Statistical mechanics of dictionary learning

AYAKA SAKATA and YOSHIYUKI KABASHIMA

Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology
Yokohama 226-8502, Japan

received 7 August 2012; accepted in final form 18 July 2013
published online 13 August 2013

PACS 89.20.Ff – Computer science and technology
PACS 75.10.Nr – Spin-glass and other random models

Abstract – Finding a basis matrix (dictionary) by which objective signals are represented sparsely is of major relevance in various scientific and technological fields. We consider a problem to learn a dictionary from a set of training signals. We employ techniques of statistical mechanics of disordered systems to evaluate the size of the training set necessary to typically succeed in the dictionary learning. The results indicate that the necessary size is much smaller than previously estimated, which theoretically supports and/or encourages the use of dictionary learning in practical situations.

Introduction. – In various fields of science and technology, such as Earth observation, astronomy, materials science, and in compiling image databases [1], it has a major relevance to recover original signals from deficient signals obtained by limited number of measurements. The Nyquist-Shannon sampling theorem [2] provides the necessary and sufficient number of measurements for recovering arbitrary band-limited signals. However, techniques based on this theorem sometimes do not match restrictions and/or demands of today’s front-line applications [3,4], and much effort is still being made to find more efficient methodologies.

The concept of sparse representations has recently drawn great attention in such research. Many real-world signals such as natural images are represented sparsely in Fourier/wavelet domains; namely, many components vanish or are negligibly small in amplitude when the signals are represented by Fourier/wavelet bases. This empirical property is exploited in the signal recovery paradigm of compressed sensing (CS) enabling the recovery of sparse signals from much fewer measurements than those the sampling theorem estimates [5–10].

However, the effectiveness of CS relies considerably on the assumption that a basis by which the objective signals look sparse is known in advance. Therefore, in applying CS to general signals of interest, whose bases for sparse representation are unknown, the primary task to accomplish is to identify an appropriate basis (dictionary) for the sparse representation from an available set of training signals. This is often termed dictionary learning (DL) [11–13].

Let us denote the training set of \( M \)-dimensional signals as an \( M \times P \) matrix \( Y = \{Y_{\mu l}\} \), where each column vector \( Y_l \) represents a sample signal and \( P \) is the number of the samples. In a simple scenario, DL is formulated as a problem to find a pair of an \( M \times N \) matrix (dictionary) \( D = \{D_{\mu i}\} \) and an \( N \times P \) sparse matrix \( X = \{X_{il}\} \) such that \( Y = DX \) holds. By DL, the characteristics/trends underlying \( \{Y_l\} \) are extracted into \( D \), and \( Y_l \) can be compactly represented as a superposition of a few dictionary columns, whose combination and strength are specified by the sparse matrix \( X \). DL suits not only efficient signal processing such as CS, but also extraction of non-trivial regularities from high-dimensional data [12,14,15].

An important question of DL is how large a sample size \( P \) is necessary to uniquely identify an appropriate dictionary \( D \), because the ambiguity of the dictionary is fatal in the use for signal/data analysis after learning. As the first answer to this question, an earlier study based on linear algebra showed that when the training set \( Y \) is generated by a pair of matrices \( D^0 \) and \( X^0 \) (planted solution) as \( Y = D^0X^0 \), one can perfectly learn these as a unique solution except for the ambiguities of signs and permutations of matrix elements if \( P > P_c = (k + 1)N_c \) and \( k \) is sufficiently small, where \( k \) is the number of non-zero elements in each column of \( X^0 \) [16]. This result is significant as
it is the first proof that guarantees the learnability with a finite size sample set for DL. At the same time, it is unfortunate that the bound becomes exponentially large in $N$ for $k \sim O(N)$, which motivates us to improve the estimation. A recent rigorous study has shown that almost all dictionaries under the uniform measure are learnable with $P_c \sim O(NM)$ samples when $k \ln M \sim O(\sqrt{N})$ [17]. However, the possibility of dictionary learning with $O(N)$ training samples is indicated by an intuitive consideration: with the knowledge of the positions of non-zero elements in $X^0$, $MN + kP$ unknown variables of $D^0$ and $X^0$ can be determined from $MP$ conditions offered by $Y$ at $P \sim O(N)$ as long as $M - k \sim O(M) > 0$.

The improvement of the estimation $P_c$ is practically significant because it may lead to a considerable reduction of the necessary cost for DL in terms of both sample and computational complexities. In this letter, we take an alternative approach to estimating $P_c$. Specifically, we examine the typical behavior of DL using the replica method in the limit of $N, M, P \to \infty$. The main result of our analysis is that the planted solution is typically learnable by $O(N)$ training samples if negligible mean square errors per element are allowed and $M/N$ is sufficiently large. This theoretically supports and/or encourages the employment of DL in practical applications.

**Problem setting.** Practically performing DL is known as a challenging task; there exists no known computationally feasible algorithm that is mathematically guaranteed to find a correct solution. One main cause of the difficulty is the invariance of $Y = D^0X^0$ under any simultaneous permutations of column and row indices of $D^0$ and $X^0$, respectively. This generally yields an $N!$ degeneracy of an intrinsically identical solution, which makes it impossible to naively formulate DL as a convex optimization problem that can be solved efficiently by various techniques of computational optimization.

Therefore, we herein focus on a major learning strategy

$$\min_{D^0} \left\| Y - \frac{DX}{\sqrt{N}} \right\|^2 \text{ subj. to } \|D\|^2 = MN, \|X\|_0 = NP\theta,$$  

and investigate $P_c$ in a non-constructive manner, where $Y(= N^{-1/2}D^0X^0)$, $\|A\|^2 = \sum_{il} A_{il}^2$ for a matrix $A = \{A_{il}\}$, and $\|A\|$ represents the number of non-zero elements in $A$. The parameter $\theta \in [0, 1]$ denotes the rate of non-zero elements assumed by the learner, and $N^{-1/2}$ is introduced for convenience in taking the large system limit. Variants of (1) are examined in the seminal paper [16], and offer bases of various heuristic learning algorithms [11,12,18-20]. For practical algorithms that achieve the performance suggested by our analysis is left to the future study.

For analytical tractability, we assume that $D^0$ and $X^0 = (X^0)_{ij} \sim \mathcal{N}(0, 1)$ of the planted solution are generated from the uniform distribution characterized by the constraint $\|D^0\|^2 = MN$ and a Gaussian-Bernoulli distribution $P(X^0) = \prod_{i,l} [(1 - \rho)\delta(X^0_{il}) + \rho \exp(- (X^0_{il})^2)/\sqrt{2\pi}],$ respectively.1 We consider that the correct non-zero density $\rho$ can differ from $\theta$ for generality, but we assume $\rho \leq \theta$; otherwise, the correct identification of $D^0$ and $X^0$ is trivially impossible. The main purpose of our study is to reveal whether $D^0$ and $X^0$ can be identified by (1) with $O(N)$ samples or not, characterizing physical picture of DL by the notion of phase transitions. For this aim, we concentrate our analysis on the case of $P \sim O(N)$.

**Statistical-mechanics approach.** Partition function

$$Z_D(D^0, X^0) = \int dD dx \exp\left( -\frac{\beta}{2N} \|DX - D^0X^0\|^2 \right) \times \delta(\|D\|^2 - NM) \delta(\|X\|_0 - NP\theta)$$

(2)

constitutes the basis of our approach since the minimized cost of (1) can be identified with the zero-temperature free energy $F = -\lim_{\beta \to \infty} \beta^{-1} \ln Z(D^0, X^0; \beta)$. This statistically fluctuates depending on $D^0$ and $X^0$. However, as $N, M, P \to \infty$, one can expect that the self-averaging property is realized: i.e., the free energy density $N^{-2}F$ converges to the typical value $f \equiv N^{-2}[F]_0$ with probability unity, where $[\cdots]_0$ stands for the average with respect to $D^0$ and $X^0$. Consequently, this property is also expected to hold for other relevant macroscopic variables of the solution of (1), $D^*$ and $X^*$. Therefore, assessing $f$ is the central issue in our analysis.

This assessment can be carried out systematically using the replica method [23] in the limit of $N \to \infty$ while keeping $\alpha = M/N \sim O(1)$ and $\gamma = P/N \sim O(1)$. Under the replica symmetric (RS) ansatz, where the solution space of (1) is assumed to be composed of at most a few pure states [24], the free energy density is given as

$$f = \text{extr}_{\Omega, \hat{\Omega}} \left\{ -\alpha \left( \hat{Q}_D - \hat{\chi}_D \hat{x}_D - \hat{m}_D m_D + \hat{x}_D + \hat{m}_D^2 \right) \right. \right.$$

$$-\gamma \left( \hat{Q}_x \hat{x}_x - \hat{\chi}_x \hat{x}_x - \hat{m}_x m_x + \lambda \theta - \langle \langle \phi(h; \hat{Q}_X, \hat{\lambda}) \rangle \hat{h} \right)$$

$$+ \alpha \gamma (Q_X - m_D m_D + \rho) + \frac{1 + \hat{Q}_x \hat{x}_x}{2},$$

(3)

where extr$_{\Omega, \hat{\Omega}} \{G(\Omega, \hat{\Omega})\}$ stands for the extremization of a function $G(\Omega, \hat{\Omega})$ with respect to a set of macroscopic variables $\Omega \equiv \{\chi_D, m_D, \chi_x, m_x\}$ and that of their conjugates $\hat{\Omega} \equiv \{\hat{Q}_D, \hat{\chi}_D, \hat{m}_D, \hat{Q}_x, \hat{\chi}_x, \hat{m}_x, \lambda\}$, and

$$\phi(h; \hat{Q}_X, \hat{\lambda}) = \min_{X} \lim_{\epsilon \to 0} \left\{ \frac{\hat{Q}_x X^2}{2} - hX + \lambda |X|^\epsilon \right\}.$$

(4)

Notation $\langle \langle \cdots \rangle \rangle_h$ represents the average with respect to $h$

appeared to the distribution $P(h) = \rho P(h|X^0 \neq 0) + (1 - \rho) P(h|X^0 = 0)$, where $P(h|X^0 \neq 0)$ and $P(h|X^0 = 0)$ are given by zero-mean Gaussian distributions with variances

1The extra knowledge on the generative distributions of $D^0$ and $X^0$ can offer the best possible learning scheme that outperforms (1). For the performance evaluation of such an optimal learning scheme, see recent studies [21,22].
A distinctive feature of $X^*$ is the divergence of the local susceptibility $\partial X^*/\partial h$ at “border” cases of $h=\pm h_{th}$ (fig. 1(d)). This affects the increase in the effective degree of freedom (ratio) as follows: $\theta_{eff}=\theta+\langle|h\delta(|h|-h_{th})\rangle\gamma$, whereas $h_{th}$ is determined so as to satisfy $\theta=\int_{|h|>h_{th}}\delta h P(h)$ indicating the sparsity condition $\|X\|_0=NP\theta$. The excess $\langle|h\delta(|h|-h_{th})\rangle\gamma$ is supposed to represent a combinatorial complexity for classifying each element of $X^*$ that corresponds to the border case $|h|=h_{th}$ into the active case, $|h|>h_{th}$ and $X^*\neq 0$, or the inactive case, $|h|<h_{th}$ and $X=0$. The divergence of $\partial X^*/\partial h|_{h=\pm h_{th}}$ is also accompanied by the instability of the RS solution against perturbations that break the replica symmetry [26]. The influence of this instability is discussed later.

**Physical implications.** At the extremum of (3), the relationships

$$m_D=[D^0\cdot D^*]_0/MN, \quad m_X=[X^0\cdot X^*]_0/NP, \quad Q_X=[X^*\cdot X^*]_0/NP$$

(5)

hold, where $A\cdot B=\sum_{ij}A_{ij}B_{ij}$ represents the inner product of two matrices of the same dimension $A$ and $B$. These provide the mean square errors (per element), which measure the performance of DL as $\epsilon_D=\langle MN\rangle^{-1}\|D^*-D^0\|_0^2=2(1-m_D)$ and $\epsilon_X=\langle NP\rangle^{-1}\|X^*-X^0\|_0^2=\rho-2m_X+Q_X$. The variables $\chi_D$ and $\chi_X$ physically mean the sensitivity of the estimates $D^*$ and $X^*$ when the cost of (1) is linearly perturbed.

Equation (4) represents the effective single-body minimization problem concerning an element of $X$ that is statistically equivalent to (1) [25]. Here, the randomness of $D^0$ and $X^0$ is effectively replaced by the random local field $h$. The first and second terms of $P(h)$ correspond to the cases where an element of $X^0$ is given as $X^0\neq 0$ and $X^0=0$, respectively. Under a given $h$, the solution $X^*$ that minimizes the cost of (4) is offered as $X^*=h/Q_X$ for $|h|>h_{th}=2Q_X\lambda$ and 0 otherwise (fig. 1(c)). We refer to the cases of $|h|>h_{th}$ and $|h|<h_{th}$ as “active” and “inactive,” respectively. When $X^0\neq 0$, $h$ is generated from a Gaussian distribution ($P(h|X^0\neq 0)$) of zero mean and variance $\tilde{\chi}_X+m^2\chi^2$, and $X^*$ is more likely to be active than when $X^0=0$, for which $h$ is characterized by another zero-mean Gaussian ($P(h|X^0=0)$) of a smaller variance $\tilde{\chi}_X$ (fig. 1(a), (b)). Therefore, one can expect that the hard-thresholding scheme based on $h_{th}$ represents proper assignment of zero/non-zero elements in $X^*$ so as to accurately estimate $X^0$ and $D^0$ if $m_X$ is sufficiently large.

Fig. 1: (Color online) (a) and (b) show distributions of the local field $h$ (a) $P(h|X^0\neq 0)$ for $X^0\neq 0$ and (b) $P(h|X^0=0)$ for $X^0=0$. (c) and (d) show $X^*$ and $\partial X^*/\partial h$ as functions of $h$, respectively.

**Actual solutions.** We found two types of solutions; the first one is characterized by $m_D=1$ and $Q_X=m_X=\rho$, while the second is characterized by $m_X=0$ and $m_D=0$. The former case provides $\epsilon_D=\epsilon_X=0$ indicating the correct identification of $D^0$ and $X^0$, and hence we call it the success solution. The latter is referred to as the failure solution since $m_D=0$ and $m_X=0$ indicate the complete failure of information extraction of $D^0$ and $X^0$.

**Success solution (S) exists when $\gamma>1$ and**

$$\alpha>\theta_{eff}(\theta, \rho)=\theta+(1-\rho)\sqrt{\frac{2}{\pi}} u \exp \left(-\frac{u^2}{2}\right)$$

(6)

hold, where $u=H^{-1}((\theta-\rho)/(2(1-\rho)))$ and $H^{-1}(x)$ is the inverse function of $H(x)=2(\pi)^{-1/2}\int_{-\infty}^x dt e^{-t^2/2}$. $S$ is further classified into two cases depending on $\gamma$.

For $\gamma<\gamma_S$, where

$$\gamma_S(\alpha, \theta, \rho)=\frac{\alpha}{\alpha-\theta_{eff}}$$

(7)

$\chi_D$ and $\chi_X$ are finite. On the other hand, for $1<\gamma<\gamma_S$, $\chi_D$ and $\chi_X$ tend to infinity, keeping $R=\chi_D/\chi_X$ finite.

To physically interpret this classification, let us take a variation around $Y=N^{-1/2}D^0X^0$, which yields

$$0=\delta(DX)|_{D^0,X^0}=D^0\delta X+\delta DX^0$$

(8)

If $\delta D=0$ and $\delta X=0$ are the unique solution of (8), the planted solution is locally stable. Otherwise, there are “marginal” modes along which the cost of (1) does not increase locally, and the solution set forms a manifold. The number of constraints of (8), $MP$, coincides with that of the degree of freedom of $\delta D$ and $\delta X$, $MN+NP\theta S_{eff}$, at $P=\gamma_S N$. Thus, the classification below/above $\gamma_S$ corresponds to the change in the number of marginal modes around the planted solution.

To confirm the validity of this interpretation, we numerically evaluated the number of marginal modes of (8) in the case of $\alpha=1/2$ and $\theta=\rho=0.1$, which is shown in fig. 2. The assessment of $\gamma_S$ when $\theta=\rho$ is conjectured to be exact since the effect of the border elements is negligible under this condition. Figure 2 indicates that the number of marginal modes scales as $O(N^2)$ for $\gamma<\gamma_S=1.25$, while
Fig. 2: (Color online) $\gamma$-dependence of the ratio of marginal modes relative to $N^2$ at $\alpha = 0.5$ and $\rho = \theta = 0.1$. The behavior at $N \to \infty$ extrapolating from the results of finite $N$ is denoted by the dashed line. Inset: $N$-dependence of the ratio of marginal modes for $\gamma = 2$. The dashed line stands for $N^{-1}$ as a guide. Each marker represents the average of 100 experiments.

it scales as $O(N)$, and the contribution of the marginal modes approaches zero, for $\gamma > \gamma_S$ (inset). This result coincides with our theoretical assessment. At the same time, this implies that identifying the planted solution without any errors by (1) is difficult as long as $\gamma \sim O(1)$, but the discrepancies per element caused by the marginal modes are negligibly small and could be allowed in many practical situations.

In the case of $\gamma < 1$, for any $N \times P$ matrix $Z$ of $\|Z\|_0 = NP\theta$, $X^* = aZ$ and $D^* = a^{-1}Y(ZZ^T)^{-1}Z^T$ minimize the cost of (1) to zero, where $a$ is determined such that $\|D^*\|^2_a = MN$. This implies that the set of solutions of (1) spreads widely, and the weight of the planted solution is negligibly small in the state space. This may be why $S$ disappears for $\gamma < 1$.

Failure solution ($F$) exists for $\forall \gamma \geq 0$. If $\alpha < \theta_{\text{eff}}^F(\theta) = \theta + \sqrt{\frac{2}{\pi}} \exp \left( -\frac{\theta^2}{2} \right)$ (9)

where $v = H^{-1}(\theta/2)$ holds, $F$ always offers $\chi_D, \chi_X \to \infty$ making the free energy $f$ vanish. For $\alpha > \theta_{\text{eff}}^F$, on the other hand, $\chi_D$ and $\chi_X$ become finite implying that a single solution of (1) is locally stable for most directions and offers $f > 0$, if $\gamma$ is greater than $\gamma_F(\alpha, \theta) = \left( \alpha^{1/2} - (\theta_{\text{eff}}^F)^{1/2} \right)^2$. (10)

The inequality $\theta_{\text{eff}}^S \geq \theta_{\text{eff}}^F$ always holds because the influence of the border elements for $F$ is stronger than that for $S$, which leads to $\gamma_S \leq \gamma_F$.

Figure 3 illustrates changes in state space that occur for sufficiently large $\alpha$ under the RS assumption. For $\gamma < 1$, $F$ is a unique solution. As $\gamma$ increases, $S$ appears at $\gamma = 1$, and the number of marginal modes changes from $O(N^2)$ to $O(N)$ at $\gamma = \gamma_S$. This implies that when negligibly small linear perturbations are added to the cost of (1), the limits $\lim_{N \to \infty} S = 0$ and $\lim_{N \to \infty} S_X = 0$ still hold for $S$ of $\gamma > \gamma_S$ while they can be boosted to $O(1)$ for $S$ of $\gamma < \gamma_S$.

For $\gamma < (\gamma_S \leq) \gamma_F$, $S$ and $F$ are degenerated providing $f = 0$. However, at $\gamma = \gamma_F$, $S$ becomes thermodynamically dominant by keeping $f = 0$, while $F$ begins to have positive $f$. This means that the planted solution is typically learnable by $P > P_c = N\gamma_F \sim O(N)$ training samples if negligible mean square errors per element are allowed.

Figure 4 plots the phase diagram on an $\alpha$-$\theta$ plane. The region above $\alpha = \theta_{\text{eff}}^F(\theta)$ (curve) represents the condition under which the planted solution is typically learnable by $O(N)$ training samples. DL is impossible in the region below $\alpha = \theta$ (straight line) because $X^0$ cannot be correctly recovered even if $D^0$ is known [7]. How $P_c$ scales with respect to $N$ in the region $\theta < \alpha < \theta_{\text{eff}}^F(\theta)$ is beyond the scope of this letter, but it is an interesting question nonetheless.

Summary and discussion. In summary, we have assessed the size of training samples required for correctly learning a planted solution in DL using the replica method. An earlier rigorous study [17] provided an upper bound of $O(N \ln N)$ for the required sample size, which is the current best estimate. Our analysis improved this estimate indicating that $O(N)$ samples are sufficient for learning a planted dictionary with allowance for negligible square errors per element when the number of non-zero signals is sufficiently small compared to that of measurements.

It was shown that the identification of the dictionary can be characterized as a phase transition with respect
to the number of training samples. The picture of phase space is shared with other solution-planted constraint satisfaction problems [27]; as γ grows, a cluster including the planted solution is separated from the other clusters at a certain critical point (γ = 1) and thermodynamically dominates for γ > γ_c (> 1). Our RS analysis probably does not predict the exact value of γ_c since the RS solutions are unstable against the replica-symmetry-breaking (RSB) disturbances. However, the RS estimate of γ_c still serves as an upper bound of improved estimates of the critical ratio γ_c obtained under the RSB framework. This is because the free energy value of F assessed under any step RSB ansatz is guaranteed to be greater than or equal to that of the RS solution due to the positivity constraint of the entropy of free energy value of F evaluated on the basis of the following formula:

\[ F \leq \frac{1}{N} \sum_i \log Z_i \]

above.

Unfortunately, assessing γ is unstable against the replica-symmetry-breaking (RSB) disturbances. However, the RS estimate of γ_c still serves as an upper bound of improved estimates of the critical ratio γ_c obtained under the RSB framework. This is because the free energy value of F assessed under any step RSB ansatz is guaranteed to be greater than or equal to that of the RS solution due to the positivity constraint of the entropy of free energy value of F evaluated on the basis of the following formula:

\[ F \leq \frac{1}{N} \sum_i \log Z_i \]

above.

This work was partially supported by a Grant-in-Aid for JSPS Fellow (No. 23-4665) (AS) and KAKENHI Nos. 22300003 and 25120013 (YK).

**Appendix:** derivation of (3). – In general, the configurational average of the free energy density could be evaluated on the basis of the following formula:

\[ f_\beta = \frac{1}{\beta N_{\text{tr}} \to \infty} \lim_{n \to 0} \frac{1}{2N} \sum_{m=0}^{\beta n} \mathcal{D}(D_0, X^0) \]

\[ Z_0^\beta(D_0, X^0) = \int \prod_{a=1}^{\beta n} \{dD_0 dX^0 \delta(||D_0||^2 - N M) \}
\]

\[ \times \delta(||X^a||_0 - N P \theta) \exp \left(-\frac{\beta}{2N} \sum_{a=1}^{\beta n} ||D_0^a X^a - D_0^a X^0||^2 \right), \]

\[ (A.2) \]

which is valid for only \( n \in \mathbb{N} \), over the distributions of the planted solutions \( D_0^a \) and \( X^0 \) that are given by

\[ P_{D_0}(D_0^a) = \frac{1}{N_D} \delta(||D_0^a||^2 - N M) \]

and

\[ P_{X_0}(X^0) = \prod_{i,a} \left\{ (1 - \rho)(X^0_{i,a}) + \frac{\rho}{\sqrt{2\pi}} \exp \left(-\frac{(X^0_{i,a})^2}{2}\right) \right\}, \]

\[ (A.4) \]

respectively, where \( N_D \) is the normalization constant. In performing the integrals of \( 2(n + 1) \) variables \( \{D_0^a, \{D^a\} \} \) (and \( \{X^a, \{X\} \} \)) that come out in this evaluation, we insert trivial identities with respect to all combinations of replicas \( \{a, b\} \in \{0, 1, 2, \ldots, n\} \),

\[ 1 = NM \int dq_D^a \delta(D_0^a \cdot D_0^b - N M q_D^a) \]

\[ \times \prod_{a < b} \delta(||D_0^a||^2 - N M q_D^a) \]

\[ (A.5) \]

and

\[ 1 = NP \int dq_X^a \delta(X^a \cdot X^b - N P q_X^a) \]

\[ \times \prod_{a \leq b} \delta(||X^a||_0 - N P \theta) \]

\[ \prod_{a \leq b} \delta(X^a \cdot X^b - N P q_X^b) \]

\[ (A.6) \]

to the integrand. Let us denote \( Q_D \equiv \{q_D^a\} \) and \( Q_X \equiv \{q_X^a\} \), and introduce two joint distributions

\[ P_D(\{D_0^a\}; Q_D) = \frac{P_{D_0}(D_0^a)}{V_D(Q_D)} \]

\[ \times \prod_{a < b} \delta(||D_0^a||^2 - N M q_D^a) \]

\[ (A.7) \]

\[ P_X(\{X^a\}; Q_X) = \frac{P_{X_0}(X^0)}{V_X(Q_X)} \]

\[ \times \prod_{a \leq b} \delta(||X^a||_0 - N P \theta) \]

\[ \prod_{a \leq b} \delta(X^a \cdot X^b - N P q_X^b) \]

\[ (A.8) \]

where \( V_D(Q_D) \) and \( V_X(Q_X) \) are the normalization constants. The above-mentioned computation provides the following expression:

\[ Z_0^\beta(D_0, X^0) = \int d(N M Q_D) d(N P Q_X) V_D(Q_D) V_X(Q_X) \]

\[ \times \prod_{a=1}^{n} \exp \left(-\frac{\beta}{2N} \sum_{\mu,l} t_{\mu,l}^a \right) \]

\[ \frac{Q_D \cdot Q_X}{Q_D \cdot Q_X} \]

\[ (A.9) \]

where \( t_{\mu,l}^a = N^{-1/2} \sum_{\nu=1}^{N} (D_{\nu \mu}^a X_{\nu \nu}^a - D_{\nu \mu}^a X_{\nu \nu}^0) \). Notation \( \{X^a, \{X\}\} \) represents the average with respect to \( \{D_0^a\} \) and \( \{X^0, \{X\}\} \) within the state space specified by \( Q_D \) and \( Q_X \), whose distributions are given by eqs. (A.7) and (A.8). Distributions (A.7) and (A.8) are independent of each other, and provide each entry of \( \{D^a\} \) and \( \{X^a\} \) with zero mean and a finite variance. This allows us to utilize the central limit theorem indicating that we can handle \( t_{\mu,l}^a \) as multivariate Gaussian random variables that follow
where $\mathcal{T}$ stands for an $n \times n$ matrix whose entries are given as $T^{ab} = \langle q^b \bar{q}^a \rangle - \langle q^b \bar{q}^a \rangle + \rho$. Utilizing this and evaluating integrals of $Q_X$ and $Q_D$ by means of the saddle point method lead to an expression

$$
\lim_{N \to \infty} \frac{1}{N^2} \langle Z^{\beta}_N(D^0, X^0) \rangle_0 = \text{extr} \left[ -\frac{\alpha \gamma}{2} \ln \det (\mathcal{I}_n + \beta \mathcal{T}) \right.
+ \gamma \left\{ \frac{1}{2} \left( \frac{\mathcal{Q}_X X}{2} + \ln \left( \prod_{a=0}^{n} dX^a \right) P_X(X^0) e^{-Z} \right) \right.
+ \frac{\alpha}{2} \left( \frac{\mathcal{Q}_D D}{2} + \ln \det \mathcal{Q}_D + n \ln(2\pi) \right) + n\lambda \theta \).
$$

(A.10)

Here, $\mathcal{I}_n$ represents the $n \times n$ identity matrix, auxiliary variables $Q_D \equiv (\hat{q}^b_D)$ and $\hat{Q}_X \equiv (\hat{q}^b_X)$ are introduced in evaluating $V_D(Q_D)$ and $V_X(Q_X)$ with use of the saddle point method, and $\Xi \equiv \frac{1}{2} \sum_{a,b=0}^{n} \hat{q}^a_X X^a X^b + \lambda \sum_{a=1}^{n} \ln \exp + |X|^2$. Extremization should be taken with respect to $\lambda$ and four kinds of macroscopic variables $Q_D$, $Q_X$, $\hat{Q}_D$, and $\hat{Q}_X$.

Exactly evaluating (A.10) should provide the correct leading-order estimate of $N^{-2} \ln \langle Z^{\beta}_N(D^0, X^0) \rangle_0$ for each of $n \in N$. However, we here restrict the candidate of the dominant saddle point to that of the replica symmetric form as

$$(q^{ab}_{D D}, q^{ab}_{D X}, q^{ab}_{X D}, q^{ab}_{X X}) = (1, Q_X, \hat{Q}_D, \hat{Q}_X), \quad a = b,$$

$$(q^{ab}_{D X}, q^{ab}_{X D}, q^{ab}_{X X}) = (q^{ab}_{D X}, q^{ab}_{X D}, q^{ab}_{X X}), \quad a \neq b, (a, b) \neq 0,$$

$$(m_{D D}, m_{D X}, m_{X D}, m_{X X}) = (0, 0, 0, 0),$$

so as to obtain an analytic expression with respect to $\rho$.

This yields

$$
\ln \det (\mathcal{I}_n + \beta \mathcal{T}) = n \ln (1 + \beta (Q_X - q_D q_X)) + \ln \left( \frac{1 + n \beta (q_D q_X - 2m_{D X} + \rho)}{1 + \beta (Q_X - q_D q_X)} \right),
$$

(A.11)

$$
\frac{1}{2} \langle \hat{Q}_X X \rangle + \ln \left( \prod_{a=0}^{n} dX^a P_X(X^0) e^{-Z} \right) - \frac{n}{2} \langle \hat{Q}_D D \rangle + \langle \hat{Q}_X X \rangle + \ln \left( \prod_{a=0}^{n} dX^a e^{-Z} \right) \rangle \rangle_h
$$

(A.12)

and

$$
\frac{1}{2} \langle \hat{Q}_D D \rangle - \ln \det (\hat{Q}_D) = \frac{n}{2} (\hat{Q}_D + \hat{q}_D q_D) - n m_{D X} m_{D X} - \frac{n^2}{2} \langle \hat{q}_D q_X \rangle + \ln \left( \prod_{a=0}^{n} dX^a e^{-Z} \right) \rangle \rangle_h
$$

(A.13)

where $\xi = (\hat{Q}_X + \hat{q}_X X)^2 / 2 - hX + \lambda \sum_{a=\infty}^{0} |X|^2$. Further, the following replacement of variables is convenient in handling our computation in the limit of $\beta \to \infty$: $\hat{Q}_D + \hat{q}_D \to \beta \hat{Q}_D$, $q_D \to \beta^2 \hat{Q}_D$, $1 - \hat{Q}_D \to \chi_D / \beta$, $\hat{Q}_X + \hat{q}_X \to \beta \hat{Q}_X$, $q_X \to \beta^2 \hat{Q}_X$, $X - q_X \to \chi_X / \beta$, and $\lambda \to \beta \lambda$. In $\beta \to \infty$, the integral form $\int dX e^{-Z}$, is replaced to $e^{-\beta_0 (\hat{Q}_X, \hat{q}_X, \lambda)}$ by applying the saddle point method. Inserting (A.11)-(A.13) and the rescaled variables into (A.10) offers the expression of the zero-temperature free energy density (3).

REFERENCES

[1] Starck J.-L. et al., Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity (Cambridge University Press) 2010.

[2] Nyquist H., Trans. AIEE, 47 (1928) 617.

[3] Maliyev S. et al., Brit. J. Anaesth., 84 (2000) 743.

[4] Liu L. Y. and Robert G. (Editors), Protein NMR Spectroscopy: Principal Techniques and Applications (John Wiley & Sons Ltd.) 2011.

[5] Donoho D., IEEE Trans. Inf. Theory, 52 (2006) 1289.

[6] Candès E. J. and Tao T., IEEE Trans. Inf. Theory, 51 (2005) 4203.

[7] Kabashima Y. et al., J. Stat. Mech., L09003. 2009.

[8] Donoho D. L. et al., Proc. Natl. Acad. Sci. U.S.A., 106 (2009) 18914.

[9] Ganguli S. and Sompolinsky H., Phys. Rev. Lett., 104 (2010) 188701.

[10] Krzakala F. et al., Phys. Rev. X, 2 (2012) 021005.

[11] Rubinstein R. et al., Proc. IEEE, 98 (2010) 1045.

[12] Elad M., Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing (Springer-Verlag) 2010.

[13] Gleichman S. and Eldar Y. C., IEEE Trans. Inf. Theor., 57 (2011) 6958.

[14] Bley O. and Elad M., J. Vis. Commun. Image Represent., 19 (2008) 270.

[15] Gong T. et al., BMC Bioinform., 12 (2011) 82.

[16] Aharon M. M. et al., Linear Algebra Appl., 416 (2006) 48.

[17] Vainscher D. et al., J. Mach. Learn. Res., 12 (2011) 3259.

[18] Olshausen B. A. and Field D. J., Vis. Res., 37 (1997) 3311.

[19] Engan K. et al., IEEE Int. Conf. Acoust. Speech Signal Process., 5 (1999) 2443.

[20] Aharon M. et al., IEEE Trans. Signal Process., 54 (2006) 4311.

[21] Sakata A. and Kabashima Y., arXiv:1301.6199.

[22] Krzakala F. et al., arXiv:1301.5898.

[23] Dotzenko V., Introduction to the Replica Theory of Statistical Systems (Cambridge University Press) 2001.

[24] Mézard M. et al., Spin Glass Theory and Beyond (World Scientific Publ.) 1987.

[25] Guo D. and Verdú S., IEEE Trans. Inf. Theory, 51 (2005) 1983.

[26] de Almeida J. R. L. and Thouless D. J., J. Phys. A: Math. Gen., 11 (1978) 983.

[27] Krzakala F. et al., Phys. Rev. Lett., 102 (2009) 238701.

[28] Mézard M. and Montanari A., Information, Physics, and Computation (Oxford University Press) 2009.

[29] Obuchi T. et al., J. Phys. A: Math. Theor., 43 (2010) 485004.