Original Paper

Asymptotics of Proximity Operator for Squared Loss and Performance Prediction of Nonconvex Sparse Signal Recovery

Ryo Hayakawa*

Institute of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan

ABSTRACT

Proximal splitting-based convex optimization is a promising approach to linear inverse problems because we can use some prior knowledge of the unknown variables explicitly. An understanding of the behavior of the optimization algorithms would be important for the tuning of the parameters and the development of new algorithms. In this paper, we first analyze the asymptotic property of the proximity operator for the squared loss function, which appears in the update equations of some proximal splitting methods for linear inverse problems. Our analysis shows that the output of the proximity operator can be characterized with a scalar random variable in the large system limit. Moreover, we apply the asymptotic result to the prediction of optimization algorithms for compressed sensing. Simulation results demonstrate that the MSE performance of the Douglas-Rachford algorithm can be well predicted in compressed sensing with the ℓ_1 optimization. We also examine the behavior of the prediction for the case with noncon-

vex smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP) regularization.

Keywords: Linear inverse problems, convex optimization, proximity operator, Douglas-Rachford algorithm, asymptotic analysis

1   Introduction

Linear inverse problems, i.e., the reconstruction of an unknown vector from its linear measurements, often appear in the field of signal processing. The linear inverse problem is called underdetermined when the number of measurements is less than that of elements of the unknown vector. In such underdetermined problems, we often use some information about the unknown vector to obtain a good reconstruction result. In compressed sensing [8, 7, 16, 9], for example, the unknown sparse vector can be effectively reconstructed by leveraging its sparsity. Another example of the underdetermined problems is the overloaded signal detection in wireless communications [33, 23, 24], where we can use the discreteness of the unknown vector.

A common approach to underdetermined linear inverse problems is to solve an optimization problem using some regularizer based on the prior knowledge of the unknown vector. In compressed sensing, one of the most widely used optimization problems is the ℓ_1 optimization. The objective function includes the ℓ_1 regularization term to promote sparsity. Even though the objective function is not differentiable because of the ℓ_1 norm, some proximal splitting methods can solve the problem with reasonable computational complexity. Examples of such methods include the iterative soft thresholding algorithm (ISTA) [15, 14, 21] and the fast ISTA (FISTA) [5]. Alternating direction method of multipliers (ADMM) [22, 6] and the Douglas-Rachford algorithm [28, 19, 13] can both be used to solve the ℓ_1 optimization problem. Despite the requirement of matrix inversion for linear inverse problems, ADMM and the Douglas-Rachford algorithm exhibit faster convergence than some gradient-based methods, such as ISTA, in certain cases [11, 12].

The asymptotic error performance of several optimization-based approaches for linear inverse problems has been analyzed by using some techniques [18, 3, 17]. Especially, the convex Gaussian min-max theorem (CGMT) [36, 35] enables us to obtain the asymptotic error in a precise manner for the optimizer of various optimization problems. For instance, the asymptotic mean-square-error (MSE) of various regularized estimators has been analyzed in [2, 35]. In [37, 27], the symbol error rate has also been analyzed for the reconstruction of discrete-valued vectors.

The above analyses have focused on the performance of the *optimizer*, whereas the application of CGMT for the performance prediction of *optimization algorithms* has been considered in [26]. In [26], we have proposed a prediction method for the error performance of the tentative estimate at each iteration of ADMM. Such performance prediction could be utilized to understand the behavior of the algorithm, tune the parameter, and develop a new algorithm. For optimization algorithms other than ADMM, however, such asymptotic performance prediction has not been discussed for linear inverse problems in the literature. Expanding the prediction method to other optimization algorithms would be important for the development of new algorithms and the understanding of the behavior of the algorithms.

In this paper, we investigate the asymptotic behavior of algorithms using the proximity operator of the squared loss function. As the target optimization problem, we consider the objective function with the squared loss function and a separable regularizer. Since the proximity operator of the squared loss function is not element-wise, we investigate its asymptotic property via the CGMT framework. Our main finding is that the optimization problem in the definition of the proximity operator can be analyzed by using CGMT for the squared loss function. Specifically, we show that the output of the proximity operator for the squared loss function can be regarded as a scalar random variable in the large system limit.

By using the analytical result, we also propose a prediction method for the evolution of the asymptotic MSE in proximal splitting methods composed of the proximity operator of the squared loss function and a separable regularizer. As an representative example of such algorithms, we here consider the Douglas-Rachford algorithm in this paper. While the derivation of the proposed method is partly non-rigorous, their simulation results for the ℓ_1 optimization demonstrate their potential as a foundation for more rigorous discussion in future work. These findings would contribute to the broader understanding of optimization algorithms in high-dimensional settings. Moreover, we also investigate the behavior for the case with nonconvex regularization in the simulation, though only the convex optimization is considered in [26]. The effectiveness of the performance prediction across various scenarios highlights its potential to extend to other optimization algorithms, including those involving nonconvex cases.

We use the following notations in this paper. The transpose and the identity matrix are denoted by $(\cdot)^\top$ and $\boldsymbol{I}$, respectively. The ℓ_1 norm and the ℓ_2 norm for a vector $\boldsymbol{u} = [u_1 \cdots u_N]^\top \in \mathbb{R}^N$ are defined as $\|\boldsymbol{u}\|_1 = \sum_{n=1}^N |u_n|$ and $\|\boldsymbol{u}\|_2 = \sqrt{\sum_{n=1}^N u_n^2}$, respectively. We denote the sign function by $\mathrm{sign}(\cdot)$. For a lower semicontinuous function $\phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, the proximity operator is defined as $\mathrm{prox}_\phi(\boldsymbol{u}) = \arg\min_{\boldsymbol{s} \in \mathbb{R}^N} \left\{ \phi(\boldsymbol{s}) + \frac{1}{2} \|\boldsymbol{s} - \boldsymbol{u}\|_2^2 \right\}$. We

denote the Gaussian distribution with mean μ and variance σ^2 by $\mathcal{N}(\mu, \sigma^2)$. If random variables $\{\Theta_n\}$ $(n = 1, 2, \dots)$ with index n converge in probability to Θ, we use the notation $\Theta_n \xrightarrow{\text{P}} \Theta$ as $n \to \infty$ or $\mathrm{plim}_{n \to \infty} \Theta_n = \Theta$.

2 Optimization for Linear Inverse Problems

2.1 Linear Inverse Problems

In this paper, we discuss the linear inverse problem, which means the reconstruction of a vector $\boldsymbol{x} \in \mathbb{R}^N$ from its measurement $\boldsymbol{y} \in \mathbb{R}^M$ given by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}. \tag{1}$$

The measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is assumed to be known and $\boldsymbol{v} \in \mathbb{R}^M$ denotes a noise vector. We mainly focus on the underdetermined case where the measurement ratio $\Delta := M/N$ is less than 1 and the solution is not unique even in the noiseless case. A common approach in this situation is the utilization of the structure of $\boldsymbol{x}$. In compressed sensing, for example, the unknown vector is assumed to be sparse, i.e., the vector has many zero elements. In some applications in wireless communications, the unknown vector has other structures such as boundedness and discreteness [34, 30, 1, 25].

2.2 Optimization-Based Approach

Mathematical optimization is a useful method for underdetermined linear inverse problems. In this approach, we can design the objective function of the problem to make the most of the structure of the unknown vector $\boldsymbol{x}$. As a simple example, we discuss the following optimization problem

$$\underset{\boldsymbol{s} \in \mathbb{R}^N}{\text{minimize}} \ \{L(\boldsymbol{s}) + f_\lambda(\boldsymbol{s})\} \tag{2}$$

in this paper. We refer to $L(\boldsymbol{s}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2$ as the squared loss function. The function $f_\lambda(\cdot) : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is a regularizer that can incorporate prior knowledge of the unknown vector $\boldsymbol{x}$. The parameter λ determines the balance between two terms in the objective function.

In the sparse vector reconstruction, for example, the ℓ_1 regularization by $f_\lambda(\boldsymbol{s}) = \lambda\|\boldsymbol{s}\|_1$ $(\lambda > 0)$ is widely used as a convex regularizer. The elastic net [40] given by $f_\lambda(\boldsymbol{s}) = \lambda_1\|\boldsymbol{s}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{s}\|_2^2$, where λ_1, λ_2 (> 0) are the parameters, is also a popular convex regularizer. Other than these convex regularizers, various nonconvex regularizers have been proposed for the sparse vector

reconstruction [38]. For example, the smoothly clipped absolute deviation (SCAD) [20] regularizer is given by $f_\lambda(\boldsymbol{s}) = \sum_{n=1}^{N} \tilde{f}_\lambda(s_n)$, where

$$
\tilde{f}_\lambda(s) = \begin{cases} \lambda \, |s| & (|s| \leq \lambda) \\[2mm] \dfrac{2a\lambda \, |s| - s^2 - \lambda^2}{2(a-1)} & (\lambda < |s| \leq a\lambda) \\[3mm] \dfrac{(a+1)\lambda^2}{2} & (|s| > a\lambda) \end{cases} \tag{3}
$$

and $a \; (> 1)$ is the parameter. For the minimax concave penalty (MCP) [39] regularizer, the function $\tilde{f}_\lambda(s)$ is given by

$$
\tilde{f}_\lambda(s) = \begin{cases} \lambda \, |s| - \dfrac{s^2}{2b} & (|s| \leq b\lambda) \\[3mm] \dfrac{b\lambda^2}{2} & (|s| > b\lambda) \end{cases}, \tag{4}
$$

where $b \; (> 0)$ is the parameter.

Various optimization algorithms have been proposed for the optimization problem in (2). For example, the Douglas-Rachford algorithm [28, 19, 13] solves the optimization problem in (2) by using the proximity operators of $L(\cdot)$ and $f_\lambda(\cdot)$. The update equations of the algorithm with the iteration index $k \; (= 0, 1, 2, \dots)$ can be written as

$$
\boldsymbol{s}^{(k+1)} = \mathrm{prox}_{\gamma L}\left(\boldsymbol{z}^{(k)}\right), \tag{5}
$$

$$
\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} + \rho_k \left(\mathrm{prox}_{\gamma f_\lambda}\left(2\boldsymbol{s}^{(k+1)} - \boldsymbol{z}^{(k)}\right) - \boldsymbol{s}^{(k+1)}\right), \tag{6}
$$

where $\gamma \; (> 0)$ and $\rho_k \in [\varepsilon, 2 - \varepsilon] \; (\varepsilon \in (0, 1))$ are the parameters in the algorithm. By definition, we can obtain the proximity operator of the function $L(\cdot)$ as

$$
\mathrm{prox}_{\gamma L}\left(\boldsymbol{z}\right) = \arg\min_{\boldsymbol{s} \in \mathbb{R}^N} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}\|_2^2 + \frac{1}{2\gamma} \|\boldsymbol{s} - \boldsymbol{z}\|_2^2 \right\} \tag{7}
$$

$$
= \left(\boldsymbol{A}^\top \boldsymbol{A} + \frac{1}{\gamma}\boldsymbol{I}\right)^{-1} \left(\boldsymbol{A}^\top \boldsymbol{y} + \frac{1}{\gamma}\boldsymbol{z}\right), \tag{8}
$$

where $\boldsymbol{z} \in \mathbb{R}^N$ and $\gamma > 0$. The proximity operator of $f_\lambda(\cdot)$ can also be computed efficiently for various regularizers. When $f_\lambda(\boldsymbol{s}) = \lambda \|\boldsymbol{s}\|_1$, for example, the proximity operator of the function $f_\lambda(\cdot)$ can be written as

$$
\left[\mathrm{prox}_{\gamma f_\lambda}(\boldsymbol{r})\right]_n = \mathrm{sign}(r_n) \max(|r_n| - \gamma\lambda, 0), \tag{9}
$$

where r_n and $\left[\mathrm{prox}_{\gamma f_\lambda}(\boldsymbol{r})\right]_n$ are the n-th element of $\boldsymbol{r}$ and $\mathrm{prox}_{\gamma f_\lambda}(\boldsymbol{r})$, respectively. For the elastic net regularizer $f_\lambda(\boldsymbol{s}) = \lambda_1 \|\boldsymbol{s}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{s}\|_2^2$, the proximity operator can be computed as

$$\left[\mathrm{prox}_{\gamma f_\lambda}(\boldsymbol{r})\right]_n = \frac{\mathrm{sign}(r_n)\max(|r_n| - \gamma\lambda_1, 0)}{1 + \gamma\lambda_2}. \tag{10}$$

As for the SCAD regularizer in (3), the proximity operator is given by

$$\begin{aligned}
&\left[\mathrm{prox}_{\gamma f_\lambda}(\boldsymbol{r})\right]_n \\
&= \begin{cases}
\mathrm{sign}(r_n)\max(|r_n| - \gamma\lambda, 0) & (|r_n| \le (1+\gamma)\lambda) \\
\dfrac{(a-1)r_n - \mathrm{sign}(r_n)a\gamma\lambda}{a - 1 - \gamma} & ((1+\gamma)\lambda < |r_n| \le a\lambda) \\
r_n & (|r_n| > a\lambda)
\end{cases}
\end{aligned} \tag{11}$$

for $a > 1 + \gamma$. Similarly, the proximity operator of the MCP regularizer in (4) can be computed as

$$\begin{aligned}
&\left[\mathrm{prox}_{\gamma f_\lambda}(\boldsymbol{r})\right]_n \\
&= \begin{cases}
0 & (|r_n| \le \gamma\lambda) \\
\dfrac{b}{b - \gamma}(r_n - \mathrm{sign}(r_n)\gamma\lambda) & (\gamma\lambda < |r_n| \le b\lambda) \\
r_n & (|r_n| > b\lambda)
\end{cases}
\end{aligned} \tag{12}$$

for $b > \gamma$. By computing (5) and (6) iteratively, we can obtain a sequence $\left\{\boldsymbol{s}^{(k)}\right\}_{k=1,2,\dots}$ converging to the solution of the optimization problem in (2) when the regularizer is convex. When the regularizer is nonconvex, the convergence of the sequence is not necessarily guaranteed.

3   Main Results

3.1   Asymptotics of Proximity Operator for Squared Loss

We firstly analyze the output of the proximity operator $\mathrm{prox}_{\gamma L}(\cdot)$ in (8). In our analysis, we assume the large system limit $M, N \to \infty$ ($M/N = \Delta$), where the sequence of problems with $\{\boldsymbol{x}, \boldsymbol{A}, \boldsymbol{v}\}$ indexed by N is considered as in several high-dimensional analyses [35]. We also assume the following conditions as in [26].

Assumption 3.1. The elements of the vector $\boldsymbol{x} \in \mathbb{R}^N$ are independent and identically distributed (i.i.d.). The distribution p_X of $\boldsymbol{x}$ is known and has some finite mean and variance. The measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ has

i.i.d. random variables with $\mathcal{N}(0, 1/N)$. The noise vector $\boldsymbol{v} \in \mathbb{R}^M$ has i.i.d. random variables with $\mathcal{N}(0, \sigma_v^2)$.

Remark 3.1. We assume that each element of the measurement matrix $\boldsymbol{A}$ follows a Gaussian distribution in Assumption 3.1. This is because we require the Gaussian assumption in the analysis with CGMT [35]. Nonetheless, the universality of random matrices discussed in [4, 32, 31] implies that the analytical result still holds for some other distributions. In [35, Section VIII-F], for example, it is expected that the analysis via CGMT holds for i.i.d. sub-Gaussian matrix. Computer simulations in [35, 26] show that the empirical performance is well predicted even when the measurement matrix is composed of i.i.d. Bernoulli distribution with $p = 0.5$.

Under Assumption 3.1, we can obtain the following result on the asymptotic behavior of the proximity operator $\text{prox}_{\gamma L}(\cdot)$.

Proposition 3.1. It is assumed that $\boldsymbol{x}$, $\boldsymbol{A}$, and $\boldsymbol{v}$ satisfy the conditions in Assumption 3.1. We then consider the output of the proximity operator given by

$$\hat{\boldsymbol{s}} = \text{prox}_{\gamma L}(\boldsymbol{z}), \tag{13}$$

where $\boldsymbol{z}$ has i.i.d. elements with a distribution $p_Z(z)$ and is assumed to be independent of $\boldsymbol{A}$. Here, we assume that the optimization problem

$$\min_{\alpha > 0} \max_{\beta \geq 0} \left\{ \frac{\alpha \beta \sqrt{\Delta}}{2} + \frac{\beta \sigma_v^2 \sqrt{\Delta}}{2\alpha} - \frac{1}{2}\beta^2 + \text{E}\left[J(\alpha, \beta; Z)\right] \right\} \tag{14}$$

has a unique optimizer (α^*, β^*)[1]. In (14),

$$
\begin{aligned}
J(\alpha, \beta; Z) = {} & \frac{\beta \sqrt{\Delta}}{2\alpha}\left(\hat{S}(\alpha, \beta; Z) - X\right)^2 \\
& - \beta H\left(\hat{S}(\alpha, \beta; Z) - X\right) \\
& + \frac{1}{2\gamma}\left(\hat{S}(\alpha, \beta; Z) - Z\right)^2,
\end{aligned} \tag{15}
$$

$$\hat{S}(\alpha, \beta; Z) = \frac{1}{\dfrac{\beta\sqrt{\Delta}}{\alpha} + \dfrac{1}{\gamma}}\left(\frac{\beta\sqrt{\Delta}}{\alpha}\left(X + \frac{\alpha}{\sqrt{\Delta}}H\right) + \frac{1}{\gamma}Z\right), \tag{16}$$

$X \sim p_X$, $H \sim \mathcal{N}(0, 1)$, and $Z \sim p_Z$. The expectation $\text{E}[\cdot]$ in (14) is calculated over all random variables X, H, and Z. Then, the following statements hold:

---

[1]The uniqueness of the solution can be established under certain conditions (e.g., the boundedness of the set of minimizers). However, in the general case, completely removing this assumption proves to be challenging. For further details, please refer to [35, Remark 19].

1. The asymptotic MSE of $\hat{\boldsymbol{s}}$ in (13) can be written as

$$\operatorname*{plim}_{M,N\to\infty} \frac{1}{N} \|\hat{\boldsymbol{s}} - \boldsymbol{x}\|_2^2 = (\alpha^*)^2 - \sigma_{\mathrm{v}}^2. \tag{17}$$

2. Let $\mu_{\hat{\boldsymbol{s}}}$ denote the empirical distribution of $\hat{\boldsymbol{s}} = [\hat{s}_1 \ \cdots \ \hat{s}_N]^\top \in \mathbb{R}^N$ that corresponds to the cumulative distribution function (CDF) given by $P_{\hat{\boldsymbol{s}}}(s) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(\hat{s}_n < s)$, where $\mathbb{I}(\hat{s}_n < s) = 1$ if $\hat{s}_n < s$ and otherwise $\mathbb{I}(\hat{s}_n < s) = 0$. Then, the distribution $\mu_{\hat{\boldsymbol{s}}}$ converges weakly in probability to the distribution μ_S of $S = \hat{S}(\alpha^*, \beta^*; Z)$, i.e., $\int g d\mu_{\hat{\boldsymbol{s}}} \xrightarrow{\mathrm{P}} \int g d\mu_S$ holds for any continuous compactly supported function $g(\cdot) : \mathbb{R} \to \mathbb{R}$.

Sketch of proof. By definition in (13), $\hat{\boldsymbol{s}}$ is characterized as the solution of the optimization problem in (7). Similarly to the discussion in [27, Remark IV.1], we can prove (17) by using the standard method with CGMT [35]. Moreover, the second statement of Proposition 3.1 is proven in the procedure of [27, Theorem III.2]. Hence, we do not include the details of the rigorous proof. For the overview of the derivation of (14), see Appendix A.1.   □

The first statement in Proposition 3.1 means that the asymptotic MSE for the output of the proximity operator can be predicted by solving the scalar optimization problem in (14). The second one implies that the distribution of the elements of the vector $\hat{\boldsymbol{s}}$ can be characterized by the random variable $S = \hat{S}(\alpha^*, \beta^*; Z)$. We can thus consider $\hat{S}(\alpha^*, \beta^*; Z)$ in (16) as a decoupled version of the proximity operator in (13) intuitively. For a similar discussion, see [26, Fig. 1].

The optimization of α and β in (14) can be performed by using searching methods like the golden section search [29]. Since it is difficult to compute the expectation in (14) exactly in general, we need to approximate it with the average of many realizations of X, H, and Z.

3.2 *Application to Performance Prediction of Optimization Algorithm*

Although the proximity operator itself is not a valid reconstruction method in general, we believe that the result in Proposition 3.1 is not only of theoretical interest but also serves as a foundation for understanding and designing optimization algorithms in linear inverse problems. By using Proposition 3.1, for example, we can derive an error prediction method for the estimate $\boldsymbol{s}^{(k)}$ in some optimization algorithm. To obtain the prediction method, we make an additional assumption that the regularizer $f_\lambda(\cdot)$ is convex and separable as follows.

Assumption 3.2. The function $f_\lambda(\cdot) : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is separable and can be written as $f_\lambda(\boldsymbol{s}) = \sum_{n=1}^{N} \tilde{f}_\lambda(s_n)$ with a convex function $\tilde{f}_\lambda(\cdot) : \mathbb{R} \to$

$\mathbb{R}\cup\{+\infty\}$. We sometimes use the notation $f_\lambda(\cdot)$ for the corresponding function $\tilde{f}_\lambda(\cdot)$ with the slight abuse of notation.

Remark 3.2. For the separable regularizer $f_\lambda(\cdot)$ in Assumption 3.2, the proximity operator $\mathrm{prox}_{\gamma f_\lambda}(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$ becomes an element-wise function. In other words, the n-th element of the output of the proximity operator is the function of the n-th element of the input only.

To predict the performance of an algorithm by using Proposition 3.1, it is necessary that the update equations except for $\mathrm{prox}_{\gamma L}(\cdot)$ are element-wise. As a representative example of such algorithms, we derive the conjecture below for Douglas-Rachford algorithm in (5) and (6). Note that the derivation of the conjecture is partly non-rigorous as [26, Claim III.1].

Conjecture 3.1. Suppose that Assumptions 3.1 and 3.2 hold. We define the stochastic process

$$S_{k+1} = \hat{S}\left(\alpha_k^*, \beta_k^*; Z_k\right) \tag{18}$$

$$Z_{k+1} = Z_k + \rho_k\left(\mathrm{prox}_{\gamma f_\lambda}\left(2S_{k+1} - Z_k\right) - S_{k+1}\right) \tag{19}$$

with the index k. We here assume that the optimization problem

$$\min_{\alpha>0}\max_{\beta\geq0}\left\{\frac{\alpha\beta\sqrt{\Delta}}{2} + \frac{\beta\sigma_{\mathrm{v}}^2\sqrt{\Delta}}{2\alpha} - \frac{1}{2}\beta^2 + \mathrm{E}\left[J(\alpha,\beta;Z_k)\right]\right\} \tag{20}$$

has a unique optimizer (α_k^*, β_k^*). Then, we have the following conjecture:

1. The asymptotic MSE of $\boldsymbol{s}^{(k+1)}$ in (5) can be written as

$$\underset{M,N\to\infty}{\mathrm{plim}}\frac{1}{N}\left\|\boldsymbol{s}^{(k+1)} - \boldsymbol{x}\right\|_2^2 = (\alpha_k^*)^2 - \sigma_{\mathrm{v}}^2. \tag{21}$$

2. The empirical distribution of $\boldsymbol{s}^{(k+1)}$ converges weakly in probability to the distribution of S_{k+1}.

Outline of derivation. Since $\boldsymbol{z}^{(k)}$ in (5) and (6) is not independent of $\boldsymbol{A}$, Proposition 3.1 cannot be used to (5) directly. We thus take the same procedure as the approach in [26, Claim III.1], where an unproven extension of CGMT has been assumed for the performance prediction of iterative algorithms. Under the assumption, when the element of $\boldsymbol{z}^{(k)}$ can be regarded as the random variable Z_k, the distribution of the element of $\boldsymbol{s}^{(k+1)}$ in (5) can be described by the distribution of the random variable S_{k+1} in (18) as in Proposition 3.1. As for (6), the update of $\boldsymbol{z}^{(k)}$ can be decoupled into (19) because the update is element-wise under Assumption 3.2. Since the procedure of the derivation is the same as [26, Claim III.1], the detailed explanation for the derivation is omitted. □

Conjecture 3.1 implies that the evolution of the asymptotic MSE in the Douglas-Rachford algorithm can be predicted by solving the optimization problem in (20) and computing (21) for each k. The updates of S_k and Z_k in (18) and (19) can be seen as a simplified version of the update of $\boldsymbol{s}^{(k)}$ and $\boldsymbol{z}^{(k)}$ in (5) and (6), respectively.

From the result in Conjecture 3.1, we can tune the parameters in the Douglas-Rachford algorithm for fast convergence. Since the relation between the parameters and the MSE is complicated, it is difficult to obtain the explicit expression of the optimal parameters. However, the result enables us to predict the error performance of the algorithm numerically and to select the parameter that achieves fast convergence in the asymptotic regime, without the empirical reconstruction.

Remark 3.3. As discussed in [19], the Douglas-Rachford algorithm is closely related to ADMM analyzed in [26]. In fact, the update equations of ADMM can be regarded as the Douglas-Rachford algorithm for the dual problem of the original optimization problem. However, the relaxation parameters ρ_k are included in the Douglas-Rachford algorithm considered in this paper, whereas the corresponding parameters are not considered in the analysis of [26]. These parameters provide additional flexibility in tuning the algorithm for faster convergence. Furthermore, since we have expressed Proposition 3.1 as the result for the proximity operator for the squared loss function, our results have the potential to be extended to other optimization algorithms that utilize the proximity operator.

4   Simulation Results

In this section, we demonstrate the validity of our approach via computer simulations. The distribution of an unknown vector $\boldsymbol{x}$ is assumed to be the Bernoulli-Gaussian distribution given by

$$p_X(x) = p_0\delta_0(x) + (1 - p_0)p_H(x). \tag{22}$$

Here, $\delta_0(\cdot)$ is the Dirac delta function, $p_H(\cdot)$ is the probability density function corresponding to $\mathcal{N}(0, 1)$, and $p_0 \in (0, 1)$ represents the probability of 0. The unknown vector $\boldsymbol{x}$ is sparse when p_0 is large. For the simulations, we assume that the measurement matrix $\boldsymbol{A}$ and the noise vector $\boldsymbol{v}$ satisfy the conditions in Assumption 3.1 unless otherwise stated. For the optimization of α and β in (14) and (20), the golden section search is used.

4.1   MSE of Output of Proximity Operator for Squared Loss

We demonstrate a simulation result for the first statement of Proposition 3.1. In Figure 1, the empirical MSE of $\hat{\boldsymbol{s}} = \text{prox}_{\gamma L}(\boldsymbol{z})$ in (13) is compared with
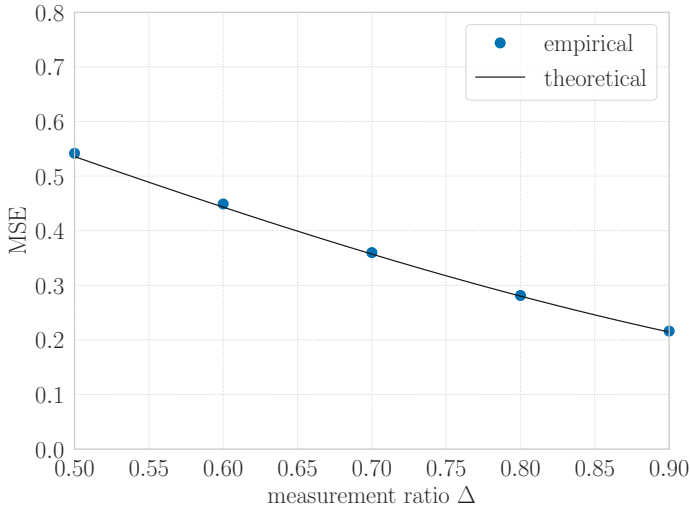
Figure 1: MSE of output of proximity operator for squared loss ($N = 1000$, $p_0 = 0.95$, $\sigma_{\mathrm{v}}^2 = 0.001$, $\gamma = 10$).

the corresponding theoretical prediction $(\alpha^*)^2 - \sigma_{\mathrm{v}}^2$ in (17). In the simulation, we set $N = 1000$, $p_0 = 0.95$, $\sigma_{\mathrm{v}}^2 = 0.001$, and $\gamma = 10$. In the figure, 'empirical' denotes the empirical MSE of $\hat{s}$ and 'theoretical' denotes the corresponding asymptotic value $(\alpha^*)^2 - \sigma_{\mathrm{v}}^2$. For the computation of the empirical MSE, we first make the realizations of $\boldsymbol{x}$, $\boldsymbol{A}$, and $\boldsymbol{v}$ to compute $\boldsymbol{y}$ from (1). We then create each element of $\boldsymbol{z}$ from the i.i.d. standard Gaussian distribution. Using these realizations, we compute the MSE of $\hat{s}$ from (8) and (13). The empirical performance in the figure is obtained by averaging the MSE for 100 independent realizations. For the theoretical prediction, we solve the scalar optimization problem in (14) to obtain α^*. Note that we approximate the expectation in the optimization problem with $100,000$ realizations of $X \sim p_X$, $H \sim \mathcal{N}(0,1)$ and $Z \sim \mathcal{N}(0,1)$. From Figure 1, we observe that the empirical performance and its theoretical prediction agree well with each other for various measurement ratios Δ.

4.2 *Distribution of Output of Proximity Operator for Squared Loss*

We then examine the second statement of Proposition 3.1. In Figure 2, we compare the probability density of the elements of $\hat{s} = \mathrm{prox}_{\gamma L}(\boldsymbol{z})$ with the theoretical result. In the simulation, we set $N = 1000$, $\Delta = 0.6$, $p_0 = 0.95$, $\sigma_{\mathrm{v}}^2 = 0.01$, and $\gamma = 10$. The empirical $\hat{s} = \mathrm{prox}_{\gamma L}(\boldsymbol{z})$ and the realizations of the corresponding random variable $S = \hat{S}(\alpha^*, \beta^*; Z)$ is obtained in the
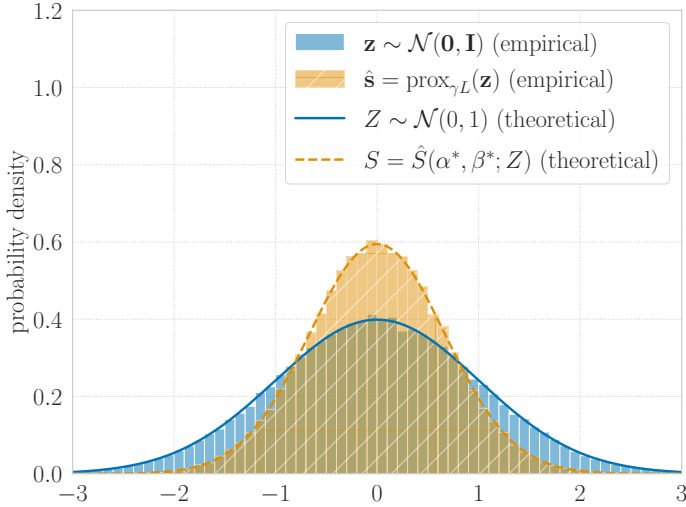
Figure 2: Comparison of probability density for $\hat{s}$ and S ($N = 1000$, $\Delta = 0.6$, $p_0 = 0.95$, $\sigma_v^2 = 0.01$, $\gamma = 10$).

same way as the first experiment. The empirical histogram is obtained from the result of 100 independent trials. From Figure 2, we can see that the empirical distribution of $\hat{s}$ agrees well with the distribution of S obtained by Proposition 3.1.

4.3 Performance Prediction of Douglas-Rachford Algorithm with Convex Regularization

Next, we investigate the validity of Conjecture 3.1. For the reconstruction of the sparse vector following (22), we first consider the ℓ_1 optimization with $f_\lambda(\boldsymbol{s}) = \lambda \left\| \boldsymbol{s} \right\|_1$, i.e.,

$$\underset{\boldsymbol{s} \in \mathbb{R}^N}{\text{minimize}} \left\{ \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{s} \right\|_2^2 + \lambda \left\| \boldsymbol{s} \right\|_1 \right\}, \tag{23}$$

which is a widely used convex optimization problem in compressed sensing and satisfies Assumption 3.2. For the ℓ_1 regularization, the proximity operator is given by (9).

We compare the empirical MSE performance of the Douglas-Rachford algorithm for (23) and its prediction obtained from Conjecture 3.1. In Figure 3, we show the MSE performance versus the number of iterations in the algorithm, where $\Delta = 0.7$, $p_0 = 0.9$, and $\sigma_v^2 = 0.001$. The parameters of the Douglas-Rachford algorithm are set as $\gamma = 10$ and $\rho_k = 1$. The initial values
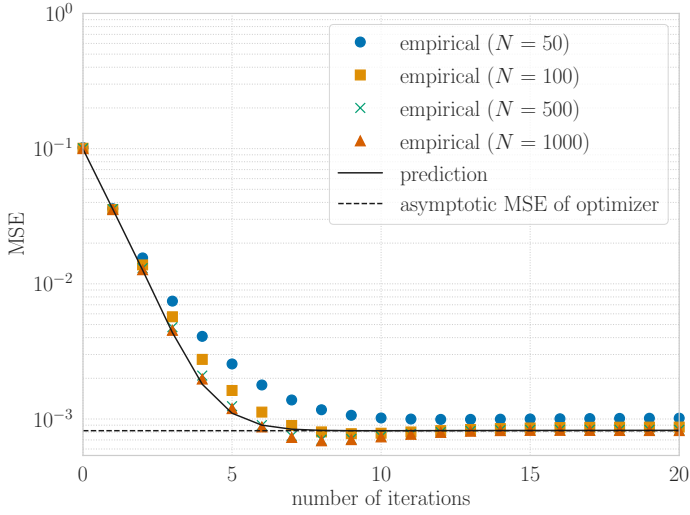
Figure 3: MSE performance of Douglas-Rachford algorithm for ℓ_1 optimization ($\Delta = 0.7$, $p_0 = 0.9$, $\sigma_{\mathrm{v}}^2 = 0.001$, $\gamma = 10$, $\rho_k = 1$).

of $\boldsymbol{s}^{(k)}$ and $\boldsymbol{z}^{(k)}$ are given by $\boldsymbol{s}^{(0)} = \boldsymbol{z}^{(0)} = \boldsymbol{0}$. In Figure 3, 'empirical' refers to the empirical MSE performance of the Douglas-Rachford algorithm in (5) and (6). The MSE curve is the average of the performance for 500 independent realizations of the reconstruction problem. Also, 'prediction' represents the predicted MSE obtained by Conjecture 3.1. To calculate the prediction, we compute (α_k^*, β_k^*) and $100,000$ realizations of (S_k, Z_k) for $k = 0, 1, \ldots$ from (18) and (19), following the similar way to [26]. Here, the expectation in (20) is approximated by using the realizations of (S_k, Z_k). In Figure 3, we also show the asymptotic MSE of the optimizer of (23) as 'asymptotic MSE of optimizer,' which is derived by using the standard CGMT approach [35]. The value of the parameter λ in (23) is selected to minimize the asymptotic MSE of the optimizer. Figure 3 shows that the empirical performance is close to the prediction when N is sufficiently large. We can also see that they approach the asymptotic MSE of the optimizer after sufficient iterations. To be precise, however, we can observe a slight difference between the prediction and the empirical performance. This is partly because we evaluate the empirical performance for finite N, while we assume $M, N \to \infty$ to obtain the asymptotic prediction. Another possible reason is that we approximate (α_k^*, β_k^*) by using many realizations of (S_k, Z_k) in the prediction.

We also investigate the performance of the Douglas-Rachford algorithm with the elastic net regularization in Figure 4. In the simulation, we set $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_{\mathrm{v}}^2 = 0.0001$, $\lambda_1 = \lambda_2 = 0.01$, $\gamma = 10$, and $\rho_k = 1$. As in
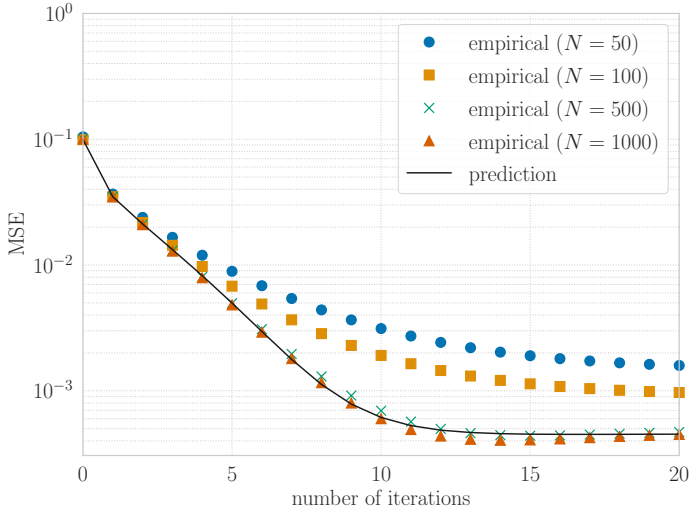
Figure 4: MSE performance of Douglas-Rachford algorithm with elastic net regularization ($\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.0001$, $\lambda_1 = \lambda_2 = 0.01$, $\gamma = 1$, $\rho_k = 1$).

the case with the ℓ_1 regularization, the empirical performance is well predicted by the prediction when N is sufficiently large.

Figure 5 shows the MSE performance of the Douglas-Rachford algorithm with the ℓ_1 regularization for the Bernoulli measurement matrix. In the evaluation of the empirical performance, the element of the measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is generated from the i.i.d. distribution given by $\Pr(a_{m,n} = 1/\sqrt{N}) = \Pr(a_{m,n} = -1/\sqrt{N}) = 0.5$, where $a_{m,n}$ denotes the (m, n) element of $\boldsymbol{A}$. For other parameters, we set the same values as in Figure 3. From Figure 5, we can see that the empirical performance is well predicted by the prediction even for the Bernoulli measurement matrix, though the Gaussian measurement matrix is assumed in derivation of the prediction.

Figure 6 shows the MSE performance of the Douglas-Rachford algorithm with the ℓ_1 regularization for different values of ρ_k. In the figure, the parameters are $N = 1000$, $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, and $\gamma = 5.0$. From the figure, we observe that the empirical performance is well predicted for every different values of ρ_k. We can also see that the convergence speed is significantly influenced by the value of ρ_k.

We then show the MSE performance of the Douglas-Rachford algorithm with the ℓ_1 regularization for different parameters γ in Figure 7. In the figure, we have $N = 500$, $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, and $\rho_k = 1$. From the figure, we can observe that the performance is improved as the iteration index k increases. The figure also implies that the value of γ significantly affects the performance of the algorithm. Since the empirical performance is well
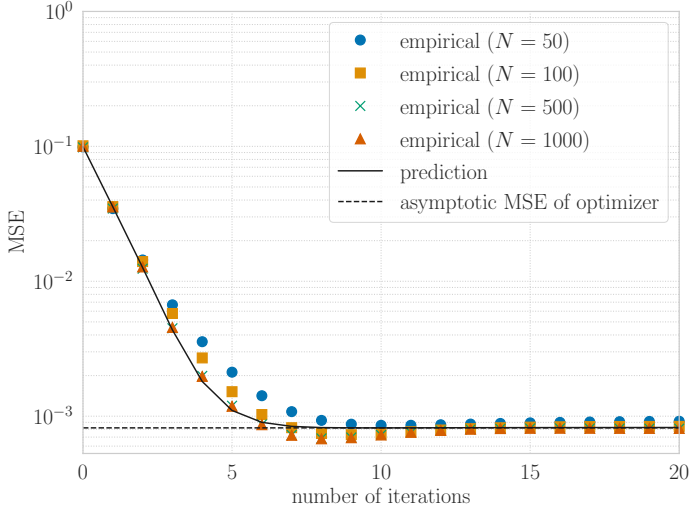
Figure 5: MSE performance of Douglas-Rachford algorithm for ℓ_1 optimization (Bernoulli measurement matrix, $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, $\gamma = 10$, $\rho_k = 1$).
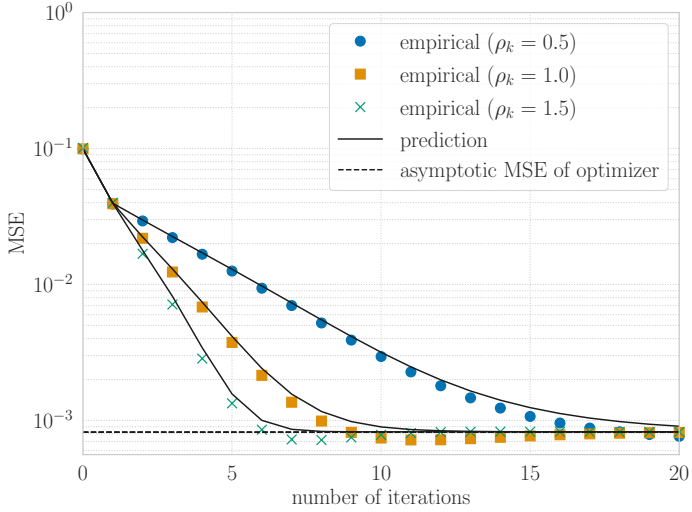

Figure 6: MSE performance of Douglas-Rachford algorithm for ℓ_1 optimization ($N = 1000$, $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, $\gamma = 5.0$).

predicted for all γ, we can tune the parameter by using the prediction. The figure shows that we should choose γ between 10 and 15 in this case.
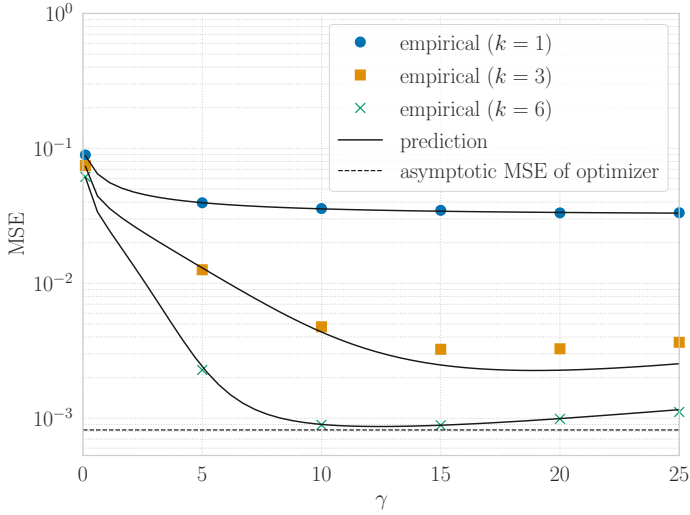
Figure 7: MSE performance of Douglas-Rachford algorithm for ℓ_1 optimization versus γ ($N = 500$, $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_{\mathrm{v}}^2 = 0.001$, $\rho_k = 1$).

We also investigate the performance of the Douglas-Rachford algorithm with the ℓ_1 regularization for different values of σ_{v}^2 in Figure 8. In the figure, we set $N = 500$, $\Delta = 0.7$, $p_0 = 0.9$, $\gamma = 10$, and $\rho_k = 1$. From Figure 8, we can see that the empirical performance is well predicted for different values of σ_{v}^2. The figure also implies that the performance is improved as the value of σ_{v}^2 decreases, though the required number of iterations for the convergence increases.

4.4   *Performance Prediction of Douglas-Rachford Algorithm with Nonconvex Regularization*

Finally, we evaluate the performance for nonconvex sparse regularization. Figure 9 shows the MSE performance of the Douglas-Rachford algorithm with SCAD regularization in (3), where $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_{\mathrm{v}}^2 = 0.001$, $\lambda = 0.1$, $\gamma = 1$, $\rho_k = 1$, and $a = 4$. The empirical performance and the prediction are obtained in the same way as Figure 3. From Figure 9, we observe that the empirical MSE and its prediction converge to almost the same value in this case. In the middle of iterations, however, the behavior of the prediction is different from Figure 3, and the empirical performance when $N = 500, 1000$ is better than the prediction. One of the possible reasons is the nonconvexity of the SCAD regularizer, and further investigation is necessary for a clear understanding. To obtain more precise results, we might need to taking the nonconvexity into account in the analysis.
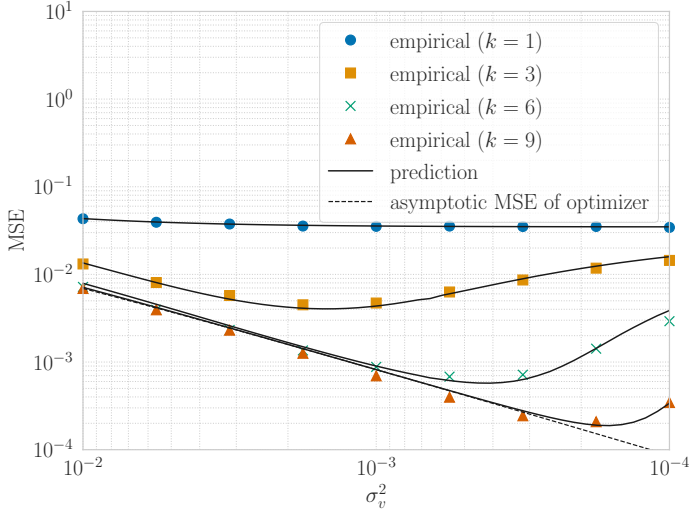
Figure 8: MSE performance of Douglas-Rachford algorithm for ℓ_1 optimization versus $\hat{\sigma}_{\mathrm{v}}^2$ ($N = 500, \Delta = 0.7, p_0 = 0.9, \gamma = 10, \rho_k = 1$).
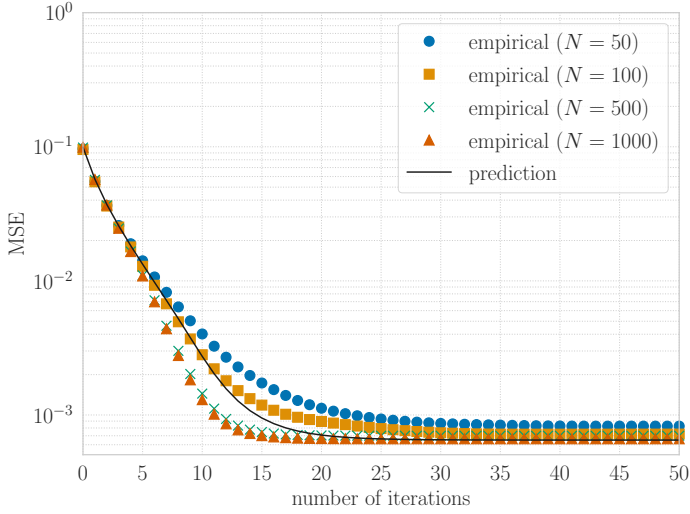

Figure 9: MSE performance of Douglas-Rachford algorithm with nonconvex SCAD regularization ($\Delta = 0.7$, $p_0 = 0.9$, $\sigma_{\mathrm{v}}^2 = 0.001$, $\lambda = 0.1$, $\gamma = 1$, $\rho_k = 1$, $a = 4$).

We also investigate the performance of the Douglas-Rachford algorithm with SCAD regularization for a larger noise variance $\sigma_{\mathrm{v}}^2 = 0.01$ in Figure 10,
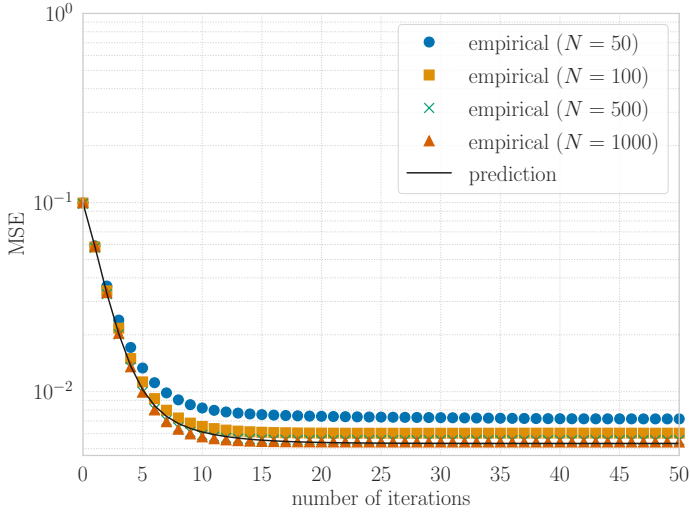
Figure 10: MSE performance of Douglas-Rachford algorithm with SCAD regularization $(\Delta = 0.7, p_0 = 0.9, \sigma_v^2 = 0.01, \lambda = 0.2, \gamma = 1, \rho_k = 1, a = 4)$.

where $\Delta = 0.7$, $p_0 = 0.9$, $\lambda = 0.2$, $\gamma = 1$, $\rho_k = 1$, and $a = 4$. From the figure, we can see that the empirical performance is well predicted in this case. The difference in behavior compared to Figure 9 is probably influenced by the value of σ_v^2 and other parameters, presenting an intriguing avenue for future research.

Figure 11 shows the performance of the Douglas-Rachford algorithm with the nonconvex MCP regularization. In the simulation, we set $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, $\lambda = 0.1$, $\gamma = 1$, $\rho_k = 1$, and $b = 4$. As in Figure 9, the empirical performance when $N = 500, 1000$ is better than the prediction in the middle of iterations. These results suggest that the nonconvexity of the MCP regularizer also affects the behavior of the prediction.

Figure 12 shows the MSE at the 50-th iteration versus the regularization parameter λ, where $\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, $\gamma = 1$, and $\rho_k = 1$. In the simulation, we use the SCAD regularization with $a = 4$. From the figure, we can see that the prediction is close to the empirical performance for $\lambda \geq 0.08$, whereas the prediction is worse for smaller values of λ. Even in this case, however, we can tune a reasonable value of $\lambda \approx 0.07$ on the basis of the prediction, without the empirical reconstruction.
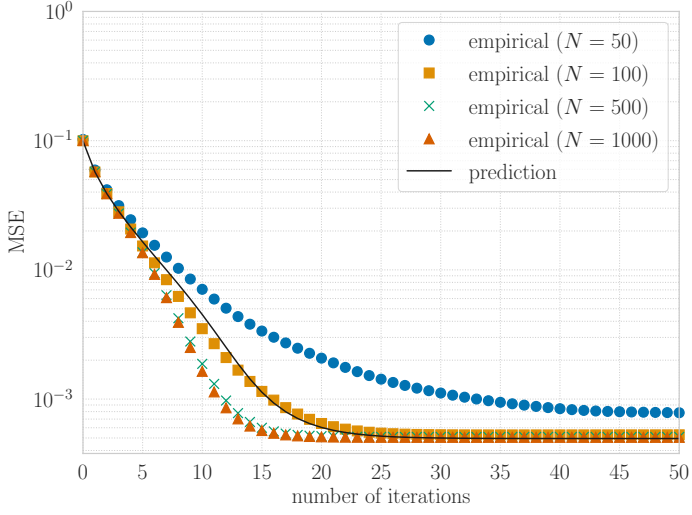
Figure 11: MSE performance of Douglas-Rachford algorithm with nonconvex MCP regularization ($\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, $\lambda = 0.1$, $\gamma = 1$, $\rho_k = 1$, $b = 4$).
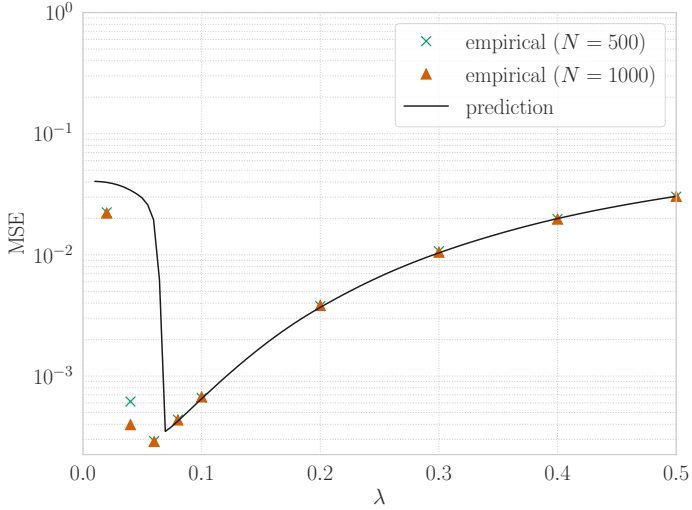

Figure 12: MSE performance of Douglas-Rachford algorithm with nonconvex SCAD regularization versus λ ($\Delta = 0.7$, $p_0 = 0.9$, $\sigma_v^2 = 0.001$, $\gamma = 1$, $\rho_k = 1$, $a = 4$, $k = 50$).

5   Conclusion

In this paper, we have analyzed the asymptotic behavior of the proximity operator for the squared loss function in linear inverse problems. We have also shown that the result in Proposition 3.1 can be used for the asymptotic performance prediction of some optimization algorithms for large-scale linear inverse problems. Simulation results show that the empirical distribution of the output of the proximity operator agrees well with the theoretical prediction by Proposition 3.1. Moreover, the empirical MSE performance of the Douglas-Rachford algorithm can be well predicted in large-scale compressed sensing via the ℓ_1 optimization. For the nonconvex SCAD regularization and MCP regularization, the prediction is also valid to some extent, although the behavior is slightly different from the convex case. These results suggest that the proposed prediction method can be a foundation of the rigorous analysis of optimization algorithms for large-scale linear inverse problems.

Future work includes the rigorous analysis of the prediction by Conjecture 3.1 and the further investigation of interesting behavior for the case with nonconvex regularization. One of possible directions is to extend or modify the approach in [10], though independent measurement matrices at each iteration of the algorithm is assumed in their analysis. Extensions of our approach beyond Assumptions 3.1 and 3.2 would also be an important research direction. Applications of our results to optimization algorithms other than the Douglas-Rachford algorithm would also be an interesting topic.

A   Appendix

A.1   Derivation of (14)

We provide an overview of the derivation of the optimization problem in (14) from (7). Since the procedure is almost the same as the CGMT-based analyses [35, 27, 26], the rigorous discussion is omitted in some parts.

We first define $\boldsymbol{e} = \boldsymbol{s} - \boldsymbol{x}$. The problem in (7) can be rewritten as the optimization problem for $\boldsymbol{e}$, i.e.,

$$\min_{\boldsymbol{e} \in \mathbb{R}^N} \frac{1}{N} \left\{ \frac{1}{2} \|\boldsymbol{A}\boldsymbol{e} - \boldsymbol{v}\|_2^2 + \frac{1}{2\gamma} \|\boldsymbol{e} + \boldsymbol{x} - \boldsymbol{z}\|_2^2 \right\}. \tag{24}$$

Note that we have normalized the objective function by N. From the property

$$\frac{1}{2} \|\boldsymbol{A}\boldsymbol{e} - \boldsymbol{v}\|_2^2 = \max_{\boldsymbol{u} \in \mathbb{R}^M} \left\{ \sqrt{N}\boldsymbol{u}^\top (\boldsymbol{A}\boldsymbol{e} - \boldsymbol{v}) - \frac{N}{2} \|\boldsymbol{u}\|_2^2 \right\}, \tag{25}$$

we can obtain the optimization problem

$$\min_{\boldsymbol{e} \in \mathbb{R}^N} \max_{\boldsymbol{u} \in \mathbb{R}^M} \left\{ \frac{1}{N} \boldsymbol{u}^\top \left(\sqrt{N} \boldsymbol{A} \right) \boldsymbol{e} - \frac{1}{\sqrt{N}} \boldsymbol{v}^\top \boldsymbol{u} \right.$$

$$\left. - \frac{1}{2} \|\boldsymbol{u}\|_2^2 + \frac{1}{N} \frac{1}{2\gamma} \|\boldsymbol{e} + \boldsymbol{x} - \boldsymbol{z}\|_2^2 \right\}. \tag{26}$$

CGMT [35] enables us to analyze the optimization problem

$$\min_{\boldsymbol{e} \in \mathbb{R}^N} \max_{\boldsymbol{u} \in \mathbb{R}^M} \left\{ \frac{1}{N} \left(\|\boldsymbol{e}\|_2 \, \boldsymbol{g}^\top \boldsymbol{u} - \|\boldsymbol{u}\|_2 \, \boldsymbol{h}^\top \boldsymbol{e} \right) - \frac{1}{\sqrt{N}} \boldsymbol{v}^\top \boldsymbol{u} \right.$$

$$\left. - \frac{1}{2} \|\boldsymbol{u}\|_2^2 + \frac{1}{N} \frac{1}{2\gamma} \|\boldsymbol{e} + \boldsymbol{x} - \boldsymbol{z}\|_2^2 \right\} \tag{27}$$

instead of (26), where the elements of $\boldsymbol{g} \in \mathbb{R}^M$ and $\boldsymbol{h} \in \mathbb{R}^N$ are i.i.d. random variables with $\mathcal{N}(0,1)$. Considering that both $\boldsymbol{g}$ and $\boldsymbol{v}$ are Gaussian, each element $\frac{\|\boldsymbol{e}\|_2}{\sqrt{N}} \boldsymbol{g} - \boldsymbol{v}$ is also Gaussian with $\mathcal{N} \left(0, \frac{\|\boldsymbol{e}\|_2^2}{N} + \sigma_{\mathrm{v}}^2 \right)$. We can thus simplify $\left(\frac{\|\boldsymbol{e}\|_2}{\sqrt{N}} \boldsymbol{g} - \boldsymbol{v} \right)^\top \boldsymbol{u}$ as $\sqrt{\frac{\|\boldsymbol{e}\|_2^2}{N} + \sigma_{\mathrm{v}}^2} \boldsymbol{g}^\top \boldsymbol{u}$, where we use the notation $\boldsymbol{g}$ to represent a vector with the elements following $\mathcal{N}(0,1)$. Hence, we can rewrite (27) as

$$\min_{\boldsymbol{e} \in \mathbb{R}^N} \max_{\boldsymbol{u} \in \mathbb{R}^M} \left\{ \frac{1}{\sqrt{N}} \sqrt{\frac{\|\boldsymbol{e}\|_2^2}{N} + \sigma_{\mathrm{v}}^2} \boldsymbol{g}^\top \boldsymbol{u} - \frac{1}{N} \|\boldsymbol{u}\|_2 \, \boldsymbol{h}^\top \boldsymbol{e} \right.$$

$$\left. - \frac{1}{2} \|\boldsymbol{u}\|_2^2 + \frac{1}{N} \frac{1}{2\gamma} \|\boldsymbol{e} + \boldsymbol{x} - \boldsymbol{z}\|_2^2 \right\}. \tag{28}$$

If we define $\beta = \|\boldsymbol{u}\|_2$, the maximum value of $\boldsymbol{g}^\top \boldsymbol{u}$ in the first term of (28) can be written as $\beta \|\boldsymbol{g}\|_2$. Moreover, we use

$$\sqrt{\frac{\|\boldsymbol{e}\|_2^2}{N} + \sigma_{\mathrm{v}}^2} = \min_{\alpha > 0} \left(\frac{\alpha}{2} + \frac{\frac{\|\boldsymbol{e}\|_2^2}{N} + \sigma_{\mathrm{v}}^2}{2\alpha} \right) \tag{29}$$

and rewrite the square root term in (28) to obtain

$$\min_{\boldsymbol{e} \in \mathbb{R}^N} \max_{\beta \geq 0} \min_{\alpha > 0} \left\{ \frac{\alpha \beta}{2} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} + \frac{1}{N} \frac{\beta}{2\alpha} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} \|\boldsymbol{e}\|_2^2 + \frac{\beta \sigma_{\mathrm{v}}^2}{2\alpha} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} \right.$$

$$\left. - \frac{\beta}{N} \boldsymbol{h}^\top \boldsymbol{e} - \frac{1}{2} \beta^2 + \frac{1}{N} \frac{1}{2\gamma} \|\boldsymbol{e} + \boldsymbol{x} - \boldsymbol{z}\|_2^2 \right\}. \tag{30}$$

We can further rewrite (30) as

$$\min_{\alpha>0} \max_{\beta\geq 0} \left\{ \frac{\alpha\beta}{2} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} + \frac{\beta\sigma_{\mathrm{v}}^2}{2\alpha} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} - \frac{1}{2}\beta^2 \right.$$
$$\left. + \min_{\boldsymbol{e}\in\mathbb{R}^N} \frac{1}{N} \sum_{n=1}^{N} J_n(e_n, \alpha, \beta) \right\}, \tag{31}$$

where $(\cdot)_n$ represents the n-th element of the corresponding bold vector and

$$J_n(e_n, \alpha, \beta) = \frac{\beta}{2\alpha} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} e_n^2 - \beta h_n e_n + \frac{1}{2\gamma} \left(e_n + x_n - z_n\right)^2. \tag{32}$$

The minimum value of $J_n(e_n, \alpha, \beta)$ over e_n is given by $J_n(\hat{s}_n(\alpha, \beta) - x_n, \alpha, \beta)$, where

$$\hat{s}_n(\alpha, \beta)$$
$$= \frac{1}{\dfrac{\beta}{\alpha}\dfrac{\|\boldsymbol{g}\|_2}{\sqrt{N}} + \dfrac{1}{\gamma}} \left(\frac{\beta}{\alpha} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} \left(x_n + \frac{\sqrt{N}}{\|\boldsymbol{g}\|_2} \alpha h_n \right) + \frac{1}{\gamma} z_n \right). \tag{33}$$

Hence, we can rewrite the optimization problem (31) as

$$\min_{\alpha>0} \max_{\beta\geq 0} \left\{ \frac{\alpha\beta}{2} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} + \frac{\beta\sigma_{\mathrm{v}}^2}{2\alpha} \frac{\|\boldsymbol{g}\|_2}{\sqrt{N}} - \frac{1}{2}\beta^2 \right.$$
$$\left. + \frac{1}{N} \sum_{n=1}^{N} J_n\left(\hat{s}_n(\alpha, \beta) - x_n, \alpha, \beta\right) \right\}. \tag{34}$$

We can show that the objective function of the optimization problem in (34) converges pointwise to (14) as $M, N \to \infty$. $J_n(\hat{s}_n(\alpha, \beta) - x_n, \alpha, \beta)$ and $\hat{s}_n(\alpha, \beta)$ correspond to (15) and (16), respectively.

Intuitively, from the definition of α in (29), the optimal value of α corresponds to $\sqrt{\frac{\|\boldsymbol{e}\|_2^2}{N} + \sigma_{\mathrm{v}}^2}$, which finally results in (17). For the precise discussion, see [35, 27].

Biography

Ryo Hayakawa received the bachelors degree in engineering, masters degree in informatics, and Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 2015, 2017, and 2020, respectively. He was a Research Fellow (DC1)

of the Japan Society for the Promotion of Science (JSPS) from 2017 to 2020, and an Assistant Professor at Osaka University from 2020 to 2023. From 2023, he is an Associate Professor at Institute of Engineering, Tokyo University of Agriculture and Technology. From 2023, he is an Associate Editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. He received the 33rd Telecom System Technology Student Award, APSIPA ASC 2019 Best Special Session Paper Nomination Award, and the 16th IEEE Kansai Section Student Paper Award. His research interests include signal processing and mathematical optimization. He is a member of IEEE and IEICE.

References

[1] A. Aïssa-El-Bey, D. Pastor, S. M. A. Sbaï, and Y. Fadlallah, "Sparsity-Based Recovery of Finite Alphabet Solutions to Underdetermined Linear Systems", *IEEE Transactions on Information Theory*, 61(4), April 2015, 2008–18.

[2] I. B. Atitallah, C. Thrampoulidis, A. Kammoun, T. Y. Al-Naffouri, M. Alouini, and B. Hassibi, "The BOX-LASSO with Application to GSSK Modulation in Massive MIMO Systems", in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2017, 1082–6.

[3] M. Bayati and A. Montanari, "The LASSO Risk for Gaussian Matrices", *IEEE Transactions on Information Theory*, 58(4), April 2012, 1997–2017.

[4] M. Bayati, M. Lelarge, and A. Montanari, "Universality in Polytope Phase Transitions and Message Passing Algorithms", *Annals of Applied Probability*, 25(2), April 2015, 753–822.

[5] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems", *SIAM Journal on Imaging Sciences*, 2(1), January 2009, 183–202.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", *Found. Trends Mach. Learn.*, 3(1), January 2011, 1–122.

[7] E. J. Candès, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information", *IEEE Transactions on Information Theory*, 52(2), February 2006, 489–509.

[8] E. J. Candès and T. Tao, "Decoding by Linear Programming", *IEEE Transactions on Information Theory*, 51(12), December 2005, 4203–15.

[9] E. J. Candès, "The Restricted Isometry Property and Its Implications for Compressed Sensing", *Comptes Rendus Mathematique*, 346(9), May 2008, 589–92.

[10] K. A. Chandrasekher, A. Pananjady, and C. Thrampoulidis, "Sharp global convergence guarantees for iterative nonconvex optimization with random data", *Ann. Stat.*, 51(1), February 2023, 179–210.

[11] P. L. Combettes and L. E. Glaudin, "Fully Proximal Splitting Algorithms in Image Recovery", in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, September 2019, 1–5.

[12] P. L. Combettes and L. E. Glaudin, "Proximal Activation of Smooth Functions in Splitting Algorithms for Convex Image Recovery", *SIAM Journal on Imaging Sciences*, 12(4), January 2019, 1905–35.

[13] P. L. Combettes and J.-C. Pesquet, "Proximal Splitting Methods in Signal Processing", in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Vol. 49, *Springer Optimization and Its Applications*, New York, NY: Springer New York, 2011, 185–212, ISBN: 978-1-4419-9569-8.

[14] P. L. Combettes and V. R. Wajs, "Signal Recovery by Proximal Forward-Backward Splitting", *Multiscale Modeling & Simulation*, 4(4), January 2005, 1168–200.

[15] I. Daubechies, M. Defrise, and C. D. Mol, "An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint", *Communications on Pure and Applied Mathematics*, 57(11), 2004, 1413–57.

[16] D. L. Donoho, "Compressed Sensing", *IEEE Transactions on Information Theory*, 52(4), April 2006, 1289–306.

[17] D. L. Donoho, I. Johnstone, and A. Montanari, "Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising", *IEEE Transactions on Information Theory*, 59(6), June 2013, 3396–433.

[18] D. L. Donoho, A. Maleki, and A. Montanari, "The Noise-Sensitivity Phase Transition in Compressed Sensing", *IEEE Transactions on Information Theory*, 57(10), October 2011, 6920–41.

[19] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators", *Mathematical Programming*, 55(1), April 1992, 293–318.

[20] J. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties", *Journal of the American Statistical Association*, 96(456), December 2001, 1348–60.

[21] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems", *IEEE Journal of Selected Topics in Signal Processing*, 1(4), December 2007, 586–97.

[22] D. Gabay and B. Mercier, "A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximation", *Computers & Mathematics with Applications*, 2(1), January 1976, 17–40.

[23] R. Hayakawa and K. Hayashi, "Convex Optimization-Based Signal Detection for Massive Overloaded MIMO Systems", *IEEE Transactions on Wireless Communications*, 16(11), November 2017, 7080–91.

[24] R. Hayakawa and K. Hayashi, "Discreteness-Aware Decoding for Overloaded Non-Orthogonal STBCs via Convex Optimization", *IEEE Communications Letters*, 22(10), October 2018, 2080–3.

[25] R. Hayakawa and K. Hayashi, "Reconstruction of Complex Discrete-Valued Vector via Convex Optimization with Sparse Regularizers", *IEEE Access*, 6, 2018, 66499–512.

[26] R. Hayakawa, "Asymptotic Performance Prediction for ADMM-based Compressed Sensing", *IEEE Transactions on Signal Processing*, 70, 2022, 5194–207.

[27] R. Hayakawa and K. Hayashi, "Asymptotic Performance of Discrete-Valued Vector Reconstruction via Box-Constrained Optimization with Sum of ℓ_1 Regularizers", *IEEE Transactions on Signal Processing*, 68, 2020, 4320–35.

[28] P. Lions and B. Mercier, "Splitting Algorithms for the Sum of Two Nonlinear Operators", *SIAM Journal on Numerical Analysis*, 16(6), December 1979, 964–79.

[29] D. G. Luenberger and Y. Ye, "Basic Descent Methods", in *Linear and Nonlinear Programming*, *International Series in Operations Research & Management Science*, New York, NY: Springer US, 2008, 215–62, ISBN: 978-0-387-74503-9.

[30] M. Nagahara, "Discrete Signal Reconstruction by Sum of Absolute Values", *IEEE Signal Processing Letters*, 22(10), October 2015, 1575–9.

[31] S. Oymak and J. A. Tropp, "Universality Laws for Randomized Dimension Reduction, with Applications", *Information and Inference: A Journal of the IMA*, 7(3), September 2018, 337–446.

[32] A. Panahi and B. Hassibi, "A Universal Analysis of Large-Scale Regularized Least Squares Solutions", in *Proc. Advances in Neural Information Processing Systems*, 2017, 3381–90.

[33] H. Sasahara, K. Hayashi, and M. Nagahara, "Multiuser Detection Based on MAP Estimation with Sum-of-Absolute-Values Relaxation", *IEEE Transactions on Signal Processing*, 65(21), November 2017, 5621–34.

[34] P. H. Tan, L. K. Rasmussen, and T. J. Lim, "Constrained Maximum-Likelihood Detection in CDMA", *IEEE Transactions on Communications*, 49(1), January 2001, 142–53.

[35] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise Error Analysis of Regularized *M*-Estimators in High Dimensions", *IEEE Transactions on Information Theory*, 64(8), August 2018, 5592–628.

[36] C. Thrampoulidis, A. Panahi, and B. Hassibi, "Asymptotically Exact Error Analysis for the Generalized ℓ_2^2-LASSO", in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2015, 2021–5.

[37] C. Thrampoulidis, W. Xu, and B. Hassibi, "Symbol Error Rate Performance of Box-Relaxation Decoders in Massive MIMO", *IEEE Transactions on Signal Processing*, 66(13), July 2018, 3377–92.

[38] F. Wen, L. Chu, P. Liu, and R. C. Qiu, "A Survey on Nonconvex Regularization-Based Sparse and Low-Rank Recovery in Signal Processing, Statistics, and Machine Learning", *IEEE Access*, 6, 2018, 69883–906.

[39] C.-H. Zhang, "Nearly Unbiased Variable Selection under Minimax Concave Penalty", *The Annals of Statistics*, 38(2), April 2010.

[40] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", en, *J. R. Stat. Soc. Series B Stat. Methodol.*, 67(2), April 2005, 301–20.