Slight deviation of symbols for the inappropriate recovery schemes indicates the necessity for knowing statistical properties of the observation matrix for constructing a theoretically proper approximation, whereas its significance becomes smaller as the compression rate α grows. The result for random DCT indicates that the performance of the row-orthogonal ensembles can be practically gained with a low computational cost, approximately $O(N)$, by random row choice of a Fourier matrix similarly to the noise free case reported in [31].

VI. Summary

We developed a computationally feasible approximate scheme of signal recovery for linear observations affected by Gaussian noises. The scheme follows the Gibbs free energy formalism of statistical mechanics and approximately overcomes the computational difficulty for evaluating the Gibbs free energy by using a Gaussian approximation for which the consistency with the true distribution is imposed for the first moment and a part of the second moment. The asymptotic consistency with the replica theory is guaranteed for a class of the measurement matrix ensembles that are characterized by rotational invariance. Experiments for the Bayesian optimal recovery for compressed sensing using the Bernoulli-Gaussian prior numerically validated the theoretically obtained results.

The combination of the developed recovery scheme and hyper-parameter estimation [14], [15] is under way. Designing a good iteration scheme to solve the recovery equation (17)–(19) is an interesting and important task.

Acknowledgments

The research was funded in part by MEXT KAKENHI Grant No. 25120013 (YK) and Swedish Research Council under VR Grant 621-2011-1024 (MV). MV acknowledges the MEXT KAKENHI Grant No. 24106008 for supporting his visit to the Tokyo Institute of Technology.

References

[1] S. Rangan, A. K. Fletcher, and V. K. Goyal, "Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing," *IEEE Trans. Inform. Theory,* vol. 58, no. 3, pp. 1902–1923, 2012.

[2] D. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *Journal of the American Mathematical Society,* vol. 22, no. 1, pp. 1–53, 2009.

[3] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on lp-norm minimization," *J. Stat. Mech.*, vol. 2009, no. 9, pp. L09003(1–12), 2009.

[4] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, pp. 18914–18919, 2009.

[5] M. Opper and O. Winther, "Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling," *Phys. Rev. E*, vol. 64, pp. 056131(1–14), 2001.

[6] M. Opper and O. Winther, "Expectation Consistent Approximate Inference," *Journal of Machine Learning Research*, vol. 6, pp. 2177–2204, 2005.

[7] T. P. Minka, "Expectation propagation for approximate Bayesian inference," *in Proceedings UAI-2001 (Ed. by J. S. Breese and D. Koller, Morgan Kaufmann)*, pp. 362–369, 2001.

[8] D. Guo, D. Baron, and S. Shamai, "A single-letter characterization of optimal noisy compressed sensing," *Proc. Annual Allerton Conf. Commun., Contr., Computing*, pp. 52–59, 2009.

[9] W. Huleihel and N. Merhav, "Asymptotic MMSE Analysis Under Sparse Representation Modeling," arXiv:1312.3417 [cs.IT]

[10] A. M. Tulino, G. Caire, S. Verdú, S. and Shamai, "Support Recovery With Sparsely Sampled Free Random Matrices," *IEEE Trans. Inform. Theory*, vol. 59, pp. 4243–4271, 2013.

[11] M. Opper and O. Winther, "A mean field algorithms for Bayes learning in large feed-forward neural networks,", *in Advances in Neural Information Processing Systems (Ed. by M. C. Mozer, M. I. Jordan, and T. Petche, The MIT Press)*, vol. 9, pp. 225–231, 1996.

[12] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A: Math. Gen.*, vol. 36, pp. 11111–11121, 2003.

[13] T. Tanaka and M. Okada, "Approximate belief propagation, density evolution, and statistical neurodynamics for CDMA multiuser detection," *IEEE Trans. on Inform. Theory,* vol. 51, pp. 700–706, 2005.

[14] J. P. Vila and P. Schniter, "Expectation-Maximization Bernoulli- Gaussian Approximate Message Passing," *in Proc. Asilomar Conf. on Signals, Systems, and Computers (Pacific Grove, CA),*, Nov. 2011.

[15] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová, "Statistical-Physics-Based Reconstruction in Compressed Sensing," *Phys. Rev. X*, vol. 2, 021005(1–18), 2012.

[16] S. Rangan, "Generalized Approximate Message Passing for Estimation with Random Linear Mixing,", arXiv:1010.5141

[17] T. Shinzato and Y. Kabashima, "Learning from correlated patterns by simple perceptrons," *J. Phys. A: Math. Theor.*, vol. 42, 015005(1–12), 2009.

[18] K. Takeda, S. Uda, and Y. Kabashima, "Analysis of CDMA systems that are characterized by eigenvalue spectrum," *Europhysics Letters*, vol. 76, pp. 1193–1199, 2006.

[19] A. Hatabu, K. Takeda, and Y. Kabashima, "Statistical mechanical analysis of the Kronecker channel model for multiple-input multiple-output wireless communication," *Phys. Rev. E*, vol. 80, 061124(1–12), 2009.

[20] Y. Kabashima, "Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels," *J. Phys. Conf. Ser.* vol. 95, 012001(1–13), 2008.

[21] M. Mézard, G. Parisi, and M. Virasoro, "Spin Glass Theory and Beyond," World Scientific, 1986.

[22] H. Nishimori, "Statistical Physics of Spin Glasses and Information Processing," Oxford University Press, 2001.

[23] T. Tanaka, "Asymptotics of Harish-Chandra-Itzykson-Zuber integrals and free probability theory," *J. Phys. Conf. Ser.*, vol. 95, 012002(1–9), 2008.

[24] R. R. Müller, D. Guo, and A. L. Moustakas, "Vector Precoding for Wireless MIMO Systems and its Replica Analysis," *IEEE JSAC*, vol. 26, no. 3, pp. 530–540, 2008

[25] M. Vehkaperä, Y. Kabashima, and S. Chatterjee, "Analysis of Regularized LS Reconstruction and Random Matrix Ensembles in Compressed Sensing," arXiv:1312.0256.

[26] H. Nishimori and D. Sherrington, "Absence of Replica Symmetry Breaking in a Region of the Phase Diagram of the Ising Spin Glass," *in Disordered and Complex Systems (Ed. by P. Sollich et al, AIP Conf. Proc. 553)*, pp. 67–72, 2001.

[27] T. Plefka, "Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model," *J. Phys. A: Math. Gen.*, vol. 15, pp. 1971–1978, 1982.

[28] K. Nakanishi and H. Takayama, "Mean-field theory for a spin-glass model of neural networks: TAP free energy and paramagnetic to spin-glass transition," *J. Phys. A: Math. Gen.*, vol. 30, pp. 8085–8094, 1997.

[29] J. Barbier, F. Krzakala, and C. Schülke, "Compressed sensing and Approximate Message Passing with spatially-coupled Fourier and Hadamard matrices," arXiv:1312.1740.

[30] T. Fabricius and O. Winther, "Improved Multistage Detector by Mean-Field Annealing in Multi-User CDMA," *In Proceedings of IEEE Intelligent Signal Processing, Applications, and Communication Systems*, 2002.

[31] D. L. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Phil. Trans. R. Soc. A*, vol. 367, 4273–4293, 2009.