Semi-analytic approximate stability selection for correlated data in generalized linear models

OPEN ACCESS

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

Semi-analytic approximate stability selection for correlated data in generalized linear models

Takashi Takahashi and Yoshiyuki Kabashima[1]

Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan
E-mail: takahashi.t.cc@m.titech.ac.jp

Received 19 March 2020
Accepted for publication 2 August 2020
Published 3 September 2020

Online at stacks.iop.org/JSTAT/2020/093402
https://doi.org/10.1088/1742-5468/ababff

Abstract. We consider the variable selection problem of generalized linear models (GLMs). Stability selection (SS) is a promising method proposed for solving this problem. Although SS provides practical variable selection criteria, it is computationally demanding because it needs to fit GLMs to many re-sampled datasets. We propose a novel approximate inference algorithm that can conduct SS without the repeated fitting. The algorithm is based on the replica method of statistical mechanics and vector approximate message passing of information theory. For datasets characterized by rotation-invariant matrix ensembles, we derive state evolution equations that macroscopically describe the dynamics of the proposed algorithm. We also show that their fixed points are consistent with the replica symmetric solution obtained by the replica method. Numerical experiments indicate that the algorithm exhibits fast convergence and high approximation accuracy for both synthetic and real-world data.

Keywords: cavity and replica method, message-passing algorithms, statistical inference

[1]Present address: Institute for Physics of Intelligence and Department of Physics, Graduate School of Science, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan

Contents

1. Introduction

Modern statistics require the handling of high-dimensional data. The term *high-dimensional* refers to the situation where the ratio of the number of measurements and the number of the parameters is of order 1. Among the many tasks in high-dimensional statistics, variable selection of statistical models is a notoriously difficult problem. In high-dimensional settings, standard sparse regression methods, including the least absolute shrinkage and selection operator (LASSO) method [1], suffer from the problem of choosing the regularization parameter. Although re-sampling methods, such as stability selection (SS) [2], can provide much more accurate variable selection criteria, these methods require substantial computational costs.

J. Stat. Mech. (2020) 093402

As an example, let us consider variable selection in logistic regression. In this regression, we have a dataset $D = \{(\boldsymbol{a}_\mu, y_\mu)\}_{\mu=1}^{M}$, where each $\boldsymbol{a}_\mu = (a_{\mu 1}, a_{\mu 2}, \ldots, a_{\mu N})^\top \in \mathbb{R}^N$ is an N-dimensional vector of features or predictors, and each $y_\mu \in \{-1, 1\}$ is the associated binary response variable. We denote by $\top$ the matrix/vector transpose. The response variables are independently generated based on a true parameter $\boldsymbol{x}_0 = (x_{0,1}, x_{0,2}, \ldots, x_{0,N})^\top \in \mathbb{R}^N$ as

$$y_\mu \sim \frac{1}{1+\mathrm{e}^{-\boldsymbol{a}_\mu^\top \boldsymbol{x}_0}}\delta(y_\mu - 1) + \frac{1}{1+\mathrm{e}^{\boldsymbol{a}_\mu^\top \boldsymbol{x}_0}}\delta(y_\mu + 1), \quad \mu = 1, 2, \ldots, M. \tag{1}$$

We denote by $\mathrm{supp}(\boldsymbol{x}_0) = \{i | x_{0,i} \neq 0, i = 1, 2, \ldots, N\}$ the support of $\boldsymbol{x}_0$. The goal of variable selection is to estimate $\mathrm{supp}(\boldsymbol{x}_0)$ from the dataset D. In high-dimensional settings, a simple strategy is to use ℓ_1 regularized logistic regression or LASSO [1]. LASSO seeks an estimator of $\boldsymbol{x}_0$ as

$$\hat{\boldsymbol{x}}(\gamma, D) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \left[-\sum_{\mu=1}^{M} \log \frac{1}{1+\mathrm{e}^{-y_\mu \boldsymbol{a}_\mu^\top \boldsymbol{x}}} + \gamma \sum_{i=1}^{N} |x_i| \right], \tag{2}$$

where $\gamma > 0$ is a parameter that controls the strength of the ℓ_1 regularizer. The ℓ_1 regularization term $\gamma \sum_{i=1}^{N} |x_i|$ allows LASSO to select variables by shrinking a part of the estimated parameters exactly to 0. For any given regularization parameter γ, LASSO estimates $\mathrm{supp}(\boldsymbol{x}_0)$ as

$$\hat{S}(\gamma, D) \equiv \{i | \hat{x}_i(\gamma, D) \neq 0, i = 1, 2, \ldots, N\}. \tag{3}$$

Unfortunately, this estimated support $\hat{S}(\gamma, D)$ depends strongly on the choice of the regularization parameter γ in real-world datasets. Hence, choosing the regularization parameter for variable selection can be more challenging than for prediction of the response variable where cross-validation is guaranteed to offer the optimal choice on average if features are generated independently from an identical distribution [3].

SS was proposed for tackling this difficulty. We denote by $D^* = \{(\boldsymbol{a}_1^*, y_1^*), (\boldsymbol{a}_2^*, y_2^*), \ldots, (\boldsymbol{a}_M^*, y_M^*)\}$ a resampled dataset of size M drawn with replacement from D. For this resampled dataset, the resampling probability $\Pi_i(\gamma)$ that the variable i is included in the estimated support is given by

$$\Pi_i(\gamma) = \mathrm{Prob}_{D^*}\left[\hat{x}_i(\gamma, D^*) \neq 0\right]. \tag{4}$$

The probability in (4) is with respect to the random resampling and it equals the relative frequency for $\hat{x}_i(\gamma, D^*) \neq 0$ over all M^M resampled dataset with size M. The probability in (4) can be approximated by B random samples $D_1^*, D_2^*, \ldots, D_B^*$ (B should be large):

$$\Pi_i(\gamma) \simeq \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(\hat{x}_i(\gamma, D_b^*) \neq 0\right), \tag{5}$$

where $\mathbb{1}(\ldots)$ is the indicator function. This probability is termed the *selection probability* and measures the *stability* of each variable. SS chooses variables that have large selection

J. Stat. Mech. (2020) 093402

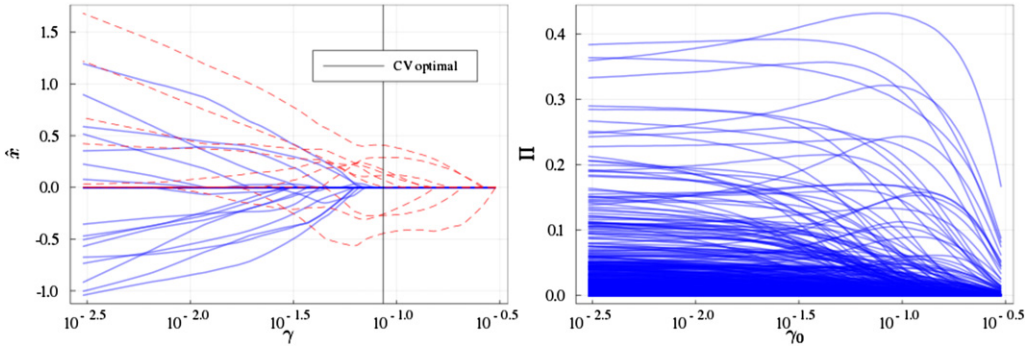

Figure 1. Left: the LASSO solutions $\hat{\boldsymbol{x}}(\gamma, D)$ based on (2) for the colon cancer dataset with $M = 62$ and $N = 2000$. The vertical line corresponds to the cross-validation optimal regularization parameter. The red-dashed lines represent variables chosen by the cross-validation procedure. The non-zero variables strongly depend on the choice of the regularization parameter γ. Right: the selection probability $\boldsymbol{\Pi}(\lambda_0)$ based on (6). The selection probability is less dependent on the choice of γ_0, indicating that choosing the regularization parameter is less critical than the naive LASSO.

probabilities. The original literature [2] combined the above resampling procedure with the randomization of the regularization parameter γ as follows

$$\Pi_i(\gamma_0) = \mathrm{Prob}_{D^*,\gamma}\left[\hat{x}_i(\boldsymbol{\gamma}, D^*) \neq 0\right], \quad i = 1, 2, \ldots, N, \tag{6}$$

$$\hat{\boldsymbol{x}}(\boldsymbol{\gamma}, D^*) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N}\left[-\sum_{\mu=1}^{M}\log\frac{1}{1+\mathrm{e}^{-y_\mu^*(\boldsymbol{a}_\mu^*)^\top\boldsymbol{x}}} + \sum_{i=1}^{N}\gamma_i|x_i|\right], \tag{7}$$

$$\gamma_i \sim \frac{1}{2}\delta(\gamma_i - \gamma_0) + \frac{1}{2}\delta(\gamma_i - 2\gamma_0), \quad i = 1, 2, \ldots, N. \tag{8}$$

Figure 1 illustrates the comparison of the LASSO solution (2) and the selection probability (6). Here we used the colon cancer dataset [4]. The task is to distinguish cancer from normal tissue using the micro-array data with $N = 2000$ features per example. The data were obtained from 22 normal ($y_\mu = -1$) and 40 ($y_\mu = 1$) cancer tissues. The total number of the samples is $M = 62$. The left panel of figure 1 shows the LASSO solutions for the various regularization parameters. Non-zero variables depend strongly on γ. Choosing the proper value of γ is difficult for the original LASSO. Although the cross-validation can optimize the prediction for the response variable, this choice often includes false positive elements [5]. The right panel of figure 1 shows the selection probability for various γ_0 in (8). This figure motivates that choosing the regularization parameter γ_0 is much less critical for the selection probability and that the selection probability approach has a better chance of selecting truly relevant variables.

A major drawback of SS is its computational cost. SS repeatedly solves the ℓ_1 regularized logistic regression in (7) for multiple resampled datasets and regularization parameters. The number of resampled datasets and regularization parameters B needs to be large so that the selection probability is reliably estimated.

J. Stat. Mech. (2020) 093402

In this study, we address the problem of this computational cost. We propose a novel approximate inference algorithm that can conduct SS without repeated fitting. The algorithm is based on the replica method [6] of statistical mechanics and *vector approximate message passing* (VAMP) [7, 8] of information theory. We term our algorithm *replicated VAMP* (rVAMP).

The rest of the paper is organized as follows. In section 2, we describe SS in generalized linear models (GLMs) that we will focus on, and in section 3, we derive the proposed algorithm using the replica method and VAMP. In section 4, we analyze the proposed algorithm in a large system limit under the assumption that the set of features is characterized by rotation-invariant matrix ensembles. There, we derive the state evolution for self-averaging rVAMP that macroscopically describes the convergence dynamics of rVAMP in an approximate manner, and show that its fixed point is consistent with the replica symmetric solution. In section 5, we apply the proposed algorithm to logistic regression. Through numerical experiments, we confirm the validity of our theoretical analysis and show that the proposed algorithm exhibits fast convergence and high approximation accuracy for both synthetic and real-world data. The final section is devoted to a summary and conclusion.

1.1. Related work

Malzahn and Opper first proposed a combination of the replica method and approximate inference to reduce the computational cost of resampling methods [9–11]. They demonstrated that employing the adaptive Thouless–Anderson–Palmer (TAP) method [12, 13], as an approximate inference algorithm, can accurately estimate the bootstrap generalization error for Gaussian process classification/regression. However, the poor convergence of this method is a major flaw of their approach. The adaptive TAP method is based on a naive iteration of TAP equations. The literature in information theory has revealed that the convergence property of such naive iteration scheme is terribly bad [8, 14, 15]. Thus it requires to find a correct choice of initial conditions. As an algorithm, the adaptive TAP method is undesirable because approximate inference aims to save computation time.

The aforementioned algorithmic problem has been significantly improved by the discovery of approximate message passing (AMP) algorithms in information theory. This type of algorithms was first introduced as an efficient signal processing algorithm [16]. [16] analyzed its convergence dynamics in a large system limit and showed its fast convergence. [16] also revealed that the fixed point of the AMP algorithm shares the same fixed point with the corresponding TAP equation, and thus, AMP can be used as an efficient algorithm to solve the TAP equation. Subsequently, [14, 17] developed its mathematically rigorous analysis. These rigorous analyses were further generalized in [18, 19]. However, the above analyses are based on the assumptions that the elements of the feature vectors are independently and identically distributed (i.i.d.) zero-mean random variables, which is not realistic in the context of statistics. To go beyond such simple distributions, VAMP and similar generalizations [7, 8, 20] were developed based on expectation propagation (EP) of machine learning [21, 22]. Under the assumption that feature matrices, whose rows are composed of each feature vectors, are drawn from rotation-invariant random matrix ensembles, VAMP algorithms were analyzed in a large

J. Stat. Mech. (2020) 093402

system limit. These analyses derived the convergence dynamics of the VAMP algorithms and revealed that their fixed points are consistent with the corresponding adaptive TAP equations [7, 8, 23–26]. In this paper, we extend such VAMP algorithms to replicated systems for approximately performing SS in GLMs.

[27] proposed an AMP-based approximate resampling algorithm for SS. However, the algorithm assumes independence between the features and was developed for linear regression only. A preliminary application of VAMP to SS in linear regression was also demonstrated [28]. In the present study, we further generalize the use of VAMP to GLMs, and also carry out a theoretical analysis of this method.

1.2. Notations

Here we introduce some shorthand notations used throughout the paper. We denote by $[\omega_i]_{1\leqslant i\leqslant N}$ a vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)^\top \in \mathbb{R}^N$. Similarly, we denote by $[\Omega_{\mu i}]_{\substack{1\leqslant\mu\leqslant M\\ 1\leqslant i\leqslant N}}$ an $M \times N$ matrix whose μith entry is $\Omega_{\mu i}$. For an integer $n = 1, 2, \ldots$, we denote by $\mathbf{1}_n = (1, 1, \ldots, 1) \in \mathbb{R}^n$ a constant vector. For integers $n \in \mathbb{N}, m \in \mathbb{Z}$, and vectors $\boldsymbol{\omega} = [\omega_i]_{1\leqslant i\leqslant n}, \boldsymbol{\psi} = [\psi_i]_{1\leqslant i\leqslant n}$, we denote by $\boldsymbol{\omega}/\boldsymbol{\psi} = [\omega_i/\psi_i]_{1\leqslant i\leqslant n}$ and $\boldsymbol{\omega}^m = [\omega_i^m]_{1\leqslant i\leqslant n}$ component-wise operations. Finally, $\langle\boldsymbol{\omega}\rangle \equiv \sum_{i=1}^n \omega_i/n$.

2. Stability selection in generalized linear models

In the following, we consider SS in generalized linear regression/classification. We have a dataset $D = \{(\boldsymbol{a}_\mu, y_\mu)\}_{\mu=1}^M$, where each $\boldsymbol{a}_\mu = (a_{\mu 1}, a_{\mu 2}, \ldots, a_{\mu N})^\top \in \mathbb{R}^N$ is an N-dimensional vector of features or predictors, and each $y_\mu \in \mathcal{Y} \subset \mathbb{R}$ is the associated response variable. The domain of the response variables $\mathcal{Y}$ includes $\mathbb{R}$ for regression and $\{-1, 1\}$ for classification. We also use matrix/vector notations $A = [a_{\mu i}]_{\substack{1\leqslant\mu\leqslant M\\ 1\leqslant i\leqslant N}} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_M)^\top \in \mathcal{Y}^M$.

Let $D^* = \{(\boldsymbol{a}_1^*, y_1^*), \ldots, (\boldsymbol{a}_M^*, y_M^*)\}$ be a resampled dataset composed of M data points drawn with replacement from D. Some data point $(\boldsymbol{a}_\mu, y_\mu)$ in D appears multiple times in D^*, and while others do not appear at all. SS in generalized linear regression/classification computes the selection probability $\boldsymbol{\Pi} \in [0, 1]^N$ by repeatedly refitting GLMs $p_{y|z}$ for multiple resampled datasets and regularization parameters:

$$\Pi_i(\gamma_0) = \mathrm{Prob}_{D^*,\boldsymbol{\gamma}}\left[\hat{x}_i(\boldsymbol{\gamma}, D^*) \neq 0\right], \quad i = 1, 2, \ldots, N, \tag{9}$$

$$\hat{\boldsymbol{x}}(\boldsymbol{\gamma}, D^*) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N}\left[-\sum_{\mu=1}^M \log\ p_{y|z}(y_\mu^*|(\boldsymbol{a}_\mu^*)^\top\boldsymbol{x}) + \sum_{i=1}^N \gamma_i|x_i|\right], \tag{10}$$

$$\gamma_i \sim \frac{1}{2}\delta(\gamma_i - \gamma_0) + \frac{1}{2}\delta(\gamma_i - 2\gamma_0), \quad i = 1, 2, \ldots, N, \tag{11}$$

where $\gamma_0 > 0$ is a control parameter that determines the amount of the regularization. The goal of this paper is to develop a computationally efficient algorithm that returns $\boldsymbol{\Pi}(\gamma_0)$ for any positive γ_0.

J. Stat. Mech. (2020) 093402

3. Replicated vector approximate message passing

To approximate the computation of the selection probability $\mathbf{\Pi}$, we will use the replica method and VAMP. This section provides a derivation of the proposed algorithm.

3.1. Occupation vector representation of sampling with replacement

For convenience, let us introduce the occupation vector representation of the resampled dataset D^*. The resampled dataset D^* is composed of M data points sampled from D with replacement. Hence, it can be represented by a vector of *occupation* numbers $\boldsymbol{c} = (c_1, c_2, \ldots, c_M)^\top \in \{0, 1, \ldots, M\}^M$ with $\sum_{\mu=1}^M c_\mu = M$, where c_μ is the number of times that the data point $(\boldsymbol{a}_\mu, y_\mu)$ appears in D^*. Although the strict distribution of $\boldsymbol{c}$ is the multinomial distribution, for large M, the correlation among $\{c_\mu\}_{\mu=1}^M$ is weak. By ignoring this correlation, we can approximate the distribution of $\boldsymbol{c}$ by a product of Poisson distribution with mean 1 [9] as:

$$p(\boldsymbol{c}) \simeq \prod_{\mu=1}^M \frac{\mathrm{e}^{-1}}{c_\mu!}. \tag{12}$$

In this way, we can rewrite the average with respect to D^* by the average over the random variable $\boldsymbol{c} \in \{0, 1, \ldots\}^M$ that follows the probability distribution (12), which is simple and easy to handle.

3.2. Statistical mechanical formulation of stability selection

The selection probability $\mathbf{\Pi}$ in (9) is defined through the optimization problem in (10). To use techniques of statistical mechanics and approximate inference algorithm, we introduce the Boltzmann distribution as

$$p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D) = \frac{1}{Z^{(\beta)}(\boldsymbol{c}, \boldsymbol{\gamma}, D)} \delta(\boldsymbol{z} - A\boldsymbol{x}) \prod_{\mu=1}^M p_{y|z}(y_\mu | z_\mu)^{\beta c_\mu} \prod_{i=1}^N \mathrm{e}^{-\beta\gamma_i |x_i|}, \tag{13}$$

$$Z^{(\beta)}(\boldsymbol{c}, \boldsymbol{\gamma}, D) = \int \delta(\boldsymbol{z} - A\boldsymbol{x}) \prod_{\mu=1}^M p_{y|z}(y_\mu | z_\mu)^{\beta c_\mu} \prod_{i=1}^N \mathrm{e}^{-\beta\gamma_i |x_i|} \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{z}, \tag{14}$$

where $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{z} \in \mathbb{R}^M$, $\beta > 0$ is the inverse temperature, and Z is the partition function. The random variables $\boldsymbol{\gamma}$ and $\boldsymbol{c}$ follow distributions (11) and (12), respectively. Then the selection probability can be written using the Boltzmann distribution at the zero-temperature limit as follows:

$$\Pi_i(\gamma_0) = \mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}} \left[\mathbb{1}(\hat{x}_i(\boldsymbol{c}, \boldsymbol{\gamma}) \neq 0) \right], \quad i = 1, 2, \ldots, N, \tag{15}$$

$$\hat{x}_i(\boldsymbol{c}, \boldsymbol{\gamma}) = \lim_{\beta \to \infty} \int x_i p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{z}. \tag{16}$$

J. Stat. Mech. (2020) 093402

In the rest of the paper, we will omit the argument D when there is no risk of confusion to avoid cumbersome notation. Still, note that we calculate the above quantities only for the *fixed* dataset D.

3.3. Replica method for semi-analytic approximate resampling method

Our purpose is to compute the selection probability $\mathbf{\Pi}(\gamma_0)$ for any $\gamma_0 > 0$. For this, we compute the distribution of $\hat{x}_i$:

$$p(m_i) = \mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\mathbb{1}\left(m_i - \hat{x}_i\left(\boldsymbol{c},\boldsymbol{\gamma}\right)\right)\right], \tag{17}$$

which is reduced to computing the moments $\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}[\hat{x}_i^r(\boldsymbol{c},\boldsymbol{\gamma})]$ for any $r = 1, 2, \ldots$. We now describe how the replica method can be used for this purpose, following the approach of [9].

We use $\mathrm{d}^r\boldsymbol{x} = \mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\ldots\mathrm{d}\boldsymbol{x}_r$ to denote a measure over $\mathbb{R}^{N\times r}$, with $\boldsymbol{x}_1 = (x_{1,1},\ldots,x_{1,N})^\top,\ldots,\boldsymbol{x}_r = (x_{r,1},\ldots,x_{r,N})^\top$. Analogously, we denote by $\mathrm{d}^r\boldsymbol{z} = \mathrm{d}\boldsymbol{z}_1\mathrm{d}\boldsymbol{z}_2\ldots\mathrm{d}\boldsymbol{z}_r$ as a measure over $\mathbb{R}^{M\times r}$, with $\boldsymbol{z}_1 = (z_{1,1},\ldots,z_{1,M})^\top,\ldots,\boldsymbol{z}_r = (z_{r,1},\ldots,z_{r,M})^\top$. Using the definition (16), the moments $\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}[\hat{x}_i^r(\boldsymbol{c},\boldsymbol{\gamma})]$ can be formally written as[2]

$$\begin{aligned}\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\hat{x}_i^r(\boldsymbol{c},\boldsymbol{\gamma})\right] &= \lim_{\beta\to\infty}\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\int\prod_{s=1}^{r}x_{s,i}\prod_{s=1}^{r}p^{(\beta)}(\boldsymbol{x}_s,\boldsymbol{z}_s)\mathrm{d}^r\boldsymbol{x}\mathrm{d}^r\boldsymbol{z}\right]\\ &= \lim_{\beta\to\infty}\int\prod_{s=1}^{r}x_{s,i}\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\prod_{s=1}^{r}\left\{\frac{1}{Z^{(\beta)}(\boldsymbol{c},\boldsymbol{\gamma})}\delta(\boldsymbol{z}_s - A\boldsymbol{x}_s)\right.\right.\\ &\quad\left.\left.\times\prod_{\mu=1}^{M}p_{y|z}(y_\mu|z_{s,\mu})^{\beta c_\mu}\prod_{i=1}^{N}\mathrm{e}^{-\beta\gamma_i|x_{s,i}|}\right\}\right]\mathrm{d}^r\boldsymbol{x}\mathrm{d}^r\boldsymbol{z},\end{aligned} \tag{18}$$

which is difficult to evaluate analytically due to the presence of the partition function that depends on $\boldsymbol{c}$ and $\boldsymbol{\gamma}$ in the denominator. The replica trick [6] bypasses this problem via an identity $\lim_{n\to 0}Z^{n-r} = Z^{-r}$. Using this identity, (18) is formally re-expressed as

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\hat{x}_i^r(\boldsymbol{c},\boldsymbol{\gamma})\right] = \lim_{n\to 0}\lim_{\beta\to\infty}\mathcal{A}_{i,n}^{(\beta)}, \tag{19}$$

where

$$\begin{aligned}\mathcal{A}_{i,n}^{(\beta)} = \int\prod_{s=1}^{r}x_{s,i}\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\left(Z^{(\beta)}(\boldsymbol{c},\boldsymbol{\gamma})\right)^{n-r}\prod_{s=1}^{r}\left\{\delta(\boldsymbol{z}_s - A\boldsymbol{x}_s)\right.\right.\\ \left.\left.\times\prod_{\mu=1}^{M}p_{y|z}(y_\mu|z_{s,\mu})^{\beta c_\mu}\prod_{i=1}^{N}\mathrm{e}^{-\beta\gamma_i|x_{s,i}|}\right\}\right]\mathrm{d}^r\boldsymbol{x}\mathrm{d}^r\boldsymbol{z}.\end{aligned} \tag{20}$$

[2] Since the aim of this paper is not to provide rigorous analysis, we assume that the exchange of limits, integrals, etc, such as $\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}[\lim_{\beta\to\infty}\ldots] = \lim_{\beta\to\infty}\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}[\ldots]$, are possible throughout the paper without further justification.

J. Stat. Mech. (2020) 093402

The advantage of this formula is that for integers $n \geqslant r$, the negative power of the partition function $(Z^{(\beta)}(\boldsymbol{c}, \boldsymbol{\gamma}))^{-r}$ is eliminated by an integral with respect to n replicated variables. More precisely, using the definition of the partition function (14), we obtain

$$\mathcal{A}_{i,n}^{(\beta)} = \Xi_n \int \prod_{s=1}^{r} x_{s,i} \frac{1}{\Xi_n} \prod_{s=1}^{n} \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s) \prod_{\mu=1}^{M} \mathbb{E}_{c_\mu} \left[\prod_{s=1}^{n} p_{y|z}(y_\mu | z_{s,\mu})^{\beta c_\mu} \right] \times \prod_{i=1}^{N} \mathbb{E}_{\gamma_i} \left[\prod_{s=1}^{n} \mathrm{e}^{-\beta\gamma_i |x_{s,i}|} \right] \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z}, \tag{21}$$

where Ξ_n is the normalization constant

$$\Xi_n = \int \prod_{s=1}^{n} \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s) \prod_{\mu=1}^{M} \mathbb{E}_{c_\mu} \left[\prod_{s=1}^{n} p_{y|z}(y_\mu | z_{s,\mu})^{\beta c_\mu} \right] \prod_{i=1}^{N} \mathbb{E}_{\gamma_i} \left[\prod_{s=1}^{n} \mathrm{e}^{-\beta\gamma_i |x_{s,i}|} \right] \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z}. \tag{22}$$

The expression (21) is much easier to evaluate than the negative power of the partition function. We call the probability density function given by

$$p^{(\beta)}(\{\boldsymbol{x}_s\}_{s=1}^n, \{\boldsymbol{z}_s\}_{s=1}^n) = \frac{1}{\Xi_n} \prod_{s=1}^{n} \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s) \prod_{\mu=1}^{M} \mathbb{E}_{c_\mu} \left[\prod_{s=1}^{n} p_{y|z}(y_\mu | z_{s,\mu})^{\beta c_\mu} \right] \times \prod_{i=1}^{N} \mathbb{E}_{\gamma_i} \left[\prod_{s=1}^{n} \mathrm{e}^{-\beta\gamma_i |x_{s,i}|} \right], \tag{23}$$

the replicated system. Note that by construction $\lim_{n\to 0} \Xi_n = 1$.

In this way, we have replaced the original problem with computing first moments of the replicated system (23). Of course, we would not expect that we could compute the moments exactly. Otherwise we should have obtained the exact solution without using the replicas. The replica method evaluates a formal expression of $\lim_{\beta\to\infty} \mathcal{A}_{i,n}^{(\beta)}$ for $n = r+1, r+2, \ldots$.under appropriate approximations, and then extrapolates it as $n \to 0$.

To obtain a formal expression of $\lim_{\beta\to\infty} \mathcal{A}_{i,n}^{(\beta)}$, the following observation is critical. Because the replicated system (23) is merely a product of the n-copied systems, it is intrinsically invariant under any permutations of $\{(\boldsymbol{x}_1, \boldsymbol{z}_1), (\boldsymbol{x}_2, \boldsymbol{z}_2), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)\}$. This property is termed the replica symmetry. From this property, de Finetti's representation theorem [29] guarantees that the replicated system (23) is expressed as

$$p^{(\beta)}(\{\boldsymbol{x}_s\}_{s=1}^n, \{\boldsymbol{z}_s\}_{s=1}^n) = \int \prod_{s=1}^{n} p^{(\beta)}(\boldsymbol{x}_s, \boldsymbol{z}_s | \boldsymbol{\eta}) p^{(\beta)}(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\eta}, \tag{24}$$

where $\boldsymbol{\eta}$ is a vector of some random variables that reflects the effects of $\boldsymbol{c}$ and $\boldsymbol{\gamma}$. This expression indicates that $\mathcal{A}_{i,n}^{(\beta)}$ is reduced to a considerably simple form

J. Stat. Mech. (2020) 093402

$$\mathcal{A}_{i,n}^{(\beta)} = \int \left(\int x_i p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z} \right)^r \left(\int p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z} \right)^{n-r} p^{(\beta)}(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\eta}$$
$$= \int \left(\int x_i p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z} \right)^r p^{(\beta)}(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\eta}, \tag{25}$$

that can be easily extrapolated as $n \to 0$. The second equality follows from the normalization condition $\int p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z} = 1$. Thus by obtaining tractable approximate densities for $p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\eta})$ and $p^{(\beta)}(\boldsymbol{\eta})$ in (24), we can obtain an arbitrary degree of the moment without refitting[3].

3.4. Replica symmetric Gaussian expectation propagation in the replicated system

To approximate the replicated system (23), we will use the Gaussian diagonal EP of machine learning [21, 22] that is used to derive VAMP in [8]. For $i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, M$, let $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{z}}_\mu \in \mathbb{R}^n$ be $(x_{1,i}, x_{2,i}, \ldots, x_{n,i})^\top \in \mathbb{R}^n$ and $(z_{1,\mu}, z_{2,\mu}, \ldots, z_{n,\mu})^\top \in \mathbb{R}^n$, respectively. The Gaussian diagonal EP recursively updates the following two approximate densities:

$$p_1^{(\beta)}(\{\boldsymbol{x}_s\}_{s=1}^n, \{\boldsymbol{z}_s\}_{s=1}^n) \propto \prod_{\mu=1}^M \mathbb{E}_{c_\mu} \left[\prod_{s=1}^n p_{y|z}(y_\mu|z_{s,\mu})^{\beta c_\mu} \right] \prod_{i=1}^N \mathbb{E}_{\gamma_i} \left[\prod_{s=1}^n \mathrm{e}^{-\beta\gamma_i|x_{s,i}|} \right]$$
$$\times \underbrace{\prod_{i=1}^N \mathrm{e}^{-\frac{1}{2}\tilde{\boldsymbol{x}}_i^\top \Lambda_{1x,i}^{(\beta)} \tilde{\boldsymbol{x}}_i + (\boldsymbol{h}_{1x,i}^{(\beta)})^\top \tilde{\boldsymbol{x}}_i} \prod_{\mu=1}^M \mathrm{e}^{-\frac{1}{2}\tilde{\boldsymbol{z}}_\mu^\top \Lambda_{1z,\mu}^{(\beta)} \tilde{\boldsymbol{z}}_\mu + (\boldsymbol{h}_{1z,\mu}^{(\beta)})^\top \tilde{\boldsymbol{z}}_\mu}}_{\tilde{p}_1^{(\beta)}(\{\boldsymbol{x}_s\}_{s=1}^n, \{\boldsymbol{z}_s\}_{s=1}^n)}, \tag{26}$$

$$p_2^{(\beta)}(\{\boldsymbol{x}_s\}_{s=1}^n, \{\boldsymbol{z}_s\}_{s=1}^n) \propto \prod_{s=1}^n \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s)$$
$$\times \underbrace{\prod_{i=1}^N \mathrm{e}^{-\frac{1}{2}\tilde{\boldsymbol{x}}_i^\top \Lambda_{2x,i}^{(\beta)} \tilde{\boldsymbol{x}}_i + (\boldsymbol{h}_{2x,i}^{(\beta)})^\top \tilde{\boldsymbol{x}}_i} \prod_{\mu=1}^M \mathrm{e}^{-\frac{1}{2}\tilde{\boldsymbol{z}}_\mu^\top \Lambda_{2z,\mu}^{(\beta)} \tilde{\boldsymbol{z}}_\mu + (\boldsymbol{h}_{2z,\mu}^{(\beta)})^\top \tilde{\boldsymbol{z}}_\mu}}_{\tilde{p}_2^{(\beta)}(\{\boldsymbol{x}_s\}_{s=1}^n, \{\boldsymbol{z}_s\}_{s=1}^n)}, \tag{27}$$

where $\Lambda_{1x,i}^{(\beta)}, \Lambda_{2x,i}^{(\beta)}, \Lambda_{1z,\mu}^{(\beta)}, \Lambda_{2z,\mu}^{(\beta)} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{h}_{1x,i}^{(\beta)}, \boldsymbol{h}_{2x,i}^{(\beta)}, \boldsymbol{h}_{1z,\mu}^{(\beta)}, \boldsymbol{h}_{2z,\mu}^{(\beta)} \in \mathbb{R}^n$ are natural parameters of the Gaussians. The first approximation is a factorized distribution but contains the original non-Gaussian factors. The second approximation is a multivariate Gaussian distribution that replaces the non-Gaussian factors by the factorized Gaussians. Both of these distributions are tractable but ignore either the interactions or non-Gaussian factors. To include both the interactions and non-Gaussian factors, EP determines the natural parameters using the following moment-matching conditions:

[3] Of course, the replica symmetry may not hold for $n \notin \mathbb{N}$. In such cases, we have to include the effect of the replica symmetry breaking [6]. However, we restrict ourselves to the replica symmetric case for simplicity.

J. Stat. Mech. (2020) 093402

$$\int x_{s,i} p_1^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int x_{s,i} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int x_{s,i} \tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z}, \tag{28}$$

$$\int z_{s,\mu} p_1^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int z_{s,\mu} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int z_{s,\mu} \tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z}, \tag{29}$$

$$\int x_{s,i} x_{t,i} p_1^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int x_{s,i} x_{t,i} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int x_{s,i} x_{t,i} \tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z}, \tag{30}$$

$$\int z_{s,\mu} z_{t,\mu} p_1^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int z_{s,\mu} z_{t,\mu} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \int z_{s,\mu} z_{t,\mu} \tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z}, \tag{31}$$

for any $i = 1, 2, \ldots, N$, $\mu = 1, 2, \ldots, M$, and $s, t = 1, 2, \ldots, n$. Schematically, the update rule of EP is depicted in algorithm 1. There, the density $\tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)}$ is used to the moment-matching condition in lines 10–17 and 25–33.

The critical issue is to choose an appropriate form of the natural parameters in (26) and (27). Based on the observations in section 3.3, we impose the replica symmetry for these parameters:

$$\Lambda_{1x,i}^{(\beta)} = \begin{pmatrix} \beta\hat{Q}_{1x,i} - \beta^2\hat{v}_{1x,i} & & -\beta^2\hat{v}_{1x,i} \\ & \ddots & \\ -\beta^2\hat{v}_{1x,i} & & \beta\hat{Q}_{1x,i} - \beta^2\hat{v}_{1x,i} \end{pmatrix}, \tag{32}$$

$$\Lambda_{2x,i}^{(\beta)} = \begin{pmatrix} \beta\hat{Q}_{2x,i} - \beta^2\hat{v}_{2x,i} & & -\beta^2\hat{v}_{2x,i} \\ & \ddots & \\ -\beta^2\hat{v}_{2x,i} & & \beta\hat{Q}_{2x,i} - \beta^2\hat{v}_{2x,i} \end{pmatrix}, \tag{33}$$

$$\Lambda_{1z,\mu}^{(\beta)} = \begin{pmatrix} \beta\hat{Q}_{1z,\mu} - \beta^2\hat{v}_{1z,\mu} & & -\beta^2\hat{v}_{1z,\mu} \\ & \ddots & \\ -\beta^2\hat{v}_{1z,\mu} & & \beta\hat{Q}_{1z,\mu} - \beta^2\hat{v}_{1z,\mu} \end{pmatrix}, \tag{34}$$

$$\Lambda_{2z,\mu}^{(\beta)} = \begin{pmatrix} \beta\hat{Q}_{2z,\mu} - \beta^2\hat{v}_{2z,\mu} & & -\beta^2\hat{v}_{2z,\mu} \\ & \ddots & \\ -\beta^2\hat{v}_{2z,\mu} & & \beta\hat{Q}_{2z,\mu} - \beta^2\hat{v}_{2z,\mu} \end{pmatrix}, \tag{35}$$

$$\boldsymbol{h}_{1x,i}^{(\beta)} = \beta h_{1x,i} \mathbf{1}_N, \tag{36}$$

$$\boldsymbol{h}_{2x,i}^{(\beta)} = \beta h_{2x,i} \mathbf{1}_N, \tag{37}$$

$$\boldsymbol{h}_{1z,\mu}^{(\beta)} = \beta h_{1z,\mu} \mathbf{1}_M, \tag{38}$$

$$\boldsymbol{h}_{2z,\mu}^{(\beta)} = \beta h_{2z,\mu} \mathbf{1}_M. \tag{39}$$

With these parameterizations, we use $\hat{\boldsymbol{Q}}_{1x} = (\hat{Q}_{1x,1}, \hat{Q}_{1x,2}, \ldots, \hat{Q}_{1x,N})^\top$ for the vector notation. $\hat{\boldsymbol{Q}}_{2x}$, $\hat{\boldsymbol{Q}}_{1z}$, $\hat{\boldsymbol{Q}}_{2z}$, $\hat{\boldsymbol{v}}_{1x}$, $\hat{\boldsymbol{v}}_{2x}$, $\hat{\boldsymbol{v}}_{1z}$, $\hat{\boldsymbol{v}}_{2z}$, $\boldsymbol{h}_{1x}$, $\boldsymbol{h}_{2x}$, $\boldsymbol{h}_{1z}$, and $\boldsymbol{h}_{2z}$ are defined similarly. These parameterizations allow the extrapolation $n \to 0$ as follows.

J. Stat. Mech. (2020) 093402

Algorithm 1. Expectation propagation.

Require: Approximate densities $p_1^{(\beta)}, p_2^{(\beta)}$ and the number of iterations T_{iter}.
1: Select initial $\Lambda_{1x,i}^{(\beta)}, \Lambda_{1z,\mu}^{(\beta)}, \boldsymbol{h}_{1x,i}^{(\beta)}$, and $\boldsymbol{h}_{1z,\mu}^{(\beta)}$
2: **for** $t = 1, 2, \ldots, T_{\mathrm{iter}}$ **do**
3: // Factorized part (moment computation for $p_1^{(\beta)}$)
4: **for** $i = 1, 2, \ldots, N, \mu = 1, 2, \ldots, M$ **do**
5: $\hat{\boldsymbol{x}}_{1,i}^{(\beta)} = \int \tilde{\boldsymbol{x}}_i p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z}$
6: $\hat{\boldsymbol{z}}_{1,\mu}^{(\beta)} = \int \tilde{\boldsymbol{z}}_\mu p_1^{(\beta)} \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{z}$
7: $V_{1x,i}^{(\beta)} = \int \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} - (\hat{\boldsymbol{x}}_{1,i}^{(\beta)})(\hat{\boldsymbol{x}}_{1,i}^{(\beta)})^\top$
8: $V_{1z,\mu}^{(\beta)} = \int \tilde{\boldsymbol{z}}_\mu \tilde{\boldsymbol{z}}_\mu^\top p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} - (\hat{\boldsymbol{z}}_{1,\mu}^{(\beta)})(\hat{\boldsymbol{z}}_{1,\mu}^{(\beta)})^\top$
9: **end for**
10: // Moment-matching $(1 \to 2)$
11: **for** $i = 1, 2, \ldots, N, \mu = 1, 2, \ldots, M$ **do**
12: update $\Lambda_{2x,i}^{(\beta)}, \Lambda_{2z,\mu}^{(\beta)}, \boldsymbol{h}_{2x,i}^{(\beta)}$ and $\boldsymbol{h}_{2z,\mu}^{(\beta)}$ so that the density $\tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)}$ has the same moment with $p_1^{(\beta)}$ calculated in line 4–9
13: $\boldsymbol{h}_{2x,i}^{(\beta)} = (V_{1x,i}^{(\beta)})^{-1} \hat{\boldsymbol{x}}_{1,i}^{(\beta)} - \boldsymbol{h}_{1x,i}^{(\beta)}$
14: $\boldsymbol{h}_{2z,\mu}^{(\beta)} = (V_{1z,\mu}^{(\beta)})^{-1} \hat{\boldsymbol{z}}_{1,\mu}^{(\beta)} - \boldsymbol{h}_{1z,\mu}^{(\beta)}$
15: $\Lambda_{2x,i}^{(\beta)} = (V_{1x,i}^{(\beta)})^{-1} - \Lambda_{1x,i}^{(\beta)}$
16: $\Lambda_{2z,\mu}^{(\beta)} = (V_{1z,\mu}^{(\beta)})^{-1} - \Lambda_{1z,\mu}^{(\beta)}$
17: **end for**
18: // Gaussian part (moment computation for $p_2^{(\beta)}$)
19: **for** $i = 1, 2, \ldots, N, \mu = 1, 2, \ldots, M$ **do**
20: $\hat{\boldsymbol{x}}_{2,i}^{(\beta)} = \int \tilde{\boldsymbol{x}}_i p_2^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z}$
21: $\hat{\boldsymbol{z}}_{2,\mu}^{(\beta)} = \int \tilde{\boldsymbol{z}}_\mu p_2^{(\beta)} \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{z}$
22: $V_{2x,i}^{(\beta)} = \int \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top p_2^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} - (\hat{\boldsymbol{x}}_{2,i}^{(\beta)})(\hat{\boldsymbol{x}}_{2,i}^{(\beta)})^\top$
23: $V_{2z,\mu}^{(\beta)} = \int \tilde{\boldsymbol{z}}_\mu \tilde{\boldsymbol{z}}_\mu^\top p_2^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} - (\hat{\boldsymbol{z}}_{2,\mu}^{(\beta)})(\hat{\boldsymbol{z}}_{2,\mu}^{(\beta)})^\top$
24: **end for**
25: // Moment-matching $(2 \to 1)$
26: **for** $i = 1, 2, \ldots, N, \mu = 1, 2, \ldots, M$ **do**
27: update $\Lambda_{1x,i}^{(\beta)}, \Lambda_{1z,\mu}^{(\beta)}, \boldsymbol{h}_{1x,i}^{(\beta)}$ and $\boldsymbol{h}_{1z,\mu}^{(\beta)}$ so that the density $\tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)}$ has the same moment with $p_2^{(\beta)}$ calculated in line 19–24
28: $\boldsymbol{h}_{1x,i}^{(\beta)} = (V_{2x,i}^{(\beta)})^{-1} \hat{\boldsymbol{x}}_{2,i}^{(\beta)} - \boldsymbol{h}_{2x,i}^{(\beta)}$
29: $\boldsymbol{h}_{1z,\mu}^{(\beta)} = (V_{2z,\mu}^{(\beta)})^{-1} \hat{\boldsymbol{z}}_{2,\mu}^{(\beta)} - \boldsymbol{h}_{2z,\mu}^{(\beta)}$
30: $\Lambda_{1x,i}^{(\beta)} = (V_{2x,i}^{(\beta)})^{-1} - \Lambda_{2x,i}^{(\beta)}$
31: $\Lambda_{1z,\mu}^{(\beta)} = (V_{2z,\mu}^{(\beta)})^{-1} - \Lambda_{2z,\mu}^{(\beta)}$
32: **end for**
33: **end for**
34: **return** $\Lambda_{1x,i}^{(\beta)}, \Lambda_{2x,i}^{(\beta)}, \Lambda_{1z,\mu}^{(\beta)}, \Lambda_{2z,\mu}^{(\beta)}$ and $\boldsymbol{h}_{1x,i}^{(\beta)}, \boldsymbol{h}_{2x,i}^{(\beta)}, \boldsymbol{h}_{1z,\mu}^{(\beta)}, \boldsymbol{h}_{2z,\mu}^{(\beta)}$.

J. Stat. Mech. (2020) 093402

For $\eta_{x,i}, \eta_{z,\mu} \in \mathbb{R}$, let $\phi_{x,i}^{(\beta)}$ and $\phi_{z,\mu}^{(\beta)}$ be

$$\phi_{x,i}^{(\beta)} = \frac{1}{\beta} \log \int \exp\left(-\beta \frac{\hat{Q}_{1x,i}}{2} x^2 + \beta(h_{1x,i} + \sqrt{\hat{v}_{1x,i}}\eta_{x,i})x - \beta\gamma_i|x| \right) \mathrm{d}x, \tag{40}$$

$$\phi_{z,\mu}^{(\beta)} = \frac{1}{\beta} \log \int \exp\left(-\beta \frac{\hat{Q}_{1z,\mu}}{2} z^2 + \beta(h_{1z,\mu} + \sqrt{\hat{v}_{1z,\mu}}\eta_{z,\mu})z + \beta c_\mu \log\ p_{y|z}(y_\mu|z) \right) \mathrm{d}z. \tag{41}$$

We also denote by $Dx = \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$ the standard Gaussian measure, and by $\mathrm{Diagm}(\boldsymbol{x})$ a diagonal matrix with $[\mathrm{Diagm}(\boldsymbol{x})]_{ii} = x_i$. The use of the replica symmetric parameterizations (32)–(39) yields the following expressions for the moments and the moment-matching conditions that are used in line 10–17 and 25–33 in algorithm 1. First, for the approximate density $p_1^{(\beta)}$, we obtain

$$\int x_{s,i} p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} = \hat{x}_{1,i}, \tag{42}$$

$$\int x_{s,i} x_{t,i} p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} = v_{1x,i} + \hat{x}_{1,i}^2, \quad s \neq t, \tag{43}$$

$$\int x_{s,i}^2 p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} = \frac{\chi_{1x,i}}{\beta} + v_{1x,i} + \hat{x}_{1,i}^2, \tag{44}$$

$$\int z_{s,\mu} p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} = \hat{z}_{1,\mu}, \tag{45}$$

$$\int z_{s,\mu} z_{t,\mu} p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} = v_{1z,\mu} + \hat{z}_{1,\mu}^2, \quad s \neq t, \tag{46}$$

$$\int z_{s,\mu}^2 p_1^{(\beta)} \mathrm{d}^n\boldsymbol{x} \mathrm{d}^n\boldsymbol{z} = \frac{\chi_{1z,\mu}}{\beta} + v_{1z,\mu} + \hat{z}_{1,\mu}^2, \tag{47}$$

where

$$\hat{x}_{1,i} = \frac{\mathbb{E}_{\gamma_i}\left[\int \frac{\partial \phi_{x,i}^{(\beta)}}{\partial h_{1x,i}} \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]}{\mathbb{E}_{\gamma_i}\left[\int \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]}, \tag{48}$$

$$\chi_{1x,i} = \frac{\mathbb{E}_{\gamma_i}\left[\int \frac{\partial^2 \phi_{x,i}^{(\beta)}}{\partial h_{1x,i}^2} \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]}{\mathbb{E}_{\gamma_i}\left[\int \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]}, \tag{49}$$

$$v_{1x,i} = \frac{\mathbb{E}_{\gamma_i}\left[\int \left(\frac{\partial \phi_{x,i}^{(\beta)}}{\partial h_{1x,i}} \right)^2 \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]}{\mathbb{E}_{\gamma_i}\left[\int \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]} - \left(\frac{\mathbb{E}_{\gamma_i}\left[\int \frac{\partial \phi_{x,i}^{(\beta)}}{\partial h_{1x,i}} \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]}{\mathbb{E}_{\gamma_i}\left[\int \mathrm{e}^{\beta n \phi_{x,i}^{(\beta)}} D\eta_{x,i} \right]} \right)^2, \tag{50}$$

J. Stat. Mech. (2020) 093402

$$\hat{z}_{1,\mu} = \frac{\mathbb{E}_{c_\mu}\left[\int \frac{\partial \phi^{(\beta)}_{z,\mu}}{\partial h_{1z,\mu}} \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]}{\mathbb{E}_{c_\mu}\left[\int \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]}, \tag{51}$$

$$\chi_{1z,\mu} = \frac{\mathbb{E}_{c_\mu}\left[\int \frac{\partial^2 \phi^{(\beta)}_{z,\mu}}{\partial h^2_{1z,\mu}} \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]}{\mathbb{E}_{c_\mu}\left[\int \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]}, \tag{52}$$

$$v_{1z,\mu} = \frac{\mathbb{E}_{c_\mu}\left[\int \left(\frac{\partial \phi^{(\beta)}_{z,\mu}}{\partial h_{1z,\mu}}\right)^2 \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]}{\mathbb{E}_{c_\mu}\left[\int \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]} - \left(\frac{\mathbb{E}_{c_\mu}\left[\int \frac{\partial \phi^{(\beta)}_{z,\mu}}{\partial h_{1z,\mu}} \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]}{\mathbb{E}_{c_\mu}\left[\int \mathrm{e}^{\beta n \phi^{(\beta)}_{z,\mu}} D\eta_{z,\mu}\right]}\right)^2, \tag{53}$$

Next, for the approximate density $p_2^{(\beta)}$, we obtain

$$\int x_{s,i} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \hat{x}_{2,i}, \tag{54}$$

$$\int x_{s,i} x_{t,i} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = v_{2x,i} + \hat{x}^2_{2,i}, \quad s \neq t, \tag{55}$$

$$\int x^2_{s,i} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \frac{\chi_{2x,i}}{\beta} + v_{2x,i} + \hat{x}^2_{2,i}, \tag{56}$$

$$\int z_{s,\mu} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \hat{z}_{2,\mu}, \tag{57}$$

$$\int z_{s,\mu} z_{t,\mu} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = v_{2z,\mu} + \hat{z}^2_{2,\mu}, \quad s \neq t, \tag{58}$$

$$\int z^2_{s,\mu} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} = \frac{\chi_{2z,\mu}}{\beta} + v_{2z,\mu} + \hat{z}^2_{2,\mu}, \tag{59}$$

where

$$\hat{\boldsymbol{x}}_2 = \left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z}) A\right)^{-1} \left(\boldsymbol{h}_{2x} + A^\top \boldsymbol{h}_{2z}\right), \tag{60}$$

$$\chi_{2x,i} = \left[\left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z}) A\right)^{-1}\right]_{ii}, \tag{61}$$

$$v_{2x,i} = \left[\left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z}) A\right)^{-1} \left(\mathrm{Diagm}(\hat{\boldsymbol{v}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{v}}_{2z}) A\right) \times \left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z}) A\right)^{-1}\right]_{ii}, \tag{62}$$

$$\hat{\boldsymbol{z}}_2 = A^\top \hat{\boldsymbol{x}}_2, \tag{63}$$

$$\chi_{2z,\mu} = \left[A\left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z}) A\right)^{-1} A^\top\right]_{\mu\mu}, \tag{64}$$

J. Stat. Mech. (2020) 093402

$$v_{2z,\mu} = \Bigg[A\Big(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z})A\Big)^{-1} \Big(\mathrm{Diagm}(\hat{\boldsymbol{v}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{v}}_{2z})A\Big) \times \Big(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z})A\Big)^{-1} A^\top \Bigg]_{\mu\mu}. \tag{65}$$

Finally, the moment-matching conditions are written as

$$h_{2x,i} = \frac{\hat{x}_{1,i}}{\chi_{1x,i}} - h_{1x,i} + \mathcal{O}(n), \qquad h_{1x,i} = \frac{\hat{x}_{2,i}}{\chi_{2x,i}} - h_{2x,i} + \mathcal{O}(n), \tag{66}$$

$$\hat{Q}_{2x,i} = \frac{1}{\chi_{1x,i}} - \hat{Q}_{1x,i} + \mathcal{O}(n), \qquad \hat{Q}_{1x,i} = \frac{1}{\chi_{2x,i}} - \hat{Q}_{2x,i} + \mathcal{O}(n), \tag{67}$$

$$\hat{v}_{2x,i} = \frac{v_{1x,i}}{\chi^2_{1x,i}} - \hat{v}_{1x,i} + \mathcal{O}(n), \qquad \hat{v}_{1x,i} = \frac{v_{2x,i}}{\chi^2_{2x,i}} - \hat{v}_{2x,i} + \mathcal{O}(n), \tag{68}$$

$$h_{2z,\mu} = \frac{\hat{z}_{1,\mu}}{\chi_{1z,\mu}} - h_{1z,\mu} + \mathcal{O}(n), \qquad h_{1z,\mu} = \frac{\hat{z}_{2,\mu}}{\chi_{2z,\mu}} - h_{2z,\mu} + \mathcal{O}(n), \tag{69}$$

$$\hat{Q}_{2z,\mu} = \frac{1}{\chi_{1z,\mu}} - \hat{Q}_{1z,\mu} + \mathcal{O}(n), \qquad \hat{Q}_{1z,\mu} = \frac{1}{\chi_{2z,\mu}} - \hat{Q}_{2z,\mu} + \mathcal{O}(n), \tag{70}$$

$$\hat{v}_{2z,\mu} = \frac{v_{1z,\mu}}{\chi^2_{1z,\mu}} - \hat{v}_{1z,\mu} + \mathcal{O}(n), \qquad \hat{v}_{1z,\mu} = \frac{v_{2z,\mu}}{\chi^2_{2z,\mu}} - \hat{v}_{2z,\mu} + \mathcal{O}(n), \tag{71}$$

In all of the above expressions, the indices i and μ run as $i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, M$, respectively. $\boldsymbol{\chi}_x$ and $\boldsymbol{\chi}_z$ are termed susceptibility. $\boldsymbol{v}_x$ and $\boldsymbol{v}_z$ are termed variance. Clearly, these equations can be easily extrapolated as $n \to 0$.

Inserting the limiting form of these quantities at $n \to 0, \beta \to \infty$ into the algorithm 1, we obtain rVAMP in algorithm 2. There, $\boldsymbol{g}_{1x}, \boldsymbol{g}_{1z}, \boldsymbol{g}'_{1x}$ and $\boldsymbol{g}'_{1z}$ are denoising functions and their derivatives. These are defined as follows:

$$\boldsymbol{g}_{1x}(\boldsymbol{h}_{1x}, \hat{\boldsymbol{Q}}_{1x}, \hat{\boldsymbol{v}}_{1x}\,; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) = [g_{1x}(h_{1x,i}, \hat{Q}_{1x,i}, \hat{v}_{1x,i}\,; \gamma_i, \eta_{x,i})]_{1\leqslant i\leqslant N}, \tag{72}$$

$$\boldsymbol{g}'_{1x}(\boldsymbol{h}_{1x}, \hat{\boldsymbol{Q}}_{1x}, \hat{\boldsymbol{v}}_{1x}\,; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) = [g'_{1x}(h_{1x,i}, \hat{Q}_{1x,i}, \hat{v}_{1x,i}\,; \gamma_i, \eta_{x,i})]_{1\leqslant i\leqslant N}, \tag{73}$$

$$\boldsymbol{g}_{1z}(\boldsymbol{h}_{1z}, \hat{\boldsymbol{Q}}_{1z}, \hat{\boldsymbol{v}}_{1z}\,; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) = [g_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}\,; c_\mu, \eta_{z,\mu}, y_\mu)]_{1\leqslant \mu\leqslant M}, \tag{74}$$

$$\boldsymbol{g}'_{1z}(\boldsymbol{h}_{1z}, \hat{\boldsymbol{Q}}_{1z}, \hat{\boldsymbol{v}}_{1z}\,; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) = [g'_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}\,; c_\mu, \eta_{z,\mu}, y_\mu)]_{1\leqslant \mu\leqslant M}, \tag{75}$$

where

$$g_{1x}(h_{1x,i}, \hat{Q}_{1x,i}, \hat{v}_{1x,i}\,; \gamma_i, \eta_{x,i}) = \frac{h_{1x,i} + \sqrt{\hat{v}}_{1x,i}\eta_{x,i} - \gamma_i\,\mathrm{sign}(h_{1x,i} + \sqrt{\hat{v}_{1x,i}}\eta_{x,i})}{\hat{Q}_{1x,i}} \times \mathbb{1}\left(\left|h_{1x,i} + \sqrt{\hat{v}_{1x,i}}\eta_{x,i}\right| > \gamma_i\right), \tag{76}$$

$$g'_{1x}(h_{1x,i}, \hat{Q}_{1x,i}, \hat{v}_{1x,i}\,; \gamma_i, \eta_{x,i}) = \frac{1}{\hat{Q}_{1x,i}} \mathbb{1}\left(\left|h_{1x,i} + \sqrt{\hat{v}_{1x,i}}\eta_{x,i}\right| > \gamma_i\right), \tag{77}$$

J. Stat. Mech. (2020) 093402

$$g_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}\,; c_\mu, \eta_{z,\mu}, y_\mu) = \arg\max_{z\in\mathbb{R}} \left[-\frac{\hat{Q}_{1z,\mu}}{2} z^2 + \Big(h_{1z,\mu} + \sqrt{\hat{v}_{1z,\mu}}\eta_{z,\mu} \Big) z + c_\mu \log p_{y|z}(y_\mu|z) \right], \tag{78}$$

$$g'_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}\,; c_\mu, \eta_{z,\mu}, y_\mu) = \frac{\partial g_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}\,; c_\mu, \eta_{z,\mu}, y_\mu)}{\partial h_{1z,\mu}}. \tag{79}$$

If the likelihood $p_{y|z}$ is differentiable with respect to z, g'_{1z} can be written as

$$g'_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}\,; c_\mu, \eta_{z,\mu}, y_\mu) = \left[\hat{Q}_{1z,\mu} - c_\mu \frac{\partial^2 \log\, p_{y|z}(y_\mu|z)}{\partial z^2}\bigg|_{z=g_{1z}} \right]^{-1}. \tag{80}$$

Because the averages with respect to $\boldsymbol{c}$ and $\boldsymbol{\gamma}$ are incorporated in line 4–9 of the algorithm 2 as the averages with respect to one-dimensional random variables, rVAMP does not require refitting.

Although the two approximate densities have the same first and second moments at a fixed point, these two densities have different characteristics. For higher-order marginal moments, we expect that $p_1^{(\beta)}$ is more precise than $p_2^{(\beta)}$ because it accurately includes the non-Gaussian factors. Similarly, $p_2^{(\beta)}$ is argued to have more accurate off-diagonal moments because it includes the interaction term correctly [22, 30]. Thus, these two distributions should be used depending on the objective. Because we are interested in the distribution of the marginal moment (17), here we use $p_1^{(\beta)}$ to compute $\Pi_i(\gamma_0)$.

3.5. Calculation of the selection probability

Using the expression

$$p_1^{(\beta)}(\{\boldsymbol{x}_s\}, \{\boldsymbol{z}_s\}) \propto \prod_{i=1}^{N} \mathbb{E}_{\gamma_i} \left[\int \prod_{s=1}^{n} \mathrm{e}^{-\frac{\beta\hat{Q}_{1x,i}}{2} x_{s,i}^2 + \beta(h_{1x,i} + \sqrt{\hat{v}_{1x,i}}\eta_{x,i})x_{s,i} - \beta\gamma_i|x_{s,i}|} D\eta_{x,i} \right] \times \prod_{\mu=1}^{M} \mathbb{E}_{c_\mu} \left[\int \prod_{s=1}^{n} \mathrm{e}^{-\frac{\beta\hat{Q}_{1z,\mu}}{2} z_{s,\mu}^2 + \beta(h_{1z,\mu} + \sqrt{\hat{v}_{1z,\mu}} z_{s,\mu})} p_{y|z}(y_\mu|z_{s,\mu})^{\beta c_\mu} D\eta_{z,\mu} \right], \tag{81}$$

we obtain the following form of the rth moment:

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\hat{x}_i^r\right] = \mathbb{E}_{\gamma_i}\left[\int g_{1x}(h_{1x,i}, \hat{Q}_{1x,i}, \hat{v}_{1x,i}\,; \gamma_i, \eta_{x,i})^r D\eta_{x,i} \right]. \tag{82}$$

To understand the meaning of $\eta_{x,i}$, suppose that we omit to take the expectations of $(\boldsymbol{c}, \boldsymbol{\gamma})$ in lines 4–9 of algorithm 2 and to run rVAMP for a fixed set of $(\boldsymbol{c}, \boldsymbol{\gamma})$. Then, one can show that $v_{1x,i} = v_{2x,i} = \hat{v}_{1x,i} = \hat{v}_{2x,i} = 0$ and $v_{1z,\mu} = v_{2z,\mu} = \hat{v}_{1z,\mu} = \hat{v}_{2z,\mu} = 0$ yield the fixed point condition for these variables, and the rest part of the algorithm exactly coincides with the VAMP algorithm for LASSO without a resampling [8]. Thus, we expect that $\sqrt{\hat{v}_{1x,i}}\eta_{x,i}$ behave as random variables that approximately reflect the effect

J. Stat. Mech. (2020) 093402

Algorithm 2. rVAMP.

Require: Denoising functions g_{1x}, g_{1z} from (72) and (74), the features $A \in \mathbb{R}^{M\times N}$, the response variable $\boldsymbol{y} \in \mathcal{Y}^M$, the convergence criterion $\epsilon_{\rm tol}$, the maximum number of iterations $T_{\rm iter}$.

Select initial $\boldsymbol{h}_{1x}^{(1)} \in \mathbb{R}^N, \boldsymbol{h}_{1z}^{(1)} \in \mathbb{R}^M$, $\hat{\boldsymbol{Q}}_{1x}^{(1)}, \hat{\boldsymbol{v}}_{1x}^{(1)} \in [0,\infty)^N$, and $\hat{\boldsymbol{Q}}_{1z}^{(1)}, \hat{\boldsymbol{v}}_{1z}^{(1)} \in [0,\infty)^M$.
for $t = 1, 2, \ldots, T_{\rm iter}$ **do**
  // Factorized part
  $\hat{\boldsymbol{x}}_1^{(t)} = \mathbb{E}_{\boldsymbol{\gamma}}[\int \boldsymbol{g}_{1x}(\boldsymbol{h}_{1x}^{(t)}, \hat{\boldsymbol{Q}}_{1x}^{(t)}, \hat{\boldsymbol{v}}_{1x}^{(t)}; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) D\boldsymbol{\eta}_x]$
  $\boldsymbol{\chi}_{1x}^{(t)} = \mathbb{E}_{\boldsymbol{\gamma}}[\int \boldsymbol{g}'_{1x}(\boldsymbol{h}_{1x}^{(t)}, \hat{\boldsymbol{Q}}_{1x}^{(t)}, \hat{\boldsymbol{v}}_{1x}^{(t)}; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) D\boldsymbol{\eta}_x]$
  $\boldsymbol{v}_{1x}^{(t)} = \mathbb{E}_{\boldsymbol{\gamma}}[\int \boldsymbol{g}_{1x}^2(\boldsymbol{h}_{1x}^{(t)}, \hat{\boldsymbol{Q}}_{1x}^{(t)}, \hat{\boldsymbol{v}}_{1x}^{(t)}; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) D\boldsymbol{\eta}_x] - (\hat{\boldsymbol{x}}_1^{(t)})^2$
  $\hat{\boldsymbol{z}}_1^{(t)} = \mathbb{E}_{\boldsymbol{c}}[\int \boldsymbol{g}_{1z}(\boldsymbol{h}_{1z}^{(t)}, \hat{\boldsymbol{Q}}_{1z}^{(t)}, \hat{\boldsymbol{v}}_{1z}^{(t)}; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) D\boldsymbol{\eta}_z]$
  $\boldsymbol{\chi}_{1z}^{(t)} = \mathbb{E}_{\boldsymbol{c}}[\int \boldsymbol{g}'_{1z}(\boldsymbol{h}_{1z}^{(t)}, \hat{\boldsymbol{Q}}_{1z}^{(t)}, \hat{\boldsymbol{v}}_{1z}^{(t)}; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) D\boldsymbol{\eta}_z]$
  $\boldsymbol{v}_{1z}^{(t)} = \mathbb{E}_{\boldsymbol{c}}[\int \boldsymbol{g}_{1z}^2(\boldsymbol{h}_{1z}^{(t)}, \hat{\boldsymbol{Q}}_{1z}^{(t)}, \hat{\boldsymbol{v}}_{1z}^{(t)}; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) D\boldsymbol{\eta}_z] - (\hat{\boldsymbol{z}}_1^{(t)})^2$
  // Moment-matching $(1 \to 2)$
  $\boldsymbol{h}_{2x}^{(t)} = \hat{\boldsymbol{x}}_1^{(t)}/\boldsymbol{\chi}_{1x}^{(t)} - \boldsymbol{h}_{1x}^{(t)}, \quad \hat{\boldsymbol{Q}}_{2x}^{(t)} = \left(\boldsymbol{\chi}_{1x}^{(t)}\right)^{-1} - \hat{\boldsymbol{Q}}_{1x}^{(t)}, \quad \hat{\boldsymbol{v}}_{2x}^{(t)} = \boldsymbol{v}_{1x}^{(t)}/\left(\boldsymbol{\chi}_{1x}^{(t)}\right)^2 - \hat{\boldsymbol{v}}_{1x}^{(t)}$
  $\boldsymbol{h}_{2z}^{(t)} = \hat{\boldsymbol{z}}_1^{(t)}/\boldsymbol{\chi}_{1z}^{(t)} - \boldsymbol{h}_{1z}^{(t)}, \quad \hat{\boldsymbol{Q}}_{2z}^{(t)} = \left(\boldsymbol{\chi}_{1z}^{(t)}\right)^{-1} - \hat{\boldsymbol{Q}}_{1z}^{(t)}, \quad \hat{\boldsymbol{v}}_{2z}^{(t)} = \boldsymbol{v}_{1z}^{(t)}/\left(\boldsymbol{\chi}_{1z}^{(t)}\right)^2 - \hat{\boldsymbol{v}}_{1z}^{(t)}$
  // Gaussian part
  $X = \left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}^{(t)}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z}) A\right)^{-1}$
  $\hat{\boldsymbol{x}}_2^{(t)} = X(\boldsymbol{h}_{2x}^{(t)} + A^\top \boldsymbol{h}_{2z}^{(t)}), \qquad \hat{\boldsymbol{z}}_2^{(t)} = A\hat{\boldsymbol{x}}_2^{(t)}$
  $\boldsymbol{\chi}_{2x}^{(t)} = \mathrm{diag}[X], \qquad \boldsymbol{\chi}_{2z}^{(t)} = \mathrm{diag}[AXA^\top]$
  $\boldsymbol{v}_{2x}^{(t)} = \mathrm{diag}\left[X\left(\mathrm{Diagm}(\hat{\boldsymbol{v}}_{2x}^{(t)}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{v}}_{2z}^{(t)}) A\right) X\right]$
  $\boldsymbol{v}_{2z}^{(t)} = \mathrm{Diagm}\left[AX\left(\mathrm{Diagm}(\hat{\boldsymbol{v}}_{2x}^{(t)}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{v}}_{2z}^{(t)}) A\right) XA^\top\right]$
  // Moment-matching $(2 \to 1)$
  $\boldsymbol{h}_{1x}^{(t+1)} = \hat{\boldsymbol{x}}_2^{(t)}/\boldsymbol{\chi}_{2x}^{(t)} - \boldsymbol{h}_{2x}^{(t)}, \quad \hat{\boldsymbol{Q}}_{1x}^{(t+1)} = \left(\boldsymbol{\chi}_{2x}^{(t)}\right)^{-1} - \hat{\boldsymbol{Q}}_{2x}^{(t)}, \quad \hat{\boldsymbol{v}}_{1x}^{(t+1)} = \boldsymbol{v}_{2x}^{(t)}/\left(\boldsymbol{\chi}_{2x}^{(t)}\right)^2 - \hat{\boldsymbol{v}}_{2x}^{(t)}$
  $\boldsymbol{h}_{1z}^{(t+1)} = \hat{\boldsymbol{z}}_2^{(t)}/\boldsymbol{\chi}_{2z}^{(t)} - \boldsymbol{h}_{2z}^{(t)}, \quad \hat{\boldsymbol{Q}}_{1z}^{(t+1)} = \left(\boldsymbol{\chi}_{2z}^{(t)}\right)^{-1} - \hat{\boldsymbol{Q}}_{2z}^{(t)}, \quad \hat{\boldsymbol{v}}_{1z}^{(t+1)} = \boldsymbol{v}_{2z}^{(t)}/\left(\boldsymbol{\chi}_{2z}^{(t)}\right)^2 - \hat{\boldsymbol{v}}_{2z}^{(t)}$
  if $\max\{\|\hat{\boldsymbol{x}}_1^{(t)} - \hat{\boldsymbol{x}}_2^{(t)}\|_2^2/N, \|\hat{\boldsymbol{z}}_1^{(t)} - \hat{\boldsymbol{z}}_2^{(t)}\|_2^2/M\} < \epsilon_{\rm tol}$ **then**
    $t \leftarrow T_{\rm iter}$
    break
  end if
end for
return $\boldsymbol{h}_{1x}^{(T_{\rm iter})}, \hat{\boldsymbol{Q}}_{1x}^{(T_{\rm iter})}, \hat{\boldsymbol{v}}_{1x}^{(T_{\rm iter})}$

of taking average of $\boldsymbol{c}$. This consideration and the expression of the rth moment in (82) yield the following form of the distribution function $p(m_i)$:

$$p(m_i) \simeq \mathbb{E}_{\gamma_i}\left[\int \mathbb{1}\left(m_i - g_{1x}(h_{1x,i}, \hat{Q}_{1x,i}, \hat{v}_{1x,i}; \gamma_i, \eta_{x,i})\right) D\eta_{x,i}\right]. \tag{83}$$

Because $g_{1x}(h_{1x,i}, \hat{Q}_{1x,i}, \hat{v}_{1x,i}\,; \gamma_i, \eta_{x,i})$ is non-zero iff $\mathbb{1}(|h_{1x,i} + \sqrt{\hat{v}_{1x,i}}\eta_{x,i}| > \gamma_i)$ is satisfied, rVAMP yields the following expression for the selection probability Π_i:

$$\Pi_i(\gamma_0) \simeq \mathbb{E}_{\gamma_i}\left[\int \mathbb{1}\left(\left|h_{1x,i} + \sqrt{\hat{v}_{1x,i}}\eta_{x,i}\right| > \gamma_i\right) D\eta_{x,i}\right], \tag{84}$$

which is easy to calculate.

3.6. Implementation details

For practical implementation, we find that it is helpful to make several small modifications to rVAMP of the algorithm 2. In this subsection, we discuss these minor modifications.

First we address the computational complexity regarding the matrix inversion. Although rVAMP requires the matrix inversion in line 14, this computational cost is reduced to $\mathcal{O}(M^3)$ from $\mathcal{O}(N^3)$ using the Woodbury identity [31]:

$$\begin{aligned}\left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z})A\right)^{-1} &= \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}^{-1}) \\ &- \mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}^{-1})A^\top\left(\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2z}^{-1}) + A\,\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}^{-1})A^\top\right)^{-1} A\,\mathrm{Diagm}(\hat{\boldsymbol{Q}}_{2x}^{-1}).\end{aligned} \tag{85}$$

Because in high-dimensional statistics, the number of the samples in the data is often one or several orders of magnitude smaller than the number of the parameters, the computational cost is drastically reduced using this identity.

Second, for a real-world dataset with a small number of samples, VAMP trajectories can show large oscillations, which lead to poor convergence. In such cases, introducing a small amount of damping factor $\eta_{\mathrm{d}} \in (0, 1]$ can improve the convergence of the algorithm. We suggest replacing line 20 and 21 with the damped versions:

$$\boldsymbol{h}_{1x}^{(t+1)} = \eta_{\mathrm{d}}\left(\frac{\hat{\boldsymbol{x}}_2^{(t)}}{\boldsymbol{\chi}_{2x}^{(t)}} - \boldsymbol{h}_{2x}^{(t)}\right) + (1-\eta_{\mathrm{d}})\boldsymbol{h}_{1x}^{(t)}, \tag{86}$$

$$\hat{\boldsymbol{Q}}_{1x}^{(t+1)} = \eta_{\mathrm{d}}\left(\frac{\mathbf{1}_N}{\boldsymbol{\chi}_{2x}^{(t)}} - \hat{\boldsymbol{Q}}_{2x}^{(t)}\right) + (1-\eta_{\mathrm{d}})\,\hat{\boldsymbol{Q}}_{1x}^{(t)}, \tag{87}$$

$$\hat{\boldsymbol{v}}_{1x}^{(t+1)} = \eta_{\mathrm{d}}\left(\frac{\boldsymbol{v}_{2x}^{(t)}}{\left(\boldsymbol{\chi}_{2x}^{(t)}\right)^2} - \hat{\boldsymbol{v}}_{2x}^{(t)}\right) + (1-\eta_{\mathrm{d}})\hat{\boldsymbol{v}}_{1x}^{(t)}, \tag{88}$$

$$\boldsymbol{h}_{1z}^{(t+1)} = \eta_{\mathrm{d}}\left(\frac{\hat{\boldsymbol{z}}_2^{(t)}}{\boldsymbol{\chi}_{2z}^{(t)}} - \boldsymbol{h}_{2z}^{(t)}\right) + (1-\eta_{\mathrm{d}})\boldsymbol{h}_{1z}^{(t)}, \tag{89}$$

$$\hat{\boldsymbol{Q}}_{1z}^{(t+1)} = \eta_{\mathrm{d}}\left(\frac{\mathbf{1}_M}{\boldsymbol{\chi}_{2z}^{(t)}} - \hat{\boldsymbol{Q}}_{2z}^{(t)}\right) + (1-\eta_{\mathrm{d}})\,\hat{\boldsymbol{Q}}_{1z}^{(t)}, \tag{90}$$

J. Stat. Mech. (2020) 093402

$$\hat{\boldsymbol{v}}_{1z}^{(t+1)} = \eta_{\mathrm{d}} \left(\frac{\boldsymbol{v}_{2z}^{(t)}}{\left(\boldsymbol{\chi}_{2z}^{(t)}\right)^2} - \hat{\boldsymbol{v}}_{2z}^{(t)} \right) + (1-\eta_{\mathrm{d}})\hat{\boldsymbol{v}}_{1z}^{(t)}. \tag{91}$$

Third, GLMs may require including an intercept term z_0 so that $y_\mu \sim p_{y|z}(y_\mu|z_0 + \boldsymbol{a}_\mu^\top \boldsymbol{x}_0)$. To incorporate the intercept term, we add an extra column in the feature matrix so that $A_{0,\mu} = 1, \mu = 1, 2, \ldots, M$, and for this component we do not require any regularization term.

The last point regards how to obtain the selection probability for various values of the regularization strength γ_0. In practice, we are often interested in finding the selection probability not only for a single fixed γ_0, but also for the various regularization parameters γ_0 (as in figure 1). A reasonable approach is to begin with the largest γ_0. Then, we decrease γ_0 by a small amount and run rVAMP until convergence. Decreasing γ_0 again and using previous parameters at the fixed point as the initial conditions (*warm start*), we then run rVAMP until convergence. Using this method, we can efficiently compute the selection probabilities over a grid of γ_0.

4. Macroscopic analysis

The salient feature of the VAMP algorithms is that we can macroscopically analyze their convergence dynamics in a large system limit under specific assumptions on the distributions of the set of feature vectors. The derived dynamics are termed state evolution (SE). In this section, we derive SE for self-averaging rVAMP (SA rVAMP), which would describe the converging dynamics of rVAMP approximately. We also show that its fixed point is consistent with the replica symmetric solution obtained by the replica method, which is believed to be exact in the large system limit under appropriate conditions. Although the procedure of the replica method has not been justified mathematically yet, many studies have rigorously validated its conjectures in the last few decades, especially in Bayes optimal settings [8, 32–34], and more recently in model-mismatched cases [35].

4.1. Setup for the macroscopic analysis

For the theoretical analysis, we assume the actual data generation process as follows. First, the true parameter vector $\boldsymbol{x}_0$ and the response variables are generated as

$$x_{0,i} \sim q_{x_0}(x_{0,i}), \quad i = 1, 2, \ldots, N, \tag{92}$$

$$y_\mu \sim q_{y|z}(y_\mu|\boldsymbol{a}_\mu^\top \boldsymbol{x}_0), \quad \mu = 1, 2, \ldots, M. \tag{93}$$

Generally, the model used for the fitting and the actual generation model may be different $p_{y|z} \neq q_{y|z}$ or $\mathrm{e}^{-\gamma|x|} \neq q_{x_0}$. Additionally, we assume that the feature matrix A is drawn from the rotation-invariant random matrix ensembles, i.e. for the singular value decomposition $A = USV^\top, U \in \mathbb{R}^{M\times M}, S \in \mathbb{R}^{M\times N}, V \in \mathbb{R}^{N\times N}$, we assume that U and V are drawn from uniform distributions over $M \times M$ and $N \times N$ orthogonal matrices.

J. Stat. Mech. (2020) 093402

We are interested in the large system limit where both of the numbers of data points and parameters diverge as $M, N \to \infty$ keeping the ratio $\alpha \equiv M/N \in (0, \infty)$. Because U and V are drawn independently from uniform distributions over $M \times M$ and $N \times N$ orthogonal matrices, for vectors $\boldsymbol{\omega} \in \mathbb{R}^N$ and $\boldsymbol{\phi} \in \mathbb{R}^M$, we expect that the empirical distributions of $V^\top \boldsymbol{\omega}$ and $U^\top \boldsymbol{\phi}$ converge to Gaussians with mean zero and variance $\|\boldsymbol{\omega}\|_2^2/N$ and $\|\boldsymbol{\phi}\|_2^2/M$ in this limit, respectively.

4.2. Self-averaging rVAMP

Our first interest is the convergence dynamics of rVAMP. Unfortunately, directly investigating the dynamics of rVAMP is difficult because the time evolution of the empirical distributions of $\boldsymbol{h}_{1x}, \boldsymbol{h}_{1z}, \boldsymbol{h}_{2x}, \boldsymbol{h}_{2z}$ may not be described by a small number of statistics, although the dynamical-functional theory [26, 36–38] might give some insights for the raw rVAMP. To detour this difficulty approximately, we consider SA rVAMP, which eliminates the site dependence of the natural parameters in the approximate densities:

$$\hat{Q}^{(t)}_{1x,i} = \hat{Q}^{(t)}_{1x}, \qquad \hat{Q}^{(t)}_{2x,i} = \hat{Q}^{(t)}_{2x}, \tag{94}$$

$$\hat{v}^{(t)}_{1x,i} = \hat{v}^{(t)}_{1x}, \qquad \hat{v}^{(t)}_{2x,i} = \hat{v}^{(t)}_{2x}, \tag{95}$$

$$\hat{Q}^{(t)}_{1z,\mu} = \hat{Q}^{(t)}_{1z}, \qquad \hat{Q}^{(t)}_{2z,\mu} = \hat{Q}^{(t)}_{2z}, \tag{96}$$

$$\hat{v}^{(t)}_{1z,\mu} = \hat{v}^{(t)}_{1z}, \qquad \hat{v}^{(t)}_{2z,\mu} = \hat{v}^{(t)}_{2z}. \tag{97}$$

Eliminating the site dependence replaces the component-wise moment-matching conditions in (30) and (31) with the macroscopic moment-matching conditions:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \int x_{s,i} x_{t,i} p_1^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} &= \frac{1}{N}\sum_{i=1}^{N} \int x_{s,i} x_{t,i} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} \\ &= \frac{1}{N}\sum_{i=1}^{N} \int x_{s,i} x_{t,i} \tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z},\end{aligned} \tag{98}$$

$$\begin{aligned}\frac{1}{M}\sum_{\mu=1}^{M} \int z_{s,\mu} z_{t,\mu} p_1^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} &= \frac{1}{M}\sum_{\mu=1}^{M} \int z_{s,\mu} z_{t,\mu} p_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z} \\ &= \frac{1}{M}\sum_{\mu=1}^{M} \int z_{s,\mu} z_{t,\mu} \tilde{p}_1^{(\beta)} \tilde{p}_2^{(\beta)} \mathrm{d}^n \boldsymbol{x} \mathrm{d}^n \boldsymbol{z}.\end{aligned} \tag{99}$$

These modifications yield SA rVAMP described in algorithm 3. We will use it in the following analysis.

4.3. State evolution

To derive the SE of SA rVAMP heuristically, we make the following assumptions following the literature [23].

J. Stat. Mech. (2020) 093402

Algorithm 3. Self averaging rVAMP.

Require: Denoising functions g_{1x}, g_{1z} from (72) and (74), the features $A \in \mathbb{R}^{M\times N}$, the response variable $\boldsymbol{y} \in \mathbb{R}^M$, the convergence criterion $\epsilon_{\rm tol}$, and the maximum number of iterations $T_{\rm iter}$.

Select initial $\boldsymbol{h}_{1x}^{(1)} \in \mathbb{R}^N, \boldsymbol{h}_{1z}^{(1)} \in \mathbb{R}^M$, $\hat{Q}_{1x}^{(1)}, \hat{v}_{1x}^{(1)}, \hat{Q}_{1z}^{(1)}$, and $\hat{v}_{1z}^{(1)} \in [0,\infty)$.
for $t = 1, 2, \ldots, T_{\rm iter}$ **do**
  // Factorized part
  $\hat{\boldsymbol{x}}_1^{(t)} = \mathbb{E}_\gamma[\int \boldsymbol{g}_{1x}(\boldsymbol{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}\mathbf{1}_N, \hat{v}_{1x}^{(t)}\mathbf{1}_N; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) D\boldsymbol{\eta}_x]$
  $\chi_{1x}^{(t)} = \langle \mathbb{E}_\gamma[\int \boldsymbol{g}'_{1x}(\boldsymbol{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}\mathbf{1}_N, \hat{v}_{1x}^{(t)}\mathbf{1}_N; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) D\boldsymbol{\eta}_x]\rangle$
  $v_{1x}^{(t)} = \langle \mathbb{E}_\gamma[\int \boldsymbol{g}_{1x}^2(\boldsymbol{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}\mathbf{1}_N, \hat{v}_{1x}^{(t)}\mathbf{1}_N; \boldsymbol{\gamma}, \boldsymbol{\eta}_x) D\boldsymbol{\eta}_x] - (\hat{\boldsymbol{x}}_1^{(t)})^2\rangle$
  $\hat{\boldsymbol{z}}_1^{(t)} = \mathbb{E}_{\boldsymbol{c}}[\int \boldsymbol{g}_{1z}(\boldsymbol{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}\mathbf{1}_M, \hat{v}_{1z}^{(t)}\mathbf{1}_M; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) D\boldsymbol{\eta}_z]$
  $\chi_{1z}^{(t)} = \langle \mathbb{E}_{\boldsymbol{c}}[\int \boldsymbol{g}'_{1z}(\boldsymbol{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}\mathbf{1}_M, \hat{v}_{1z}^{(t)}\mathbf{1}_M; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) D\boldsymbol{\eta}_z]\rangle$
  $v_{1z}^{(t)} = \langle \mathbb{E}_{\boldsymbol{c}}[\int \boldsymbol{g}_{1z}^2(\boldsymbol{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}\mathbf{1}_M, \hat{v}_{1z}^{(t)}\mathbf{1}_M; \boldsymbol{c}, \boldsymbol{\eta}_z, \boldsymbol{y}) D\boldsymbol{\eta}_z] - (\hat{\boldsymbol{z}}_1^{(t)})^2\rangle$
  // Moment-matching $(1 \to 2)$
  $\boldsymbol{h}_{2x}^{(t)} = \hat{\boldsymbol{x}}_1^{(t)}/(\chi_{1x}^{(t)}\mathbf{1}_N) - \boldsymbol{h}_{1x}^{(t)}, \quad \hat{Q}_{2x}^{(t)} = \left(\chi_{1x}^{(t)}\right)^{-1} - \hat{Q}_{1x}^{(t)}, \quad \hat{v}_{2x}^{(t)} = v_{1x}^{(t)}/\left(\chi_{1x}^{(t)}\right)^2 - \hat{v}_{1x}^{(t)}$
  $\boldsymbol{h}_{2z}^{(t)} = \hat{\boldsymbol{z}}_1^{(t)}/(\chi_{1z}^{(t)}\mathbf{1}_M) - \boldsymbol{h}_{1z}^{(t)}, \quad \hat{Q}_{2z}^{(t)} = \left(\chi_{1z}^{(t)}\right)^{-1} - \hat{Q}_{1z}^{(t)}, \quad \hat{v}_{2z}^{(t)} = v_{1z}^{(t)}/\left(\chi_{1z}^{(t)}\right)^2 - \hat{v}_{1z}^{(t)}$
  // Gaussian part
  $X = \left(\hat{Q}_{2x}^{(t)} I_N + \hat{Q}_{2z}^{(t)} A^\top A\right)^{-1}$
  $\hat{\boldsymbol{x}}_2^{(t)} = X(\boldsymbol{h}_{2x} + A^\top \boldsymbol{h}_{2z}), \quad \hat{\boldsymbol{z}}_2^{(t)} = A\hat{\boldsymbol{x}}_2^{(t)}$
  $\chi_{2x}^{(t)} = N^{-1}\,\mathrm{Tr}[X], \quad \chi_{2z}^{(t)} = M^{-1}\,\mathrm{Tr}[AXA^\top]$
  $\boldsymbol{v}_{2x}^{(t)} = N^{-1}\,\mathrm{Tr}\left[X\left(\mathrm{Diagm}(\hat{\boldsymbol{v}}_{2x}^{(t)}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{v}}_{2z}^{(t)})A\right)X\right]$
  $\boldsymbol{v}_{2z}^{(t)} = M^{-1}\,\mathrm{Tr}\left[AX\left(\mathrm{Diagm}(\hat{\boldsymbol{v}}_{2x}^{(t)}) + A^\top \mathrm{Diagm}(\hat{\boldsymbol{v}}_{2z}^{(t)})A\right)XA^\top\right]$
  // Moment-matching $(2 \to 1)$
  $\boldsymbol{h}_{1x}^{(t+1)} = \hat{\boldsymbol{x}}_2^{(t)}/(\chi_{2x}^{(t)}\mathbf{1}_N) - \boldsymbol{h}_{2x}^{(t)}, \hat{Q}_{1x}^{(t+1)} = \left(\chi_{2x}^{(t)}\right)^{-1} - \hat{Q}_{2x}^{(t)}, \hat{v}_{1x}^{(t+1)} = v_{2x}^{(t)}/\left(\chi_{2x}^{(t)}\right)^2 - \hat{v}_{2x}^{(t)}$
  $\boldsymbol{h}_{1z}^{(t+1)} = \hat{\boldsymbol{z}}_2^{(t)}/(\chi_{2z}^{(t)}\mathbf{1}_M) - \boldsymbol{h}_{2z}^{(t)}, \hat{Q}_{1z}^{(t+1)} = \left(\chi_{2z}^{(t)}\right)^{-1} - \hat{Q}_{2z}^{(t)}, \hat{v}_{1z}^{(t+1)} = v_{2z}^{(t)}/\left(\chi_{2z}^{(t)}\right)^2 - \hat{v}_{2z}^{(t)}$
  if $\max\{\|\hat{\boldsymbol{x}}_1^{(t)} - \hat{\boldsymbol{x}}_2^{(t)}\|_2^2/N, \|\hat{\boldsymbol{z}}_1^{(t)} - \hat{\boldsymbol{z}}_2^{(t)}\|_2^2/M\} < \epsilon_{\rm tol}$ **then**
    $t \leftarrow T_{\rm iter}$
    break
  end if
end for
return $\boldsymbol{h}_{1x}^{(T_{\rm iter})}, \hat{Q}_{1x}^{(T_{\rm iter})}, \hat{v}_{1x}^{(T_{\rm iter})}$

Assumption: at each iteration $t = 1, 2, \ldots, T_{\rm iter}$, positive constants $\hat{m}_{kx}^{(t)}$, $\hat{m}_{kz}^{(t)}$, $\hat{\chi}_{kx}^{(t)}$, $\hat{\chi}_{kz}^{(t)} \in \mathbb{R}$, $(k = 1, 2)$ exist such that for the singular value decomposition $A = USV^\top$,

$$\boldsymbol{h}_{1x}^{(t)} - \hat{m}_{1x}^{(t)}\boldsymbol{x}_0 \doteq \sqrt{\hat{\chi}_{1x}^{(t)}}\boldsymbol{\xi}_{1x}^{(t)}, \tag{100}$$

$$\boldsymbol{h}_{1z}^{(t)} - \hat{m}_{1z}^{(t)}\boldsymbol{z}_0 \doteq \sqrt{\hat{\chi}_{1z}^{(t)}}\boldsymbol{\xi}_{1z}^{(t)}, \tag{101}$$

J. Stat. Mech. (2020) 093402

$$V^\top(\boldsymbol{h}_{2x}^{(t)} - \hat{m}_{2x}^{(t)}\boldsymbol{x}_0) \doteq \sqrt{\hat{\chi}_{2x}^{(t)}}\boldsymbol{\xi}_{2x}^{(t)}, \tag{102}$$

$$U^\top(\boldsymbol{h}_{2z}^{(t)} - \hat{m}_{2z}^{(t)}\boldsymbol{z}_0) \doteq \sqrt{\hat{\chi}_{2z}^{(t)}}\boldsymbol{\xi}_{2z}^{(t)}, \tag{103}$$

hold, where $\doteq$ denotes the equality of empirical distributions, $\boldsymbol{z}_0$ is $A\boldsymbol{x}_0$, and $\boldsymbol{\xi}_{kx}^{(t)}, \boldsymbol{\xi}_{kz}^{(t)}, (k = 1, 2, t = 1, 2, \ldots, T_{\rm iter})$ are mutually independent standard Gaussian variables.

The equations (102) and (103) are expected from the mixing by randomly sampled orthogonal matrices $V^\top$ and $U^\top$. The equations (100) and (101) are expected from the Onsager correction terms $-\boldsymbol{h}_{2x}^{(t)}, -\boldsymbol{h}_{2z}^{(t)}$ that appears in the moment-matching conditions in line 20–21.

To characterize macroscopic behavior of rVAMP, we introduce the following macroscopic order parameters for $t = 1, 2, \ldots, T_{\rm iter}$:

$$m_{1x}^{(t)} = \frac{1}{N}\boldsymbol{x}_0^\top\hat{\boldsymbol{x}}_1^{(t)}, \qquad m_{1z}^{(t)} = \frac{1}{M}\boldsymbol{z}_0^\top\hat{\boldsymbol{z}}_1^{(t)}, \tag{104}$$

$$q_{1x}^{(t)} = \frac{1}{N}\left\|\hat{\boldsymbol{x}}_1^{(t)}\right\|_2^2, \qquad q_{1z}^{(t)} = \frac{1}{M}\left\|\hat{\boldsymbol{z}}_1^{(t)}\right\|_2^2, \tag{105}$$

$$m_{2x}^{(t)} = \frac{1}{N}\boldsymbol{x}_0^\top\hat{\boldsymbol{x}}_2^{(t)}, \qquad m_{2z}^{(t)} = \frac{1}{M}\boldsymbol{z}_0^\top\hat{\boldsymbol{z}}_2^{(t)}, \tag{106}$$

$$q_{2x}^{(t)} = \frac{1}{N}\left\|\hat{\boldsymbol{x}}_2^{(t)}\right\|_2^2, \qquad q_{2z}^{(t)} = \frac{1}{M}\left\|\hat{\boldsymbol{z}}_2^{(t)}\right\|_2^2, \tag{107}$$

$$T_x = \frac{1}{N}\left\|\boldsymbol{x}_0\right\|_2^2, \qquad T_z = \frac{1}{M}\left\|\boldsymbol{z}_0\right\|_2^2. \tag{108}$$

These order parameters and the susceptibilities have limiting expressions in the limit $N \to \infty$. First, $q_{1x}^{(t)}$ can be written as

$$\begin{aligned} q_{1x}^{(t)} &\simeq \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{\gamma_i}\left[\int g_{1x}(h_{1x,i}^{(t)}, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \gamma_i, \eta_{x,i}) D\eta_{x,i}\right]\right)^2 \\ &\overset{N\to\infty}{\to} \mathbb{E}_{x_0}\left[\int\left(\mathbb{E}_\gamma\left[\int g_{1x}(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \gamma, \eta_x) D\eta_x\right]\right)^2 D\xi_x\right]. \end{aligned} \tag{109}$$

Here, the summation is replaced with the average in the limit $N \to \infty$. The average $\mathbb{E}_\gamma[\ldots]$ is with respect to the density $p(\gamma) = \delta(\gamma - \gamma_0)/2 + \delta(\gamma - 2\gamma_0)/2$. Similar results can be obtained for $m_{1x}^{(t)}, m_{1z}^{(t)}, \chi_{1x}^{(t)}, v_{1x}^{(t)}, q_{1z}^{(t)}, \chi_{1z}^{(t)}$ and $v_{1z}^{(t)}$. Next, for the singular value decomposition $A = USV^\top$, we denote by $\{\sqrt{\lambda_i}\}$ the diagonal elements of S. Then, $q_{2x}^{(t)}$ can be written as follows:

$$q_{2x}^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\hat{m}_{2x}^{(t)} + S^\top S\hat{m}_{2z}^{(t)}\right)(V^\top\boldsymbol{x}_0) + \left(\sqrt{\hat{\chi}_{2x}^{(t)}}\boldsymbol{\xi}_{2x}^{(t)} + \sqrt{\hat{\chi}_{2z}^{(t)}}S^\top\boldsymbol{\xi}_{2z}^{(t)}\right)\right)^\top\left(\hat{Q}_{2x}^{(t)}I_N + S^\top S\hat{Q}_{2z}^{(t)}\right)^{-2}$$

$$\times \left(\left(\hat{m}_{2x}^{(t)} + S^\top S \hat{m}_{2z}^{(t)} \right) (V^\top \boldsymbol{x}_0) + \left(\sqrt{\hat{\chi}_{2x}^{(t)}} \boldsymbol{\xi}_{2x}^{(t)} + \sqrt{\hat{\chi}_{2z}^{(t)}} S^\top \boldsymbol{\xi}_{2z}^{(t)} \right) \right)$$

$$\simeq \frac{1}{N} \sum_{i=1}^{N} \frac{(\hat{m}_{2x}^{(t)} + \lambda_i \hat{m}_{2z}^{(t)})^2 (V^\top \boldsymbol{x}_0)_i^2}{(\hat{Q}_{2x}^{(t)} + \lambda_i \hat{Q}_{2z}^{(t)})^2} + \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{\chi}_{2x}^{(t)} \xi_{2x,i}^2 + \lambda_i \hat{\chi}_{2z} \xi_{2z,i}^2}{(\hat{Q}_{2x}^{(t)} + \lambda_i \hat{Q}_{2z}^{(t)})^2}$$

$$\overset{N\to\infty}{\to} T_x \mathbb{E}_\lambda \left[\frac{(\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)})^2} \right] + \mathbb{E}_\lambda \left[\frac{(\hat{\chi}_{2x}^{(t)} + \lambda \hat{\chi}_{2z}^{(t)})}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)})^2} \right], \tag{110}$$

where we used the independence between $\boldsymbol{\xi}_{2x}^{(t)}, \boldsymbol{\xi}_{2z}^{(t)}, \boldsymbol{x}_0$ and $\{\lambda_i\}$, and we denoted by $\mathbb{E}_\lambda[\ldots]$ an average with respect to the limiting eigenvalue spectrum $\rho(\lambda)$ of $A^\top A$. The calculations for $m_{2x}^{(t)}, m_{2z}^{(t)}, \chi_{2x}^{(t)}, v_{2x}^{(t)}, q_{2z}^{(t)}, \chi_{2z}^{(t)}$ and $v_{2z}^{(t)}$ are similar. Finally, using the singular value decomposition $A = USV^\top$, T_x and T_z are written as

$$T_x = \frac{1}{N} \sum_{i=1}^{N} x_{0,i}^2 \overset{N\to\infty}{\to} \int x_0^2 q_{x_0}(x_0) \mathrm{d}x_0, \tag{111}$$

$$T_z \overset{N\to\infty}{\to} \mathbb{E}_{z_0}[z_0^2] = \frac{\mathbb{E}_\lambda[\lambda]}{\alpha} T_x, \tag{112}$$

where the average of z_0 is taken with respect to a Gaussian measure

$$\exp\left(-\frac{\hat{T}_z}{2} z_0^2 \right) \sqrt{\frac{\hat{T}_z}{2\pi}} \mathrm{d}z, \quad \hat{T}_z = \frac{\alpha}{\mathbb{E}_\lambda[\lambda] T_x}, \tag{113}$$

based on the observation in [39]; for a vector $\boldsymbol{\omega} \in \mathbb{R}^N$ that is independent of A, the empirical distribution of $A\boldsymbol{\omega}$ is a Gaussian with mean zero and variance $\mathbb{E}_\lambda[\lambda] \|\boldsymbol{\omega}\|_2^2 / (\alpha N)$ in the large system limit.

The moment-matching conditions also have the following limiting expressions. First, $\hat{m}_{2x}^{(t)}$ can be written as

$$\hat{m}_{2x}^{(t)} \overset{\text{(a)}}{\longrightarrow} \frac{1}{\|x_0\|_2^2} \boldsymbol{x}_0^\top \boldsymbol{h}_{2x}^{(t)}$$

$$\overset{\text{(b)}}{=} \frac{1}{\|x_0\|_2^2} \boldsymbol{x}_0^\top \left(\frac{\hat{\boldsymbol{x}}_1^{(t)}}{\chi_{1x}^{(t)}} - \boldsymbol{h}_{1x}^{(t)} \right)$$

$$\overset{\text{(c)}}{=} \frac{m_{1x}^{(t)}}{T_x \chi_{1x}^{(t)}} - \hat{m}_{1x}^{(t)}, \tag{114}$$

where the limit (a) follows from the definition of $\hat{m}_{2x}^{(t)}$; (b) follows from the moment-matching condition of SA rVAMP; (c) follows from the definitions of $m_{1x}^{(t)}$ and $\hat{m}_{1x}^{(t)}$. For $\hat{\chi}_{2x}^{(t)}$, its update rule can be written as

J. Stat. Mech. (2020) 093402

$$\begin{aligned}
\hat{\chi}_{2x}^{(t)} &\xrightarrow{\text{(a)}} \frac{1}{N}\|\boldsymbol{h}_{2x}^{(t)} - \hat{m}_{2x}^{(t)}\boldsymbol{x}_0\|_2^2 \\
&\overset{\text{(b)}}{=} \frac{1}{N}\left\| \frac{\hat{\boldsymbol{x}}_1^{(t)}}{\chi_{1x}^{(t)}} - \frac{m_{1x}^{(t)}}{T_x\chi_{1x}^{(t)}}\boldsymbol{x}_0 - \sqrt{\hat{\chi}_{1x}^{(t)}}\boldsymbol{\xi}_{1x}^{(t)}\right\|_2^2 \\
&\overset{\text{(c)}}{=} \frac{q_{1x}^{(t)}}{(\chi_{1x}^{(t)})^2} - \frac{(m_{1x}^{(t)})^2}{T_x(\chi_{1x}^{(t)})^2} + \chi_{1x}^{(t)} - 2\frac{\sqrt{\hat{\chi}_{1x}^{(t)}}}{\chi_{1x}^{(t)}}\frac{1}{N}(\hat{\boldsymbol{x}}_{1x}^{(t)})^\top\boldsymbol{\xi}_{1x}^{(t)}, \\
&\overset{\text{(d)}}{=} \frac{q_{1x}^{(t)}}{(\chi_{1x}^{(t)})^2} - \frac{(m_{1x}^{(t)})^2}{T_x(\chi_{1x}^{(t)})^2} - \chi_{1x}^{(t)},
\end{aligned} \tag{115}$$

where (a) follows from the definition of $\hat{\chi}_{2x}^{(t)}$; (b) follows from the moment-matching condition of SA rVAMP and the assumption 2; (c) uses the independence between $\boldsymbol{x}_0$ and $\boldsymbol{\xi}_{1x}^{(t)}$, and the definition of $m_{1x}^{(t)}$; (d) can be obtained from the following integration by parts according to

$$\begin{aligned}
\frac{1}{N}(\hat{\boldsymbol{x}}_{1x}^{(t)})^\top\boldsymbol{\xi}_{1x}^{(t)} &\to \mathbb{E}_{x_0}\left[\int \mathbb{E}_\gamma\left[\int g_{1x}(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \lambda, \eta_x)D\eta_x\right]\xi_x D\xi_x\right] \\
&= \sqrt{\hat{\chi}_{1x}^{(t)}}\mathbb{E}_{x_0}\left[\int \mathbb{E}_\gamma\left[\int g'_{1x}(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \lambda, \eta_x)D\eta_x\right]D\xi_x\right] \\
&= \sqrt{\hat{\chi}_{1x}^{(t)}}\chi_{1x}^{(t)}.
\end{aligned} \tag{116}$$

Similarly, $\hat{m}_{1x}^{(t+1)}$ and $\hat{\chi}_{1x}^{(t+1)}$ are obtained as follows. For $\hat{m}_{1x}^{(t+1)}$, its update rule is derived exactly same way as in (114). For $\hat{\chi}_{1x}^{(t)}$,

$$\begin{aligned}
\hat{\chi}_{1x}^{(t)} &\xrightarrow{\text{(a)}} \frac{1}{N}\|\boldsymbol{h}_{1x}^{(t+1)} - \hat{m}_{1x}^{(t+1)}\boldsymbol{x}_0\|_2^2 \\
&\overset{\text{(b)}}{=} \frac{1}{N}\left\| \frac{\hat{\boldsymbol{x}}_2}{\chi_{2x}^{(t)}} - \frac{m_{2x}^{(t)}}{T_x\chi_{2x}^{(t)}}\boldsymbol{x}_0 - \sqrt{\hat{\chi}_{2x}^{(t)}}V\boldsymbol{\xi}_{2x}^{(t)}\right\|_2^2 \\
&\overset{\text{(c)}}{=} \frac{q_{2x}^{(t)}}{(\chi_{2x}^{(t)})^2} - \frac{(m_{2x}^{(t)})^2}{T_x(\chi_{2x}^{(t)})^2} + \hat{\chi}_{2x}^{(t)} - 2\frac{\sqrt{\hat{\chi}_{2x}^{(t)}}}{\chi_{2x}^{(t)}}\frac{1}{N}(V^\top\hat{\boldsymbol{x}}_2^{(t)})^\top\boldsymbol{\xi}_{2x}^{(t)} \\
&\overset{\text{(d)}}{=} \frac{q_{2x}^{(t)}}{(\chi_{2x}^{(t)})^2} - \frac{(m_{2x}^{(t)})^2}{T_x(\chi_{2x}^{(t)})^2} - \hat{\chi}_{2x}^{(t)},
\end{aligned} \tag{117}$$

where (a) follows from the definition of $\hat{\chi}_{1x}^{(t+1)}$; (b) follows from the moment-matching condition and the assumption 2; (c) uses the independence between $V^\top\boldsymbol{x}_0$ and $\boldsymbol{\xi}_{2x}^{(t)}$, and the definition of $m_{2x}^{(t)}$; (d) can be obtained from the independence between $V^\top\boldsymbol{x}_0, S^\top\boldsymbol{\xi}_{2z}^{(t)}$

J. Stat. Mech. (2020) 093402

and $\boldsymbol{\xi}_{2x}^{(t)}$:

$$\begin{aligned}\frac{1}{N}(V^\top \hat{\boldsymbol{x}}_2^{(t)})^\top \boldsymbol{\xi}_{2x}^{(t)} &= \frac{1}{N}\sum_{i=1}^{N} \frac{\left((\hat{m}_{2x}^{(t)} + \lambda_i \hat{m}_{2z}^{(t)})[V^\top \boldsymbol{x}_0]_i + \sqrt{\hat{\chi}_{2x}^{(t)}}\xi_{2x,i} + \sqrt{\hat{\chi}_{2z}^{(t)}}[S^\top(\boldsymbol{\xi}_{2z}^{(t)})]_i\right)\xi_{2x,i}^{(t)}}{\hat{Q}_{2x}^{(t)} + \lambda_i \hat{Q}_{2z}^{(t)}} \\ &\to \sqrt{\hat{\chi}_{2x}^{(t)}}\mathbb{E}_\lambda\left[\frac{1}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)}}\right]\int \xi_{2x}^2 D\xi_{2x} \\ &= \sqrt{\hat{\chi}_{2x}^{(t)}}\chi_{2x}^{(t)}. \end{aligned} \tag{118}$$

Similar results can be obtained for $\hat{m}_{2z}^{(t)}, \hat{\chi}_{2z}^{(t)}, \hat{m}_{1z}^{(t+1)}$ and $\hat{\chi}_{1z}^{(t+1)}$.

The above observations yield the SE of SA rVAMP as follows:

Initialization: select initial $\hat{m}_{1x}^{(1)}, \hat{\chi}_{1x}^{(1)}, \hat{Q}_{1x}^{(1)}, \hat{v}_{1x}^{(1)}, \hat{m}_{1z}^{(1)}, \hat{\chi}_{1z}^{(1)}, \hat{Q}_{1z}^{(1)}$, and $\hat{v}_{1z}^{(1)} \in [0, \infty)$.

Iteration: for $t = 1, 2, \ldots, T_{\text{iter}}$, update the parameters as follows:

factorized part:

$$q_{1x}^{(t)} = \mathbb{E}_{x_0}\left[\int\left(\mathbb{E}_\gamma\left[\int g_{1x}(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \gamma, \eta_x)D\eta_x\right]\right)^2 D\xi_x\right], \tag{119}$$

$$\chi_{1x}^{(t)} = \mathbb{E}_{x_0}\left[\int \mathbb{E}_\gamma\left[\int g'_{1x}(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \gamma, \eta_x)D\eta_x\right] D\xi_x\right], \tag{120}$$

$$\begin{aligned} v_{1x}^{(t)} &= \mathbb{E}_{x_0}\left[\int \mathbb{E}_\gamma\left[\int g_{1x}^2(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \gamma, \eta_x)D\eta_x\right] D\xi_x\right] \\ &\quad - \mathbb{E}_{x_0}\left[\int\left(\mathbb{E}_\gamma\left[\int g_{1x}(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \gamma, \eta_x)D\eta_x\right]\right)^2 D\xi_x\right], \end{aligned} \tag{121}$$

$$m_{1x}^{(t)} = \mathbb{E}_{x_0}\left[\int x_0 \mathbb{E}_\gamma\left[\int g_{1x}(\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_x, \hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}; \gamma, \eta_x)D\eta_x\right] D\xi_x\right], \tag{122}$$

$$q_{1z}^{(t)} = \mathbb{E}_{z_0}\left[\int\left(\mathbb{E}_c\left[\int g_{1z}(\hat{m}_{1z}^{(t)}z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_z, \hat{Q}_{1z}^{(t)}, \hat{v}_{1z}^{(t)}; c, \eta_z, y)D\eta_x\right]\right)^2 q_{y|z}(y|z_0)\mathrm{d}y D\xi_z\right], \tag{123}$$

$$\chi_{1z}^{(t)} = \mathbb{E}_{z_0}\left[\int \mathbb{E}_c\left[\int g'_{1z}(\hat{m}_{1z}^{(t)}z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_z, \hat{Q}_{1z}^{(t)}, \hat{v}_{1z}^{(t)}; c, \eta_z, y)D\eta_z\right] q_{y|z}(y|z_0)\mathrm{d}y D\xi_z\right], \tag{124}$$

$$v_{1z}^{(t)} = \mathbb{E}_{z_0}\left[\int \mathbb{E}_c\left[\int g_{1z}^2(\hat{m}_{1z}^{(t)}z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_z, \hat{Q}_{1z}^{(t)}, \hat{v}_{1z}^{(t)}; c, \eta_z, y)D\eta_z\right] q_{y|z}(y|z_0)\mathrm{d}y D\xi_z\right]$$

J. Stat. Mech. (2020) 093402

$$- \mathbb{E}_{z_0}\left[\int \left(\mathbb{E}_c\left[\int g_{1z}(\hat{m}_{1z}z_0+\sqrt{\hat{\chi}_{1z}^{(t)}}\xi_z, \hat{Q}_{1z}^{(t)}, \hat{v}_{1z}^{(t)}; c, \eta_z, y)D\eta_z\right]\right)^2 q_{y|z}(y|z_0)\mathrm{d}yD\xi_z\right], \tag{125}$$

$$m_{1z}^{(t)} = \mathbb{E}_{z_0}\left[\int z_0 \mathbb{E}_c\left[\int g_{1z}(\hat{m}_{1z}z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_z, \hat{Q}_{1z}^{(t)}, \hat{v}_{1z}^{(t)}; c, \eta_z, y)D\eta_z\right] q_{y|z}(y|z_0)\mathrm{d}yD\xi_z\right]. \tag{126}$$

Moment-matching:

$$\hat{Q}_{2x}^{(t)} = \frac{1}{\chi_{1x}^{(t)}} - \hat{Q}_{1x}^{(t)}, \qquad \hat{Q}_{2z}^{(t)} = \frac{1}{\chi_{1z}^{(t)}} - \hat{Q}_{1z}^{(t)}, \tag{127}$$

$$\hat{v}_{2x}^{(t)} = \frac{v_{1x}^{(t)}}{(\chi_{1x}^{(t)})^2} - \hat{v}_{1x}^{(t)}, \qquad \hat{v}_{2z}^{(t)} = \frac{v_{1z}^{(t)}}{(\chi_{1z}^{(t)})^2} - \hat{v}_{1z}^{(t)}, \tag{128}$$

$$\hat{m}_{2x}^{(t)} = \frac{m_{1x}^{(t)}}{T_x\chi_{1x}^{(t)}} - \hat{m}_{1x}^{(t)}, \qquad \hat{m}_{2z}^{(t)} = \frac{m_{1z}^{(t)}}{T_z\chi_{1z}^{(t)}} - \hat{m}_{1z}^{(t)}, \tag{129}$$

$$\hat{\chi}_{2x}^{(t)} = \frac{q_{1x}^{(t)}}{(\chi_{1x}^{(t)})^2} - \frac{(m_{1x}^{(t)})^2}{T_x(\chi_{1x}^{(t)})^2} - \hat{\chi}_{1x}^{(t)}, \qquad \hat{\chi}_{2z}^{(t)} = \frac{q_{1z}^{(t)}}{(\chi_{1z}^{(t)})^2} - \frac{(m_{1z}^{(t)})^2}{T_z(\chi_{1z}^{(t)})^2} - \hat{\chi}_{1z}^{(t)}. \tag{130}$$

Gaussian part:

$$q_{2x}^{(t)} = T_x\mathbb{E}_\lambda\left[\frac{(\hat{m}_{2x}^{(t)} + \lambda\hat{m}_{2z}^{(t)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right] + \mathbb{E}_\lambda\left[\frac{(\hat{\chi}_{2x}^{(t)} + \lambda\hat{\chi}_{2z}^{(t)})}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right], \tag{131}$$

$$\chi_{2x}^{(t)} = \mathbb{E}_\lambda\left[\frac{1}{\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)}}\right], \tag{132}$$

$$v_{2x}^{(t)} = \mathbb{E}_\lambda\left[\frac{\hat{v}_{2x}^{(t)} + \lambda\hat{v}_{2z}^{(t)}}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right], \tag{133}$$

$$m_{2x}^{(t)} = T_x\mathbb{E}_\lambda\left[\frac{\hat{m}_{2x}^{(t)} + \lambda\hat{m}_{2z}^{(t)}}{\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)}}\right], \tag{134}$$

$$q_{2z}^{(t)} = \frac{T_x}{\alpha}\mathbb{E}_\lambda\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda\hat{m}_{2z}^{(t)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right] + \mathbb{E}_\lambda\left[\frac{\lambda(\hat{\chi}_{2x}^{(t)} + \lambda\hat{\chi}_{2z}^{(t)})}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right], \tag{135}$$

$$\chi_{2z} = \frac{1}{\alpha}\mathbb{E}_\lambda\left[\frac{\lambda}{\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)}}\right], \tag{136}$$

$$v_{2z}^{(t)} = \frac{1}{\alpha}\mathbb{E}_\lambda\left[\frac{\lambda(\hat{v}_{2x}^{(t)} + \lambda\hat{v}_{2z}^{(t)})}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right], \tag{137}$$

J. Stat. Mech. (2020) 093402

$$m_{2z}^{(t)} = \frac{T_x}{\alpha}\mathbb{E}_\lambda\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda\hat{m}_{2z}^{(t)})}{\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)}}\right]. \tag{138}$$

Moment-matching:

$$\hat{Q}_{1x}^{(t+1)} = \frac{1}{\chi_{2x}^{(t)}} - \hat{Q}_{2x}^{(t)}, \qquad \hat{Q}_{1z}^{(t+1)} = \frac{1}{\chi_{2z}^{(t)}} - \hat{Q}_{2z}^{(t)}, \tag{139}$$

$$\hat{v}_{1x}^{(t+1)} = \frac{v_{2x}^{(t)}}{(\chi_{2x}^{(t)})^2} - \hat{v}_{1x}^{(t)}, \qquad \hat{v}_{1z}^{(t+1)} = \frac{v_{2z}^{(t)}}{(\chi_{2z}^{(t)})^2} - \hat{v}_{1z}^{(t)}, \tag{140}$$

$$\hat{m}_{1x}^{(t+1)} = \frac{m_{2x}^{(t)}}{T_x\chi_{2x}^{(t)}} - \hat{m}_{1x}^{(t)}, \qquad \hat{m}_{1z}^{(t+1)} = \frac{m_{2z}^{(t)}}{T_z\chi_{2z}^{(t)}} - \hat{m}_{1z}^{(t)}, \tag{141}$$

$$\hat{\chi}_{1x}^{(t+1)} = \frac{q_{2x}^{(t)}}{(\chi_{2x}^{(t)})^2} - \frac{(m_{2x}^{(t)})^2}{T_x(\chi_{2x}^{(t)})^2} - \hat{\chi}_{2x}^{(t)}, \qquad \hat{\chi}_{1z}^{(t+1)} = \frac{q_{2z}^{(t)}}{(\chi_{2z}^{(t)})^2} - \frac{(m_{2z}^{(t)})^2}{T_z(\chi_{2z}^{(t)})^2} - \hat{\chi}_{2z}^{(t)}, \tag{142}$$

where $\mathbb{E}_c[\dots]$ is the average with respect to the probability function $p(c) = \mathrm{e}^{-1}/c!, c = 0, 1, \dots$.

At the fixed point, $q_{1x}^{(t)} = q_{2x}^{(t)}, \chi_{1x}^{(t)} = \chi_{2x}^{(t)}, v_{1x}^{(t)} = v_{2x}^{(t)}$, and $m_{1x}^{(t)} = m_{2x}^{(t)}$ are approximate values of the following quantities:

$$q_x \simeq \lim_{\beta\to\infty, N\to\infty} \frac{1}{N}\left\|\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\int \boldsymbol{x}p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\right]\right\|_2^2, \tag{143}$$

$$\chi_x \simeq \lim_{\beta\to\infty, N\to\infty} \frac{\beta}{N}\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\int \|\boldsymbol{x}\|_2^2 p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\right] - \left\|\int \boldsymbol{x}p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\right\|_2^2\Bigg], \tag{144}$$

$$v_x \simeq \lim_{\beta\to\infty, N\to\infty} \frac{1}{N}\left(\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\left\|\int \boldsymbol{x}p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\right\|_2^2\right) - \left\|\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\int \boldsymbol{x}p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\right]\right\|_2^2\Bigg], \tag{145}$$

$$m_x \simeq \lim_{\beta\to\infty, N\to\infty} \mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\boldsymbol{x}_0^\top \int \boldsymbol{x}p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\gamma}, D)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\right]. \tag{146}$$

A similar interpretation is also possible for $q_{1z}^{(t)} = q_{2z}^{(t)}, \chi_{1z}^{(t)} = \chi_{2z}^{(t)}, v_{1z}^{(t)} = v_{2z}^{(t)}$, and $m_{1z}^{(t)} = m_{2z}^{(t)}$.

J. Stat. Mech. (2020) 093402

4.4. Replica analysis

Generally, typical values of the macroscopic order parameters introduced in the last section can be obtained by calculating the Helmholtz free energy f using the replica method [6]:

$$f = \mathbb{E}_D\left[f(D)\right] \equiv -\lim_{N,\beta\to\infty,n\to 0} \frac{1}{Nn\beta}\mathbb{E}_D[\log \Xi_n(D)] \tag{147}$$

$$= -\lim_{\substack{N,\beta\to\infty\\ n,\tilde{l}\to 0}} \frac{1}{Nn\tilde{l}\,\beta}\mathbb{E}_D\left[\Xi_n(D)^{\tilde{l}}\right]. \tag{148}$$

Although the above formula contains the nested replicas, its replica symmetric computation is formally analogous to the standard one-step replica symmetry breaking (one-RSB) computation by treating $\tilde{l}$ as the Parisi's breaking parameter. Because the one-RSB computation was already described in appendix C of reference [23], we only show the final result. By rescaling the replica number as $\tilde{l} = l/\beta$, we obtain the following expression:

$$f = -\lim_{\beta\to\infty,l\to 0} \operatorname*{extr}_{\substack{m_x,q_x,v_x,\chi_x,\\ m_z,q_z,v_z,\chi_z}} \left[g_{\mathrm{F}} + g_{\mathrm{G}} - g_{\mathrm{S}}\right], \tag{149}$$

$$\begin{aligned} g_{\mathrm{F}} = \operatorname*{extr}_{\substack{\hat{m}_{1x},\hat{\chi}_{1x},\hat{v}_{1x},\hat{Q}_{1x},\\ \hat{m}_{1z},\hat{\chi}_{1z},\hat{v}_{1z},\hat{Q}_{1z}}} \Bigg[& -m_x\hat{m}_{1x} + \frac{1}{2}\left(q_x + v_x + \frac{\chi_x}{\beta}\right)\hat{Q}_{1x} \\ & - \frac{l}{2}\left((q_x+v_x)(\hat{\chi}_{1x}+\hat{v}_{1x}) - q_x\hat{\chi}_{1x}\right) - \frac{1}{2}\chi_x(\hat{\chi}_{1x}+\hat{v}_{1x}) - \alpha m_z\hat{m}_{1z} \\ & + \frac{\alpha}{2}\left(q_z + v_z + \frac{\chi_z}{\beta}\right)\hat{Q}_{1z} - \frac{l\alpha}{2}\left((q_z+v_z)(\hat{\chi}_{1z}+\hat{v}_{1z}) - q_z\hat{\chi}_{1z}\right) \\ & - \frac{\alpha}{2}\chi_z(\hat{\chi}_{1z}+\hat{v}_{1z}) + \frac{1}{l}\int\left\{\log \mathbb{E}_\gamma\left[\int \mathrm{e}^{l\phi_x^{(\beta)}} D\eta_x\right]\right\} q_{x_0}(x_0)\mathrm{d}x_0 D\xi_x \\ & + \frac{1}{l}\int\left\{\log\ \mathbb{E}_c\left[\int \mathrm{e}^{l\phi_z^{(\beta)}} D\eta_z\right]\right\}\sqrt{\frac{\hat{T}_z}{2\pi}}\mathrm{e}^{-\frac{\hat{T}_z}{2}z_0^2} q_{y|z}(y|z_0)\,\mathrm{d}z_0 D\xi_z\,\mathrm{d}y\Bigg], \end{aligned} \tag{150}$$

$$\begin{aligned} g_G = \operatorname*{extr}_{\substack{\hat{m}_{2x},\hat{\chi}_{2x},\hat{v}_{2x},\hat{Q}_{2x},\\ \hat{m}_{2z},\hat{\chi}_{2z},\hat{v}_{2z},\hat{Q}_{2z}}} \Bigg[& -m_x\hat{m}_{2x} + \frac{1}{2}\left(q_x + v_x + \frac{\chi_x}{\beta}\right)\hat{Q}_{2x} \\ & - \frac{l}{2}\left((q_x+v_x)(\hat{\chi}_{2x}+\hat{v}_{2x}) - q_x\hat{\chi}_{2x}\right) - \frac{1}{2}\chi_x(\hat{\chi}_{2x}+\hat{v}_{2x}) - \alpha m_z\hat{m}_{2z} \\ & + \frac{\alpha}{2}\left(q_z + v_z + \frac{\chi_z}{\beta}\right)\hat{Q}_{2z} - \frac{\alpha l}{2}\left((q_z+v_z)(\hat{\chi}_{2z}+\hat{v}_{2z}) - q_z\hat{\chi}_{2z}\right) \\ & - \frac{\alpha}{2}\chi_z(\hat{\chi}_{2z}+\hat{v}_{2z}) - \frac{1}{2}\left(\frac{1}{\beta} - \frac{1}{l}\right)\mathbb{E}_\lambda\left[\log\left(\hat{Q}_{2x} + \lambda\hat{Q}_{2z}\right)\right] \end{aligned}$$

$$- \frac{1}{2l}\mathbb{E}_\lambda\left[\log\left(\hat{Q}_{2x}+\lambda\hat{Q}_{2z}-l(\hat{v}_{2x}+\lambda\hat{v}_{2z})\right)\right] + \frac{1}{2}\mathbb{E}_\lambda\left[\frac{\hat{\chi}_{2x}+\lambda\hat{\chi}_{2z}}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}-l(\hat{v}_{2x}+\lambda\hat{v}_{2z})}\right] + \frac{T_x}{2}\mathbb{E}_\lambda\left[\frac{(\hat{m}_{2x}+\lambda\hat{m}_{2z})^2}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}-l(\hat{v}_{2x}+\lambda\hat{v}_{2z})}\right]\Bigg], \tag{151}$$

$$g_{\mathrm{S}} = \frac{1}{2}\left(\frac{1}{\beta}-\frac{1}{l}\right)\log\ \chi_x + \frac{1}{2l}\log(\chi_x+lv_x) + \frac{1}{2}\frac{q_x}{\chi_x+lv_x} - \frac{1}{2}\frac{m_x^2}{T_x(\chi_x+lv_x)} + \frac{\alpha}{2}\left(\frac{1}{\beta}-\frac{1}{l}\right)\log\ \chi_z + \frac{\alpha}{2l}\log(\chi_z+lv_z) + \frac{\alpha}{2}\frac{q_x}{\chi_z+lv_z} - \frac{\alpha}{2}\frac{m_z^2}{T_z(\chi_z+lv_z)}, \tag{152}$$

where

$$\phi_x^{(\beta)} = \frac{1}{\beta}\log\int \mathrm{e}^{-\beta\frac{\hat{Q}_{1x}}{2}x^2+\beta(\hat{m}_{1x}x_0+\sqrt{\hat{\chi}_{1x}}\xi_x+\sqrt{\hat{v}_{1x}}\eta_x)x-\beta\gamma|x|}\mathrm{d}x, \tag{153}$$

$$\phi_z^{(\beta)} = \frac{1}{\beta}\log\int \mathrm{e}^{-\beta\frac{\hat{Q}_{1z}}{2}z^2+\beta(\hat{m}_{1z}z_0+\sqrt{\hat{\chi}_{1z}}\xi_z+\sqrt{\hat{v}_{1z}}\eta_z)z+\beta c\ \log p_{y|z}(y|z)}\mathrm{d}z. \tag{154}$$

In the limit $l\to 0, \beta\to\infty$, the extreme condition yields the same form of the equations that appear in the fixed point condition of the SE equations (127)–(142). Additionally, at the extremum, the variational parameters q_x, χ_x, v_x and m_x are in accordance with the right-hand side of the equations (143)–(146). Similar accordance also holds for q_z, χ_z, v_z and m_z. Thus, the fixed point of SE of SA rVAMP is consistent with the replica symmetric calculation.

5. Application to logistic regression

For checking the validity of the results obtained so far, we applied rVAMP to logistic regression and conducted numerical experiments in order to (i) validate our SE, (ii) obtain insights about the convergence speed from SE, and (iii) test the applicability of rVAMP to real-world problems.

In logistic regression, the domain of the response variables $\mathcal{Y}$ is $\{-1,1\}$, and the likelihood is given as

$$p_{y|z}(y|z) = \delta(y-1)\frac{1}{1+\mathrm{e}^{-z}} + \delta(y+1)\frac{1}{1+\mathrm{e}^{z}}. \tag{155}$$

Additionally, g'_{1z} in (80) can be written as

$$g'_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}; c_\mu, \eta_{z,\mu}, y_\mu) = \left[\hat{Q}_{1z,\mu} + \frac{c_\mu}{4\cosh^2\left(\frac{1}{2}g_{1z}(h_{1z,\mu}, \hat{Q}_{1z,\mu}, \hat{v}_{1z,\mu}; c_\mu, \eta_{z,\mu}, y_\mu)\right)}\right]^{-1}. \tag{156}$$

J. Stat. Mech. (2020) 093402

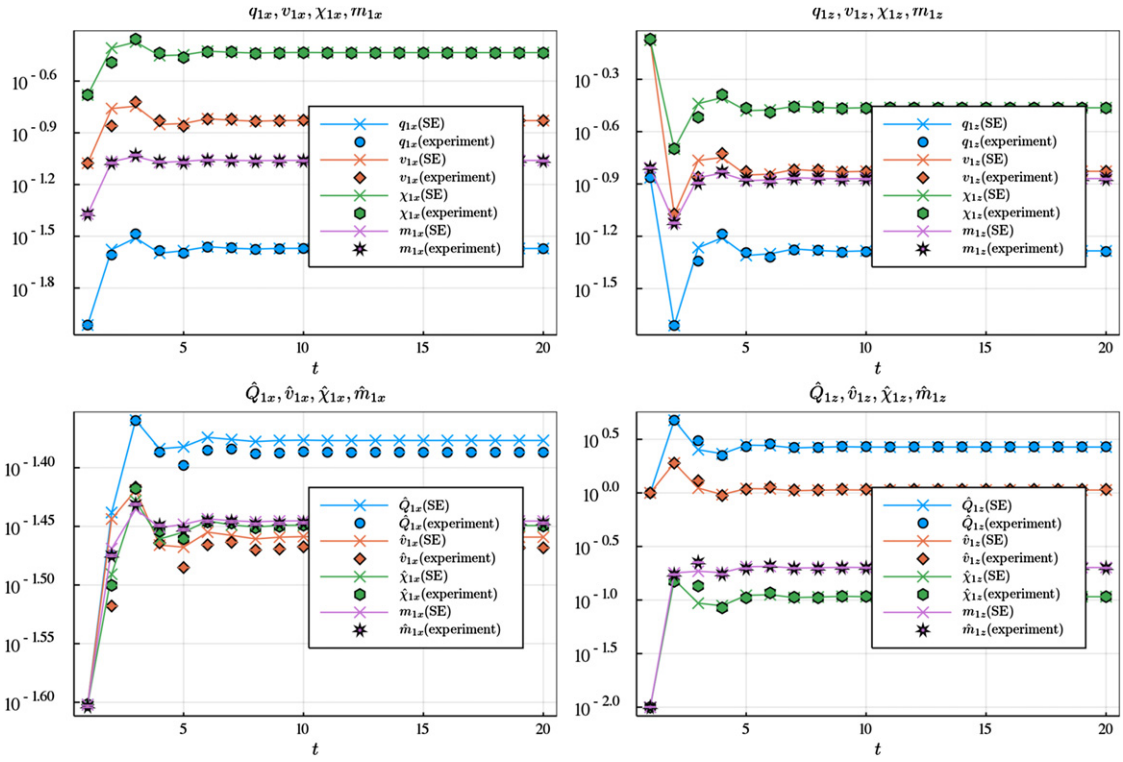

Figure 2. Comparison between the iteration dynamics of SA rVAMP in the algorithm 3 and in the SE equations defined in (127)–(142). The solid lines show the SE trajectories. The symbols represent the median of SA rVAMP trajectories that are obtained from 1000 experiments. Top left: macroscopic variables $q_{1x}^{(t)}, \chi_{1x}^{(t)}, v_{1x}^{(t)}$, and $m_{1x}^{(t)}$ versus algorithm iteration. Top right: macroscopic variables $q_{1z}^{(t)}, \chi_{1z}^{(t)}, v_{1z}^{(t)}$, and $m_{1z}^{(t)}$ versus algorithm iteration. Bottom left: parameters $\hat{Q}_{1x}^{(t)}, \hat{v}_{1x}^{(t)}, \hat{\chi}_{1x}^{(t)}$ and $\hat{m}_{1x}^{(t)}$ versus algorithm iteration. Bottom right: parameters $\hat{Q}_{1z}^{(t)}, \hat{v}_{1z}^{(t)}, \hat{\chi}_{1z}^{(t)}$ and $\hat{m}_{1z}^{(t)}$ versus algorithm iteration.

All the experiments were conducted on a single Intel(R) Core(TM) i7-8700B (3.20 GHz) CPU.

5.1. Comparing with SE using synthetic data

Synthetic data were generated under the settings described in section 4.1. The actual data generation process are described by

$$q_{x_0}(x_{0,i}) = \rho \mathcal{N}(x_{0,i}\,; 0, \rho^{-1}) + (1-\rho)\delta(x_{0,i}), \tag{157}$$

$$q_{y|z}(y_\mu | \boldsymbol{a}_\mu^\top \boldsymbol{x}_0) = \delta(y_\mu - 1)\frac{1}{1+\mathrm{e}^{-\boldsymbol{a}_\mu^\top \boldsymbol{x}_0}} + \delta(y_\mu + 1)\frac{1}{1+\mathrm{e}^{\boldsymbol{a}_\mu^\top \boldsymbol{x}_0}}, \tag{158}$$

where $\mathcal{N}(x_{0,i}\,; \mu, \sigma^2)$ is the Gaussian measure with mean μ and variance σ^2, and $\rho \in [0,1]$ is the sparsity. The system size N, the measurement ratio $\alpha = M/N$, and the sparsity ρ were specified as $N = 10000, \alpha = 0.2$, and $\rho = 0.01$, respectively. Additionally, the

J. Stat. Mech. (2020) 093402

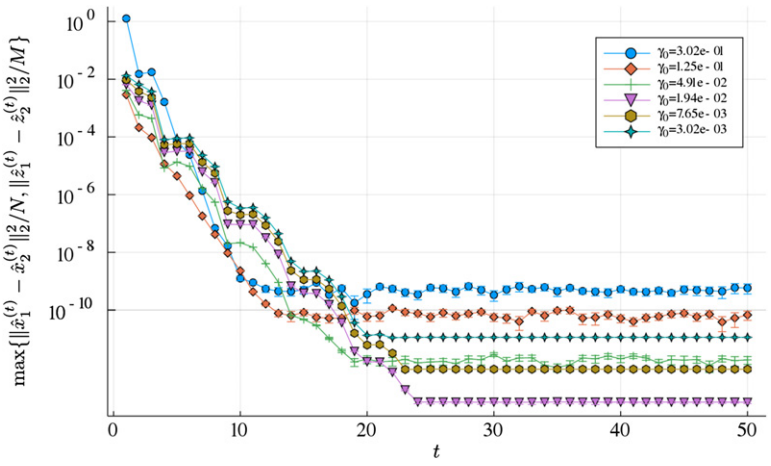

Figure 3. Time evolution of the convergence criterion $\max\{\|\hat{\boldsymbol{x}}_1^{(t)} - \hat{\boldsymbol{x}}_2^{(t)}\|_2^2/N, \|\hat{\boldsymbol{z}}_1^{(t)} - \hat{\boldsymbol{z}}_2^{(t)}\|_2^2/M\}$ is plotted versus the iteration step t. The error bars represent the standard errors. The symbols represent the median of rVAMP trajectories obtained from 1000 experiments.

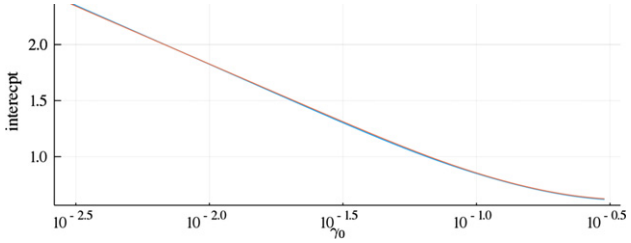

Figure 4. Intercept term of logistic regression model plotted versus γ_0. The red line is obtained by the naive refitting procedure, while the blue line is obtained using rVAMP.

feature matrix A was drawn from the row-orthogonal ensemble [40] for which the limiting eigenvalue distribution of $A^\top A$ was $\rho(\lambda) = \alpha\delta(\lambda - 1) + (1 - \alpha)\delta(\lambda)$.

To validate SE, we compared the iteration dynamics of SA rVAMP to those of SE. Figure 2 plots the order parameters and the parameters of $p_1^{(\beta)}$ versus the iteration index t. The data of SA rVAMP were obtained from 1000 random trials. The error bars are smaller than the size of the markers. Although some systematic disagreements are present in $\hat{Q}_{1x}^{(t)}$ and $\hat{v}_{1x}^{(t)}$ possibly due to the finite-size effect, most of the experimental values are in good agreement with the predictions of SE. This shows the validity of our SE.

The iteration dynamics of SE suggest that rVAMP converges in a few dozens of iterations, guaranteeing the fast convergence of rVAMP for the synthetic data.

J. Stat. Mech. (2020) 093402

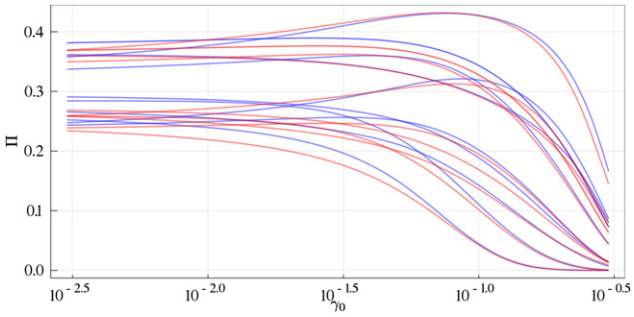

Figure 5. Comparison of the selection probability plotted for various values of the regularization strength γ_0. For ease of viewing, the selection probabilities are shown only for 10 features that had the largest selection probability for the smallest γ_0. Red lines are obtained using the naive refitting procedure, while blue lines are obtained using rVAMP.

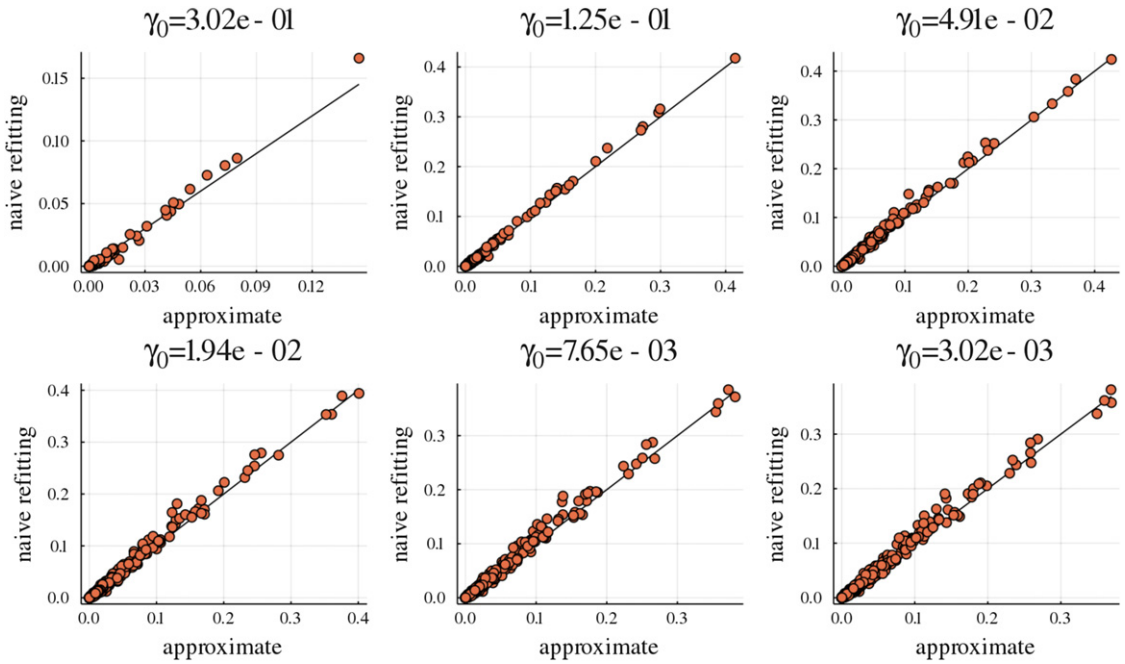

Figure 6. Naive refitting estimates of the selection probability $\Pi_i, i = 1, 2, \ldots, N$ plotted versus those computed by rVAMP for various regularization strengths.

5.2. Applicability of rVAMP in real world data

We explored the performance of rVAMP on the colon cancer dataset [4], which is also used in the introduction. The data is publicly available at http://genomics-pubs.princeton.edu/oncology/. The task is to distinguish cancer from normal tissues using micro-array data with $N = 2000$ features per example. The data were derived from 22 normal ($y_\mu = -1$) and 40 ($y_\mu = 1$) cancer tissues. The total number of samples is $M = 62$. We pre-processed the data by carrying out base 10 logarithmic

J. Stat. Mech. (2020) 093402

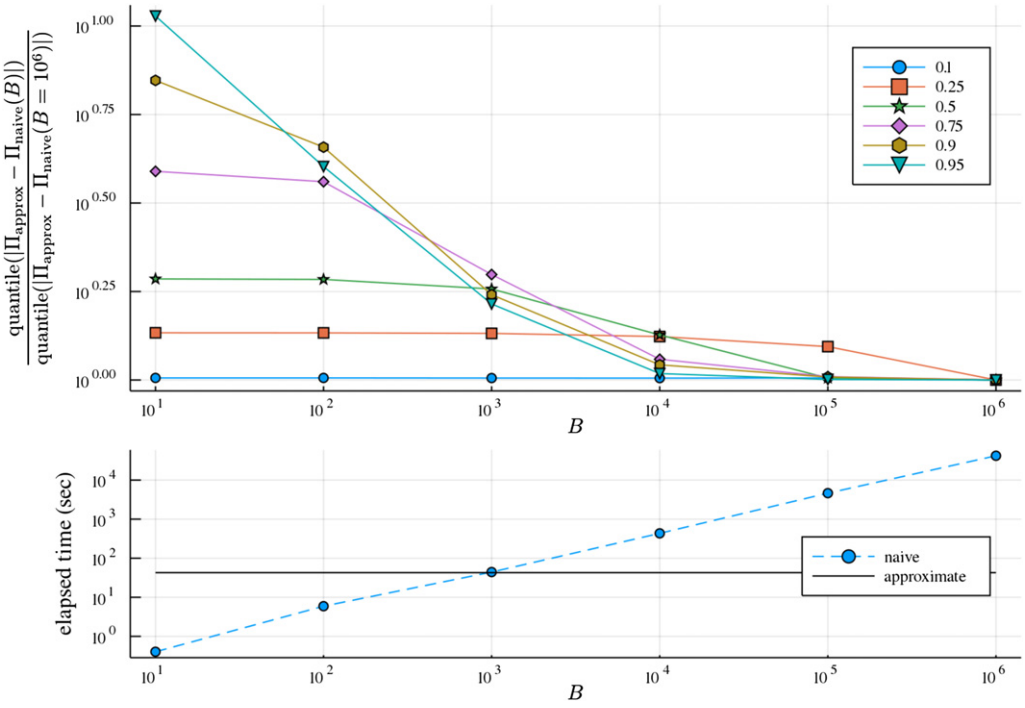

Figure 7. Upper panel: the difference between approximated and naively calculated selection probabilities plotted versus the number of resampled datasets B. We denote by $\Pi_{\mathrm{approximate}}$ the selection probability obtained by rVAMP, and by Π_{naive} that obtained by naive resampling procedure using B resampled datasets. The difference is measured as a q-quantile of the difference for all of the selection probabilities in the grid of γ_0. Lower panel: elapsed time is plotted versus the size of the resampled dataset B.

transformation and standardizing each feature to zero mean and unit variance. Because the class labels are biased, we included the intercept term. To obtain the selection probabilities for a grid of γ_0, we used the warm start procedure. Finally, the damping factor η_{d} was set to 0.85.

First, we examined the convergence speed of rVAMP. Figure 3 shows the time evolution of the convergence criterion $\max\{\|\hat{\boldsymbol{x}}_1^{(t)} - \hat{\boldsymbol{x}}_2^{(t)}\|_2^2/N, \|\hat{\boldsymbol{z}}_1^{(t)} - \hat{\boldsymbol{z}}_2^{(t)}\|_2^2/M\}$ by plotting its value versus the iteration step t. For various regularization strengths, regular exponential decay is observed, This demonstrating the fast convergence of rVAMP in a real-world dataset.

Next, we examine the accuracy of rVAMP. To compare the estimate of rVAMP with that of the naive refitting procedure of SS, the naive refitting on 1000000 resampled datasets was conducted using GLMNet [41]. Figure 4 shows the intercept term plotted versus the regularization strength. For a wide range of γ_0, rVAMP accurately estimated the intercept term. Figure 5 plots the comparison between the selection probabilities estimated by rVAMP and by the naive reffiting for the entire grid of γ_0. For ease of viewing, we only plot these values for the 10 features that had the largest selection probabilities for the smallest γ_0. Figure 6 plots the same comparison of all of the features

J. Stat. Mech. (2020) 093402

for a selected set of γ_0. Although the accuracy decreases slightly as we weaken the regularization, rVAMP successfully approximate the selection probability. The upper panel of ignore 7 plots the difference between approximated and naively calculated selection probabilities as a function of the number of resampled datasets B. These results also provide evidence for the accuracy of rVAMP. The lower panel of figure 7 plots the elapsed time used to obtain all of the selection probabilities for various γ_0. Although the actual computation time depends on the implementation, this figure suggests that rVAMP can provide accurate estimate of $\mathbf{\Pi}$ in a much shorter time than the naive SS. These observations demonstrate the accuracy of rVAMP.

6. Summary and conclusion

We developed an approximate SS algorithm that enables SS without the use of the repeated fitting procedure. The key concept is to use the combination of the replica method of statistical mechanics and the VAMP algorithm of information theory. The derivation of the algorithm was based on the EP of machine learning. We also derived the SE that macroscopically describes the dynamics of the proposed algorithm, and showed that its fixed point is consistent with the replica symmetric solution. Through numerical experiments, we confirmed that the SE equation is valid and that the proposed algorithm converges in a few dozens of iterations. We applied the proposed algorithm to logistic regression and demonstrated its application to a real-world dataset through numerical experiments. Although the real-world dataset has statistical correlations among the features, the proposed algorithm achieved fast convergence and high-estimation accuracy, demonstrating its utility for real-world problems.

A possible drawback of our algorithm is its computational complexity, even though it was not significant for the experiments described in section 5. Because the algorithm requires the computation of matrix inversion at each iteration, the computational burden may increase significantly with the increasing number of samples in the datasets. This shortcoming may be addressed by the self-averaging version of the proposed algorithm or the dual-decomposition-like variable augmentation used in the alternating direction method of multipliers [42, 43].

A promising future research direction includes analyzing the variable selection performance of the SS algorithm using SE. Generally, theoretical analysis of resampling techniques is difficult in general because we cannot explicitly write down the analytical form of the estimators. This difficulty prevents the obtaining of useful insights from quantitative theoretical analysis. Thus, the replica theory [6] may provide a promising analytical tool in this area. Because our framework can treat only synthetic settings, we believe that the goal is to investigate precise asymptotic properties for a comprehensive range of parameters and to find some phenomena that would hold universally, such as novel phase transitions. However, this kind of exhaustive analysis is quite involved in practice, although obtaining an order parameter for one specific setting is not difficult. Thus we postpone this analysis as future work. Another research direction is the investigation of the dynamics of raw rVAMP using techniques such as the dynamical-functional theory [26, 36–38].

J. Stat. Mech. (2020) 093402

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 19J10711, 17H00764, and JST CREST Grant Number JPMJCR1912, Japan.

References

[1] Tibshirani R 1996 *J. R. Stat. Soc.* B **58** 267–88
[2] Meinshausen N and Bühlmann P 2010 *J. R. Stat. Soc. B Stat. Methodol.* **72** 417–73
[3] Homrighausen D and McDonald D J 2014 *Mach. Learn.* **97** 65–78
[4] Alon U, Barkai N, Notterman D A, Gish K, Ybarra S, Mack D and Levine A J 1999 *Proc. Natl Acad. Sci.* **96** 6745–50
[5] Bühlmann P and Van De Geer S 2011 *Statistics for High-Dimensional Data: Methods, Theory and Applications* (Berlin: Springer)
[6] Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and beyond: An Introduction to the Replica Method and its Applications* vol 9 (Singapore: World Scientific)
[7] Schniter P, Rangan S and Fletcher A K 2016 Vector approximate message passing for the generalized linear model *2016 50th Asilomar Conf. on Signals, Systems and Computers* (IEEE) pp 1525–9
[8] Rangan S, Schniter P and Fletcher A K 2019 *IEEE Trans. Inf. Theory* **65** 6664–84
[9] Malzahn D and Opper M 2003 *J. Mach. Learn. Res.* **4** 1151–73
[10] Malzahn D and Opper M 2003 A statistical mechanics approach to approximate analytical bootstrap averages *Advances in Neural Information Processing Systems* ed S Thrun, L K Saul and B Schölkopf (Cambridge, MA: The MIT Press) vol 16 pp 343–50
[11] Malzahn D and Opper M 2004 Approximate analytical bootstrap averages for support vector classifiers *Advances in Neural Information Processing Systems* ed L K Saul, Y Weiss and L Bottou (Cambridge, MA: The MIT Press) vol 17 pp 1189–96
[12] Opper M and Winther O 2001 *Phys. Rev. Lett.* **86** 3695
[13] Opper M and Winther O 2001 *Phys. Rev.* E **64** 056131
[14] Bolthausen E 2014 *Commun. Math. Phys.* **325** 333–66
[15] Cakmak B, Winther O and Fleury B H 2014 S-amp: approximate message passing for general matrix ensembles *2014 IEEE Information Theory Workshop (ITW 2014)* (IEEE) pp 192–6
[16] Kabashima Y 2003 *J. Phys. A: Math. Gen.* **36** 11111
[17] Donoho D L, Maleki A and Montanari A 2009 *Proc. Natl Acad. Sci.* **106** 18914–9
[18] Bayati M and Montanari A 2011 *IEEE Trans. Inf. Theory* **57** 764–85
[19] Javanmard A and Montanari A 2013 *Inf. Inference* **2** 115–44
[20] Ma J and Ping L 2017 *IEEE Access* **5** 2020–33
[21] Minka T P 2001 Expectation propagation for approximate bayesian inference *Proc. of the 17th Conf. on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers) pp 362–9
[22] Opper M and Winther O 2005 *J. Mach. Learn. Res.* **6** 2177–204
[23] Takahashi T and Kabashima Y 2020 arXiv:2001.02824
[24] Çakmak B and Opper M 2018 Expectation propagation for approximate inference: free probability framework *2018 IEEE Int. Symp. on Information Theory (ISIT)* (IEEE) pp 1276–80
[25] Çakmak B and Opper M 2019 Convergent dynamics for solving the tap equations of ising models with arbitrary rotation invariant coupling matrices *2019 IEEE Int. Symp. on Information Theory (ISIT)* (IEEE) pp 1297–301
[26] Çakmak B and Opper M 2019 *Phys. Rev.* E **99** 062140
[27] Obuchi T and Kabashima Y 2019 *J. Mach. Learn. Res.* **20** 1–33
[28] Takahashi T and Kabashima Y 2019 arXiv:1905.09545
[29] Hewitt E and Savage L J 1955 *Trans. Am. Math. Soc.* **80** 470–501
[30] Opper M and Winther O 2004 Variational linear response *Advances in Neural Information Processing Systems* ed L K Saul, Y Weiss and L Bottou (Cambridge, MA: The MIT Press) vol 17 pp 1157–64
[31] Golub G H and Van Loan C F 1996 *Matrix Computations* 3rd edn vol 3 (Baltimore, MD: John Hopkins University Press)
[32] Barbier J, Macris N, Maillard A and Krzakala F 2018 The mutual information in random linear estimation beyond iid matrices *2018 IEEE Int. Symp. on Information Theory (ISIT)* (IEEE) pp 1390–4
[33] Barbier J, Krzakala F, Macris N, Miolane L and Zdeborová L 2019 *Proc. Natl Acad. Sci.* **116** 5451–60

J. Stat. Mech. (2020) 093402

[34] Reeves G and Pfister H D 2016 The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact *2016 IEEE Int. Symp. on Information Theory (ISIT)* (IEEE) pp 665–9
[35] Gerbelot C, Abbara A and Krzakala F 2020 arXiv:2006.06581
[36] Cakmak B, Opper M, Winther O and Fleury B H 2017 Dynamical functional theory for compressed sensing *2017 IEEE Int. Symp. on Information Theory (ISIT)* (IEEE) pp 2143–7
[37] Martin P C, Siggia E and Rose H 1973 *Phys. Rev.* A **8** 423
[38] Eissfeller H and Opper M 1992 *Phys. Rev. Lett.* **68** 2094
[39] Kabashima Y 2008 Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels *J. Phys.: Conf. Ser.* **95** 012001
[40] Kabashima Y and Vehkaperä M 2014 Signal recovery using expectation consistent approximation for linear observations *2014 IEEE Int. Symp. on Information Theory* (IEEE) pp 226–30
[41] Qian J, Hastie T, Friedman J, Tibshirani R and Simon N 2013 Glmnet for Matlab http://www.stanford.edu/&tnqx223c;hastie/glmnet_matlab/
[42] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J *et al* 2011 *Found. Trends® Mach. Learn.* **3** 1–122
[43] Boyd S, Boyd S P and Vandenberghe L 2004 *Convex Optimization* (Cambridge: Cambridge University Press)

J. Stat. Mech. (2020) 093402