Cross validation in sparse linear regression with piecewise continuous nonconvex penalties and its acceleration

OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **52** (2019) 414003 (30pp)
https://doi.org/10.1088/1751-8121/ab3e89

Cross validation in sparse linear regression with piecewise continuous nonconvex penalties and its acceleration

Tomoyuki Obuchi[1] **and Ayaka Sakata**[2,3]

[1] Department of Mathematical & Computing Science, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo 152-8552, Japan
[2] Department of Statistical Inference & Mathematics, The Institute of Statistical Mathematics, Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[3] Department of Statistical Science, The Graduate University for Advanced Science (SOKENDAI), Hayama-cho, Kanagawa 240-0193, Japan

E-mail: obuchi@c.titech.ac.jp and ayaka@ism.ac.jp

Received 27 February 2019, revised 30 July 2019
Accepted for publication 27 August 2019
Published 18 September 2019

Abstract

We investigate the signal reconstruction performance of sparse linear regression in the presence of noise when piecewise continuous nonconvex penalties are used. Among such penalties, we focus on the smoothly clipped absolute deviation (SCAD) penalty. The contributions of this study are three-fold: we first present a theoretical analysis of a typical reconstruction performance, using the replica method, under the assumption that each component of the design matrix is given as an independent and identically distributed (i.i.d.) Gaussian variable. This clarifies the superiority of the SCAD estimator compared with ℓ_1 in a wide parameter range, although the nonconvex nature of the penalty tends to lead to solution multiplicity in certain regions. This multiplicity is shown to be connected to replica symmetry breaking in the spin-glass theory, and associated phase diagrams are given. We also show that the global minimum of the mean square error between the estimator and the true signal is located in the replica symmetric phase. Second, we develop an approximate formula efficiently computing the cross-validation error without actually conducting the cross-validation, which is also applicable to the non-i.i.d. design matrices. It is shown that this formula is only applicable to the unique solution region and tends to be unstable in the multiple solution region. We implement instability detection procedures, which allows the approximate formula to stand alone and resultantly enables

us to draw phase diagrams for any specific dataset. Third, we propose an annealing procedure, called nonconvexity annealing, to obtain the solution path efficiently. Numerical simulations are conducted on simulated datasets to examine these results to verify the consistency of the theoretical results and the efficiency of the approximate formula and nonconvexity annealing. The characteristic behaviour of the annealed solution in the multiple solution region is addressed. Another numerical experiment on a real-world dataset of Type Ia supernovae is conducted; its results are consistent with those of earlier studies using the ℓ_0 formulation. A MATLAB package of numerical codes implementing the estimation of the solution path using the annealing with respect to λ in conjunction with the approximate CV formula and the instability detection routine is distributed in Obuchi (2019 https://github.com/T-Obuchi/SLRpackage_AcceleratedCV_matlab).

Keywords: sparse linear regression, compressed sensing, replica method, cross-validation, nonconvex penalty

(Some figures may appear in colour only in the online journal)

1. Introduction

Variable selection problems ubiquitously appear in statistics and machine learning tasks. Although traditional statistical approaches to variable selection work well in principle [2], difficulties in computational efficiency and stability emerge owing to the largeness and high dimensionality of the datasets. To overcome this, the possibility of sparse estimation has been pursued for decades. Naive methods for sparse estimation yet require solving discrete optimisation problems, involving a serious computational difficulty, even in the simplest case of linear models [3]. Hence, certain relaxations or approximations are required for handling such large high-dimensional datasets.

A breakthrough was made with the least absolute shrinkage and selection operator (LASSO) [4]. Its basic idea is to relax the sparsity constraint by using ℓ_1 regularisation. The success of LASSO motivated the usage of ℓ_1 regularisation in many different contexts and models [5–7], leading to an ongoing innovation in signal and information processing [8–10].

Although LASSO has many attractive properties, the shrinkage introduced by the ℓ_1 regularisation results in a significant bias in regression coefficients. To solve this, some nonconvex penalties, such as the smoothly clipped absolute deviation (SCAD) penalty [11] and the minimax concave penalty (MCP) [12], have recently been proposed. Although the estimators under these regularisations have desirable properties such as unbiasedness and continuity [11], there exist some concerns about the stability and interpretability of the estimators because of the potential local minimums owing to the lack of convexity. Hence, investigations of typical performance of those estimators are desired.

Under this situation, one of the present authors recently analyzed the SCAD estimator performance in [13], using the replica method and the message passing technique [14–16]. It was shown that there exist two regions in the space of regularization parameters, and in one of them the estimator is uniquely and stably obtained. It was also shown that the estimator outperforms LASSO in the fit quality, and that the emergence of the two regions can be viewed as a phase transition involving replica symmetry breaking (RSB). The latter finding yields a nontrivial insight to the behaviour of local search algorithms, and it was demonstrated that

the convergence limit of the coordinate descent (CD) algorithm is closely related to the RSB transition and that a sufficient condition of the convergence, derived in [17], is not tight.

The above-mentioned analysis was limited to the data compression context, in which only the fit quality to a given dataset was considered important. However, with regard to certain applications of sparse estimation, such as compressed sensing [18, 19], the reconstruction performance of the true signal embedded in the data generation process is more important. The current study addresses this problem and conducts a quantitative analysis for the case where noise exists, while the noiseless limit is investigated in a separate study [20]. We provide phase diagrams with respect to regularization parameters derived using the replica method and discuss their implications to the reconstruction performance and the behaviour of local search algorithms. Moreover, we develop an approximate formula for efficiently computing the cross-validation (CV) error, which can be identified with the reconstruction error in our setting. The key results are summarised as follows:

(i) In the replica symmetric (RS) phase, a unique solution is stably obtained also in the signal reconstruction context.
(ii) The global minimum of the CV error is (presumably always) obtained in the RS phase.
(iii) Our approximate formula efficiently estimates the CV error, without actually conducting CV.

These imply that we need not to care about the RSB phase as long as our purpose is to obtain the model best reconstructing the true signal, and in the RS phase we can benefit from the proposed approximate CV formula enabling an efficient estimation of the reconstruction error. Below, we show the theoretical results supporting these messages.

The remaining of the paper is organised as follows: in the next section, our problem setting and an overview of the SCAD penalty are given; in section 3, the replica analysis result is shown without the derivation because the essential part is already given in [13, 20], and phase diagrams and plots of relevant quantities are shown; in section 4, the approximate formula of the CV error is derived; in section 5, numerical experiments are carried out on both simulated and real-world datasets to check the accuracy of the replica result and the approximate formula. The last section concludes the paper.

2. Problem settings

Suppose a data vector $\boldsymbol{y} \in \mathbb{R}^M$ is generated by the following linear process with a design matrix $A \in \mathbb{R}^{M\times N}$ and a signal vector $\boldsymbol{x}^0 \in \mathbb{R}^N$:

$$\boldsymbol{y} = A\boldsymbol{x}^0 + \boldsymbol{\Delta}, \tag{1}$$

where $\boldsymbol{\Delta}$ is a noise vector, the component of which is assumed to be an independent and identically distributed (i.i.d.) variable from the normal distribution with zero mean and variance σ_Δ^2, $\mathcal{N}(0,\sigma_\Delta^2)$. We denote our dataset as $D_M = \{\boldsymbol{y}, A\}$. In the context of compressed sensing, the design matrix A represents the measurement process, and we try to infer $\boldsymbol{x}^0$ given A and $\boldsymbol{y}$. The inference is herein formulated as a regularised linear regression, and the concrete form of our estimator is given by:

$$\hat{\boldsymbol{x}}(\eta, D_M) = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}||\boldsymbol{y} - A\boldsymbol{x}||_2^2 + J(\boldsymbol{x};\eta) \right\}, \tag{2}$$

where $J(\boldsymbol{x};\eta) = \sum_{i=1}^N J(x_i;\eta)$ is the regularisation inducing the estimator sparsity, and η is a set of regularisation parameters with a concrete form shown below. To quantify the fit quality of the estimator $\hat{\boldsymbol{x}}$ to the data $\boldsymbol{y}$, we introduce:

$$\epsilon_y(\hat{\boldsymbol{x}}|D_M) = \frac{1}{2M}||\boldsymbol{y} - A\hat{\boldsymbol{x}}||_2^2, \tag{3}$$

and call it the output mean squared error (MSE). We also introduce a MSE between the estimator and the true signal as:

$$\epsilon_x(\hat{\boldsymbol{x}}|\boldsymbol{x}^0) = \frac{1}{2N}||\hat{\boldsymbol{x}} - \boldsymbol{x}^0||_2^2, \tag{4}$$

which is termed input MSE, and these characterise the goodness of fit of our estimator $\hat{\boldsymbol{x}}$.

The purpose of this study is to compute the typical behaviour of ϵ_x and ϵ_y to obtain insights into the estimator behaviour; meanwhile, some other relevant quantities are also evaluated. The analytical techniques for achieving this purpose are explained in section 3, with a more detailed description on $\boldsymbol{x}^0$ and A.

2.1. SCAD regularisation

As a representative piecewise continuous nonconvex penalty, we investigate SCAD regularisation in this study. The parameter set consists of $\eta = \{\lambda, a\}$ $(a > 1)$, and the functional form is:

$$J(\theta;\eta) = \begin{cases} \lambda|\theta| & (|\theta| \leqslant \lambda) \\ -\frac{\theta^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)} & (\lambda < |\theta| \leqslant a\lambda) \\ \frac{(a+1)\lambda^2}{2} & (|\theta| > a\lambda) \end{cases}. \tag{5}$$

An illustration of this form is given as the left panel of figure 1. In the limit $a \to \infty$, the SCAD regularisation tends to be the ℓ_1 regularisation $J(\theta;\{\lambda, a\to\infty\}) \to \lambda|\theta|$, and correspondingly, the SCAD estimator converges to the LASSO one, allowing the comparison in a continuous manner. For later convenience, we term a the switching parameter, and λ the amplitude parameter.

To obtain an intuitive view for the SCAD estimator behaviour, we compute the one-dimensional case:

$$\hat{\theta}(w;\sigma_w^2,\eta) = \arg\min_{\theta}\left\{\frac{1}{2\sigma_w^2}(\theta - w)^2 + J(\theta;\eta)\right\}. \tag{6}$$

The solution is given by:

$$\hat{\theta}(w;\sigma_w^2,\eta) = V_{\mathrm{SCAD}}(w/\sigma_w^2;\sigma_w^2,\eta)S_{\mathrm{SCAD}}(w/\sigma_w^2;\sigma_w^2,\eta), \tag{7}$$

where

$$S_{\mathrm{SCAD}}(x;\sigma^2,\eta) = \begin{cases} x - \mathrm{sgn}(x)\lambda & \text{for } \lambda(1+\sigma^{-2}) \geqslant |x| > \lambda \\ x - \mathrm{sgn}(x)\frac{a\lambda}{a-1} & \text{for } a\lambda\sigma^{-2} \geqslant |x| > \lambda(1+\sigma^{-2}) \\ x & \text{for } |x| > a\lambda\sigma^{-2} \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

$$V_{\mathrm{SCAD}}(x;\sigma^2,\eta) = \begin{cases} \sigma^2 & \text{for } \lambda(1+\sigma^{-2}) \geqslant |x| > \lambda \\ \left(\sigma^{-2} - \frac{1}{a-1}\right)^{-1} & \text{for } a\lambda\sigma^{-2} \geqslant |x| > \lambda(1+\sigma^{-2}) \\ \sigma^2 & \text{for } |x| > a\lambda\sigma^{-2} \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

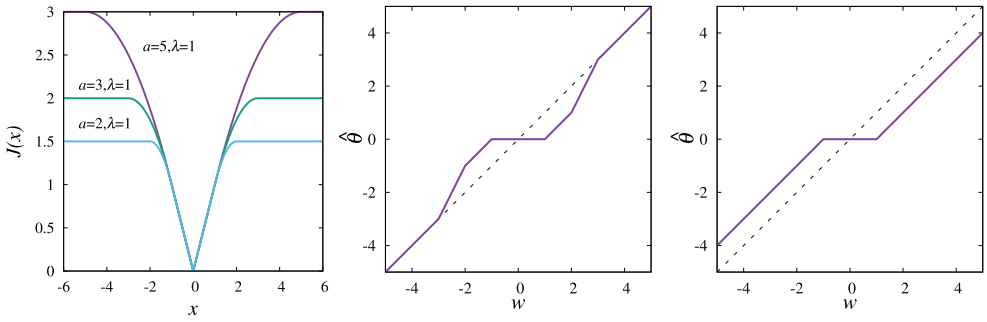

Figure 1. (Left) Shapes of the SCAD regularisations for some parameters. (Middle) Behaviour of the SCAD estimator (7) at $a = 3, \lambda = 1, \sigma_w^2 = 1$; the diagonal dashed line represents the OLS estimator. (Right) Behaviour of the LASSO estimator at $\lambda = 1$ for comparison with the SCAD; a shrinkage bias is clearly seen for a large w.

The middle panel of figure 1 presents an illustration of the estimator at $a = 3, \lambda = 1, \sigma_w^2 = 1$, which behaves as the LASSO estimator when $\lambda(1 + \sigma_w^{-2}) \geqslant |w| > \lambda$, and as the ordinary least square (OLS) estimator when $|w| > a\lambda\sigma_w^{-2}$. In the region $a\lambda\sigma_w^{-2} \geqslant |w| > \lambda(1 + \sigma_w^{-2})$, the estimator linearly transits between LASSO and OLS estimators.

The one-dimensional case estimator plays a key role in our analysis, because our original problem with high dimensionality is, eventually, reduced to an effective one-dimensional problem in the limit $N \to \infty$, termed *decoupling principle* in [21].

3. Macroscopic analysis

In this section, we provide order parameters and their determining equations of state (EOS). Associated phase diagrams are shown, and their implications on the performance and the computational stability of the SCAD estimator are also discussed.

For proceeding with the analysis, the ensemble of A and $\boldsymbol{x}^0$ is required to be fixed. We assume that A is a random matrix whose component is i.i.d. from $\mathcal{N}(0, M^{-1})$. The true signal $\boldsymbol{x}^0$ is also assumed to be a random number drawn from the independent Bernoulli–Gaussian distribution:

$$P(\boldsymbol{x}^0) = \prod_{i=1}^{N} \left\{ (1 - \rho_0)\delta(x_i^0) + \frac{\rho_0}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x_i^0)^2}{2\sigma_x^2} \right) \right\}. \tag{10}$$

We note that the i.i.d. assumption on A is crucial for completing the computation, while the choice of the distribution of $\boldsymbol{x}^0$ does not matter for the analytical tractability. We admit that this i.i.d. assumption on A is not necessarily realistic, but it provides a sufficiently nontrivial setup for our purpose. Although it is possible to extend the present analysis to certain other ensembles [22–29], we leave this as a future study.

In the following discussion, we consider the so-called thermodynamic limit $N \to \infty$, while keeping $\alpha \equiv M/N = O(1)$.

3.1. Outline of analysis

In order to avoid duplication with [13, 20], an outline of the analysis is presented here, instead of the EOS derivation.

Our analysis starts from defining Hamiltonian $\mathcal{H}$, partition function Z, and free energy density f as follows:

$$\mathcal{H}(\boldsymbol{x}|D_M) \equiv \frac{1}{2}||\boldsymbol{y} - A\boldsymbol{x}||_2^2 + J(\boldsymbol{x};\eta), \tag{11}$$

$$Z(\beta|D_M) \equiv \int \mathrm{d}\boldsymbol{x}\ \mathrm{e}^{-\beta\mathcal{H}(\boldsymbol{x}|D_M)}, \tag{12}$$

$$f(\beta|D_M) \equiv -\frac{1}{N\beta}\ln Z(\beta|D_M). \tag{13}$$

As seen from (2), the minimiser or the ground state of $\mathcal{H}$ corresponds to our estimator and, hence, we are interested in the $\beta \to \infty$ limit of the free energy density. The input and output MSEs can be computed from the free energy in this limit, by following a standard prescription. The free energy density becomes the primary object to be computed, and enjoys the self-averaging property, and the typical value thus converges to the averaged one $f(\beta|D_M) \to E_{\boldsymbol{y},A}\left[f(\beta|D_M)\right] \equiv f(\beta)$, where $E_{\boldsymbol{y},A}\left[\cdots\right]$ denotes the average over $\boldsymbol{y}$ and A.

Unfortunately, the average density $E_{\boldsymbol{y},A}\left[f(\beta|D_M)\right]$ is not analytically tractable. To overcome this, we employ the following identity:

$$E_{\boldsymbol{y},A}[\ln Z(\beta|D_M)] = \lim_{n\to 0}\frac{E_{\boldsymbol{y},A}[Z^n(\beta|D_M)] - 1}{n}. \tag{14}$$

However, the computation for general $n \in \mathbb{R}$ is still intractable; thus, we additionally assume that n is a positive integer, because $E_{\boldsymbol{y},A}[Z^n(\beta|D_M)]$ is analytically computable for $n \in \mathbb{N}$. Then, using an analytically continuable expression of $E_{\boldsymbol{y},A}[Z^n(\beta|D_M)]$ from $\mathbb{N}$ to $\mathbb{R}$, we evaluate $\lim_{n\to 0} E_{\boldsymbol{y},A}[Z^n(\beta|D_M)]$ at the final step. These procedures are termed replica method.

The final expression of $f(\beta)$ is given as an extremisation problem with respect to a number of parameters, called order parameters. The extremisation condition appears because of the limit $N \to \infty$, and yields EOS determining the values of the order parameters. The explicit formulas are given below. It should be noted that the following analysis is conducted only under the RS assumption, although RSB occurs in some parameter regions. This is because the RS analysis is sufficient for the present purpose of obtaining insights on the stability of the SCAD estimator. Beyond this purpose, the RSB analysis will provide further quantitative information about the estimator when many local minimums exist, which will be an interesting future direction.

3.2. Order parameters, equations of state, and stability condition

Here, we summarise the order parameters and EOS. In the RS level, our system is characterised by the following three-order parameters:

$$m = \frac{1}{N}\sum_i E_{\boldsymbol{y},A}\left[\left\langle x_i^0 x_i\right\rangle\right], \tag{15}$$

$$Q = \frac{1}{N}\sum_i E_{\boldsymbol{y},A}\left[\left\langle x_i^2\right\rangle\right], \tag{16}$$

$$q = \frac{1}{N}\sum_i E_{\boldsymbol{y},A}\left[\left\langle x_i\right\rangle^2\right], \tag{17}$$

where the angular brackets, $\langle\cdots\rangle$, denote the average over the Boltzmann distribution $P(\boldsymbol{x}|\beta, D_M) = \mathrm{e}^{-\beta\mathcal{H}(\boldsymbol{x}|D_M)}/Z(\beta|D_M)$. m is the overlap with the true signal $\boldsymbol{x}^0$ and is relevant to the reconstruction performance. Q and q both describe the powers (per element) of the estimator, but the latter takes into account the 'thermal' fluctuation that results from the introduction of β. These two quantities fall within the limit $\beta \to \infty$, but their infinitesimal difference yields an important contribution:

$$\chi = \beta(Q - q). \tag{18}$$

This is $O(1)$, even in the limit $\beta \to \infty$. Besides, we introduce the conjugate parameters of Q, χ, m as $\tilde{Q}, \tilde{\chi}, \tilde{m}$, respectively, and denote their sets as $\Omega = \{Q, \chi, m\}$ and $\tilde{\Omega} = \{\tilde{Q}, \tilde{\chi}, \tilde{m}\}$. The RS free-energy density in the limit $\beta \to \infty$ takes the following extremisation problem, with respect to Ω and $\tilde{\Omega}$:

$$f(\beta \to \infty) = \underset{\Omega,\tilde{\Omega}}{\mathrm{Extr}} \left\{ \frac{Q - 2m + \rho_0\sigma_x^2 + \alpha\sigma_\Delta^2}{2(1 + \chi/\alpha)} + m\tilde{m} - \frac{\tilde{Q}Q - \tilde{\chi}\chi}{2} + \overline{\frac{\xi(\sigma; \tilde{Q})}{2}} \right\}, \tag{19}$$

where:

$$L(h; \tilde{Q}) \equiv \min_x \left\{ \frac{\tilde{Q}}{2}x^2 - hx + J(x; \eta) \right\}. \tag{20}$$

$$\int Dz(\cdots) \equiv \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)(\cdots), \tag{21}$$

$$\xi(\sigma; \tilde{Q}) \equiv 2\int Dz\, L(\sigma z; \tilde{Q}), \tag{22}$$

and $\overline{\cdots}$ represents the average over σ, whose distribution is:

$$P_\sigma(\sigma) = (1 - \rho)\delta(\sigma - \sigma_-) + \rho\delta(\sigma - \sigma_+), \tag{23}$$

$$\sigma_- = \sqrt{\tilde{\chi}}, \tag{24}$$

$$\sigma_+ = \sqrt{\tilde{\chi} + \tilde{m}^2\sigma_x^2}. \tag{25}$$

The minimiser of (20) is the solution of the one-dimensional problem (7), with $\sigma_w^2 \to \tilde{Q}^{-1}$ and $w \to h/\tilde{Q}$, thus we can denote it as:

$$x^*(h; \tilde{Q}^{-1}) = V_{\mathrm{SCAD}}(h; \tilde{Q}^{-1}, \eta) S_{\mathrm{SCAD}}(h; \tilde{Q}^{-1}, \eta). \tag{26}$$

The extremisation condition in (19) yields EOS as:

$$\chi = \int Dz \overline{\left.\frac{\partial x^*(h; \tilde{Q}^{-1})}{\partial h}\right|_{h=\sigma z}} = \frac{1}{\tilde{Q}} \left\{ \hat{\rho} + \frac{\frac{1}{a-1}}{\tilde{Q} - \frac{1}{a-1}} \overline{\xi_4(\sigma)} \right\}, \tag{27a}$$

$$Q = \int Dz \overline{(x^*(\sigma z; \tilde{Q}^{-1}))^2} = \overline{\left\{ \frac{\xi_1(\sigma)}{\tilde{Q}} + \frac{\xi_2(\sigma)}{\tilde{Q} - \frac{1}{a-1}} + \frac{\xi_3(\sigma)}{\tilde{Q}} \right\}}, \tag{27b}$$

$$m = \rho\tilde{m}\sigma_x^2 \int Dz \frac{\partial x^*(h;\tilde{Q}^{-1})}{\partial h}\bigg|_{h=\sigma_+ z} = \rho_0\sigma_x^2 \left\{ \mathrm{erfc}(\theta_1(\sigma_+)) + \frac{\frac{1}{a-1}\xi_4(\sigma_+)}{\tilde{Q}-\frac{1}{a-1}} \right\}, \tag{27c}$$

$$\tilde{\chi} = \frac{1}{\alpha}\frac{Q - 2m + \rho_0\sigma_x^2 + \alpha\sigma_\Delta^2}{(1+\chi/\alpha)^2}, \tag{27d}$$

$$\tilde{Q} = \frac{1}{1+\chi/\alpha}, \tag{27e}$$

$$\tilde{m} = \frac{1}{1+\chi/\alpha}, \tag{27f}$$

where

$$\theta_1(\sigma) = \lambda/(\sqrt{2}\sigma), \tag{28a}$$

$$\theta_2(\sigma) = \lambda(1+\tilde{Q})/(\sqrt{2}\sigma), \tag{28b}$$

$$\theta_3(\sigma) = a\lambda\tilde{Q}/(\sqrt{2}\sigma), \tag{28c}$$

$$\hat{\rho} = \overline{\mathrm{erfc}(\theta_1(\sigma))}, \tag{28d}$$

$$\xi_1(\sigma) = \frac{\sigma^2}{\tilde{Q}}\Big[-\frac{2\theta_1(\sigma)}{\sqrt{\pi}}\Big(\mathrm{e}^{-\theta_1^2(\sigma)} + (\tilde{Q}-1)\mathrm{e}^{-\theta_2^2(\sigma)}\Big) + (1+2\theta_1^2(\sigma))\{\mathrm{erfc}(\theta_1(\sigma)) - \mathrm{erfc}(\theta_2(\sigma))\}\Big], \tag{28e}$$

$$\xi_2(\sigma) = \frac{\sigma^2}{\tilde{Q}-\frac{1}{a-1}}\Big[\frac{2}{\sqrt{\pi}}\Big\{\theta_2(\sigma)\mathrm{e}^{-\theta_2^2(\sigma)} - \theta_3(\sigma)\mathrm{e}^{-\theta_3^2(\sigma)} - \frac{2\theta_3(\sigma)}{\tilde{Q}(a-1)}\Big(\mathrm{e}^{-\theta_2^2(\sigma)} - \mathrm{e}^{-\theta_3^2(\sigma)}\Big)\Big\} + \Big\{1 + 2\Big(\frac{\theta_3(\sigma)}{\tilde{Q}(a-1)}\Big)^2\Big\}\xi_4(\sigma)\Big], \tag{28f}$$

$$\xi_3(\sigma) = \frac{\sigma^2}{\tilde{Q}}\Big[\frac{2\theta_3(\sigma)}{\sqrt{\pi}}\mathrm{e}^{-\theta_3^2(\sigma)} + \mathrm{erfc}(\theta_3(\sigma))\Big], \tag{28g}$$

$$\xi_4(\sigma) = \mathrm{erfc}(\theta_2(\sigma)) - \mathrm{erfc}(\theta_3(\sigma)). \tag{28h}$$

The SCAD regularisation divides the domain of definition into some analytic components, and $\{\theta_i\}_{i=1}^3$ are the corresponding boundary values for z for the integration $\int Dz(\cdots)$. The parameter $\hat{\rho}$ is the density of non-zero components in the estimate.

Using the solution of EOS, the input and output MSEs can be expressed as:

$$\epsilon_x = \frac{1}{2}\left(\rho_0\sigma_x^2 - 2m + Q\right), \tag{29}$$

$$\epsilon_y = \frac{1}{2}\tilde{\chi}. \tag{30}$$

Furthermore, we additionally quantify the reconstruction performance of the support of the true signal. Denoting the support or *active set* of $\boldsymbol{x}$ as $S_A(\boldsymbol{x}) = \{i|x_i \neq 0\}$, we introduce the true positive rate $TP(\boldsymbol{x}|\boldsymbol{x}_0) = \frac{|S_A(\boldsymbol{x}) \cap S_A(\boldsymbol{x}^0)|}{|S_A(\boldsymbol{x}^0)|}$ and the false positive rate $FP(\boldsymbol{x}|\boldsymbol{x}_0) = \frac{|S_A(\boldsymbol{x}) \cap S_A^c(\boldsymbol{x}^0)|}{|S_A^c(\boldsymbol{x}^0)|}$, where S^c denotes the complement set of S. These are expressed by using the solution of EOS as:

$$TP = \int Dz \ \left|x^*(\sigma_+ z; \tilde{Q}^{-1})\right|_0 = \mathrm{erfc}(\theta_1(\sigma_+)), \tag{31}$$

$$FP = \int Dz \ \left|x^*(\sigma_- z; \tilde{Q}^{-1})\right|_0 = \mathrm{erfc}(\theta_1(\sigma_-)), \tag{32}$$

where $|x|_0$ expresses ℓ_0 operator giving 0 if $x = 0$ and 1 otherwise. Following the standard analysis [13], we can derive the stability condition of the RS solution, called de Almeida–Thouless (AT) condition [30]. The derivation of our specific case is already given in [13] and we just quote the resultant expression:

$$\begin{aligned} &\frac{1}{\alpha\left(1+\chi/\alpha\right)^2} \overline{\int Dz \left(\left.\frac{\partial x^*(h;\tilde{Q}^{-1})}{\partial h}\right|_{h=\sigma z}\right)^2} \\ &= \frac{1}{\alpha\left(1+\chi/\alpha\right)^2}\left[\frac{\hat{\rho}}{\tilde{Q}^2} + \left\{\left(\frac{1}{\tilde{Q}-\frac{1}{a-1}}\right)^2 - \frac{1}{\tilde{Q}^2}\right\}\overline{\xi_4(\sigma)}\right] < 1. \end{aligned} \tag{33}$$

Apart from the AT condition, we also notice that the RS solution does not exist when the switching nonconvexity parameter a is small. This is because (20) tends to have no solution in the small a limit, leading to the following existence condition:

$$\tilde{Q} - \frac{1}{a-1} \geqslant 0. \tag{34}$$

These provide sufficient information for the following analyses.

3.3. Phase diagram

In this subsection, we show the phase diagrams in the λ–a plane for a wide range of parameters. We introduce three boundaries: the first one, derived from (33), is the AT line $a_{\mathrm{AT}}(\lambda)$ below which the RS solution is unstable; the second, derived from (34), is the existence limit of the RS solution $a_{\mathrm{RS}}(\lambda)$, below which the RS solution does not exist; and the third, $a_{\mathrm{IMSE}}(\lambda)$ represents the minimum point of the input MSE ϵ_x, when sweeping λ given a. For clarity, the variance of the non-zero components of $\boldsymbol{x}^0$ is fixed as $\sigma_x^2 = 1/\rho_0$, setting the signal power per component unity, in average, $\sum_{i=1}^N \left(x_i^0\right)^2 / N \approx 1$.

First, we compare the phase diagrams for different noise strengths σ_Δ^2 at $\alpha = 0.5$ and $\rho_0 = 0.2$ in figure 2. We plot $a_{\mathrm{AT}}, a_{\mathrm{RS}}$, and a_{IMSE} by blue, red, and green lines, respectively. The green diamond represents the location of the global minimum of ϵ_x under the RS assumption. A useful finding concerning this is that the location is above the AT line, which is always the case as far as we have examined, and some additional evidences are later given in figures 3 and 4. In the right panel of figure 2, the green diamond is not shown, because the input MSE continuously decreases as a grows, implying that the global minimum of ϵ_x is obtained at the LASSO limit $a \to \infty$. These imply that the best reconstruction performance of the true signal is always obtained in the RS phase, which is one of main claims of this study. Admittedly,

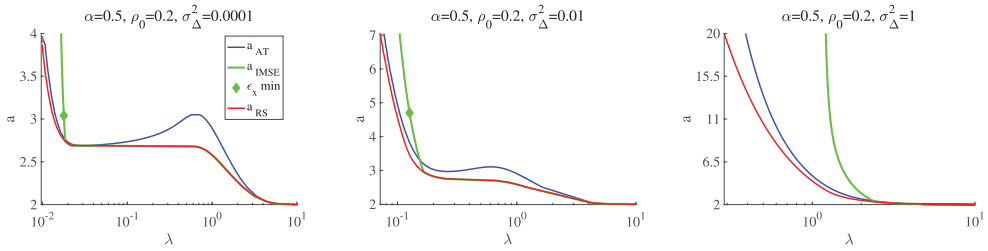

Figure 2. Phase diagrams in the λ–a plane for different noise strengths ($\sigma_\Delta^2 = 10^{-4}$ (left), 10^{-2} (middle), 1 (right)) at $\alpha = 0.5$ and $\rho_0 = 0.2$. The blue, green, and red lines denote $a_{\rm AT}$, $a_{\rm IMSE}$, and $a_{\rm RS}$, respectively. The green diamond represents the location of the global minimum of the input MSE ϵ_x. For the right panel of the largest noise case $\sigma_\Delta^2 = 1$, there seems to be no global minimum of ϵ_x for finite a (located in the LASSO limit $a \to \infty$).

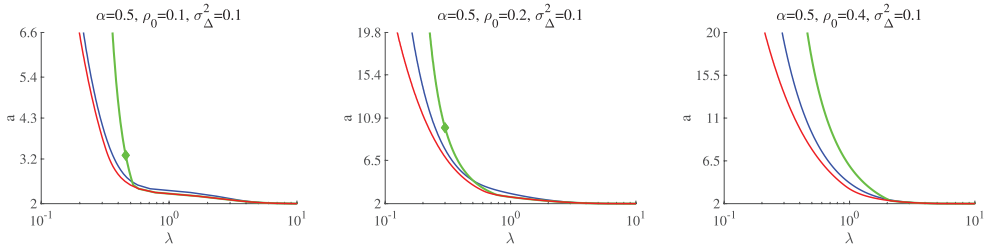

Figure 3. λ–a phase diagrams for different densities of non-zero components ($\rho_0 = 0.1$ (left), 0.2 (middle), 0.4 (right)) at $\alpha = 0.5$ and $\sigma_\Delta^2 = 0.1$. The lines and markers have the same meaning as figure 2. As ρ_0 increases, the location of the minimum of ϵ_x tends to be at larger values of a.

there is a possibility that the true global minimum exists in the RSB phase, and the green diamond just represents a local minimum. Our present analysis does not exclude this possibility. To clarify this point, further quantitative analysis in the RSB framework is required, but this is beyond the scope of this study.

Another interesting observation in figure 2 is the re-entrant phase transition concerning λ in relatively small a regions for the weak noise cases (left and middle panels). For example at $a = 2.8$ in the left panel, when decreasing λ from a large enough value, we first go across the rightmost branch of $a_{\rm AT}$ around $\lambda \approx 1$ and enter into the RSB phase from the RS phase; further decreasing λ we meet the middle branch of $a_{\rm AT}$ around $\lambda \approx 0.1$ and thus re-enter into the RS phase; still decreasing λ we hit the leftmost branch of $a_{\rm AT}$ around $\lambda \approx 0.01$ and we are eventually in the RSB phase. Although the physical reason of the emergence of the re-entrance is not clear, it seems to only exist in the weak noise region. We also note that the AT line, $a_{\rm AT}$, is always located above $a_{\rm RS}$. The solution vanishment in the low λ region is thus an artefact of the RS assumption, and the corresponding parameter regions should be described by the RSB solution.

Next, we check the ρ_0 dependence of the phase structures. Phase diagrams at $\alpha = 0.5$ and a moderate noise level $\sigma_\Delta^2 = 0.1$ are shown in figure 3. Although the basic structure does not change from figure 2, the re-entrant transitions in the weak noise cases disappear. As ρ_0 increases, the minimum location of ϵ_x increases along the a-axis, and for the large ρ_0 (right

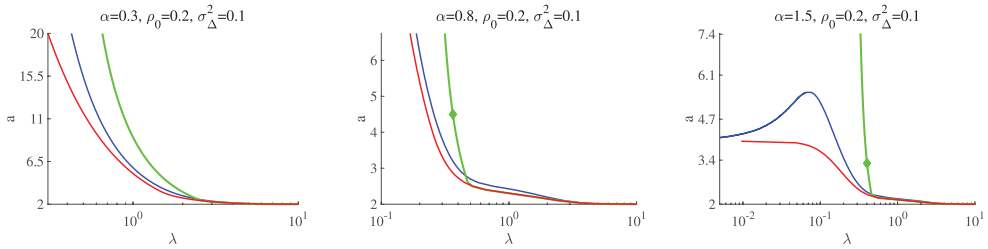

Figure 4. Phase diagrams for different ratios of the dataset size to the model dimensionality ($\alpha = 0.3$ (left), 0.8 (middle), 1.5 (right)) at $\rho_0 = 0.2$ and $\sigma_\Delta^2 = 0.1$. The lines and markers have the same meaning as figure 2.

panel) the green diamond tends to disappear at finite values of a, implying the LASSO limit yields the minimum of ϵ_x as the strong noise case.

The last phase diagrams are given for checking the α dependence. Phase diagrams for $\alpha = 0.3, 0.8, 1.5$ at $\rho_0 = 0.2$ and $\sigma_\Delta^2 = 0.1$ are shown in figure 4. As seen in the left panel, if the value of α is close to that of ρ_0, the larger a tends to give smaller values of the input MSE, as in the right panel of figure 3. In contrast to the other diagrams of the underdetermined case ($\alpha < 1$), the right panel of the $\alpha = 1.5$ case shows a particular behaviour, as both the AT line and the RS existence limit tend to converge to certain finite values of a in the $\lambda \to 0$ limit. Hence, the whole λ region becomes RS at sufficiently large but finite a values.

In all the phase diagrams shown above, the minimum value of a is fixed to be 2. This is because the RS solution cannot describe the region $a < 2$. There is a simple reason for this. According to the argument of the approximate message passing technique [13, 20], the effective one-dimensional problem (20) corresponds to the following marginal distribution:

$$P_i(x_i) \propto \lim_{\beta\to\infty} \mathrm{e}^{-\beta\left\{\frac{1}{2}\left(\sum_{\mu=1}^{M}\frac{A_{\mu i}^2}{1+\chi_\mu}\right)x_i^2 - h_i x_i + J(x_i;\eta)\right\}}, \tag{35}$$

where the minimiser of (20) corresponds to the location of x_i, at which the measure concentrates in the limit $\beta \to \infty$. The factor $\left(\sum_{\mu=1}^{M}\frac{A_{\mu i}^2}{1+\chi_\mu}\right)$ corresponds to $\tilde{Q}$, while χ_μ is a non-negative quantity related to χ in the RS solution. This means that, under the assumption of $A_{\mu i} \sim \mathcal{N}(0, 1/M)$, $\tilde{Q}$ is bounded as:

$$\tilde{Q} = \left(\sum_{\mu=1}^{M}\frac{A_{\mu i}^2}{1+\chi_\mu}\right) \leqslant \sum_{\mu=1}^{M} A_{\mu i}^2 \approx 1. \tag{36}$$

Combining this with (34), we find that the condition $a \geqslant 2$ is necessary for the existence of the RS solution. The merging behaviour of the three lines to the $a = 2$ line, as λ grows in the phase diagrams, well matches to this condition. Although $a = 2$ is a known critical value [11, 31], the above argument provides another perspective from a different viewpoint. We also note that this non-existence of the RS solution does not mean the non-existence of the SCAD estimators. Actually, numerical experiments easily show that the estimators take non-trivial values in the region $a < 2$, and they tend to show strong multiplicity and dependency on the initial condition. To analyse the behaviour of those estimators, we need to consider the RSB solution, but it is beyond the present purpose as already declared.

3.4. Receiver operating characteristic curve

To characterise the reconstruction performance of the true signal's support, we employ the so-called receiver operating characteristic (ROC) curve. The ROC curve is a plot of TP (31) against FP (32). The best ROC curve goes through the point $(TP, FP) = (1, 0)$. Accordingly, to quantify 'optimality' of the points on a ROC curve, we use the following quantity:

$$R = (TP - 1)^2 + (FP - 0)^2. \tag{37}$$

Thus, the smallest value of R defines the 'optimal' point of the ROC curve. This easy-to-use quantity is commonly applied as a criterion, followed here.

First, ROC curves when sweeping λ at given values of a are plotted in the left panel of figure 5. The other parameters are $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.2, 0.1)$. The curves are not monotonic and tend to change in the small λ region sharply, but the locations of the minimums of R, depicted by filled magenta circles, tend to be in the monotonic region. To compare the values of the R minimums, we plot them against a in the right panel. The global minimum is located at $a \approx 10$, which matches to the minimum location of ϵ_x, depicted by the green diamond in the middle panel of figure 3. This suggests a possibility that minimising ϵ_x also approximately minimises the error in the variable selection.

To scrutinise the possibility, we show ROC curves when adaptively changing the non-convexity parameters along the $a_{\mathrm{IMSE}}(\lambda)$ line in the λ–a phase diagrams: the upper panels of figure 6 are the ROC curves for $(\alpha, \sigma_\Delta^2) = (0.5, 0.0001), (0.5, 0.1)$, and $(1.5, 0.1)$ at $\rho_0 = 0.2$. The corresponding plots of R and ϵ_x against a are also shown in the lower panels. These figures show that the minimums of ϵ_x are actually close to that of R. As far as we have searched, similar tendency holds in other parameters. These fully support the above-mentioned possibility. Such a nice property is absent in LASSO [32], and the minimum point of ϵ_x in LASSO tends to give a solution with rather large FP[4]. Hence in the reconstruction performance of the true model, the SCAD estimator is superior to the LASSO one.

Readers may doubt the effectiveness of this statement, because the input MSE ϵ_x cannot be computed for realistic settings with unknown true signals. As explained later, the input MSE has a simple linear relation to the generalisation error estimated by CV, when rows of the design matrix are uncorrelated with each other. Hence, we may minimise the CV error instead of the input MSE.

Figures 2–4 show that in some parameter regions there seems to be no global minimum of the input MSE at finite a, as for the strong noise case of $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.2, 1)$ in figure 2 and the dense signal case $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.4, 0.1)$ in figure 3. To examine those cases, we plot relevant quantities when changing a and λ again along the $a_{\mathrm{IMSE}}(\lambda)$ line in figure 7. The left panels show the plots of TP and FP against a, the middle panels display the plots of R and ϵ_x against a, and the right panels give the associated ROC curves. All the quantities of TP, FP, R, and ϵ_x show monotonic behaviours with respect to a, and seem to converge to finite values in the LASSO limit $a \to \infty$. The minimums of R and ϵ_x would be thus obtained by LASSO, with the optimised λ. These observations imply that LASSO is sufficient for difficult cases with strong noises or dense signals. This also implies that it is difficult to determine a good value of a to find the least ϵ_x solution given a dataset prior to actual analyses, because it strongly depends on the noise strength or the signal density.

[4] In [32], essentially the same analysis is done for LASSO, but a wrong terminology is used. The quantity R is termed Youden's index in that study, but it is contradictory to the conventional terminology. Youden's index is another similar but different criterion for choosing an 'optimal' point on ROC curve.

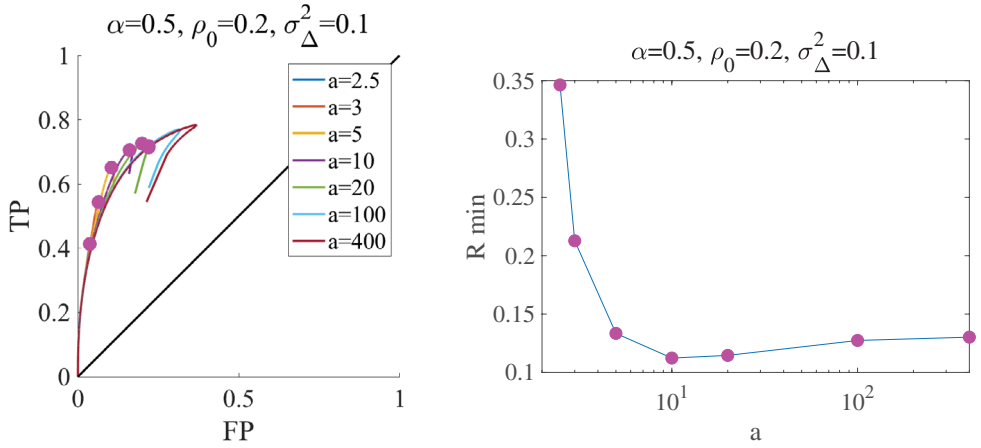

Figure 5. (Left) ROC curves when sweeping λ at different values of a for $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.2, 0.1)$. The minimums of R on the curves, for given values of a are plotted by filled magenta circles. (Right) The minimum value of R, when sweeping λ given a, is plotted against a. The global minimum tends to be located around $a \approx 10$, which matches to the minimum location of ϵ_x depicted by the green diamond in the middle panel of figure 3.

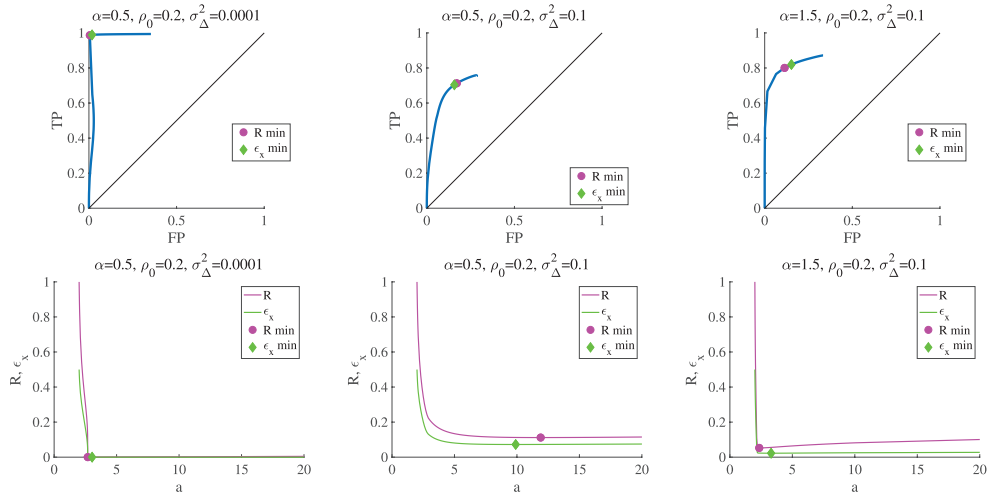

Figure 6. (Upper) ROC curves when changing the nonconvexity parameters along the $a_{\mathrm{IMSE}}(\lambda)$ line in the λ–a phase diagrams for $(\alpha, \sigma_\Delta^2) = (0.5, 0.0001)$ (left), $(0.5, 0.1)$ (middle), and $(1.5, 0.1)$ (right) at $\rho_0 = 0.2$. The other parameters are $(\rho_0, \sigma_\Delta^2) = (0.2, 0.1)$. The minimum values of R and ϵ_x are depicted by filled magenta circle and green diamond, respectively. (Lower) Plots of R and ϵ_x against a along the a_{IMSE} line. The parameters are identical to the corresponding upper panels.

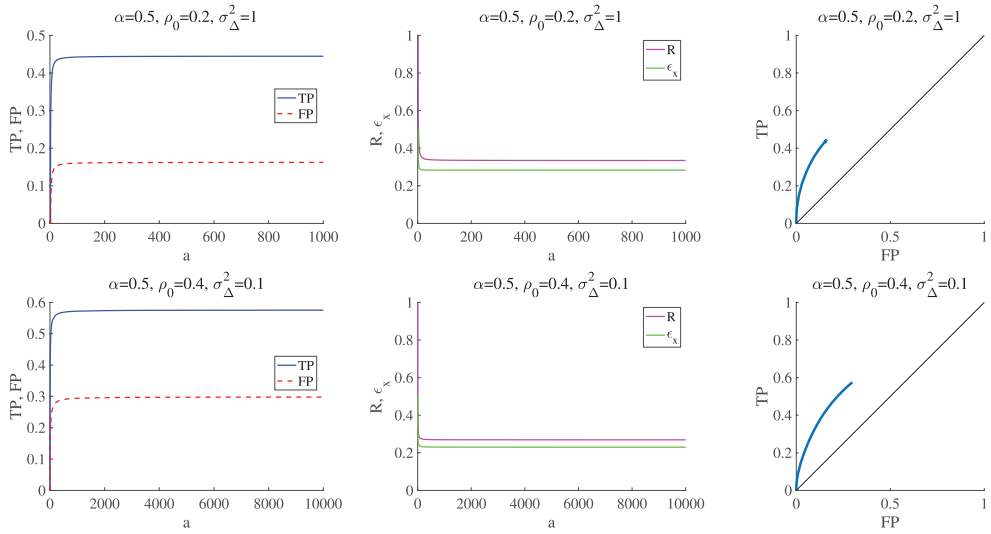

Figure 7. (Left) Plots of TP and FP against a along the $a_{\mathrm{IMSE}}(\lambda)$ line. (Middle) Plots of R and ϵ_x against a along the $a_{\mathrm{IMSE}}(\lambda)$ line. (Right) The associated ROC curves. The upper row is for the strong noise case of $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.2, 1)$ corresponding to the right panel of figure 2, while the lower row is for the dense signal case $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.4, 0.1)$ corresponding to the right panel of figure 3. In both the cases, all the quantities of TP, FP, R, and ϵ_x behave monotonically with respect to a, and seem to converge to finite values in the LASSO limit $a \to \infty$.

4. Approximate formula for cross-validation

In this section, we derive an approximate formula for the leave-one-out (LOO) CV error. If the dataset size M is large enough, the difference between the estimators of the full and LOO datasets is considered to be small, and it is expected that those two estimators can be connected in a perturbative manner. We concretise this idea below.

The estimator without the μth data in (2) is, hereafter, termed μth LOO estimator, and the explicit formula is given by:

$$\hat{\boldsymbol{x}}^{\backslash\mu}(\eta, D_M) = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2} \sum_{\nu(\neq\mu)} \left(y_\nu - \sum_i A_{\nu i} x_i \right)^2 + J(\boldsymbol{x}; \eta) \right\}. \tag{38}$$

The LOO CV error (LOOE) is accordingly defined as:

$$\epsilon_{\mathrm{LOO}}(\eta, D_M) = \frac{1}{2M} \sum_{\mu=1}^{M} (y_\mu - \boldsymbol{a}_\mu^\top \hat{\boldsymbol{x}}^{\backslash\mu}(\eta, D_M))^2, \tag{39}$$

where $\boldsymbol{a}_\mu^\top = (A_{\mu 1}, \cdots, A_{\mu N})$ is the μth row vector of A. The LOOE is an estimator for the generalisation error or extra-sample error, defined as:

$$\epsilon_{\mathrm{g}}(\eta, D_M) \equiv \int \mathrm{d}y_{\mathrm{new}} \mathrm{d}\boldsymbol{a}_{\mathrm{new}} P(y_{\mathrm{new}}, \boldsymbol{a}_{\mathrm{new}}) \frac{1}{2} (y_{\mathrm{new}} - \boldsymbol{a}_{\mathrm{new}}^\top \hat{\boldsymbol{x}}(\eta, D_M))^2, \tag{40}$$

where $\{y_{\rm new}, \boldsymbol{a}_{\rm new}\}$ represents a new data sample, and $P(y_{\rm new}, \boldsymbol{a}_{\rm new})$ denotes its distribution. In our setting, the distribution corresponds to the i.i.d. process described around (1), and it is analytically shown that $\epsilon_{\rm g}$ has a direct connection to ϵ_x as:

$$\epsilon_{\rm g}(\eta, D_M) = \frac{1}{\alpha}\epsilon_x\left(\hat{\boldsymbol{x}}\left(\eta, D_M\right)|\boldsymbol{x}^0\right) + \frac{1}{2}\sigma_\Delta^2. \tag{41}$$

Hence, we can estimate the input MSE from the LOOE. Note that the sufficient condition for (41) is that both the noise components and the rows of the design matrix are zero-mean and uncorrelated; correlations in the signal vector $\boldsymbol{x}^0$ may exist because they do not affect (41).

Owing to sparse priors, the variables in the estimator are separated in two types. Some variables are set to zero and the others take non-zero values. We call the former *inactive* variables and the latter *active* variables. The index set of the inactive variables, or inactive set, is denoted by $S_I = \{i|\hat{x}_i = 0\}$, while one of the active variables, or active set, is $S_A = \{i|\hat{x}_i \neq 0\}$. The active (inactive) components of a vector $\boldsymbol{x}$ are formally expressed as $\boldsymbol{x}_{S_A}(\boldsymbol{x}_{S_I})$. For any matrix X, we use double subscripts in the same manner and introduce the symbol $*$ meaning all the components in the respective dimension. For example, for an $N \times N$ matrix X, $X_{S_A S_I}$ and X_{*S_I} denote X's sub-matrices having row components of S_A and all, respectively, while their column components are commonly of S_I.

4.1. Derivation

The basic assumption to derive the approximate formula is that the active set is 'common' between the full and LOO estimators. Although this assumption is literally not true, we numerically confirmed that this approximately holds in the RS region. In other words, the change of the active set is small enough compared to the size of the active set itself, when considering the LOO operation under the situation with large N and M. Moreover, in the LASSO case, it has been shown that the contribution of the active set change vanishes in a limit $N, M \to \infty$, keeping $\alpha = M/N = O(1)$ [32]. It is expected that the same holds in the present problem. Hence, we adopt this assumption in the following derivation. We also note that this assumption is not applicable to the RSB region.

Once assuming the active set is known and common between the full and LOO systems, we can easily get the determining equations for the coefficients in the active set, by differentiating the cost function with respect to $\boldsymbol{x}_{S_A}$, yielding:

$$\left(\left(A_{*S_A}\right)^\top A_{*S_A}\right)\hat{\boldsymbol{x}}_{S_A} - \left(A_{*S_A}\right)^T \boldsymbol{y} + \nabla J(\hat{\boldsymbol{x}}_{S_A}; \eta) = 0, \tag{42}$$

for the full system and:

$$\left(\left(A^{\setminus\mu}_{*S_A}\right)^\top A^{\setminus\mu}_{*S_A}\right)\hat{\boldsymbol{x}}^{\setminus\mu}_{S_A} - \left(A^{\setminus\mu}_{*S_A}\right)^T \boldsymbol{y}^{\setminus\mu} + \nabla J(\hat{\boldsymbol{x}}^{\setminus\mu}_{S_A}; \eta) = 0, \tag{43}$$

for the LOO system.

Let us denote $\boldsymbol{d} = \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{\setminus\mu}$ and expand (43) with respect to $\boldsymbol{d}$ up to the first order. Erasing some terms using (42), and solving the remaining expression with respect to $\boldsymbol{d}$, we obtain an equation of $\boldsymbol{d}_{S_A}$ as:

$$\boldsymbol{d}_{S_A} \approx (y_\mu - \boldsymbol{a}_\mu^\top \hat{\boldsymbol{x}})\left(\left(A^{\setminus\mu}_{*S_A}\right)^\top A^{\setminus\mu}_{*S_A} + \left(\partial^2 J(\hat{\boldsymbol{x}}_{S_A}; \eta)\right)_{S_A S_A}\right)^{-1} \boldsymbol{a}_\mu. \tag{44}$$

Using (44), we can connect the residuals of the LOO and full systems as:

$$\begin{aligned} y_\mu - \boldsymbol{a}_\mu^\top \hat{\boldsymbol{x}}^{\setminus\mu} &= y_\mu - \boldsymbol{a}_\mu^\top (\hat{\boldsymbol{x}} - \boldsymbol{d}) \\ &\approx \left(1 + (\boldsymbol{a}_\mu)_{S_A}^\top \left(\left(A^{\setminus\mu}_{*S_A} \right)^\top A^{\setminus\mu}_{*S_A} + \left(\partial^2 J(\hat{\boldsymbol{x}}_{S_A}; \eta) \right)_{S_A S_A} \right)^{-1} (\boldsymbol{a}_\mu)_{S_A} \right) (y_\mu - \boldsymbol{a}_\mu^\top \hat{\boldsymbol{x}}). \end{aligned} \tag{45}$$

Using the relation $\left(\left(A^{\setminus\mu} \right)^\top A^{\setminus\mu} \right) = \left(A^\top A - \boldsymbol{a}_\mu \boldsymbol{a}_\mu^\top \right)$ and the Woodbury matrix inversion formula, we finally get

$$\epsilon_{\mathrm{LOO}} \approx \frac{1}{2M} \sum_{\mu=1}^{M} \Theta_\mu \left(y_\mu - \boldsymbol{a}_\mu^\top \hat{\boldsymbol{x}} \right)^2, \tag{46}$$

where

$$\Theta_\mu = \left(1 - (\boldsymbol{a}_\mu)_{S_A}^\top \left((A_{*S_A})^\top A_{*S_A} + \left(\partial^2 J(\hat{\boldsymbol{x}}_{S_A}; \eta) \right)_{S_A S_A} \right)^{-1} (\boldsymbol{a}_\mu)_{S_A} \right)^{-2}. \tag{47}$$

The righthand side of (46) can be computed only from the full solution, enabling an approximate evaluation of the LOOE, without literally conducting CV. The error bar can be put as the standard deviation among all the terms in (46) divided by $\sqrt{M}$[5]. This is convincing because each term of (46) gives an independent estimator to the generalisation error (40) and hence its error bar can be given as the standard error. Numerical experiments below show that this definition gives a reasonable error bar.

In the case of LASSO, the Hessian of the penalty term is identically zero, $\partial^2 J_{\mathrm{LASSO}} = 0$, meaning that (46) comes back to the 'approximation 1' in [32]. For SCAD, the Hessian takes the following form:

$$\left(\partial^2 J_{\mathrm{SCAD}}(\hat{\boldsymbol{x}}; \eta = \{\lambda, a\}) \right)_{ij} = \frac{1}{1-a} \delta_{ij} I(\lambda < |\hat{x}_i| \leqslant a\lambda), \tag{48}$$

where $I(\text{statement})$ denotes the indicator function, giving 1 if the statement is true and 0 otherwise.

5. Numerical experiments and numerical codes

Here we present numerical experiments. To obtain the SCAD estimator, we use the CD algorithm because it is common and stable. We implement this using C language, while the approximate CV formula is implemented as a raw code in MATLAB®. Hence, it is not necessarily fair to compare the computational time in the literal and approximate CVs, which are computed by (39) and (46) respectively, conducted below as a part of experiments. However, even in this comparison there is a meaningful difference in the computational time. When showing the computational time, we fix our experimental environment which uses a single CPU of 3.3 GHz Intel Core i7.

In a single step of the CD algorithm, we update all the components of $\boldsymbol{x}$ in a random order. To judge the convergence of the CD algorithm, we monitor the difference between the estimate $\hat{\boldsymbol{x}}^{(t)}$ at the step t and the previous one $\hat{\boldsymbol{x}}^{(t-1)}$. If all the component-wise differences

[5] Note that the main text description about the error bar of the approximate CV error in [32] is inconsistent with this, but this one is the correct one. Although the text description is incorrect, the experimentally reported error bars in [32] are correct and consistent with this.

$\{d_i = |\hat{x}_i^{(t)} - \hat{x}_i^{(t-1)}|\}_i$ are smaller than a threshold value δ, then the algorithm stops; otherwise it continues. We set the threshold value as $\delta = 10^{-10}$ in all the experiments below.

5.1. Simulated dataset

In this subsection, we conduct experiments using simulated datasets. The main purpose is to confirm analytical predictions and to examine the accuracy of the approximate formula. Our simulated datasets are generated by the process described around (1) and (10). The signal power is set to be $\sigma_x^2 = 1/\rho_0$, matching with section 3.3.

5.1.1. Consistency check of the replica solution. To check the accuracy of the replica result and to examine the finite-size effect, we first plot the input and output MSEs against λ, given a for different sizes. The plots for $(\alpha, \rho_0, \sigma_\Delta^2, a) = (0.5, 0.2, 0.1, 3)$ and $(1.5, 0.2, 0.1, 4)$ are given in figure 8 as example cases. The black thick curves and the colour markers denote the analytical and numerical results, respectively. For the numerical data, the average over different samples of the set $\{A, \boldsymbol{x}^0, \boldsymbol{\Delta}\}$ is taken; the sample numbers are $200, 100, 100, 50$ for $N = 50, 100, 200, 400$, respectively; the error bars are given as the standard error among the samples. The vertical dashed line represents the AT instability point below which the RS solution is unstable. Focusing only on the RS region, we can find that the finite-size effect is quite weak, and the numerical results show an excellent agreement with the analytical ones, justifying our analytical solutions. In this experiment, even below the AT point, the numerical results show a strong regularity. This is because the solutions are obtained by gradually changing λ from large to small values, and hence these solutions below the AT point are, in some sense, continuously connected to the ones above the AT point. We term this scheme λ *annealing*, which can be considered as a part of the *nonconvexity control* proposed in [20]. We warn that the solution path obtained by the λ annealing is very atypical below the AT point, as implied in section 5.1.2.

As another check of the RS solution's consistency, we also draw ROC curves by numerical experiments. In figure 9, we give the ROC curves along the a_{IMSE} line for $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.2, 0.1)$ and $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.4, 0.1)$, which correspond to the upper middle panel of figure 6 and the lower middle panel of figure 7, respectively. The numerical result is displayed as the scatter plots (orange cross points) of *TP* and *FP*, for the experiments of 10 different samples at $N = 1000$. The numerical plots show a fairly good agreement with the analytical curve, which again justifies our analytical solutions.

5.1.2. Accuracy of the approximate CV formula. To check the accuracy of the approximate CV formula, in figure 10 we compare the CV errors between the literal (by (39)) and approximate (by (46)) CVs for two specific samples of the system size $N = 100$. The other parameters are $(\alpha, \rho_0, \sigma_\Delta^2, a) = (0.5, 0.2, 0.1, 3)$ and $(\alpha, \rho_0, \sigma_\Delta^2, a) = (1.5, 0.2, 0.1, 4)$, which correspond to figure 8. Here, all the results are obtained using the λ annealing and they show regular behaviours, even below the AT point. In all the cases, the approximate result reproduces well the literal one up to the AT point, even for the error bars given by the way explained at section 4.1. The uncontrolled behaviour of the approximate formula below the AT point is owing to the singular behaviour in the factor $\left(\left(A^{\setminus\mu}_{*S_A} \right)^\top A^{\setminus\mu}_{*S_A} + \left(\partial^2 J(\hat{\boldsymbol{x}}_{S_A}; \eta) \right)_{S_A S_A} \right)^{-1}$ in (44). This is natural because this factor is nothing but the susceptibility, which is known to involve diverging modes when the AT instability occurs [14]. These considerations mean that our approximate formula is only applicable above the AT point or in the RS phase.

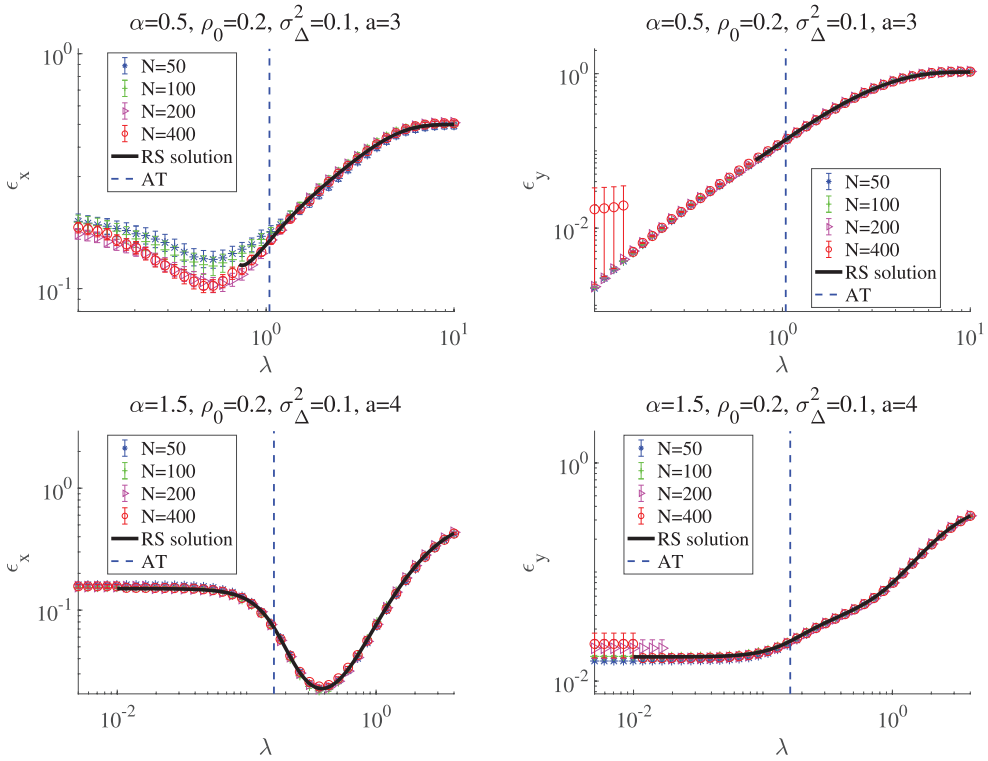

Figure 8. Plots of the input MSE (left) and the output MSE (right) against λ for $(\alpha, \rho_0, \sigma_\Delta^2, a) = (0.5, 0.2, 0.1, 3)$ (upper) and $(\alpha, \rho_0, \sigma_\Delta^2, a) = (1.5, 0.2, 0.1, 4)$ (lower). The black thick curves and the colour markers denote the analytical and numerical results, respectively. The left end point of the analytical curve corresponds to the existence limit of the RS solution. The vertical blue dashed line represents the AT instability point below which the RS solution is unstable. The agreement between the analytical and numerical results is excellent in the RS region. The numerical results are obtained by the annealing with respect to the amplitude parameter λ, explaining the regularity of the numerical results even below the AT instability point.

Can we detect the instability point only from the approximate CV result without referring to the replica computation? Figure 10 speaks for this, because the approximate CV error tends to show uncontrolled behaviours at and below the AT point. As a trial, assuming a combination use with the λ annealing, we examine the following procedures to detect the uncontrolled behaviours:

(i) Detect 'irregular' datapoints by locally comparing each datapoint with neighbouring points along the λ path (here datapoints mean the approximate CV result, red circles in figure 10).
(ii) Find the maximum value of λ whose corresponding datapoint is irregular. Regard all the λ region below it as 'instability region'.

To obtain a concrete result, we need to implement the first step (i) as an algorithm. The actual implementation is:

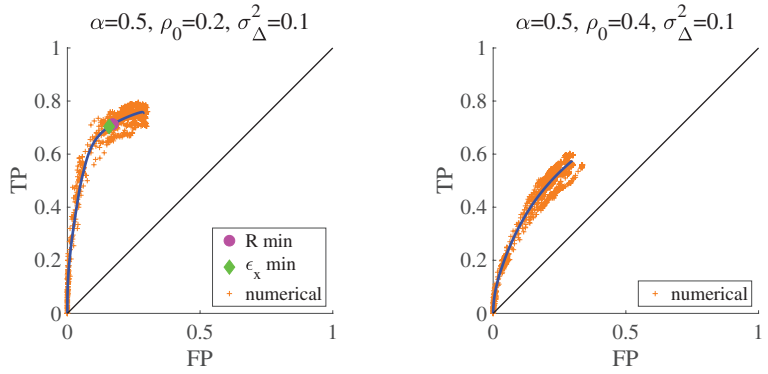

Figure 9. ROC curves evaluated by the analytical (blue thick curve) and numerical (orange cross point) for $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.2, 0.1)$ (left) and $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.4, 0.1)$ (right). The numerical data is obtained by the experiments of 10 different samples at $N = 1000$. The agreement between the two are fairly good.

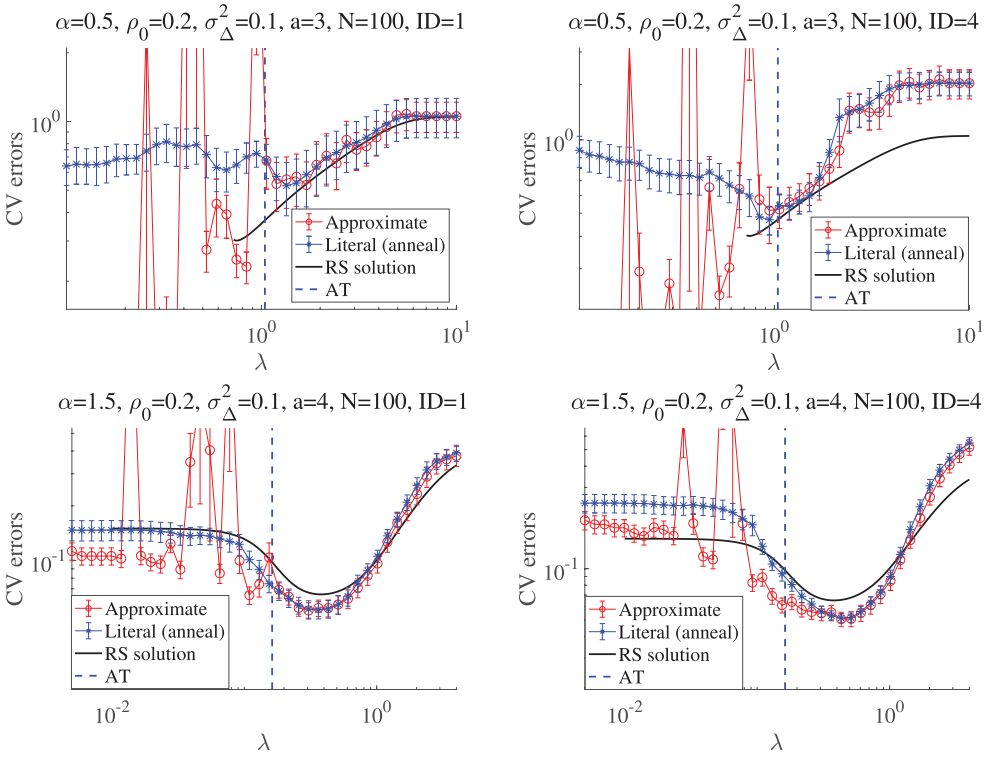

Figure 10. Plots of the literal (by (39)) and approximate (by (46)) CV errors against λ at $N = 100$ for given $(\alpha, \rho_0, \sigma_\Delta^2, a) = (0.5, 0.2, 0.1, 3)$ (upper) and $(\alpha, \rho_0, \sigma_\Delta^2, a) = (1.5, 0.2, 0.1, 4)$ (lower). The results of two specific different samples (left and right) are shown. The black thick curve and the vertical blue dashed line represent the RS solution and the AT point, respectively. The approximate results are well matching to the literal ones up to the AT point.

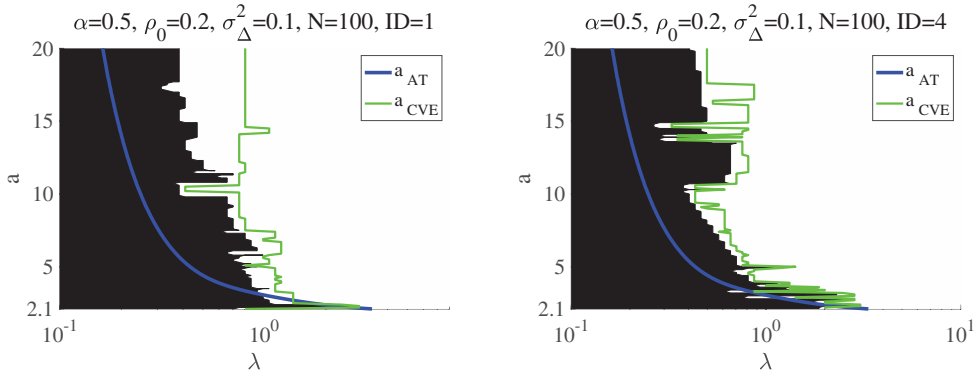

Figure 11. 'Phase diagrams' drawn for two specific samples used in figure 10 using the instability detection procedures described in the main text. The white and black regions represent the stable and unstable regions which correspond to the RS and RSB phases. The blue thick curve denotes the AT line $a_{\rm AT}$ computed by the replica method, while the green curve $a_{\rm CVE}$ shows the λ location of the approximate CV error minimum given a in the stable region which is supposedly related to $a_{\rm IMSE}$.

(i)-1 If the CV error difference between the irregular point candidate and the compared datapoint is larger enough in reference to the error bar of the compared point, then the candidate is regarded as 'irregular'.

By these procedures, we can separate all the parameter regions into two parts: Stable and unstable regions corresponding to the RS and RSB phases, respectively. By employing this, in figure 11 we draw 'phase diagrams' of two specific samples, used also in figure 10. This figure shows that the boundary between the white and black regions behaves similarly to the AT line $a_{\rm AT}(\lambda)$, although there is a gap supposedly owing to the sample fluctuation and the finite-size effect. As a reference, we also compute the minimum location of the approximate CV error with respect to λ given a, defining $a_{\rm CVE}(\lambda)$, which is denoted by the green curve in figure 11. According to (41), this corresponds to $a_{\rm IMSE}$ appearing in the phase diagrams in section 3.3. Note that these procedures are somewhat overcautious, and can miss some stable regions possibly existing in the small λ region. As typically seen in the left panel of figure 2, the re-entrant transition can emerge in the weak noise case but the present procedures cannot detect this re-entrancy, because these procedures detect the first RS-RSB transition corresponding to the rightmost branch of $a_{\rm AT}$ and all the region below this first transition point is regarded as 'instability region'. However, for practical use, it is more important to avoid giving wrong estimates by our approximate formula. Hence, we do not aim to improve the above instability detection procedures in this study.

Apart from the re-entrancy, it is worthwhile to investigate the cause of the gap between the AT line and the instability points detected by our procedures. To this end, we compute the boundary value between the black and white regions in figure 11 given a, $\lambda_c(a)$, for many samples and different system sizes. The results at $a = 5$ and $a = 10$ are shown in figure 12. The left panels provide the histograms of λ_c from $N_{\rm samp} = 100$ samples for different sizes $N = 100, 200, 400, 800, 1600$ discriminated by different colors. As the system size grows, the width of the histogram shrinks and the mode value tends to approach the λ value at the AT point. Here the number of bins $N_{\rm bin}$ for the histogram is determined by the so-called Sturges rule [33] as $N_{\rm bin} = \lceil 1 + \log_2 N_{\rm samp} \rceil$, enabling us to define the mode value without ambiguity. To quantify the convergence behavior of the mode value, we plot the mode against the inverse

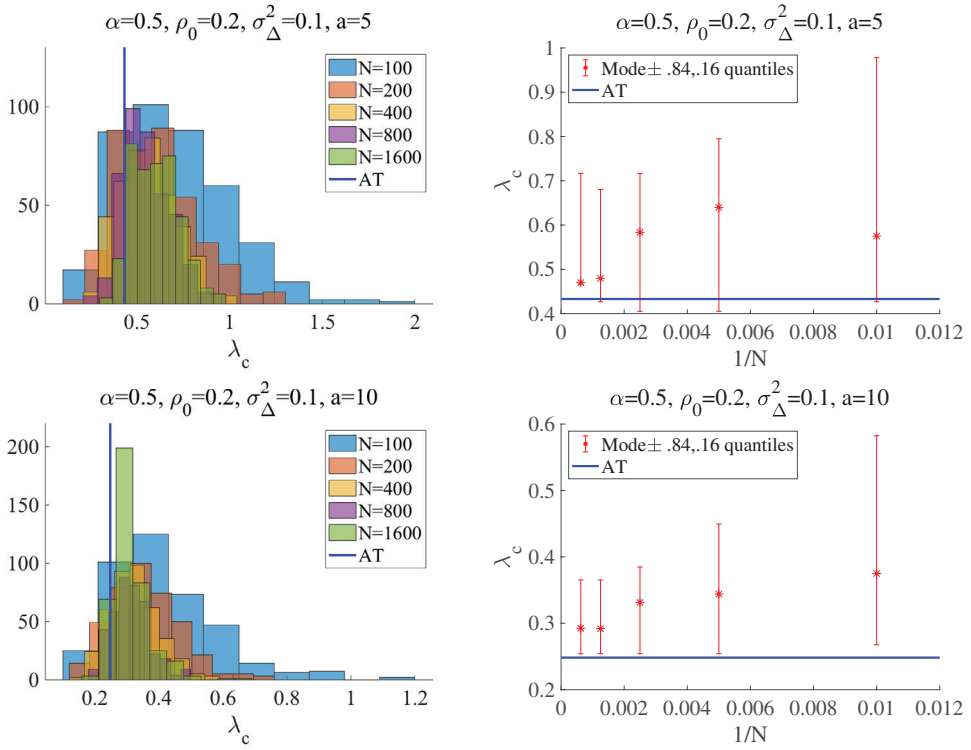

Figure 12. Histograms of λ_c (left) and plots of the mode values against the inverse system size (right) at $a = 5$ (upper) and $a = 10$ (lower) computed from $N_{\text{samp}} = 400$ samples. The blue straight line represents the λ value at the AT point commonly in all the panels. The error bar of the mode is put as the 0.86 and 0.14 quantiles of the histogram. The examined sizes are $N = 100, 200, 400, 800, 1600$ and the different colors of the histograms correspond to different sizes as shown in the legend. The other parameters are set to $(\alpha, \rho_0, \sigma_\Delta^2) = (0.5, 0.2, 0.1)$. To unambiguously define the mode value, we set the number of bins N_{bin} by Sturges rule as $N_{\text{bin}} = \lceil 1 + \log_2 N_{\text{samp}} \rceil$. As the system size grows, the width of the histogram shrinks and the mode value approaches the AT point.

system size $1/N$ in the right panels. Here the mode value shows a clear tendency of approaching to the AT value as N grows. This indicates that the gap is actually due to the sample fluctuation and the finite-size effect, and also implies that our instability detection procedures are reasonably connected to the AT instability.

The AT instability is known to be connected to the emergence of many local minimums [14]. To directly check this, we conduct the literal CV without the λ annealing. For each point of λ, the estimator is computed from ten different randomly initialized $\boldsymbol{x}$, each component of which is i.i.d. from $\mathcal{N}(0, 1)$, by the CD algorithm. In figure 13, the resultant CV errors are given as scatter plots in combination with the CV error using the annealing. The experimental setup of each panel is again identical to the corresponding one in figure 10. This figure gives a clear evidence of the multiple solutions below the AT point. Figure 13 also implies that the solution obtained by the λ annealing is rather atypical: solutions obtained from random initial conditions tend to give rather different values of CV error from the annealed solution. To give a better theoretical background to this statement, we have to construct the full-step RSB

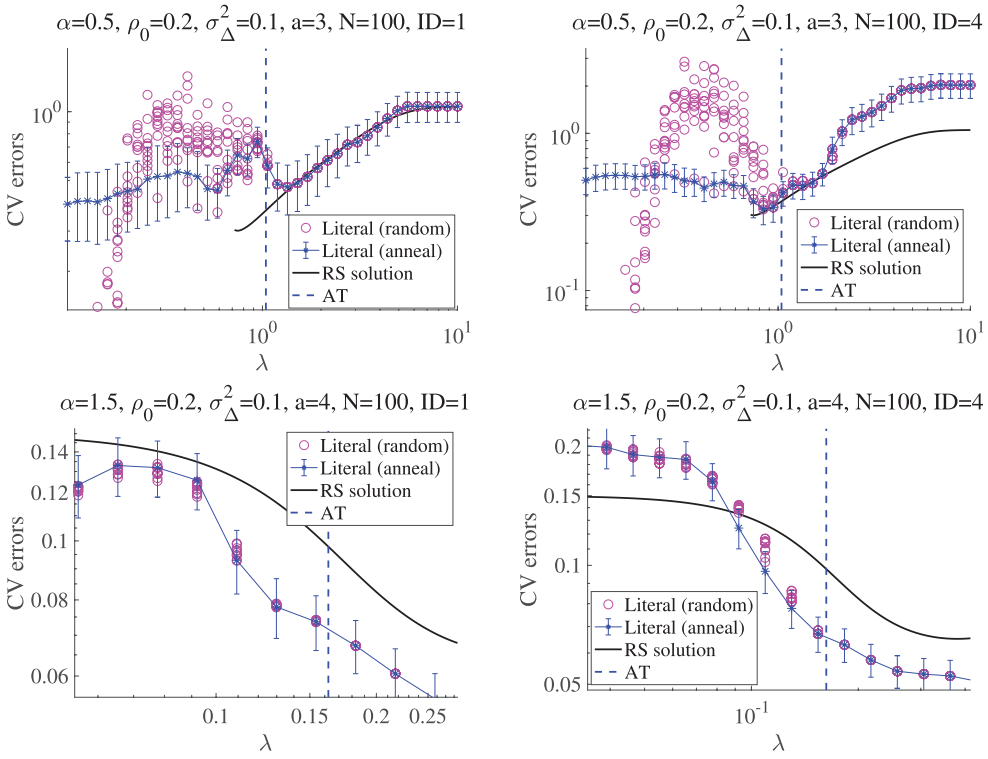

Figure 13. Comparison of the literal CV errors with and without the λ annealing in the same experimental condition as the corresponding panel of figure 10. The result without the annealing is shown as scatter plots (magenta circles) for ten different random initial conditions and it exhibits visible differences from the annealing result (blue asterisks, identical result to figure 10) below the AT point, while no difference exists sufficiently above the AT point. For the lower panels, the region around the AT point is magnified because the difference is small, although it exists.

solution and to figure out the characterisation of the annealed solution in the ensemble of all the local minimums. This is beyond the present purpose but will be an interesting future work.

Our present attitude to the multiplicity of the solutions is to avoid it. This is reasonable because the global minimum of the generalisation error is in the RS region without the multiplicity, as clarified by our analytical computation. Once accepting this attitude, we can use the proposed approximate formula efficiently estimating the generalisation error and, fortunately, the formula also enables us to avoid the multiple solution region by using the above instability detection procedures. This is the main outcome and contribution of this study.

Before closing this subsection, we check the computational time and the approximate accuracy of the proposed formula more quantitatively. Here we quantify the error difference between the literal and approximate CVs by a *normalised MSE* defined as:

$$\text{normalised MSE} = \left(\frac{\epsilon_{\text{CV,approx.}} - \epsilon_{\text{CV,literal}}}{\epsilon_{\text{CV,literal}}} \right)^2, \tag{49}$$

where $\epsilon_{\text{CV,approx.}}$ and $\epsilon_{\text{CV,literal}}$ are the CV errors evaluated by the approximate and literal CV procedures, respectively. According to the derivation of the formula in section 4, the accuracy is considered to be better as N and M increase. Thus, we plot the normalised MSE against N

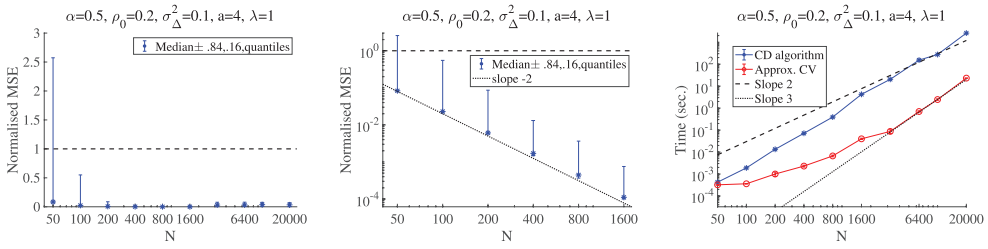

Figure 14. (Left) The normalised MSE of the CV error difference is plotted against the system size N in the log-linear scale. The number of samples is $\{1000, 1000, 200, 200, 100, 50, 10, 10, 10, 2\}$ for the system size $\{50, 100, 200, 400, 800, 1600, 3200, 6400, 10\,000, 20\,000\}$, respectively. The marker denotes the median and the error bars consist of the 0.86 and 0.14 quantiles among the samples. The dashed horizontal line denotes unity, given as a reference. For $N \geqslant 3200$, the literal CV is conducted by the ten-fold CV instead of the LOO CV, to save the computational cost. (Middle) The same plot as the left panel in the double log scale for small sizes $N \leqslant 1600$. The normalised MSE decreases in the scale N^{-2} as the system size grows, which is clearly indicated by the dotted line representing slope -2. (Right) The computational time for the CD algorithm convergence (blue asterisk) and for the approximate CV formula (red circle), in the same experiment as the left panel. The error bars are smaller than the marker sizes and hard to see. The dashed line denotes the slope 2 while the dotted one represents the slope 3, both of which are the expected size scaling of the computational time of the CD algorithm and the approximate CV formula, respectively.

as the left panel of figure 14. The parameters are set to $(\alpha, \rho_0, \sigma_\Delta^2, a, \lambda) = (0.5, 0.2, 0.1, 4, 1)$ as an example. For each N, we compute the normalised MSE for several different samples of $\{A, \boldsymbol{x}_0, \boldsymbol{\Delta}\}$, and the marker (blue asterisk) denotes the median among the samples, and the upper and lower error bars correspond to the 0.86 and 0.14 quantiles, respectively. The number of samples is $\{1000, 1000, 200, 200, 100, 50, 10, 10, 10, 2\}$ for the system size $\{50, 100, 200, 400, 800, 1600, 3200, 6400, 10\,000, 20\,000\}$, respectively. As expected, the normalised MSE is quite small for large sizes of $N \geqslant 200$, although at the smallest size $N = 50$ there is a non-negligible difference. This difference is dominated by a few percent of samples giving accidentally large values of $\epsilon_{\mathrm{CV,approx.}}$. The probability of the accidents seems to become smaller rapidly as the system size grows. In the middle panel, the same plot in the double log scale for small sizes $N \leqslant 1600$ is given, showing a clear decay of the normalized MSE in the scale N^{-2} as the system size grows. This is naturally understood from the scaling argument presented in section 5.1.1 in [32]. The corresponding computational time of the CD algorithm convergence and the approximate formula are given in the right panel. The approximate formula requires to take the inverse of the Hessian, leading to the computational cost of $O(|S_A|^3)$, which is scaled as the third order polynomial of N if $|S_A| = O(N)$. This computational cost can be more expensive than the optimisation cost by the CD algorithm in the large N limit, because the total computational time of the CD algorithm is considered to be scaled as $O(N^2)$, under the assumption such that the convergence of the CD algorithm takes place in constant computational steps independent of the system size N, although the $O(N^2)$ behaviour is hard to see in figure 14. Despite this inconvenience in the limiting case, figure 14 shows that the computational time of the approximate formula is much smaller than the CD algorithm convergence, in all the investigated range of the system size. We note that the computational time of the CD algorithm shown in figure 14 is just for one-time optimisation, and hence, for conducting the literal k-fold CV, the required computational time becomes approximately k times

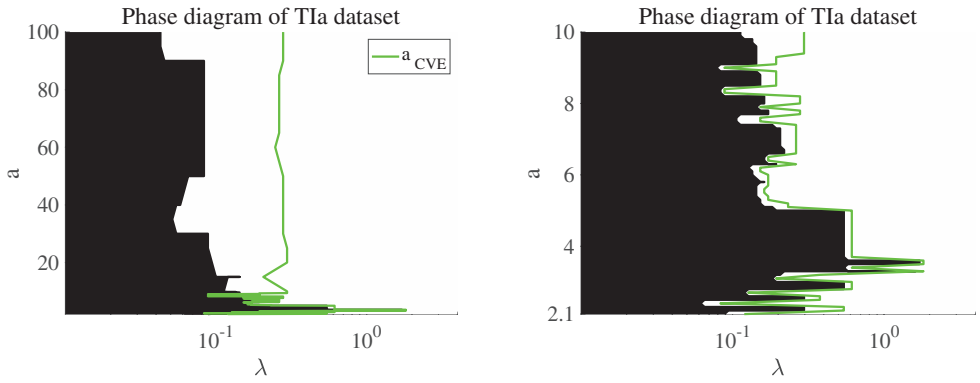

Figure 15. λ–a phase diagram of the Type Ia supernovae dataset from [36]. The right panel is the magnified view of the left one in the small a region. The black region represents the instability region for which the approximate formula cannot be applied, while the white one is the stable region in which the approximate formula gives a reliable estimate. The minimum point of the CV error in the stable region is given by $a_{\mathrm{CVE}}(\lambda)$, depicted by the green line.

larger than that. Overall, although there is no superiority in the large N limit, our approximate formula practically works very efficiently in a wide range of system sizes.

5.2. A real-world dataset: Type Ia supernovae

Here we apply the proposed approximate method to a dataset of Type Ia supernovae. Our dataset is a part of the data from [34, 35], which is screened by a certain criterion [36]. This dataset was treated by a number of sparse estimation techniques recently, and a set of important variables, which is known to be empirically significant, has been reproduced [32, 36–38]. In those studies, the LASSO and ℓ_0 cases are treated, and the CV is employed for hyperparameter estimation. We reanalyse this dataset by using the SCAD penalty and compute the CV error by using the approximate formula. The parameters of the screened data are $M = 78$ and $N = 276$, and an appropriate standardisation is employed as pre-processing.

Again, we use the λ annealing to obtain the SCAD estimators for this dataset, and the CV error is computed by our approximate formula. The instability region is detected by the procedures explained in section 5.1.2. As figure 11, the instability detection gives a phase diagram which is in figure 15. The overall shape of this phase diagram is similar to the ones in section 3.3 or figure 11, supporting the practical relevance of our results so far. To directly check the approximation accuracy, we also conducted the literal CV at a number of values of a. The results for $a = 4$ and 50 are given in figure 16. The approximate error well matches to the literal one up to the instability point, determined by the procedures explained in section 5.1.2, which justifies our instability detection procedures. For the left panel of $a = 4$, however, even below the instability point, there exists a region in which two CV errors agree well. This implies the presence of re-entrant transition, and it is probably related to a protruding black region around $a \in (3, 5)$ in the right panel of figure 15. As declared in section 5.1.2, we do not try to detect the re-entrancy in the present study, but there must be some ways. For example, the annealing with respect to a instead of λ would be able to identify the re-entrancy with respect to λ. We found that this strategy can actually detect the re-entrancy, but the strategy itself is far from perfect. There are some reasons for this. One reason is that the switching parameter a has no upper bound in contrast to λ (λ has an effective upper bound

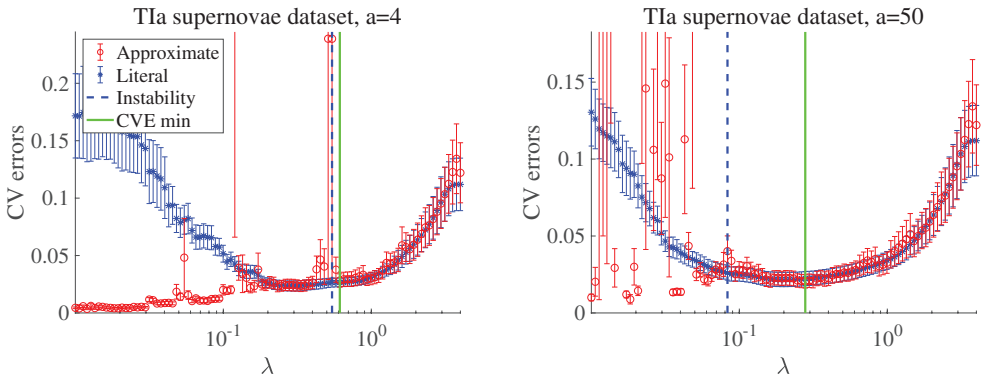

Figure 16. Plot of the CV errors by the approximate (red circle) and the literal (blue asterisk) CV for $a = 4$ (left) and $a = 50$ (right) against λ for the Type Ia supernovae dataset. The blue dashed vertical line indicates the instability point obtained by the procedures described in the main text and well matches to the point at which the literal and approximate CV errors deviate from each other.

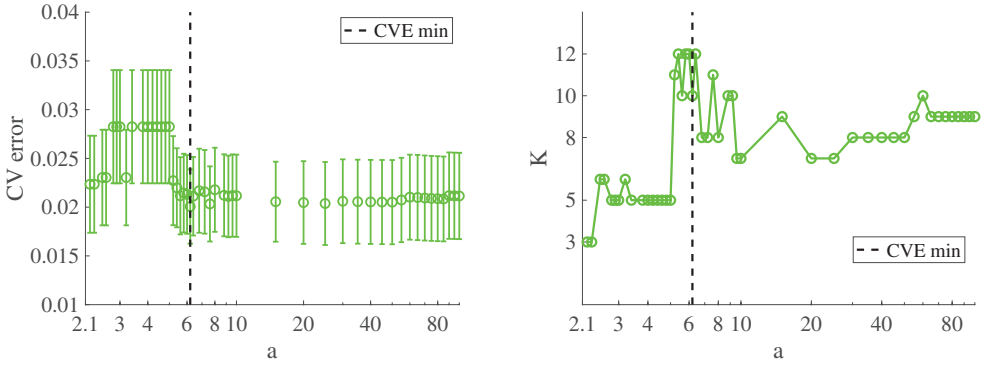

Figure 17. Plots of the approximate CV error (left) and the number of non-zero components (right) along the a_{CVE} line for the Type Ia supernovae dataset. The black vertical dashed line represents the minimum location of the CV error. Almost all datapoints are within the error bar of the minimum error point, and the sparsest solution within the one-standard error is obtained at $a = 2.2$ and 2.3 with $K = 3$.

as explained in section 5.3) and hence the initialization becomes nontrivial. Another reason is that some instability 'islands' seem to exist at unexpected regions on the parameter space for this specific dataset: some compact parameter regions exhibiting the instability seem to be able to exist, in contrast to the theoretically derived phase diagrams in section 3.3, and hence isolating the instability regions becomes nontrivial even if the annealing with respect to a is correctly performed. Due to these difficulties, we leave further exploration of better ways of nonconvexity control as a future work.

To extract relevant values of the parameters, we plot the approximate CV error and the number of non-zero components $K = ||\hat{\boldsymbol{x}}||_0$ along the a_{CVE} line in figure 17. Here, some outliers exhibiting extraordinary small CV errors are omitted. At the CV error minimum, the solution with $K = 10$ is obtained, which is comparable with $K = 9$ of the LASSO solution at the minimum CV error [32, 36]. In the case of LASSO, it is common to select a sparser solution than the one at the CV error minimum according to the one-standard error rule [10, 39].

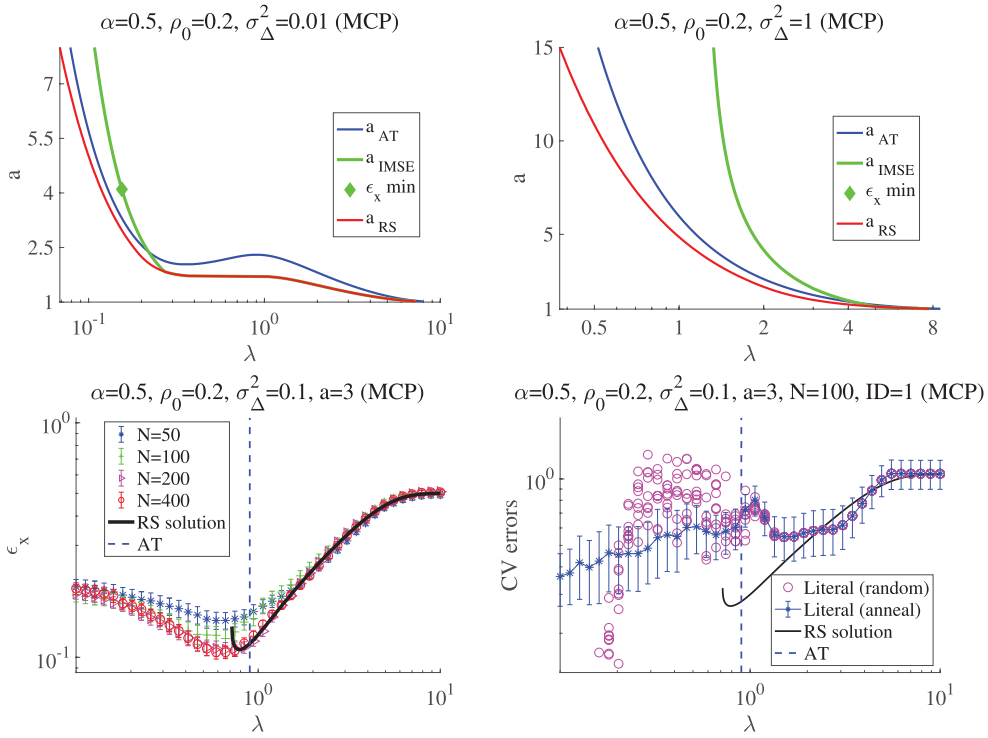

Figure 18. The result for the MCP case. (Upper) Phase diagrams for $\sigma_\Delta^2 = 0.01$ (left) $\sigma_\Delta^2 = 1$ (right) at $(\alpha, \rho_0) = (0.5, 0.2)$, corresponding to the middle and right panels of figure 2. (Lower) Plots of the input MSE (left) and of the literal CV errors with and without the λ annealing (right) against λ at $(\alpha, \rho_0, \sigma_\Delta^2, a) = (0.5, 0.2, 0.1, 3)$, corresponding to the left upper panel of figure 8 and that of figure 13, respectively. Qualitatively similar results to the SCAD case are obtained.

Although it is unclear if the application of this rule to the SCAD estimators is appropriate or not, we here try to apply to our case. As a result, we have the solution with $K = 3$ obtained at $a = 2.2$ or 2.3 as seen in figure 17. We globally examined all the datapoints within the one-standard error in the stable phase, and confirmed that the $K = 3$ solution is the sparsest. This sparsest solution consists of variables whose IDs are 1, 2 and 233. This is accurately matching to the result of [37, 38, 40], in which the ℓ_0 formalism is treated, while the LASSO estimator tends to give a denser solution with $K = 6$ even under the one-standard error rule [32, 36]. These demonstrate the effectiveness of the SCAD estimator, and the presented analysis and approximate formula resolve its disadvantages of the multiplicity of solutions and the computational cost in hyperparameter estimation. The effect of the one-standard error rule on the SCAD estimator seems to be also good, though further exemplifications would be needed.

5.3. Numerical codes

In [1], a MATLAB package of numerical codes implementing the estimation of the solution path using the λ annealing in conjunction with the approximate CV formula is distributed; the optimization is performed by the CD algorithm as the experiments so far. In the package three regularizations, LASSO, SCAD, and MCP, are treated in a unified manner. All the

parameters are tunable in the codes, but the minimally required quantities to run the codes are the data vector $\boldsymbol{y}$, the design matrix A[6], and the switching parameter a. In the default setting, the $L = 100$ values of λ are chosen as to be a descending order $\lambda_1 > \lambda_2 > \cdots > \lambda_L$, and the largest is set to be $\lambda_1 = \lceil \max_{1 \leqslant j \leqslant N} \left(|\boldsymbol{a}_j^\top \boldsymbol{y}| \right) \rceil$ where $\boldsymbol{a}_j$ is the jth column vector of A, because only the trivial solution $\hat{\boldsymbol{x}} = 0$ exists for $\lambda > \max_{1 \leqslant j \leqslant N} \left(|\boldsymbol{a}_j^\top \boldsymbol{y}| \right)$; the smallest is set to be $\lambda_L = \epsilon \lambda_1$ with $\epsilon = 0.01$ and the intermediate values are given to interpolate these two values by the geometric progression with a constant rate. This way follows that of a commonly-used package *glmnet* [10, 41]. The λ annealing is basically the same as *warm starts* explained in [10], but it has a stronger meaning in the nonconvex penalties because it inevitably picks out a certain solution path as exemplified in the numerical experiments so far. On each point of λ_k, the CD algorithm finds the solution $\hat{\boldsymbol{x}}(\lambda_k)$ from the initial condition $\hat{\boldsymbol{x}}(\lambda_{k-1})$ (for $k = 1$ the initial condition is the zero vector), and hence $\hat{\boldsymbol{x}}(\lambda_k)$ and $\hat{\boldsymbol{x}}(\lambda_{k-1})$ are expected to be close each other. In the default setting, after obtaining the whole solution path over $\{\lambda_k\}_{k=1}^L$, the approximate CV formula is subsequently applied and it is followed by the instability detection routine, yielding the approximate values of CV error and its reliable region. In the package, demonstration codes are also included and some experiments in section 5.1 can be easily re-obtained; readers who are interested in the experiments are thus encouraged to try to use them. The details of usage are more explained in [1].

6. Conclusion

In this study, using the replica method, we analysed the macroscopic properties of the SCAD estimator in the context of the signal reconstruction in the presence of noise, under the assumption that the design matrix is the i.i.d. random matrix. We derived the phase diagrams involving the RSB phase, and showed the superiority of the SCAD estimator to the LASSO one based on ROC curves. We also provided an analytical evidence that the global minimum of the input MSE or the generalisation error is located in the RS phase. Furthermore, we derived an approximate formula for the CV error, although it is applicable only for the RS phase. We implemented procedures detecting the AT instability or the approximation instability, enabling to clarify the applicable limit of the approximate formula and making the formula stand-alone.

To examine the analytical results, numerical experiments on simulated datasets and a real-world dataset of Type Ia supernovae were conducted. On the simulated datasets, the replica prediction was well reproduced. The accuracy and the computational time of the approximate CV formula were examined, and its effectiveness was demonstrated in a wide range of the system size. For the real-world dataset, the application of the SCAD penalty reproduced the variables known to be empirically important. By using the approximate formula, we could globally search the parameters efficiently, and find that the SCAD estimator can provide a very sparse solution giving a reasonable value of the CV error. This solution is matching to the one of the earlier studies using the ℓ_0 formulation [37, 38, 40], and cannot be found by LASSO. These experiments demonstrate the effectiveness of the SCAD estimator, and the presented analysis and approximate formula resolve its disadvantages of the multiplicity of solutions and the computational cost in hyperparameter estimation.

As an efficient strategy to obtain a solution path, we proposed nonconvexity annealing as a part of nonconvexity control proposed in [20], and especially focused on the usage of the annealing with respect to λ, termed λ annealing in this paper. It was shown that this strategy works well also in combination with our approximate CV formula, but it further raised up

[6] In the package, the design matrix is denoted as X and the regression coefficients are given as β, following the statistics convention.

a question related to RSB. In the RSB phase exhibiting the multiplicity of solutions, what solution is obtained by the annealing? Our numerical experiments showed that the annealed solution tends to give a smaller CV error compared to the solutions computed from random initial conditions, and in this sense the annealing is a nice strategy even in the multiple solution region. A similar observation was obtained in an inference in Gaussian mixture model [42, 43]. To make a more accurate and quantitative analysis about these findings, it is needed to construct the full-step RSB solution and to figure out the characterisation of the annealed solution in the ensemble of all the solutions. This will be an interesting future work.

The present instability detection procedures for the approximate CV formula are rather ad hoc and have some ambiguity, especially in specifying irregular datapoints along the solution path with respect to λ. This ambiguity is related to which points of λ should be sampled when computing the solution path. In the case of LASSO, the change points of active set, usually called knots, can be efficiently computed [44], which provides a clear criterion to the above ambiguity problem. It is expected that a similar technique computing knots for SCAD will be useful for improving the instability detection procedures.

As a final remark, we mention about the MCP penalty defined by:

$$J_{\mathrm{MCP}}(\theta;\eta) = \begin{cases} \lambda|\theta| - \frac{\theta^2}{2a} & (|\theta| \leqslant a\lambda) \\ \frac{a\lambda^2}{2} & (|\theta| > a\lambda) \end{cases}, \tag{50}$$

where $\eta = \{\lambda, a\}$. If we use this instead of the SCAD penalty, the effective one-dimensional estimator, (26) in the SCAD case, is replaced as:

$$x^*(h;\tilde{Q}^{-1}) = V_{\mathrm{MCP}}(h;\tilde{Q}^{-1},\eta) S_{\mathrm{MCP}}(h;\tilde{Q}^{-1},\eta), \tag{51}$$

where

$$S_{\mathrm{MCP}}(x;\sigma^2,\eta) = \begin{cases} x - \mathrm{sgn}(x)\lambda & \text{for } a\lambda\sigma^{-2} \geqslant |x| > \lambda \\ x & \text{for } |x| > a\lambda\sigma^{-2} \\ 0 & \text{otherwise} \end{cases}, \tag{52}$$

$$V_{\mathrm{MCP}}(x;\sigma^2,\eta) = \begin{cases} (\sigma^{-2} - a^{-1})^{-1} & \text{for } a\lambda\sigma^{-2} \geqslant |x| > \lambda \\ \sigma^{-2} & \text{for } |x| > a\lambda\sigma^{-2} \\ 0 & \text{otherwise} \end{cases}. \tag{53}$$

Replacing x^* in equations (27*a*)–(27*c*) by (51), we can get EOS for the MCP penalty, and the AT condition (33) can be replaced by the same way. Corresponding to (34), the RS existence limit of the MCP case is also given as

$$\tilde{Q} - \frac{1}{a} \geqslant 0. \tag{54}$$

Using these replacements, it is easy to obtain the result for the MCP case. As far as we searched, the MCP result is qualitatively similar to the SCAD one. For illustration of this, we give some phase diagrams, ϵ_x plots, and plots of literal CV errors with and without λ annealing in figure 18. We see qualitatively similar results to the SCAD case: the re-entrancy for the weak noise region; the no global minimum of the input MSE in the RS phase at finite a for the strong noise or dense signal cases; the accurate accordance between the RS and numerical results above the AT point; the solution multiplicity below the AT point. Although there can be a difference between the SCAD and MCP penalties in a quantitative level as reported in [17], such a comparative study requires more detailed quantitative analyses and we also leave it as a future work. Note that the lower existence limit of the RS phase of the MCP case is given as

$a = 1$, which is derived from (36) and (54), and hence the RS stable region tends to be wider than the SCAD case. However, the direct comparison of two parameter spaces is not necessarily meaningful, and another systematic way of comparison is desired.

Acknowledgments

The authors would like to thank Yoshiyuki Kabashima, Satoshi Takabe, Takashi Takahashi, and Yingying Xu for their helpful discussions and comments. This work is partially supported by JSPS KAKENHI No. 16K16131 (AS) and Nos. 18K11463 and 17H00764 (TO). TO is also supported by a Grant for Basic Science Research Projects from the Sumitomo Foundation.

ORCID iDs

Tomoyuki Obuchi https://orcid.org/0000-0003-1216-489X
Ayaka Sakata https://orcid.org/0000-0003-1660-0222

References

[1] Obuchi T 2019 Matlab package of sparse linear regression with accelerated cross-validation under L1 or continuous nonconvex penalties https://github.com/T-Obuchi/SLRpackage_AcceleratedCV_matlab
[2] Breiman L *et al* 1996 *Ann. Stat.* **24** 2350–83
[3] Natarajan B K 1995 *SIAM J. Comput.* **24** 227–34
[4] Tibshirani R 1996 *J. R. Stat. Soc.* B **58** 267–88 (https://www.jstor.org/stable/2346178)
[5] Meinshausen N and Bühlmann P 2004 *Consistent Neighbourhood Selection for Sparse High-Dimensional Graphs with the Lasso* (Zürich: Seminar für Statistik, Eidgenössische Technische Hochschule (ETH))
[6] Banerjee O, Ghaoui L E, d'Aspremont A and Natsoulis G 2006 *Proc. 23rd Int. Conf. on Machine Learning* (ACM) pp 89–96
[7] Friedman J, Hastie T and Tibshirani R 2008 *Biostatistics* **9** 432–41
[8] Rish I and Grabarnik G 2014 *Sparse Modeling: Theory, Algorithms, and Applications* 1st edn (Boca Raton, FL: CRC Press)
[9] Mairal J, Bach F and Ponce J 2014 *Found. Trends Comput. Graph. Vis.* **8** 85–283
[10] Hastie T, Tibshirani R and Wainwright M 2015 *Statistical Learning with Sparsity: the Lasso and Generalizations* (London: Chapman and Hall)
[11] Fan J and Li R 2001 *J. Am. Stat. Assoc.* **96** 1348–60
[12] Zhang C H 2010 *Ann. Stat.* **38** 894–942
[13] Sakata A and Xu Y 2018 *J. Stat. Mech.* **2018** 033404
[14] 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)
[15] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: an Introduction* vol 111 (Oxford: Clarendon)
[16] Dotsenko V 2005 *Introduction to the Replica Theory of Disordered Statistical Systems* vol 4 (Cambridge: Cambridge University Press)
[17] Breheny P and Huang J 2011 *Ann. Appl. Stat.* **5** 232–53
[18] Donoho D L 2006 *IEEE Trans. Inf. Theory* **52** 1289–306
[19] Donoho D L, Maleki A and Montanari A 2011 *IEEE Trans. Inf. Theory* **57** 6920–41
[20] Sakata A and Obuchi T 2019 (arXiv:1902.07436)
[21] Guo D and Verdú S 2005 *IEEE Trans. Inf. Theory* **51** 1983–2010
[22] Opper M and Winther O 2001 *Phys. Rev.* E **64** 056131
[23] Opper M and Winther O 2001 *Phys. Rev. Lett.* **86** 3695
[24] Opper M and Winther O 2005 *J. Mach. Learn. Res.* **6** 2177–204 (http://www.jmlr.org/papers/v6/opper05a.html)

[25] Çakmak B, Winther O and Fleury B H 2014 *IEEE Information Theory Workshop* (IEEE) pp 192–6
[26] Kabashima Y and Vehkaperä M 2014 *IEEE Int. Symp. on Information Theory* (IEEE) pp 226–30
[27] Cespedes J, Olmos P M, Sánchez-Fernández M and Perez-Cruz F 2014 *IEEE Trans. Commun.* **62** 2840–9
[28] Rangan S, Schniter P and Fletcher A 2016 (arXiv:1610.03082)
[29] Ma J and Ping L 2017 *IEEE Access* **5** 2020–33
[30] De Almeida J and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
[31] Lee S and Breheny P 2015 *J. Comput. Graph. Stat.* **24** 1074–91
[32] Obuchi T and Kabashima Y 2016 *J. Stat. Mech.* 053304
[33] Sturges H A 1926 *J. Am. Stat. Assoc.* **21** 65–6
[34] Filippenko A V, Ganeshalingam M, Li W and Silverman J M 2012 *Mon. Not. R. Astron. Soc.* **425** 1889–916
[35] The SNDB http://heracles.astro.berkeley.edu/sndb/info
[36] Uemura M, Kawabata K S, Ikeda S and Maeda K 2015 *Publ. Astron. Soc. Japan* 67 (https://doi.org/10.1093/pasj/psv031)
[37] Kabashima Y, Obuchi T and Uemura M 2016 *54th Annual Allerton Conf. on Communication, Control, and Computing (Allerton)* pp 596–600
[38] Obuchi T and Kabashima Y 2016 *24th European Signal Processing Conf.* pp 1247–51
[39] John Lu Z 2010 *J. R. Stat. Soc.* A **173** 693–4
[40] Igarashi Y, Takenaka H, Nakanishi-Ohno Y, Uemura M, Ikeda S and Okada M 2018 *J. Phys. Soc. Japan* **87** 044802
[41] Friedman J, Hastie T, Simon N, Qian J and Tibshirani R Glmnet https://cran.r-project.org/web/packages/glmnet/index.html
[42] Barkai N, Seung H and Sompolinsky H 1993 *Phys. Rev. Lett.* **70** 3167
[43] Barkai N and Sompolinsky H 1994 *Phys. Rev.* E **50** 1766
[44] Efron B *et al* 2004 *Ann. Stat.* **32** 407–99