
Average case analysis of Lasso under ultra-sparse conditions

Koki Okajima

Xiangming Meng

Takashi Takahashi

Yoshiyuki Kabashima

Department of Physics, The University of Tokyo

Abstract

We analyze the performance of the least absolute shrinkage and selection operator (Lasso) for the linear model when the number of regressors N grows larger keeping the true support size d finite, i.e., the ultra-sparse case. The result is based on a novel treatment of the non-rigorous replica method in statistical physics, which has been applied only to problem settings where N , d and the number of observations M tend to infinity at the same rate. Our analysis makes it possible to assess the average performance of Lasso with Gaussian sensing matrices without assumptions on the scaling of N and M , the noise distribution, and the profile of the true signal. Under mild conditions on the noise distribution, the analysis also offers a lower bound on the sample complexity necessary for partial and perfect support recovery when M diverges as $M = O(\log N)$. The obtained bound for perfect support recovery is a generalization of that given in previous literature, which only considers the case of Gaussian noise and diverging d . Extensive numerical experiments strongly support our analysis.

1 Introduction

An important objective of high dimensional statistics is to extract information in situations where the signal's dimension N is overwhelmingly large compared to the accumulated sample size M . It is crucial to incorporate prior knowledge on the signal structure to reduce the signal space dimension for reliable estimation. A particularly common assumption is *sparsity*, which postulates that the true signal has few nonzero entries. Exploiting this property allows one to obtain robust and interpretable results specifying the few relevant variables explaining the retrieved data (Donoho, 2006).

For instance, consider the sparse linear regression problem where measurements $\mathbf{y} \in \mathbb{R}^M$ of the signal $\mathbf{x}^0 \in \mathbb{R}^N$ with d non-zero components are given by the linear model

$$\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \boldsymbol{\xi}, \quad (1)$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the sensing matrix, and $\boldsymbol{\xi} \in \mathbb{R}^M$ is the noise vector distributed according to $p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$. The most fundamental yet popular sparse signal estimation method is the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), which offers the estimator by solving the following convex program:

$$\hat{\mathbf{x}}_{\lambda}(\mathbf{A}, \mathbf{y}) := \arg \min_{\mathbf{x}} \left(\frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + M\lambda \|\mathbf{x}\|_1 \right), \quad (2)$$

where λ is a regularization parameter. Since its introduction, this simple ℓ_1 -regularization scheme has been successfully adapted as a backbone technique for solving a wide variety of sparse estimation problems. A particularly interesting question to ask is if one can make any guarantees on the performance of Lasso under general scalings of (N, M, d) , its dependence on λ , and statistical properties of the noise and true signal.

A sheer amount of research has been devoted to assessing the performance of Lasso. Traditionally, research based on the irrepresentability condition (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006) has been popular in establishing guarantees in terms of support recovery of the sparse signal (Wainwright, 2009b; Dossal et al., 2012; Meinshausen and Bühlmann, 2006; Zhang and Huang, 2008; Candès and Plan, 2009; Zhao and Yu, 2006). A different approach based on approximate message-passing (AMP) theory (Donoho et al., 2009), and the heuristical replica method (Mézard et al., 1986) from statistical physics has focused on assessing the sharp, asymptotic properties of Lasso in the large N and M limit under random sensing matrix designs. Despite the previous works, the understanding of the Lasso estimator is still limited. Analysis based on the irrepresentability condition often offers only scaling guarantees with respect to (N, M, d) , or statements with strong assumptions on the regularization parameter. Besides, the AMP/replica-based analysis has been only limited to linear sparsity, i.e. $d/N = O(1)$ and $M/N = O(1)$ as $N \rightarrow \infty$, which may be somewhat unrealistic compared to real-world situations.

1.1 Contributions

In this work, we complement the drawbacks in both the irrepresentability condition approach and AMP / replica approach by theoretically analyzing the average performance of Lasso when $d = O(1)$, i.e. the *ultra-sparsity* case (Donoho et al., 1992; Bhadra et al., 2017), which is a more typical situation in certain applications such as materials informatics (Ghiringhelli et al., 2015; Kim et al., 2016; Piliand et al., 2016). Moreover, our result offers a necessary condition for support recovery in the limit $N, M \rightarrow \infty$. Specifically, our contributions are summarized as follows:

- We provide a new way to apply the replica method in the ultra-sparsity regime. This is done by explicitly handling the correlations and finite-size effects acting on the active set $\text{supp}(\mathbf{x}^0) = \{i \mid x_i^0 \neq 0, i = 1, \dots, N\}$, which is otherwise ignored in conventional analysis (Section 2.1, Claim 1).
- Using this enhanced replica method, we precisely evaluate the average property of Lasso under ultra-sparsity and standard Gaussian matrix design, i.e. each element of \mathbf{A} is i.i.d. according to a standard Gaussian distribution. This provides an extension to previous results derived from the AMP theory and the replica method, where linear sparsity is necessary for the analysis (Section 2.2, Claim 2).
- We derive a necessary condition for partial support recovery $\text{supp}(\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y})) \subseteq \text{supp}(\mathbf{x}^0)$ under some mild conditions (Assumption 1). Specifically, the number of false positives, and subsequently the model misselection probability vanishes only if $M > \alpha_C \log N$ for $N \rightarrow \infty$. This constant α_C is determined by the mean prediction error of an oracle (Section 2.3, Claim 3, 4).
- In addition to partial support recovery, the analysis also provides a necessary condition for perfect support recovery $\text{supp}(\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y})) = \text{supp}(\mathbf{x}^0)$, which generalizes the sample complexity bound given by Wainwright (2009b) for i.i.d. Gaussian noise distributions in the limit $d \rightarrow \infty$ to more general noise distributions under constant d (Section 2.3, Claim 5).
- We demonstrate that our theory agrees well with experiment by conducting extensive numerical simulations (Section 3).

Note that all of the results are derived from the enhanced replica method, which is yet to be proven rigorously; hence the statements are presented as claims.

1.2 Related Work

Irrepresentability Condition. As aforementioned, the irrepresentability condition, first introduced by Mein-

shausen and Bühlmann (2006) and Zhao and Yu (2006), has been an important cornerstone, as it establishes a sufficient condition for perfect support recovery. This condition indicates whether the covariates, i.e. the columns of \mathbf{A} , are linearly independent enough to be distinguishable from one another, and hence variable selection is relatively feasible. It has been revealed that Lasso is an “optimal” support estimator in the sublinear regime $d = o(N)$, i.e. Lasso has its success/failure threshold for sample complexity in the same order as the informational-theoretical one (Fletcher et al., 2009; Wainwright, 2009a). However, little is known about the constants involved in these conditions. Wainwright (2009b) provided necessary and sufficient conditions for perfect support recovery under random Gaussian matrices for diverging d . This is a simple and explicit bound which depends on the regularization parameter and intensity of the noise, which is restricted to i.i.d. Gaussian. Focusing on the case $d = O(1)$, Dossal et al. (2012) derived sufficient conditions for partial and perfect support recovery under deterministic noise, whose bound is similar to the one given in Wainwright (2009b).

AMP theory. A particular line of work has aimed in assessing the properties of Lasso under general random matrix designs via careful analysis of the dynamical behavior of the AMP algorithm (Kabashima, 2003; Donoho et al., 2009; Takahashi and Kabashima, 2022), whose convergence point coincides with (2) in the large N limit. Rather than establishing inequality bounds or conditions, the objective is to establish sharp results on the Lasso for a random instance of (\mathbf{A}, \mathbf{y}) . Although analysis is limited to linear sparsity regime, powerful and precise results have been proven rigorously under this framework (Bayati and Montanari, 2012). For instance, Su et al. (2017) and Wang et al. (2020) determine the possible rate of false positives and true positives achievable under certain settings, which can be obtained by solving a small set of nonlinear equations. Nevertheless, the analysis does not give insight on support recovery, since this is impossible in the linear sparsity regime (Fletcher et al., 2009; Wainwright, 2009a).

Replica method. Results similar to those from AMP theory have also been derived by using the non-rigorous replica method in statistical mechanics. Unlike AMP theory, which is based on a convergence analysis of a particular algorithm, the replica method aims at directly calculating the average over (\mathbf{A}, \mathbf{y}) of a cumulant generating function for some probability distribution, i.e. of the form $K_\phi(t) = \mathbb{E}_{\mathbf{A}, \mathbf{y}} \log \int d\mathbf{x} e^{t\phi(\mathbf{x})} p(\mathbf{x}|\mathbf{A}, \mathbf{y})$. This calculation is often encountered in the field of statistics, where one is interested in the average behavior of a statistical model. While lacking a complete proof, this method has been successful in predicting the average performance of machine learning and optimization methods under general random designs in the linear sparsity regime (Vehkaperä

et al., 2016; Zdeborová and Krzakala, 2016). In fact, under certain assumptions, the average predictions given by the replica method have been proven to be consistent with the asymptotic results obtained from AMP theory and other rigorous methods (Stojnic, 2013; Thrampoulidis et al., 2018). Similar to AMP theory, however, reliable adaptations of this method outside linear sparsity are still open problems. Previous research such as Abbara et al. (2020), Meng et al. (2021a) and Meng et al. (2021b) analyzed the performance of sparse Ising model selection using a variation of the replica method. However, this was accomplished through a series of ansatzes which are generally difficult to justify theoretically.

1.3 Preliminaries

Here we summarize the notations used in this paper. The expression $\|\cdot\|$ denotes the ℓ_2 norm. The active set S is defined as the support of the d -sparse true signal \mathbf{x}^0 , $S := \text{supp}(\mathbf{x}^0) = \{i \mid x_i^0 \neq 0, i = 1, \dots, N\}$. Define $\tilde{N} := N - d$, the size of the inactive set. The matrix \mathbf{A}_S denotes the submatrix constructed by concatenating the columns of \mathbf{A} with indices in S . The vector \mathbf{x}_S^0 denotes the subvector of \mathbf{x}^0 with indices in S . For simplicity, \mathbf{x}^0 is assumed to be a deterministic, although this can be extended to random signals trivially. The expression $\mathbb{E}_{\mathbf{A}, \mathbf{y}}$ denotes the average over the joint probability with respect to the pair (\mathbf{A}, \mathbf{y}) , i.e.

$$\begin{aligned} & \mathbb{E}_{\mathbf{A}, \mathbf{y}}(\dots) \\ &= \int d\mathbf{y} d\mathbf{A} d\xi p_\xi(\xi)(\dots) \frac{e^{-\frac{1}{2}\text{Tr}\mathbf{A}^\top \mathbf{A}}}{(2\pi)^{(NM/2)}} \delta(\mathbf{y} - \mathbf{A}\mathbf{x}^0 - \xi), \end{aligned}$$

where $\delta(\cdot)$ denotes the Dirac delta function. The definition of $\mathbb{E}_{\mathbf{A}_S, \mathbf{y}}$ follows straightforwardly from the above. Also, define $D\mathbf{z}$ as the standard Gaussian measure, $D\mathbf{z} = d\mathbf{z} e^{-\|\mathbf{z}\|^2/2} / (2\pi)^{n/2}$ for $\mathbf{z} \in \mathbb{R}^n$. Given $(\mathbf{A}, \mathbf{x}^0, \mathbf{y})$ and regularization parameter λ , the oracle Lasso estimator is defined as $\hat{\mathbf{x}}_\lambda(\mathbf{A}_S, \mathbf{y})$, which is the Lasso estimator with the true support identified beforehand. It is also convenient to define the oracle Lasso fit, defined by $\gamma_\lambda(\mathbf{y}) := \mathbf{A}_S \hat{\mathbf{x}}_\lambda(\mathbf{A}_S, \mathbf{y})$, with its dependence on \mathbf{A}_S suppressed for convenience. Given configuration $(\mathbf{A}, \mathbf{x}^0, \mathbf{y})$, and regularization parameter λ , the number of false positives FP and the number of true positives TP of the lasso estimator is defined as

$$\text{FP}(\mathbf{A}, \mathbf{y}) = \#\{S^C \cap \text{supp}(\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y}))\}, \quad (3)$$

$$\text{TP}(\mathbf{A}, \mathbf{y}) = \#\{S \cap \text{supp}(\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y}))\}, \quad (4)$$

where S^C denotes the complement of set S from $\{1, \dots, N\}$. Without confusion, the dependence on $(\mathbf{A}_S, \mathbf{y})$ is suppressed for convenience.

We say that an event A holds with asymptotically high probability (w.a.h.p.) if there exists a constant $c > 0$ such

that $\Pr[A] > 1 - O(N^{-c})$. We also say that A holds with probability approaching one (w.p.a.1) if $\Pr[A] > 1 - o(1)$ as $N \rightarrow \infty$.

2 Replica analysis

Define the Boltzmann distribution as

$$\begin{aligned} & P_\beta(\mathbf{x} \mid \mathbf{A}, \mathbf{y}) \\ &:= Z_\beta^{-1}(\mathbf{A}, \mathbf{y}) \exp\left(-\frac{\beta}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \beta M \lambda \|\mathbf{x}\|_1\right), \end{aligned} \quad (5)$$

where $Z_\beta(\mathbf{A}, \mathbf{y})$ is the normalization constant. Note that in the limit $\beta \rightarrow \infty$, (5) converges to a point-wise distribution concentrated on the Lasso estimator $\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y})$. The main objective of our analysis is to calculate the average of the logarithm of $Z_\beta(\mathbf{A}, \mathbf{y})$ over the random variables (\mathbf{A}, \mathbf{y}) in the limit $\beta \rightarrow \infty$, which is called the free energy or the cumulant generating function

$$\mathcal{F} = - \lim_{\beta \rightarrow \infty} \beta^{-1} \mathbb{E}_{\mathbf{A}, \mathbf{y}} \log Z_\beta(\mathbf{A}, \mathbf{y}). \quad (6)$$

The properties of $\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y})$ averaged over the population of (\mathbf{A}, \mathbf{y}) can then be assessed by taking appropriate derivatives of \mathcal{F} .

Although (6) is difficult to calculate straightforwardly, this can be resolved by using the replica method (Mézard and Montanari, 2009; Mézard et al., 1986), which is based on the following equality

$$\mathbb{E}_{\mathbf{A}, \mathbf{y}} \log Z_\beta(\mathbf{A}, \mathbf{y}) = \lim_{n \rightarrow +0} \frac{\mathbb{E}_{\mathbf{A}, \mathbf{y}} Z_\beta^n(\mathbf{A}, \mathbf{y}) - 1}{n}. \quad (7)$$

Instead of handling the cumbersome log expression in (6) directly, one calculates the average of the n -th power of Z_β for $n \in \mathbb{N}$, analytically continues this expression to $n \in \mathbb{R}$, and finally takes the limit $n \rightarrow +0$. Based on this replica "trick", it suffices to calculate

$$\begin{aligned} & \mathbb{E}_{\mathbf{A}, \mathbf{y}} Z_\beta^n(\mathbf{A}, \mathbf{y}) = \mathbb{E}_{\mathbf{A}, \mathbf{y}} \int \prod_{a=1}^n d\mathbf{x}^a \\ & \exp\left(-\frac{\beta}{2} \sum_{a=1}^n \|\mathbf{A}\mathbf{x}^a - \mathbf{y}\|^2 - \beta M \lambda \sum_{a=1}^n \|\mathbf{x}^a\|_1\right). \end{aligned} \quad (8)$$

up to the first order of n to take the $n \rightarrow +0$ limit in the right hand side of (7).

2.1 Outline of the derivation

Here, we only give a brief outline of the derivation; for details, see Supplementary Materials. Rewriting $\Delta_i^a := x_i^a$ ($i \notin S$), it is convenient to introduce the auxiliary variable $h_\mu^a \equiv \sum_{i \notin S} A_{\mu i} \Delta_i^a$ ($\mu = 1, \dots, M$), which accounts for the effect from the variables not in the true support in each replica a . A crucial observation

is that $\{A_{\mu i}\}_{1 \leq \mu \leq M, i \notin S}$ is statistically independent from $(\mathbf{A}_S, \mathbf{y})$, which allows the average to be taken individually. By taking the average over the Gaussian variables $\{A_{\mu i}\}_{1 \leq \mu \leq M, i \notin S}$ first, we find that h_μ^a is Gaussian with zero mean and covariance $\mathbb{E}h_\mu^a h_\nu^b = \delta_{\mu\nu} \sum_{i \notin S} \Delta_i^a \Delta_i^b$. By assuming the *replica symmetric* (RS) ansatz (Mézard et al., 1986)

$$\sum_{i \notin S} \Delta_i^a \Delta_i^b := \begin{cases} Q & a = b \\ Q - \chi/\beta & \text{otherwise} \end{cases}, \quad (9)$$

the integral for the replicated vectors $\{\Delta^a\}_{a=1}^n$ over the whole $\mathbb{R}^{\tilde{N} \times n}$ space is restricted to a subspace satisfying the constraints (9). More explicitly, one can rewrite (8) as

$$\mathbb{E}_{\mathbf{A}_S, \mathbf{y}} \int \prod_{a=1}^n d\Delta^a \int dQ d\chi e^{-\beta M \lambda \sum_{a=1}^n \|\Delta^a\|_1} \mathcal{I} \mathcal{L}, \quad (10)$$

where \mathcal{I} corresponds to the contribution from the RS constraint: i.e.

$$\mathcal{I} := \prod_{a=1}^n \delta\left(Q - \sum_{i \notin S} (\Delta_i^a)^2\right) \prod_{a < b} \delta\left(Q - \frac{\chi}{\beta} - \sum_{i \notin S} \Delta_i^a \Delta_i^b\right), \quad (11)$$

and \mathcal{L} is the contribution from the second line of (8), albeit simplified as a result of replica symmetry:

$$\mathcal{L} := \int D\mathbf{z} \left(\int d\mathbf{x}_S e^{-M\beta G(\mathbf{x}_S; \mathbf{z})} \right)^n, \quad (12)$$

$$G(\mathbf{x}_S; \mathbf{z}) := \frac{\|\mathbf{A}_S \mathbf{x}_S + \sqrt{Q} \mathbf{z} - \mathbf{y}\|^2}{2M(1 + \chi)} + \lambda \|\mathbf{x}_S\|_1.$$

By using the Fourier representation of the delta function, (11) can be further rewritten as

$$\mathcal{I} = \int_{-i\infty}^{+i\infty} d\hat{Q} d\hat{\chi} e^{\frac{Mn\beta}{2}(Q\hat{Q} + (n-1)\chi\hat{\chi} - n\beta Q\hat{\chi})} \times \int D\hat{\mathbf{z}} e^{-\frac{M\beta\hat{Q}}{2} \sum_{a=1}^n \|\Delta^a\|^2 + \beta\sqrt{M\hat{\chi}} \mathbf{z}^T \Delta^a + o(\beta)}. \quad (13)$$

Using this expression, the integral with respect to $\{\Delta^a\}_{1 \leq a \leq n}$ in (10) can be calculated analytically. Performing the saddle point approximation for large M to the integrals with respect to $(Q, \hat{Q}, \chi, \hat{\chi})$, and finally taking the limit $\beta \rightarrow \infty$ after $n \rightarrow +0$ in (7) yields the following expression for \mathcal{F} .

Claim 1. *The free energy is given by*

$$\mathcal{F} = \mathbb{E}_{\mathbf{A}_S, \mathbf{y}} \text{Extr}_{\Theta} \left\{ -\frac{Q\hat{Q} - \chi\hat{\chi}}{2} - \frac{\tilde{N}}{2\hat{Q}} \left[(\Lambda + \hat{\chi}) \text{erfc} \left(\sqrt{\frac{\Lambda}{2\hat{\chi}}} \right) - \sqrt{\frac{2\Lambda\hat{\chi}}{\pi}} e^{-\Lambda/2\hat{\chi}} \right] + \int D\mathbf{z} \min_{\mathbf{x}_S} G(\mathbf{x}_S; \mathbf{z}) \right\}. \quad (14)$$

Here, $\Lambda := (M\lambda)^2$, erfc is the complementary error function $\text{erfc}(x) := 2/\sqrt{\pi} \int_x^\infty dt e^{-t^2}$, and Extr refers to the extremum condition with respect to $\Theta := (Q, \hat{Q}, \chi, \hat{\chi})$, which are random variables dependent on $(\mathbf{A}_S, \mathbf{y})$.

Straightforward calculation shows that the extremum conditions are given by

$$Q = \frac{\tilde{N}}{\hat{Q}^2} \left[(\Lambda + \hat{\chi}) \text{erfc} \left(\sqrt{\frac{\Lambda}{2\hat{\chi}}} \right) - \sqrt{\frac{2\Lambda\hat{\chi}}{\pi}} e^{-\frac{\Lambda}{2\hat{\chi}}} \right], \quad (15)$$

$$\chi = \frac{\tilde{N}}{\hat{Q}} \text{erfc} \left(\sqrt{\frac{\Lambda}{2\hat{\chi}}} \right), \quad (16)$$

$$\hat{Q} = \frac{M}{1 + \chi} - \frac{1}{1 + \chi} \int D\mathbf{z} \nabla \cdot \gamma_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y}) = \frac{M - \int D\mathbf{z} \|\hat{\mathbf{x}}_{(1+\chi)\lambda}(\mathbf{A}_S, \sqrt{Q}\mathbf{z} + \mathbf{y})\|_0}{1 + \chi}, \quad (17)$$

$$\hat{\chi} = \frac{\int D\mathbf{z} \|\gamma_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y}) - \sqrt{Q}\mathbf{z} - \mathbf{y}\|^2}{(1 + \chi)^2}, \quad (18)$$

where the second equality in (17) is from Theorem 1 in Tibshirani and Taylor (2012). Note that the dependence of Θ on $(\mathbf{A}_S, \mathbf{y})$ is not explicitly written for sake of simplicity. This evaluation of \mathcal{F} reduces the high-dimensional integral over \mathbf{A} and \mathbf{y} to an average over a four-dimensional extremum problem involving a M -dimensional integral with respect to \mathbf{z} , which can be numerically computed via iterative substitution and Monte Carlo sampling over $(\mathbf{A}_S, \mathbf{y})$ and \mathbf{z} .

It is interesting to compare our replica analysis in the large N and M limit to the ones considering linear sparsity (Kabashima et al., 2009; Vekkerperä et al., 2016). In linear sparsity, the lasso estimator's statistical property can effectively be described by a population of N decoupled, independent scalar estimators under Gaussian noise with identical intensity as $N \rightarrow \infty$. This is often referred to as the *decoupling principle* in information theory; see Guo and Verdú (2005) and Bayati and Montanari (2011) for details. In the ultra-sparse case, the elements of the Lasso estimator in the active set, consisting of $d = O(1)$ terms, cannot be expected to decouple, as finite-size effects of non-Gaussian and correlated nature are expected to be significant to describe its profile. This is why a d -body optimization procedure and the average with respect to $(\mathbf{A}_S, \mathbf{y})$ appears explicitly in (14). On the other hand, the decoupling principle is implicitly employed for the \tilde{N} non-active variables conditioned on $(\mathbf{A}_S, \mathbf{y})$. More explicitly, for each configuration of $(\mathbf{A}_S, \mathbf{y})$, each element of the non-active Lasso estimator is statistically equivalent to

$$\begin{aligned} (\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y}))_{i \notin S} &\sim g_\lambda(\hat{Q}, \sqrt{\hat{\chi}} z_i) \\ &= \min_x \left(\frac{\hat{Q}}{2} x^2 - \sqrt{\hat{\chi}} z_i x + M\lambda|x| \right), \end{aligned} \quad (19)$$

where z_i are i.i.d. according to $\mathcal{N}(0, 1)$. Note that the decoupling principle, rigorously proven under AMP theory, does not necessarily need N and M to diverge at the same rate (Rush and Venkataramanan, 2018).

2.2 Performance assessment of Lasso

The free energy allows convenient evaluation of averages of certain functions of the estimator. More explicitly, for a function $\Psi : \mathbb{R}^N \rightarrow \mathbb{R}$, its average with respect to the Boltzmann distribution (5) and (\mathbf{A}, \mathbf{y}) is given by

$$\begin{aligned} \langle \Psi(\mathbf{x}) \rangle &:= \lim_{\beta \rightarrow \infty} \mathbb{E}_{\mathbf{A}, \mathbf{y}} \int d\mathbf{x} P_\beta(\mathbf{x} | \mathbf{A}, \mathbf{y}) \Psi(\mathbf{x}) \\ &= - \lim_{\beta \rightarrow \infty} \lim_{h \rightarrow 0} \frac{\partial}{\partial h} \beta^{-1} \mathbb{E}_{\mathbf{A}, \mathbf{y}} \log Z_\beta(\mathbf{A}, \mathbf{y}; h\Psi), \end{aligned} \quad (20)$$

where

$$Z_\beta(\mathbf{A}, \mathbf{y}; h\Psi) := \int d\mathbf{x} e^{-\frac{\beta}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \beta M \lambda \|\mathbf{x}\|_1 - \beta h \Psi(\mathbf{x})}. \quad (21)$$

For a class of functions Ψ , the above can be calculated trivially, which we state in the following claim:

Claim 2 (Average with respect to active and inactive sets). For arbitrary functions $\psi : \mathbb{R} \rightarrow \mathbb{R}$ and $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}$, we have

$$\left\langle \sum_{i \notin S} \psi(x_i) \right\rangle = \tilde{N} \mathbb{E}_{\mathbf{A}_S, \mathbf{y}} \int D\mathbf{z} \psi(g_\lambda(\hat{Q}, \sqrt{\hat{\chi}}z)), \quad (22)$$

and

$$\langle \Psi(\mathbf{x}_S) \rangle = \mathbb{E}_{\text{eff}} \Psi(\hat{\mathbf{x}}_{(1+\chi)\lambda}(\mathbf{A}_S, \sqrt{Q}\mathbf{z} + \mathbf{y})), \quad (23)$$

where $\mathbb{E}_{\text{eff}} := \mathbb{E}_{\mathbf{A}_S, \mathbf{y}} \int D\mathbf{z}$, and $(Q, \hat{Q}, \hat{\chi}, \chi)$ is given by the solution of the extremum conditions (15)–(18) for each $(\mathbf{A}_S, \mathbf{y})$. In particular, performance measures such as the average of true positives (TP), false positives (FP) and ℓ_2 error $\epsilon_x := \|\mathbf{x}_\lambda(\mathbf{A}, \mathbf{y}) - \mathbf{x}^0\|^2$ is given by

$$\langle \text{TP} \rangle = \mathbb{E}_{\text{eff}} \left\| \hat{\mathbf{x}}_{(1+\chi)\lambda}(\mathbf{A}_S, \sqrt{Q}\mathbf{z} + \mathbf{y}) \right\|_0, \quad (24)$$

$$\langle \text{FP} \rangle = \tilde{N} \mathbb{E}_{\mathbf{A}_S, \mathbf{y}} \text{erfc} \left(\sqrt{\frac{\Lambda}{2\hat{\chi}}} \right), \quad (25)$$

$$\langle \epsilon_x \rangle = \mathbb{E}_{\text{eff}} \left(Q + \left\| \hat{\mathbf{x}}_{(1+\chi)\lambda}(\mathbf{A}_S, \sqrt{Q}\mathbf{z} + \mathbf{y}) - \mathbf{x}_S^0 \right\|^2 \right). \quad (26)$$

2.3 Necessary condition for support recovery

A particular topic of interest is partial support recovery, and the minimum number of samples M necessary for the false positives to vanish in the limit $N \rightarrow \infty$. Although the fixed point equations (15)–(18) do not admit a closed form solution, a necessary condition in terms of the sample complexity can be derived under the following mild conditions:

Assumption 1.

A: (Uniqueness of fixed point) The solutions of the fixed point equations (15)–(18) are unique and satisfy $(Q, \hat{Q}, \chi, \hat{\chi}) \in (0, \infty)^4$.

B: (Concentration of the oracle Lasso estimator) The random variable

$$s_\lambda^{(M)} := \frac{1}{M} \|\gamma_\lambda(\mathbf{y}) - \mathbf{y}\|^2$$

has finite mean $\bar{s}_\lambda^{(M)}$ and variance converging to zero.

C: (Bounded variance of noise distribution) The distribution p_ξ satisfies

$$\Gamma^{(M)} := \frac{1}{M} \int d\xi p_\xi(\xi) \|\xi\|^2 < C$$

for some constant C .

Claim 3 (Necessary sample complexity for asymptotically zero false positives). Let M diverge with N with scaling $M = \alpha \log N$ ($\alpha > 0$). Under Claim 1 and Assumption 1, if there exists a constant $c > 0$ such that $\langle \text{FP} \rangle < O(N^{-c})$ in the limit $N \rightarrow \infty$, then

$$\alpha(1 + \epsilon) > \alpha_C = \frac{\bar{s}_\lambda}{2\lambda^2}, \quad (27)$$

holds for any constant $\epsilon > 0$, where $\bar{s}_\lambda = \lim_{M \rightarrow \infty} \bar{s}_\lambda^{(M)}$.

The proof is postponed to Section 4. From this claim, the necessary sample complexity for partial support recovery follows immediately:

Claim 4 (Necessary sample complexity for partial support recovery). Under the settings in Claim 3, if $\text{supp}(\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y})) \subseteq \text{supp}(\mathbf{x}^0)$ w.a.h.p., then $\alpha > \alpha_C$.

By definition, \bar{s}_λ is the prediction error of the oracle, which is given the sensing submatrix \mathbf{A}_S and observation vector \mathbf{y} . This is reminiscent of the primal-dual witness construction in Wainwright (2009b), where sufficient conditions for asymptotically zero FPs are derived by solving the oracle Lasso first, and observing whether the oracle solution concatenated with $N - d$ zero elements is a unique solution of the original Lasso problem (2).

Furthermore, the necessary condition for perfect support recovery can also be derived using Claim 3.

Claim 5 (Necessary sample complexity for perfect support recovery). Under the settings in Claim 3, suppose $\text{supp}(\hat{\mathbf{x}}_\lambda(\mathbf{A}, \mathbf{y})) = \text{supp}(\mathbf{x}^0)$ holds w.a.h.p. Then

$$\alpha(1 + \epsilon) > 2 \left(d + \frac{\Gamma}{\lambda^2} \right), \quad (28)$$

holds for any constant $\epsilon > 0$, where $\Gamma = \lim_{M \rightarrow \infty} \Gamma^{(M)}$.

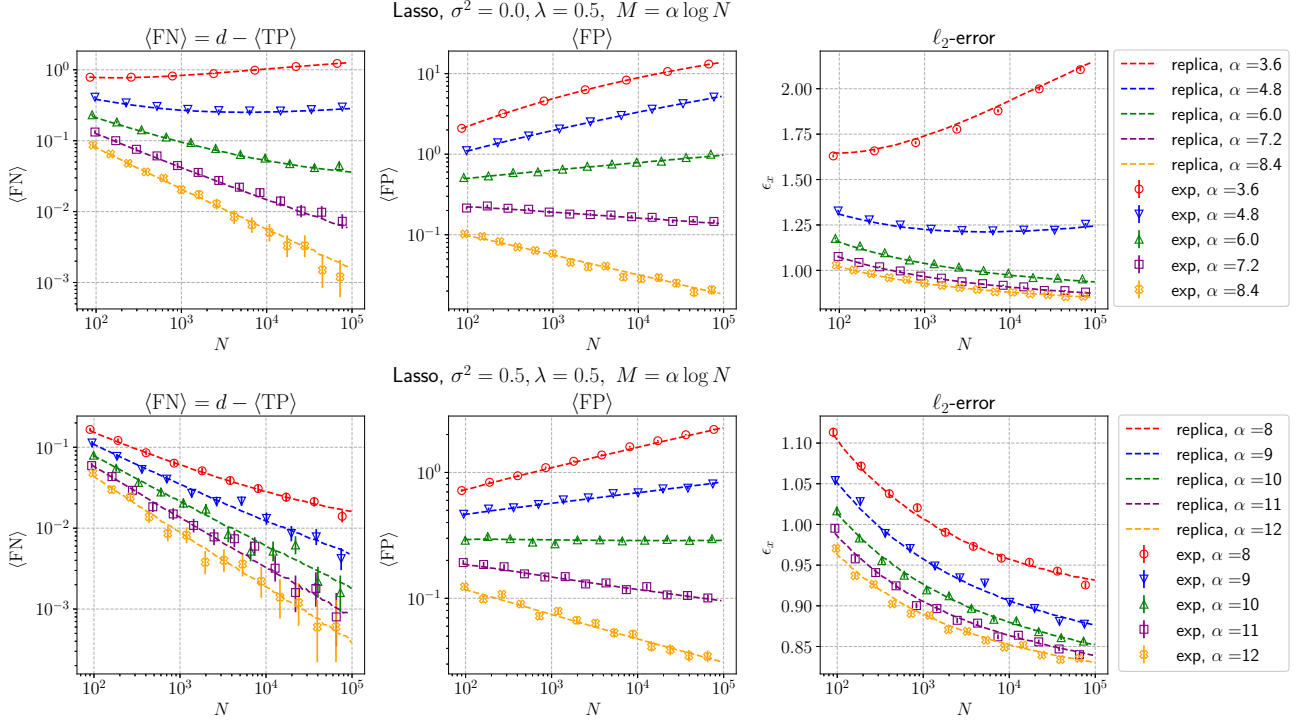


Figure 1: Average values of false negatives ($d - \text{TP}$), false positives and ℓ_2 error for $(\lambda, \sigma^2) = (0.5, 0.0)$ (upper panels) and $(\lambda, \sigma^2) = (0.5, 0.5)$ (lower panels) with M given by $M = \alpha \log N$. Error bars for $\langle \text{FN} \rangle$ and $\langle \text{FP} \rangle$ represent the 95% interval of the mean, assuming that the samples from the 10^4 experimental runs follow a binomial distribution. Error bars for ℓ_2 -error represent the standard error obtained from 10^4 experimental runs.

Note that in the special case of Gaussian noise with variance σ^2 , we have $\Gamma = \sigma^2$, which extends the result of Wainwright (2009b), Theorem 4 to the case $d = O(1)$. Moreover, our result can be applied to any noise distribution satisfying Assumption 1.C.

3 Numerical experiments

3.1 Non-asymptotic results

To verify the derived results based on Claim 1, numerical experiments were conducted. For simplicity, we consider the case where the active set has size $d = 3$ with $\mathbf{x}_S^0 = \mathbf{1}_3$, and $\boldsymbol{\xi}$ is generated from a Gaussian distribution with variance σ^2 . Here, the value of d is taken to be small enough such that finite-size effects are nonignorable. The values of $\langle \text{TP} \rangle$, $\langle \text{FP} \rangle$ and ϵ_x obtained from our replica predictions (24)-(26) are compared with the average over 10^4 experimental runs. The average with respect to $(\mathbf{A}_S, \mathbf{y})$ for obtaining the replica prediction was approximated using a Monte Carlo procedure over 10^6 samples.

Figure 1 shows that all three values from theory and experiment are in good agreement for parameters $(\lambda, \sigma^2) = (0.5, 0.0)$ and $(0.5, 0.5)$.

3.2 Asymptotic results

Claims 3 and 4 are also verified via numerical experiments; see Supplementary Materials for numerical experiments on Claim 5. In order to access the critical point α_C in (27), Monte Carlo experiments were conducted to evaluate $s_\lambda^{(M)}$ for different values of M . Figure 2 shows the value of $s_\lambda^{(M)}$ at $(\lambda, \sigma^2) = (0.5, 0.0)$ and $(0.5, 0.5)$ for both $\mathbf{x}_S^0 = \mathbf{1}_3$ and $\mathbf{x}_S^0 = [\frac{1}{3}, \frac{2}{3}, 1]$. From its asymptotic behavior, α_C can be evaluated as the values given in Table 1. Interestingly, for the case $\mathbf{x}_S^0 = \mathbf{1}_3$, s_λ approaches 6 and 10 for $\sigma^2 = 0$ and 0.5 respectively, which is equivalent to $2(d + \Gamma/\lambda^2)$ given in Claim 5.

Table 1: Values of α_C evaluated from figure 2.

\mathbf{x}_S^0	(λ, σ^2)	α_C
[1, 1, 1]	(0.5, 0.0)	6.00
[1, 1, 1]	(0.5, 0.5)	10.0
[1/3, 2/3, 1]	(0.5, 0.0)	4.89
[1/3, 2/3, 1]	(0.5, 0.5)	8.89

Figure 3 shows the average number of FP and partial support recovery probability over 10,000 experimental runs for α in the vicinity of the numerically evaluated α_C for different values of N . We observe that for $\alpha < \alpha_C$, the average

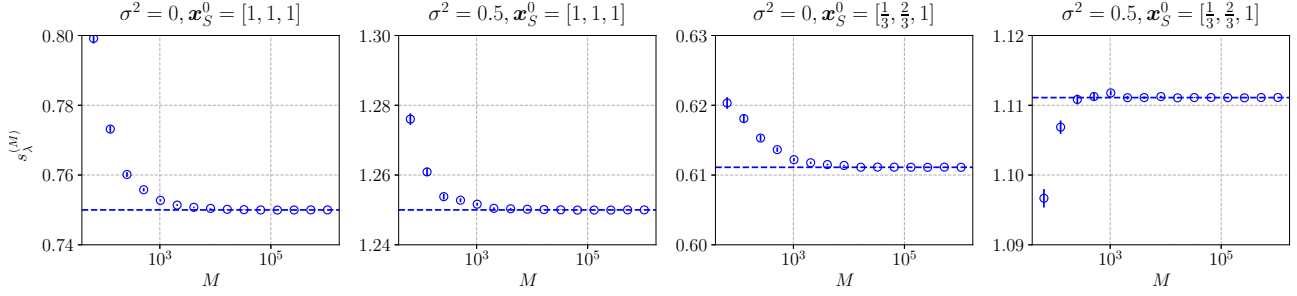


Figure 2: Value of $\bar{s}_\lambda^{(M)}$ for M . The asymptotic values of $\bar{s}_\lambda^{(M)}$ (blue dashed lines) are evaluated as 0.750, 1.250, 0.611 and 1.111 from left to right. Error bars represent the standard error obtained from 10,000 Monte Carlo samples.

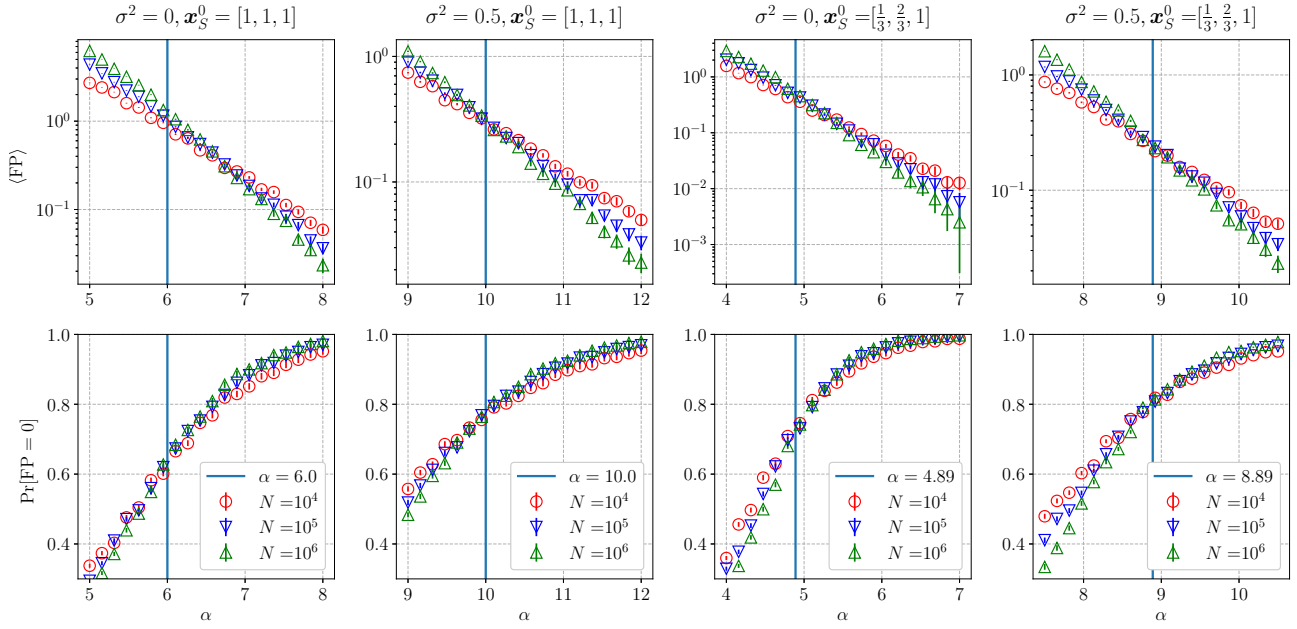


Figure 3: Average number of false positives (upper panels) and partial support recovery probability (lower panels) near complexity $\alpha = \alpha_C$ (blue vertical lines). Error bars represent the standard error obtained from 10,000 experimental runs. For $\alpha < \alpha_C$, the number of false positives is consistently nondecreasing with respect to N , while the partial support recovery probability is consistently nonincreasing with respect to N , which is in agreement with Claims 3 and 4.

FP is consistently nondecreasing with respect to N , while partial support recovery probability is consistently nonincreasing with respect to N .

4 Proofs

4.1 Proof of Claim 3

The following lemmas will be useful in the proof.

Lemma 1 (Lemma 1, Dossal et al. (2012)). *There is a finite increasing sequence $(\lambda_t)_{t \leq K}$ with $\lambda_0 = 0$ such that for all $t < K$, the sign and support of $\hat{\mathbf{x}}_\lambda(\mathbf{A}_S, \mathbf{y})$ are constant on each interval $(\lambda_t, \lambda_{t+1})$.*

Lemma 2 (Lemma 1, Tibshirani and Taylor (2012)). *The Lasso fit is 1-Lipschitz continuous with respect to ℓ_2 norm.*

Lemma 3 (Theorem II.13, Davidson and Szarek (2001)). *Let $\mathbf{A} \in \mathbb{R}^{M \times d}$ be a random matrix with i.i.d standard Gaussian entries. The largest and smallest eigenvalue of $\mathbf{B} = \mathbf{A}^\top \mathbf{A}$ satisfy*

$$\Pr \left[\lambda_{\max}(\mathbf{B}) \geq \left(\sqrt{M} + \sqrt{d} + t \right)^2 \right] \leq e^{-\frac{t^2}{2}} \quad (29)$$

for $t > 0$ and

$$\Pr \left[\lambda_{\min}(\mathbf{B}) \leq \left(\sqrt{M} - \sqrt{d} - t \right)^2 \right] \leq e^{-\frac{t^2}{2}} \quad (30)$$

for $0 < t < \sqrt{M} - \sqrt{d}$.

We now prove Claim 3. Define

$$s_{\lambda, Q}^{(M)} := \frac{1}{M} \int D\mathbf{z} \left\| \gamma_\lambda(\mathbf{y} + \sqrt{Q}\mathbf{z}) - (\mathbf{y} + \sqrt{Q}\mathbf{z}) \right\|^2.$$

Let us evaluate the difference between $s_{(1+\chi)\lambda, Q}^{(M)}$ and $s_{\lambda, 0}^{(M)} = s_{\lambda}^{(M)}$ when $\langle \text{FP} \rangle < O(N^{-c})$. Using the Cauchy Schwartz inequality and symmetry $\gamma_{\lambda}(\mathbf{y}) = -\gamma_{\lambda}(-\mathbf{y})$,

$$\begin{aligned} & M(s_{(1+\chi)\lambda, Q}^{(M)} - s_{\lambda}^{(M)}) \\ & \geq - \int D\mathbf{z} \left\| \gamma_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y}) - \gamma_{\lambda}(\mathbf{y}) - \sqrt{Q}\mathbf{z} \right\| \\ & \quad \times \left\| \gamma_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y}) - \gamma_{\lambda}(-\mathbf{y}) - 2\mathbf{y} - \sqrt{Q}\mathbf{z} \right\|. \end{aligned} \quad (31)$$

The triangle inequality and Lemma 2 implies that

$$\begin{aligned} & \left\| \gamma_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y}) - \gamma_{\lambda}(\mathbf{y}) - \sqrt{Q}\mathbf{z} \right\| \\ & \leq \sqrt{Q}\|\mathbf{z}\| + \left\| \gamma_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y}) - \gamma_{(1+\chi)\lambda}(\mathbf{y}) \right\| \\ & \quad + \left\| \gamma_{(1+\chi)\lambda}(\mathbf{y}) - \gamma_{\lambda}(\mathbf{y}) \right\| \\ & \leq 2\sqrt{Q}\|\mathbf{z}\| + \left\| \gamma_{(1+\chi)\lambda}(\mathbf{y}) - \gamma_{\lambda}(\mathbf{y}) \right\|, \end{aligned} \quad (32)$$

and similarly,

$$\begin{aligned} & \left\| \gamma_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y}) - \gamma_{\lambda}(-\mathbf{y}) - 2\mathbf{y} - \sqrt{Q}\mathbf{z} \right\| \\ & \leq 2\sqrt{Q}\|\mathbf{z}\| + 4\|\mathbf{y}\| + \left\| \gamma_{(1+\chi)\lambda}(\mathbf{y}) - \gamma_{\lambda}(\mathbf{y}) \right\|. \end{aligned} \quad (33)$$

To derive a bound for the last term in (32) and (33), Lemma 1 is employed. Let the support and sign of $\hat{\mathbf{x}}_{\lambda}(\mathbf{A}_S, \mathbf{y})$ be constant in intervals $(M\lambda_t, M\lambda_{t+1})$ ($t = 0, \dots, K-1$), where $\lambda = \lambda_0 < \dots < \lambda_K = (1+\chi)\lambda$. Let the support set in interval $(\lambda_t, \lambda_{t+1})$ be given by I_t , and define $\mathbf{s}_t \in \{-1, 0, 1\}^{|I_t|}$ be the sign vector of $\hat{\mathbf{x}}_{\lambda'}(\mathbf{A}_S, \mathbf{y})$ ($\lambda' \in (\lambda_t, \lambda_{t+1})$) restricted to I_t . From the KKT conditions, the Lasso fit is expressed as

$$\gamma_{\lambda_t}(\mathbf{y}) = \mathbf{A}_{SI_t} \mathbf{A}_{SI_t}^+ \mathbf{y} - M\lambda_t \mathbf{A}_{SI_t} (\mathbf{A}_{SI_t}^T \mathbf{A}_{SI_t})^{-1} \mathbf{s}_t,$$

where \mathbf{M}^+ denotes the pseudoinverse of matrix \mathbf{M} . We deduce

$$\begin{aligned} & \left\| \gamma_{(1+\chi)\lambda}(\mathbf{y}) - \gamma_{\lambda}(\mathbf{y}) \right\| \leq \sum_{t=0}^{K-1} \left\| \gamma_{\lambda_t}(\mathbf{y}) - \gamma_{\lambda_{t+1}}(\mathbf{y}) \right\| \\ & \leq M\lambda \sum_{t=0}^{K-1} (\lambda_{t+1} - \lambda_t) \left\| \mathbf{A}_{SI_t} (\mathbf{A}_{SI_t}^T \mathbf{A}_{SI_t})^{-1} \mathbf{s}_t \right\| \\ & \leq \chi M \sqrt{d\lambda^2} \max_t \sqrt{\rho((\mathbf{A}_{SI_t}^T \mathbf{A}_{SI_t})^{-1})}. \end{aligned}$$

Lemma 3, with the inclusion principle $\rho((\mathbf{A}_{SI_t}^T \mathbf{A}_{SI_t})^{-1}) \leq \rho((\mathbf{A}_S^T \mathbf{A}_S)^{-1})$ implies that w.a.h.p., $\left\| \gamma_{(1+\chi)\lambda}(\mathbf{y}) - \gamma_{\lambda}(\mathbf{y}) \right\| \leq 2\chi\sqrt{d\lambda^2 M}$. The relations (31) – (33), and inequality $\int D\mathbf{z} \|\mathbf{z}\| = \sqrt{2}\Gamma((M+1)/2)/\Gamma(M/2) < \sqrt{M}$ then leads to the following holding w.a.h.p.

$$\begin{aligned} & s_{(1+\chi)\lambda, Q}^{(M)} - s_{\lambda}^{(M)} \\ & \geq -4(\sqrt{Q} + \chi\sqrt{d\lambda^2}) \left(\sqrt{Q} + \chi\sqrt{d\lambda^2} + \frac{2\|\mathbf{y}\|}{\sqrt{M}} \right). \end{aligned} \quad (34)$$

We now use the following lemma which shows that Q and χ are negligible almost surely.

Lemma 4. *Under the assumptions of Claim 3, $\chi < N^{-c/2}$ and $Q < N^{-c/4}$ holds w.a.h.p.*

The proof is given in Supplementary Materials. Since $\|\mathbf{y}\| < \|\mathbf{A}_S \mathbf{x}^0\| + \|\boldsymbol{\xi}\|$ is bounded by M^2 w.p.a.1 from Lemma 3 and Assumption 1.C, the right hand side of eq. (34) is of $O(N^{-c/8})$ w.p.a.1. We therefore have

$$\Pr \left[\frac{\hat{\chi}}{M} = \frac{s_{(1+\chi)\lambda, Q}^{(M)}}{(1+\chi)^2} \geq s_{\lambda}^{(M)} - O(N^{-c/8}) \right] > 1 - o(1). \quad (35)$$

On the other hand, the extremum conditions (16) and (17) imply that $\hat{\chi}$ is always bounded.

Lemma 5. *Suppose the extremum conditions (15)–(18) are satisfied. Then, the variable $\hat{\chi}$ satisfies*

$$\frac{\hat{\chi}}{M} \leq \frac{1}{2} \alpha \lambda^2 \left(1 - (2\alpha + 1) \frac{\log M}{M} \right)^{-1}. \quad (36)$$

Combined with (35), for sufficiently large M

$$\Pr \left[s_{\lambda}^{(M)} \leq \frac{1}{2} \alpha \lambda^2 (1 + \epsilon) \right] > 1 - o(1), \quad (37)$$

holds for arbitrary constant $\epsilon > 0$. This implies that $\frac{1}{2} \alpha \lambda^2 (1 + \epsilon)$ must be larger than the median of $s_{\lambda}^{(M)}$. Now, the difference between the median and average is no larger than one standard deviation, which is negligible from Assumption 1.B. This yields the statement of the claim in the limit $M \rightarrow \infty$.

4.2 Proof of Claim 4

From Theorem 6 in Osborne et al. (2000), the number of false positives is bounded by $\min(M, N)$. Hence, we have $\langle \text{FP} \rangle < M \times \Pr[\text{FP} \neq 0] = O(N^{-c})$ for some $c > 0$. The statement of Claim 4 then follows from Claim 3.

4.3 Proof of Claim 5

From Claim 3 and 4, it suffices to show that

$$\mathbb{E} s_{\lambda}^{(M)} > d\lambda^2 + \Gamma_M - o(1), \quad (38)$$

The KKT conditions imply that w.p.a.1, $\mathbf{A}_S \hat{\mathbf{x}} = \mathbf{A}_S \mathbf{A}_S^+ (\mathbf{y} - M\lambda (\mathbf{A}_S^T)^T \mathbf{s})$, where we abbreviated $\hat{\mathbf{x}} := \hat{\mathbf{x}}_{\lambda}(\mathbf{A}_S, \mathbf{y})$, and $\mathbf{s} = \text{sgn}(\hat{\mathbf{x}})$. Therefore, $\mathbf{y} - \mathbf{A}_S \hat{\mathbf{x}}$ can be decomposed into a sum of two linearly independent vectors

$$\mathbf{y} - \mathbf{A}_S \hat{\mathbf{x}} = \mathbf{v} + \mathbf{v}_{\perp}, \quad (39)$$

where $\mathbf{v} := M\lambda \mathbf{A}_S (\mathbf{A}_S^T \mathbf{A}_S)^{-1} \mathbf{s}$, $\mathbf{v}_{\perp} := \mathcal{P}_{\ker(\mathbf{A}_S)}(\mathbf{y}) = \mathcal{P}_{\ker(\mathbf{A}_S)}(\boldsymbol{\xi})$, and $\mathcal{P}_{\ker(\mathbf{A}_S)}$ is the projection onto the kernel of \mathbf{A}_S . The average of the squared norm of \mathbf{v} can be

evaluated as

$$\mathbb{E}_{\text{eff}} \|\mathbf{v}\|^2 \geq \mathbb{E}_{\text{eff}} \frac{\Lambda d}{\lambda_{\min}(\mathbf{A}_S^T \mathbf{A}_S)} \geq \frac{M\lambda^2}{(1 + \sqrt{d/M})^2}, \quad (40)$$

where the last inequality follows from Jensen's inequality and $\mathbb{E}_{\text{eff}} \lambda_{\min}(\mathbf{A}_S^T \mathbf{A}_S) \geq (\sqrt{M} - \sqrt{d})^2$ (Davidson and Szarek, 2001).

To obtain a lower bound on the squared norm of \mathbf{v}_\perp , fix the vector $\boldsymbol{\xi}$. Noticing that entries of $\mathbf{A}_S^T \boldsymbol{\xi} / \|\boldsymbol{\xi}\|^2$ are i.i.d. standard Gaussian, the tail bound for χ^2 -random variables (Laurent and Massart, 2000) implies that for some constant $C > 0$,

$$\Pr \left[\|\mathbf{A}_S^T \boldsymbol{\xi}\|^2 \leq C \|\boldsymbol{\xi}\|^2 \log M \right] \geq 1 - \frac{1}{M}. \quad (41)$$

Using this inequality, (30) with $t = \sqrt{2 \log M}$ and the union bound, we have that

$$\begin{aligned} \mathbb{E}_{\text{eff}} \|\mathbf{v}_\perp\|^2 &\geq \mathbb{E}_{\text{eff}} \|\boldsymbol{\xi}\|^2 - \mathbb{E}_{\text{eff}} \frac{\|\mathbf{A}_S^T \boldsymbol{\xi}\|^2}{\lambda_{\min}(\mathbf{A}_S^T \mathbf{A}_S)} \\ &\geq \left(1 - \frac{2}{M}\right) \left(1 - \frac{C \log M}{(\sqrt{M} - O(\sqrt{\log M}))^2}\right) \mathbb{E}_{\boldsymbol{\xi}} \|\boldsymbol{\xi}\|^2 \\ &= M \Gamma_M(1 - o(1)). \end{aligned} \quad (42)$$

Equation (38) immediately follows from (40) and (42), which completes the proof.

5 Conclusion

In this paper, we provided an analysis based on an enhanced replica method for assessing the average performance of the Lasso estimator under ultra-sparse conditions. Besides, we deduced conditions necessary for support recovery which are derived from the oracle Lasso estimator. Numerical experiments strongly support the validity of our analysis.

The methodological novelty originates from an observation of finite-size effects and correlations within the active set, which is implicitly assumed to be negligible in the conventional replica analysis. We anticipate that this framework is applicable to analysis of other machine learning or optimization problems where finite-size effects are nonnegligible. Extending this method further to more general sensing matrix ensembles is also another exciting direction for future work.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Nos. 22J21581 (KO), 21K21310 (TT), 17H00764, 19H01812, 22H05117 (YK) and JST CREST Grant Number JPMJCR1912 (YK).

References

- Abbara, A., Kabashima, Y., Obuchi, T., and Xu, Y. (2020). Learning performance in inverse Ising problems with sparse teacher couplings. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(7):073402.
- Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.
- Bayati, M. and Montanari, A. (2012). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.
- Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Analysis*, 12(4):1105 – 1131.
- Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145 – 2177.
- Chang, S.-H., Cosman, P. C., and Milstein, L. B. (2011). Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944.
- Davidson, K. R. and Szarek, S. J. (2001). Chapter 8 - local operator theory, random matrices and banach spaces. In *Handbook of the Geometry of Banach Spaces*, volume 1, pages 317–366. Elsevier Science B.V.
- Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.
- Donoho, D. L., Johnson, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum Entropy and the Nearly Black Object. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):41 – 81.
- Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.
- Dossal, C., Chabanol, M.-L., Peyré, G., and Fadili, J. (2012). Sharp support recovery from noisy random measurements by ℓ_1 -minimization. *Applied and Computational Harmonic Analysis*, 33(1):24–43.
- Fletcher, A. K., Rangan, S., and Goyal, V. K. (2009). Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772.
- Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C., and Scheffler, M. (2015). Big data of materials science: Critical role of the descriptor. *Physical Review Letters*, 114:105503.
- Guo, D. and Verdú, S. (2005). Randomly spread CDMA: asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010.

- Kabashima, Y. (2003). A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111–11121.
- Kabashima, Y., Wadayama, T., and Tanaka, T. (2009). A typical reconstruction limit for compressed sensing based on ℓ_p -norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003.
- Kim, C., Pilania, G., and Ramprasad, R. (2016). From organized high-throughput data to phenomenological theory using machine learning: The example of dielectric breakdown. *Chemistry of Materials*, 28(5):1304–1311.
- Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436 – 1462.
- Meng, X., Obuchi, T., and Kabashima, Y. (2021a). Ising model selection using ℓ_1 -regularized linear regression: A statistical mechanics analysis. In *Advances in Neural Information Processing Systems*, volume 34, pages 6290–6303.
- Meng, X., Obuchi, T., and Kabashima, Y. (2021b). Structure learning in inverse Ising problems using ℓ_2 -regularized linear estimator. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(5):053403.
- Mézard, M. and Montanari, A. (2009). *Information, Physics, and Computation*. Oxford University Press, Inc., USA.
- Mézard, M., Parisi, G., and Virasoro, M. (1986). *Spin Glass Theory and Beyond*. WORLD SCIENTIFIC.
- Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337.
- Pilania, G., Mannodi-Kanakithodi, A., Uberuaga, B. P., Ramprasad, R., Gubernatis, J. E., and Lookman, T. (2016). Machine learning bandgaps of double perovskites. *Scientific Reports*, 6(1):19375.
- Rush, C. and Venkataramanan, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286.
- Stojnic, M. (2013). A framework to characterize performance of lasso algorithms. *arXiv*, <https://arxiv.org/abs/1303.7291>.
- Su, W., Bogdan, M., and Candès, E. (2017). False discoveries occur early on the lasso path. *The Annals of Statistics*, 45(5):2133–2150.
- Takahashi, T. and Kabashima, Y. (2022). Macroscopic analysis of vector approximate message passing in a model-mismatched setting. *IEEE Transactions on Information Theory*, 68(8):5579–5600.
- Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.
- Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198 – 1232.
- Vehkaperä, M., Kabashima, Y., and Chatterjee, S. (2016). Analysis of regularized ls reconstruction and random matrix ensembles in compressed sensing. *IEEE Transactions on Information Theory*, 62(4):2100–2124.
- Wainwright, M. J. (2009a). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741.
- Wainwright, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.
- Wang, H., Yang, Y., Bu, Z., and Su, W. (2020). The complete lasso tradeoff diagram. In *Advances in Neural Information Processing Systems*, volume 33, pages 20051–20060. Curran Associates, Inc.
- Zdeborová, L. and Krzakala, F. (2016). Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552.
- Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567 – 1594.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90):2541–2563.

Supplementary Materials

A Detailed derivation of Claim 1

Here, we derive the expression in Claim 1; see Figure 4 for an outline of the calculation. For simplicity, we abbreviate $\mathbb{E}_{\mathbf{A}_{\setminus S}}$, the average over \mathbf{A} excluding the submatrix acting on S , as \mathbb{E} , and $\mathbf{A}_{\setminus S}$ as the submatrix of \mathbf{A} excluding \mathbf{A}_S . Using the shorthand expression $d\mathbf{x}_S^a := \prod_{i \in S} dx_i^a$ and $d\Delta^a := \prod_{i \neq S} d\Delta_i^a$, $\mathbb{E} Z_\beta^n(\mathbf{A}, \mathbf{y})$ can be written as

$$\mathbb{E} Z_\beta^n(\mathbf{A}, \mathbf{y}) = \int \left(\prod_{a=1}^n d\mathbf{x}_S^a d\Delta^a e^{-\beta M \lambda \|\Delta^a\|} \right) \mathbb{E} \left[\prod_{a=1}^n \int d\mathbf{h}^a \delta(\mathbf{h}^a - \mathbf{A}_{\setminus S} \Delta^a) e^{-\frac{\beta}{2} \|\mathbf{A}_S \mathbf{x}_S^a - \mathbf{h}^a - \mathbf{y}\|^2 - \beta M \lambda \|\mathbf{x}_S^a\|_1} \right] \quad (43)$$

Using the Fourier representation, the average of the delta functions over $\mathbf{A}_{\setminus S} = (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_M)^\top$ is given by

$$\begin{aligned} \mathbb{E} \prod_{a=1}^n \delta(\mathbf{h}^a - \mathbf{A}_{\setminus S} \Delta^a) &= \int \prod_{\mu=1}^M \frac{d\tilde{\mathbf{a}}_\mu e^{-\frac{1}{2} \|\tilde{\mathbf{a}}_\mu\|^2}}{(2\pi)^{(N-d)/2}} \prod_{a=1}^n \int \frac{d\tilde{h}_\mu^a}{2\pi} e^{-i\tilde{h}_\mu^a (\mathbf{h}_\mu^a - \tilde{\mathbf{a}}_\mu^\top \Delta^a)} \\ &= \int \prod_{\mu=1}^M \left\{ \frac{d\tilde{\mathbf{a}}_\mu}{(2\pi)^{(N-d)/2}} \int \left(\prod_{a=1}^n \frac{d\tilde{h}_\mu^a}{2\pi} \right) \exp \left[-i \sum_{a=1}^n \tilde{h}_\mu^a \tilde{\mathbf{a}}_\mu^\top \Delta^a - \frac{1}{2} \left\| \tilde{\mathbf{a}}_\mu - i \sum_{a=1}^n \tilde{h}_\mu^a \Delta^a \right\|^2 - \frac{1}{2} \sum_{a,b} \tilde{h}_\mu^a \tilde{h}_\mu^b (\Delta^a)^\top \Delta^b \right] \right\} \\ &= \int \left(\prod_{a=1}^n \prod_{\mu=1}^M \frac{d\tilde{h}_\mu^a}{2\pi} \right) \exp \left[-\frac{1}{2} \sum_{\mu=1}^M \sum_{a,b} \tilde{h}_\mu^a \tilde{h}_\mu^b (\Delta^a)^\top \Delta^b - i \sum_{a=1}^n \sum_{\mu=1}^M \tilde{h}_\mu^a \tilde{\mathbf{a}}_\mu^\top \Delta^a \right] = \prod_{\mu=1}^M \frac{1}{\sqrt{(2\pi)^n \det \mathbf{Q}}} \exp \left(-\frac{1}{2} \mathbf{h}_\mu^\top \mathbf{Q}^{-1} \mathbf{h}_\mu \right), \end{aligned}$$

where we defined the matrix \mathbf{Q} as $(\mathbf{Q})_{ab} := (\Delta^a)^\top \Delta^b$ and used the notation $\mathbf{h}_\mu := (h_\mu^1, \dots, h_\mu^n) \in \mathbb{R}^n$ without confusion. This implies that the vector \mathbf{h}_μ is Gaussian with covariance matrix \mathbf{Q} . Now, the replica symmetric ansatz (9) implies that the integral over $\{\Delta^a\}_{a=1}^n$ is dominated by the subspace of the form

$$\int dQ d\chi \prod_{a=1}^n \delta(Q - \|\Delta^a\|^2) \prod_{a < b} \delta \left(Q - \frac{\chi}{\beta} - (\Delta^a)^\top \Delta^b \right), \quad (44)$$

which allows us to simplify the profile of h_μ^a as

$$h_\mu^a = \sqrt{Q - \frac{\chi}{\beta}} z_\mu + \sqrt{\frac{\chi}{\beta}} v_\mu^a, \quad (45)$$

where z_μ and v_μ^a ($a = 1, \dots, n$) are all i.i.d. standard Gaussian variables. Using (43)–(45) yields the expression

$$\mathbb{E} Z_\beta^n(\mathbf{A}, \mathbf{y}) = \int \prod_{a=1}^n d\Delta^a \int dQ d\chi e^{-\beta M \lambda \sum_{a=1}^n \|\Delta^a\|} \mathcal{I} \mathcal{L}, \quad (46)$$

where \mathcal{L} is given by

$$\begin{aligned} \log \mathcal{L} &= \log \int Dz \prod_{a=1}^n \left\{ \int D\mathbf{v}^a d\mathbf{x}_S^a \exp \left(-\frac{\beta}{2} \left\| \mathbf{A}_S \mathbf{x}_S^a - \sqrt{Q - \frac{\chi}{\beta}} \mathbf{z} - \sqrt{\frac{\chi}{\beta}} \mathbf{v}^a - \mathbf{y} \right\|^2 - \beta M \lambda \|\mathbf{x}_S^a\|_1 \right) \right\} \\ &= \log \int Dz \left\{ \int d\mathbf{x}_S \exp \left(-\frac{\beta}{2(1+\chi)} \left\| \mathbf{A}_S \mathbf{x}_S - \sqrt{Q} \mathbf{z} - \mathbf{y} \right\|^2 - \beta M \lambda \|\mathbf{x}_S\|_1 + o(\beta) \right) \right\}^n \\ &= n \int Dz \log \int d\mathbf{x}_S \exp \left(-\frac{\beta}{2(1+\chi)} \left\| \mathbf{A}_S \mathbf{x}_S - \sqrt{Q} \mathbf{z} - \mathbf{y} \right\|^2 - \beta M \lambda \|\mathbf{x}_S\|_1 \right) + O(n^2) + o(\beta). \end{aligned}$$

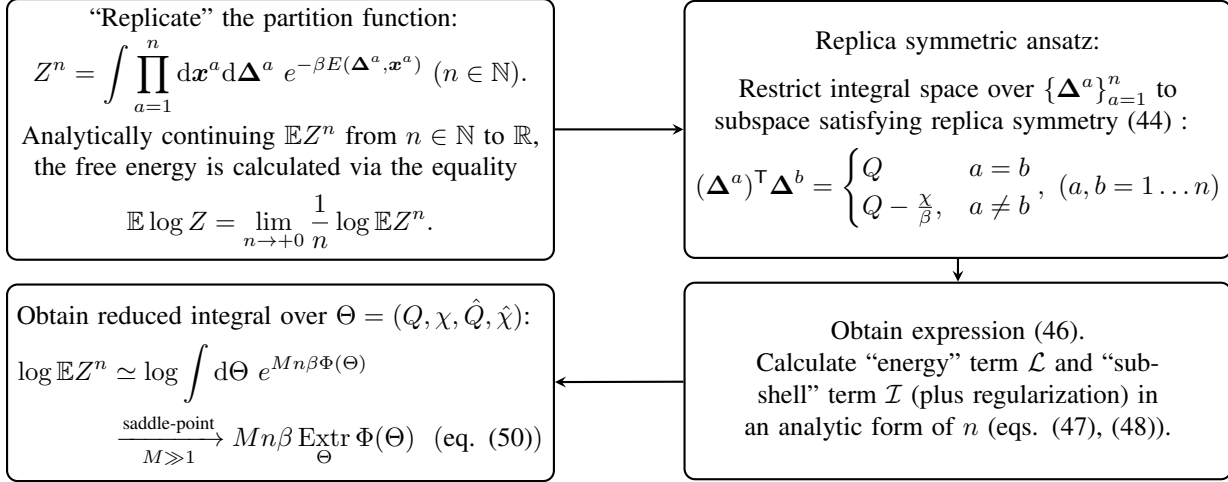


Figure 4: Outline of the replica calculation for Claim 1.

The integral with respect to \mathbf{x}_S can be evaluated using Laplace's method for large β , yielding

$$\mathcal{L} \simeq \exp \left[-Mn\beta \int D\mathbf{z} \min_{\mathbf{x}} \left(\frac{\|\mathbf{A}_S \mathbf{x} - \sqrt{Q}\mathbf{z} - \mathbf{y}\|^2}{2M(1+\chi)} + \lambda \|\mathbf{x}\|_1 \right) \right], \quad (47)$$

where the subleading terms are ignored. Similarly, \mathcal{I} is given by

$$\begin{aligned} \mathcal{I} &= \prod_{a=1}^n \delta(Q - \|\Delta^a\|^2) \prod_{a<b} \delta\left(Q - \frac{\chi}{\beta} - (\Delta^a)^\top \Delta^b\right) \\ &= \int_{-i\infty}^{+i\infty} d\hat{Q} d\hat{\chi} \exp M \left[\frac{1}{2} (\beta\hat{Q} - \beta^2\hat{\chi}) \sum_{a=1}^n (Q - \|\Delta^a\|^2) - \frac{1}{2} \beta^2 \hat{\chi} \sum_{a \neq b} \left(Q - \frac{\chi}{\beta} - (\Delta^a)^\top \Delta^b \right) + o(\beta) \right] \\ &= \int_{-i\infty}^{+i\infty} d\hat{Q} d\hat{\chi} e^{\frac{n\beta M}{2} (Q\hat{Q} + (n-1)\chi\hat{\chi} - n\beta Q\hat{\chi})} \exp M \left[-\frac{\beta\hat{Q}}{2} \sum_{a=1}^n \|\Delta^a\|^2 + \frac{\beta^2\hat{\chi}}{2} \left\| \sum_{a=1}^n \Delta^a \right\|^2 + o(\beta) \right] \\ &= \int_{-i\infty}^{+i\infty} d\hat{Q} d\hat{\chi} e^{\frac{n\beta M}{2} (Q\hat{Q} + (n-1)\chi\hat{\chi} - n\beta Q\hat{\chi})} \int D\hat{\mathbf{z}} \exp \beta \left[\left(-\frac{M\hat{Q}}{2} \sum_{a=1}^n \|\Delta^a\|^2 + \sqrt{M\hat{\chi}} \sum_{a=1}^n \hat{\mathbf{z}}^\top \Delta^a \right) + o(\beta) \right]. \end{aligned}$$

Therefore, ignoring the subleading term with respect to β ,

$$\begin{aligned} &\int \prod_{a=1}^n d\Delta^a e^{-\beta M \lambda \sum_{a=1}^n \|\Delta^a\|} \mathcal{I} \\ &\simeq \int_{-i\infty}^{+i\infty} d\hat{Q} d\hat{\chi} e^{\frac{n\beta M}{2} (Q\hat{Q} + (n-1)\chi\hat{\chi} - n\beta Q\hat{\chi})} \int D\hat{\mathbf{z}} \left\{ \int d\Delta \exp \beta \left[-\frac{M\hat{Q}}{2} \|\Delta\|^2 + \sqrt{M\hat{\chi}} \hat{\mathbf{z}}^\top \Delta - M\lambda \|\Delta\| \right] \right\}^n \quad (48) \end{aligned}$$

The log of the integral with respect to $D\hat{z}$ can be expanded as

$$\begin{aligned}
 & \log \int D\hat{z} \left\{ \int d\Delta \exp \beta \left[-\frac{M\hat{Q}}{2} \|\Delta\|^2 + \sqrt{M\hat{\chi}} \hat{z}^\top \Delta - M\lambda \|\Delta\|_1 \right] \right\}^n \\
 &= n \sum_{i=1}^{\tilde{N}} \int D\hat{z}_i \log \int d\Delta \exp \beta \left[-\frac{M\hat{Q}}{2} \Delta^2 + \sqrt{M\hat{\chi}} \hat{z}_i \Delta - M\lambda |\Delta| \right] + O(n^2) \\
 &\simeq -n\beta \tilde{N} \int D\hat{z} \min_{\Delta} \left(\frac{M\hat{Q}}{2} \Delta^2 - \sqrt{M\hat{\chi}} \hat{z} \Delta + M\lambda |\Delta| \right) + O(n^2) \\
 &= Mn\beta \frac{\tilde{N}}{2\hat{Q}} \left[\left(\lambda^2 + \frac{\hat{\chi}}{M} \right) \operatorname{erfc} \left(\sqrt{\frac{\Lambda}{2M\hat{\chi}}} \right) - \sqrt{\frac{2\lambda^2 \hat{\chi}}{\pi M}} e^{-\Lambda/2M\hat{\chi}} \right] + O(n^2), \tag{49}
 \end{aligned}$$

where Laplace's approximation was used for large β to obtain the third line. Substituting (47), (48) and (49) into (46), using the saddle point method for large M results in

$$\begin{aligned}
 \log \mathbb{E} Z_\beta^n(\mathbf{A}, \mathbf{y}) &= Mn\beta \operatorname{E}_{Q, \hat{Q}, \chi, \hat{\chi}} \left\{ \frac{Q\hat{Q} + (n-1)\chi\hat{\chi} - n\beta Q\hat{\chi}}{2} \right. \\
 &\left. - \int D\mathbf{z} \min_{\mathbf{x}} \left(\frac{\|\mathbf{A}_S \mathbf{x} - \sqrt{Q}\mathbf{z} - \mathbf{y}\|^2}{2M(1+\chi)} + \lambda \|\mathbf{x}\|_1 \right) + \frac{\tilde{N}}{2\hat{Q}} \left[\left(\lambda^2 + \frac{\hat{\chi}}{M} \right) \operatorname{erfc} \left(\sqrt{\frac{\Lambda}{2M\hat{\chi}}} \right) - \sqrt{\frac{2\lambda^2 \hat{\chi}}{\pi M}} e^{-\Lambda/2M\hat{\chi}} \right] \right\}. \tag{50}
 \end{aligned}$$

Noticing that

$$\lim_{n \rightarrow +0} \frac{\mathbb{E} Z_\beta^n(\mathbf{A}, \mathbf{y}) - 1}{n} = \lim_{n \rightarrow +0} \frac{\log \mathbb{E} Z_\beta^n(\mathbf{A}, \mathbf{y})}{n}, \tag{51}$$

and finally rescaling $\hat{Q} \leftarrow M\hat{Q}$ and $\hat{\chi} \leftarrow M\hat{\chi}$, one obtains (14).

B Proof of auxiliary lemmas

B.1 Proof of Lemma 4

From (16) and (17), we have $\chi = f\left(\frac{\tilde{N}}{M-d} \operatorname{erfc}\left(\sqrt{\frac{\Lambda}{2\hat{\chi}}}\right)\right)$, where $f(x) = x/(1-x)$, and $\bar{d} := \int D\mathbf{z} \|\hat{\mathbf{x}}_{(1+\chi)\lambda}(\sqrt{Q}\mathbf{z} + \mathbf{y})\|_0$. From $\chi > 0$, we see that f is a increasing function. By using the Markov inequality, it can be deduced that

$$\Pr[\chi > N^{-c_1}] = \Pr[f^{-1}(\chi) > f^{-1}(N^{-c_1})] \leq \frac{\mathbb{E} f^{-1}(\chi)}{f^{-1}(M^{-1})} < \frac{1 + N^{c_1}}{M-d} \langle \text{FP} \rangle < O(N^{-(c-c_1)}),$$

which proves the first part of the lemma with $c_1 = c/2$. For the probability bound on Q , using $\operatorname{erfc}(x) < \frac{1}{x\sqrt{\pi}} e^{-x^2}$,

$$Q < \frac{\hat{\chi}(1+\chi)}{(M-d)^2} \tilde{N} \operatorname{erfc}\left(\sqrt{\frac{\Lambda}{2\hat{\chi}}}\right) = g\left(\frac{\tilde{N}}{M-d} \operatorname{erfc}\left(\sqrt{\frac{\Lambda}{2\hat{\chi}}}\right)\right),$$

where

$$g(x) = \frac{M^2 \lambda^2}{2(M-d)} \frac{x}{(1-x)^2} \left[\operatorname{erfc}^{-1}\left(\frac{Mx}{N}\right) \right]^{-2}.$$

Using $\operatorname{erfc}^{-1}(Mx/N) > (1-x)^2$ for $M/N < 1$, we have for large enough M ,

$$g(N^{-c_1}) < \frac{\lambda^2}{2} N^{-c_1} M^2 (M-d)^{-1} (1-N^{-3c_1})^{-6}.$$

Since both g and g^{-1} are nonnegative and increasing, for large enough M ,

$$g^{-1}(N^{-c_1/2}) > g^{-1}\left(\frac{\lambda^2}{2} N^{-c_1} M^2 (M-d)^{-1} (1-N^{-3c_1})^{-6}\right) > N^{-c_1}.$$

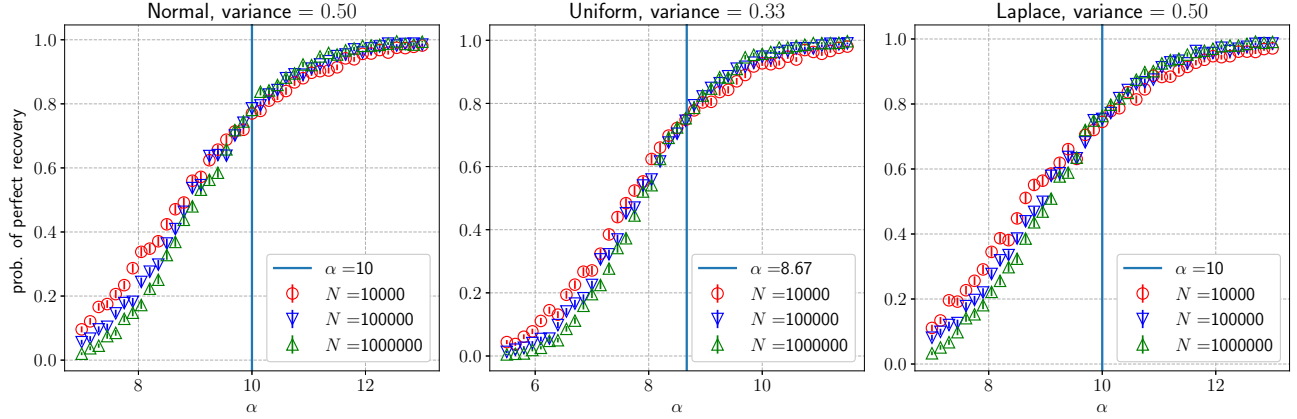


Figure 5: Probability of perfect recovery for noise distributed according to (1) Gaussian distribution with mean 0 and variance 0.5 (left), (2) Uniform distribution defined on the interval $[-1, 1]$ (middle), and (3) Laplace distribution with mean 0 and variance 0.5 (right) at $N = 10^4, 10^5$, and 10^6 . Error bars represent the standard error obtained from 1,000 experimental runs. The horizontal blue line depicts the necessary sample complexity given by Claim 5.

The Markov inequality then implies the second part of the lemma with $c_1 = c/2$:

$$\Pr[Q > N^{-c_1/2}] = \Pr[g^{-1}(Q) > g^{-1}(N^{-c_1/2})] < \frac{\langle \text{FP} \rangle}{(M-d)g^{-1}(N^{-c_1/2})} < \frac{1}{M-d} N^{-c_1} \langle \text{FP} \rangle < O(N^{-(c-c_1)}).$$

B.2 Proof of Lemma 5

Equations (16) and (17) imply that $N \text{erfc}\left(\frac{\Lambda}{\sqrt{2\hat{\chi}}}\right) = \hat{Q}\chi \leq M \frac{\chi}{1+\chi} \leq M$ holds for any $(\mathbf{A}_S, \mathbf{y})$. Thus, $\hat{\chi}$ is deterministically upper-bounded as

$$\hat{\chi} \leq \frac{1}{2} \frac{\Lambda^2}{[\text{erfc}^{-1}(M/N)]^2}. \quad (52)$$

Now, erfc satisfies (Chang et al., 2011) for $0 < \epsilon < 1/3$,

$$\text{erfc}(x) \geq \exp[-(1+\epsilon)x^2 + \log \epsilon]. \quad \therefore [\text{erfc}^{-1}(x)]^{-2} \leq (1+\epsilon)(-\log x + \log \epsilon)^{-1}.$$

Applying this inequality to (52) with $\epsilon = M^{-1}$ and $N = \exp(M/\alpha)$ for $M > 3$ yields

$$\hat{\chi} \leq \frac{1}{2} \alpha M \lambda^2 \left(1 + \frac{1}{M}\right) \left(1 - 2\alpha \frac{\log M}{M}\right)^{-1} < \frac{1}{2} \alpha M \lambda^2 \left(1 - (2\alpha + 1) \frac{\log M}{M}\right)^{-1}. \quad (53)$$

C Additional numerical experiments : Necessary condition for perfect support recovery

To verify the necessary sample complexity for perfect recovery given by Claim 5, numerical experiments were conducted. The profile of \mathbf{x}^0 is the same as that of Section 3.1, and the regularization parameter is taken as $\lambda = 0.5$. Figure 5 shows the perfect support recovery probability for noise distributed according to the Gaussian, uniform, and Laplace distribution. Clearly, for all three cases, perfect recovery fails with finite probability as N tends to infinity when α is less than the value indicated by Claim 5.