Machine Learning (2019) 108:1019–1056
https://doi.org/10.1007/s10994-018-5772-7

A simple homotopy proximal mapping algorithm for compressive sensing

Tianbao Yang[1] · Lijun Zhang[2] · Rong Jin[3] · Shenghuo Zhu[3] · Zhi-Hua Zhou[2]

Received: 23 February 2018 / Accepted: 7 November 2018 / Published online: 16 November 2018

Abstract
In this paper, we present novel yet simple homotopy proximal mapping algorithms for reconstructing a sparse signal from (noisy) linear measurements of the signal or for learning a sparse linear model from observed data, where the former task is well-known in the field of compressive sensing and the latter task is known as model selection in statistics and machine learning. The algorithms adopt a simple proximal mapping of the ℓ_1 norm at each iteration and gradually reduces the regularization parameter for the ℓ_1 norm. We prove a global linear convergence of the proposed homotopy proximal mapping (HPM) algorithms for recovering the sparse signal under three different settings (i) sparse signal recovery under noiseless measurements, (ii) sparse signal recovery under noisy measurements, and (iii) nearly-sparse signal recovery under sub-Gaussian noisy measurements. In particular, we show that when the measurement matrix satisfies restricted isometric properties (RIP), one of the proposed algorithms with an appropriate setting of a parameter based on the RIP constants converges linearly to the optimal solution up to the noise level. In addition, in setting (iii), a practical variant of the proposed algorithms does not rely on the RIP constants and our results for sparse signal recovery are better than the previous results in the sense that our recovery error bound is smaller. Furthermore, our analysis explicitly exhibits that more observations lead to not only more accurate recovery but also faster convergence. Finally our empirical studies provide further support for the proposed homotopy proximal mapping algorithm and verify the theoretical results.

Keywords Compressive sensing · Sparse signal recovery · Proximal mapping · Linear convergence

1 Introduction

The problem of sparse signal recovery is to reconstruct a sparse signal given a number of linear measurements of the signal. The problem has been studied extensively in the literature

Editor: Jean-Philippe Vert.

✉ Tianbao Yang
tianbao-yang@uiowa.edu

Extended author information available on the last page of the article

related to compressive sensing (Candès and Wakin 2008; Donoho 2006) and model selection in statistics and machine learning (Tibshirani 1996; Efron et al. 2004; Kyrillidis and Cevher 2012). Let $\mathbf{x}_* \in \mathbb{R}^d$ denote a target (nearly) sparse signal and $\mathbf{y} = U\mathbf{x}_* + \mathbf{e} \in \mathbb{R}^n$ denote $n < d$ measurements of $\mathbf{x}_*$, where $U \in \mathbb{R}^{n\times d}$ is a measurement matrix and $\mathbf{e}$ encodes potential noise in the observations. The task is to recover $\mathbf{x}_*$ from $\mathbf{y}$ and U (maybe with some knowledge of noise $\mathbf{e}$ and the sparsity of $\mathbf{x}_*$). In this paper, we will use the terminologies from the field of compressive sensing for our presentation. One can easily map the terminologies to the ones that are common in statistics and machine learning, e.g., sparse signal is also called sparse linear model, the measurement matrix is also known as input data matrix, the observations in $\mathbf{y}$ are also called output.

To facilitate the presentation and discussion below, we first introduce some notations. A vector $\mathbf{x}_* \in \mathbb{R}^d$ is said to be an s-sparse signal if the number of non-zero elements in $\mathbf{x}_*$ is s. Let $|\mathcal{S}|$ denote the cardinality of a subset $\mathcal{S} \subseteq \{1, \ldots, d\}$, and let $\mathbf{x}^s \in \mathbb{R}^d$ denote the vector $\mathbf{x} \in \mathbb{R}^d$ with all but the s largest entries (in magnitude) set to zero. Denote by $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_\infty$ and $\|\mathbf{x}\|_0$ the ℓ_2, ℓ_1, ℓ_∞ and ℓ_0 norm, respectively.

Numerous algorithms and results have been developed for sparse signal recovery under different settings and different conditions. In the earliest studies of compressive sensing (Candès and Tao 2005; Candès 2008; Chen et al. 2001; Donoho and Tsaig 2008), the sparse signal recovery is cast into a convex programming problem:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \quad & \|\mathbf{x}\|_1 \\ s.t. \quad & \|U\mathbf{x} - \mathbf{y}\|_2 \le \epsilon. \end{aligned} \tag{1}$$

It was shown that when the measurement matrix U satisfies RIP with small RIP constants (c.f. Definition 1), the solution to (1) denoted by $\bar{\mathbf{x}}$ can recover the sparse signal $\mathbf{x}_*$ up to the noise level $\|\mathbf{e}\|_2$. In their seminal work (Candès and Tao 2005), Candès and Tao proved that when $\mathbf{e} = 0$, i.e, there is no noise in the observations, $\mathbf{x}_*$ is the unique solution to (1) provided that RIP constants of U satisfy $\delta_s + \delta_{2s} + \delta_{3s} < 1$, where s is the number of non-zero elements in $\mathbf{x}_*$. The recovery result was later generalized to a more general setting of nearly-sparse signal recovery with noisy observations, under the condition $\delta_{2s} \le \sqrt{2} - 1$ and $\epsilon \ge \|\mathbf{e}\|_2$ (Candès 2008). Similar recovery results have been obtained for the Dantzig selector (Candès and Tao 2007)

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \quad & \|\mathbf{x}\|_1 \\ s.t. \quad & \|U^\top(U\mathbf{x} - \mathbf{y})\|_\infty \le \lambda \end{aligned} \tag{2}$$

with $\lambda \ge \|U^\top \mathbf{e}\|_\infty$. The sparse signal recovery is also closely related to the basis pursuit denoising problem (BPDN) (Chen et al. 1998), which aims to solve the following unconstrained ℓ_1 regularized least-squares minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad \underbrace{\frac{1}{2}\|U\mathbf{x} - \mathbf{y}\|_2^2}_{f(\mathbf{x})} + \lambda\|\mathbf{x}\|_1, \tag{3}$$

where λ is a regularization parameter. Various properties of the optimal solution $\bar{\mathbf{x}}$ to (3) have been investigated (Meinshausen and Bühlmann 2006; Tropp 2006b; Zhao and Yu 2006; Zhang and Huang 2008; Zhang 2009; Bickel et al. 2009; van de Geer and Bühlmann 2009; Wainwright 2009). In particular, it is known that under RIP for U, as long as $\lambda > c\|U^\top \mathbf{e}\|_\infty$, where c is a universal constant, the optimal solution $\bar{\mathbf{x}}$ to (3) can recover an s-sparse signal $\mathbf{x}_*$ up to the noise level.

In this paper, we study the problem of sparse signal recovery by directly analyzing the convergence of a new design of optimization algorithms. The algorithms adopt a proximal mapping for the ℓ_1 norm regularization at each iteration:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})\right)\right\|_2^2 + \lambda_t\|\mathbf{x}\|_1,$$

with a gradually reduced regularization parameter λ_t. It is not difficult to show that the proximal mapping above is one proximal gradient step (Nesterov 2007) for solving (3) with λ_t, i.e.,

$$\begin{aligned}\mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2 + \left[f(\mathbf{x}_t) + (\mathbf{x} - \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t)\right] + \lambda_t\|\mathbf{x}\|_1 \\ &= \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2 + \mathbf{x}^\top U^\top(U\mathbf{x}_t - \mathbf{y}) + \lambda_t\|\mathbf{x}\|_1,\end{aligned}$$

where the terms in the square bracket can be considered as a first-order Taylor expansion of $f(\mathbf{x})$ around $\mathbf{x}_t$. We name the proposed algorithms as homotopy proximal mapping algorithms, where the term "homotopy" is similarly used in Xiao and Zhang (2013) to name their method, which refers to the gradually decreasing strategy of λ_t.[1]

We prove that under RIP conditions for U the solution $\mathbf{x}_t$ will converge *linearly* to a solution $\bar{\mathbf{x}}$ that recovers the sparse signal up to the noise level. The definition of involved RIP constant is given below.

Definition 1 (*s-restricted isometry constant*) Let $\delta_s \geq 0$ be the smallest constant such that for any subset $\mathcal{T} \subseteq \{1, \ldots, d\}$ with $|\mathcal{T}| \leq s$ and $\mathbf{x} \in \mathbb{R}^{|\mathcal{T}|}$,

$$(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|U_{\mathcal{T}}\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2,$$

where $U_{\mathcal{T}}$ denotes a sub-matrix of U with column indices from $\mathcal{T}$.

In particular, we establish the convergence results in three settings.

Setting I Sparse signal recovery under noiseless observations. For any s-sparse vector $\mathbf{x}_*$, if $\mathbf{e} = 0$ and U satisfies the RIP such that

$$\gamma = \delta_s + \sqrt{2}\delta_{2s} + \delta_{3s} < 1, \tag{4}$$

then with an appropriate setting of λ_t the proposed algorithm (Algorithm 1) produces a sequence of solutions $\mathbf{x}_{t+1}$, which converges linearly to $\mathbf{x}_*$, i.e.,

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq \gamma^t \Delta_1,$$

where Δ_1 is an upper bound of $\|\mathbf{x}_1 - \mathbf{x}_*\|_2$, here and in Settings II and III.

Setting II Sparse signal recovery under noisy observations. For any s-sparse vector $\mathbf{x}_*$, if U satisfies the RIP such that (4) holds, then with an appropriate setting of λ_t the proposed algorithm (Algorithm 1) produces a sequence of solutions $\mathbf{x}_{t+1}$, which converges linearly to a region in which any point recovers $\mathbf{x}_*$ up to the noise level, i.e.,

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq \gamma^t \Delta_1 + \frac{1 + \sqrt{2}}{1 - \gamma}\sqrt{s}\|U^\top \mathbf{e}\|_\infty,$$

where γ is given in (4).

[1] It is notable that the decreasing strategy of the homotopy parameter in this paper uses an exogenously specified sequence unlike that in other homotopy-type methods that is adaptively changed (Asif and Romberg 2014; Brauer et al. 2018).

Setting III Nearly sparse signal recovery under a sub-Gaussian measurement matrix U. For a fixed vector $\mathbf{x}_*$, then with an appropriate setting of λ_t the proposed algorithm (Algorithm 2) produces a sequence of solutions $\mathbf{x}_{t+1}$, which converges linearly to a solution $\bar{\mathbf{x}}$ that recovers $\mathbf{x}_*^s$ up to the noise level under a condition that the number of measurements n is large enough. In particular, assuming that there exists $\eta > 0$ such that $c\sqrt{(\tau + s\log[d/s])/n} \leq \eta < 1/(1+\sqrt{2})$ holds for some universal constant $c > 0$ (from the JL lemma of sub-Gaussian measurement matrix U, e.g., 16 for a Gaussian measurement matrix) and some $\tau > 0$, then with high probability $1 - 2te^{-\tau}$ the following inequality holds

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq ((1+\sqrt{2})\eta)^t \Delta_1 + \frac{1+\sqrt{2}}{1-\gamma}\Lambda, \tag{5}$$

where Λ is defined as

$$\Lambda = \sqrt{s}\|U^\top \mathbf{e}\|_\infty + c\sqrt{\frac{\tau + s\log[d/s]}{n}}\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + c\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2.$$

In addition, in all three settings we show that $|\mathrm{supp}(\mathbf{x}_t)\backslash\mathrm{supp}(\mathbf{x}_*^s)| \leq s$, where $\mathrm{supp}(\mathbf{x})$ denotes the support set of $\mathbf{x}$, which implies that the number of non-zero elements beyond $\mathrm{supp}(\mathbf{x}_*^s)$ is no more than s.

We give few remarks about differences of theoretical/algorithmic results for Setting I, II and Setting III. The results in Settings I and II of the proposed algorithm hinge on appropriately setting the sequence of regularization parameters λ_t that depend on the RIP constants. In Setting III, we develop a more practical algorithm (Algorithm 3) with no algorithmic dependence on the RIP constants. As a tradeoff, the results in Setting III are developed for a random matrix U that satisfies the JL lemma (Johnson and Lindenstrauss 1984). In contrast, the results for the first two settings are deterministic by assuming U satisfies the RIP conditions. As a consequence, the results in Setting I and II hold for any sparse vector $\mathbf{x}_*$ and the result in Setting III only holds for a fixed $\mathbf{x}_*$ with high probability. In Sect. 2, we briefly discuss the above results in comparison with previous work.

2 Related work

We first compare our recovery results with state of the art results for (nearly) sparse signal recovery and then discuss the optimization algorithms for sparse signal recovery. We note that there exist studies focusing on the phase transition phenomenon of compressive sensing (Donoho et al. 2013, 2011; Donoho and Tanner 2009; Maleki and Donoho 2010). However, we will focus on sparse recovery under RIP of the measurement matrix and the convergence analysis of the proposed algorithms.

2.1 Sparse signal recovery with noiseless observations

Candès and Tao (2005) analyzed the recovery result for solving the ℓ_1 minimization problem (1) with noiseless observations $\mathbf{y} = U\mathbf{x}_*$, and showed that for any s-sparse signal $\mathbf{x}_*$ when U satisfies RIP[2] such that

[2] Using the restricted orthogonality constant $\theta_{s,s'}$ defined in **Definition** 2, a better condition on RIP constants can be established in their result as well as in our analysis. We use the restricted isometry constant δ_s in order to compare with other works and benefit from previous methods that estimate δ_s.

$$\delta_s + \delta_{2s} + \delta_{3s} < 1, \tag{6}$$

then the optimal solution to (1) with $\epsilon = 0$ is unique and is equal to $\mathbf{x}_*$. Comparing the inequality (4) and (6), our condition for exact recovery is close to the above condition. The exact recovery was also indicated in Candès' later work (Candès 2008) but with a slightly different RIP condition $\delta_{2s} < \sqrt{2} - 1$.

2.2 Sparse signal recovery with noisy observations

Candès (2008) also proved a recovery result for noisy observations. For any s-sparse vector $\mathbf{x}_*$, when U satisfies RIP such that $\delta_{2s} < \sqrt{2} - 1$, the optimal solution $\bar{\mathbf{x}}$ to (1) with $\epsilon \geq \|\mathbf{e}\|_2$ obeys

$$\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2 \leq C_2 \epsilon,$$

where C_2 is a constant depending on δ_{2s}. In comparison, our recovery error in **Setting II** depends on $\sqrt{s}\|U\mathbf{e}\|_\infty$ which could be smaller than $\|\mathbf{e}\|_2$ (e.g., when the entries in U are sub-Gaussian as stated in Proposition 6 in the "Appendix").

2.3 Nearly sparse signal recovery with noisy observations

A more general recovery result was also established in Candès (2008). For any vector $\mathbf{x}_*$, when U satisfies RIP such that $\delta_{2s} < \sqrt{2} - 1$, the optimal solution $\bar{\mathbf{x}}$ to (1) with $\epsilon \geq \|\mathbf{e}\|_2$ obeys

$$\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2 \leq C_0 \frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + C_2 \epsilon,$$

where C_0 is a constant depending on δ_{2s}. Similar results have also been developed for the Dantzig selector (2) (Candès and Tao 2007). Namely, when the RIP constant δ_{2s} of U satisfies $\delta_{2s} < \sqrt{2} - 1$, the optimal solution $\bar{\mathbf{x}}$ to (2) with $\lambda \geq \|U^\top \mathbf{e}\|_\infty$ satisfies

$$\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2 \leq C_0 \frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + C_3 \sqrt{s}\lambda,$$

where C_3 is a constant depending on δ_{2s}. A recovery result for nearly sparse signal under noisy observations in this paper is presented under our **Setting III**. There are two major differences between our result in **Setting III** and the previous recovery results (Candès 2008; Candès and Tao 2007). First, our result is probabilistic, which is directly established for a sub-Gaussian measurement matrix. In contrast, the mentioned previous results are deterministic by assuming RIP conditions of the measurement matrix. Second, our recovery error bound (5) could be smaller than that mentioned above (cf. the discussion below at the end of the second paragraph on page 7).

It is worth mentioning that there exist some studies on establishing sharper conditions on the RIP constants for exact or accurate recovery [see Cai and Zhang (2014) and references therein], which, however, is not the focus of this paper.

2.4 Instance-level recovery result

A weaker recovery result is that given a fixed signal $\mathbf{x}_*$, we can draw a random measurement matrix U and with a high probability expect certain performance for the recovery of the signal $\mathbf{x}_*$. We refer to this type of guarantee as instance-level recovery result (Eldar and

Kutyniok 2012). An advantage of the instance-level recovery is that we can achieve a recovery error in the form of $\|\bar{\mathbf{x}} - \mathbf{x}_*^s\|_2 \leq C\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ with C being a constant and $\bar{\mathbf{x}}$ being the recovered signal. However, such a result is impossible for any signal $\mathbf{x}_*$ without using a large number of observations, or in other words, such a result is only possible for any signal $\mathbf{x}_*$ when $n \geq cd$ for a constant $c > 0$ (i.e., $n = \Omega(d)$). In Eldar and Kutyniok (2012), it was shown that when the observations are free of noise and $U \in \mathbb{R}^{n\times d}$ is a sub-Gaussian random matrix with $n = O(s\log(d/s)/\delta_{2s}^2)$, then for a fixed signal $\mathbf{x}_*$, with probability $1 - 2\exp(-c_1\delta_{2s}^2 n) - \exp(c_0 n)$, the optimal solution $\bar{\mathbf{x}}$ to (1) with $\epsilon = 2\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ obeys

$$\|\bar{\mathbf{x}} - \mathbf{x}_*^s\|_2 \leq 2C_2\|\mathbf{x}_* - \mathbf{x}_*^s\|_2, \tag{7}$$

$$\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2 \leq (2C_2 + 1)\|\mathbf{x}_* - \mathbf{x}_*^s\|_2, \tag{8}$$

where $C_2 > 4$ is a constant depending on δ_{2s}. In contrast, our sparse signal recovery result for $\|\bar{\mathbf{x}} - \mathbf{x}_*^s\|_2$ in **Setting III** (considering no noise) could be much better than that in (7) since our error is dominated by $c\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2 + c\sqrt{s\log\lceil d/s\rceil/n}\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ and $\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2 \leq \|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ and $\sqrt{s\log\lceil d/s\rceil/n}\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 \leq \|\mathbf{x}_* - \mathbf{x}_*^s\|_2$, where c is a universal constant[3] and $\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2$ is the ℓ_2 norm of the largest s elements in $\mathbf{x}_* - \mathbf{x}_*^s$. To the best of our knowledge, this is the first such result in the literature.

There are also many studies on analyzing the properties of the optimal solution $\bar{\mathbf{x}}$ to the ℓ_1 regularized minimization problem in (3) (Meinshausen and Bühlmann 2006; Tropp 2006b; Zhao and Yu 2006; Zhang and Huang 2008; Zhang 2009; Bickel et al. 2009; van de Geer and Bühlmann 2009; Wainwright 2009). It is known that under RIP condition for U and $\lambda > c\|U^\top \mathbf{e}\|_\infty$ (for some universal constant c), we can obtain a recovery bound for any s-sparse signal $\mathbf{x}_*$

$$\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2 \leq O(\sqrt{s}\lambda).$$

In comparison, our analysis also exhibits that the final value of λ_t is $\Omega(\|U^\top \mathbf{e}\|_\infty)$ for sparse signal recovery (cf. Thereom 2). More literature on sparse signal recovery can be found in Eldar and Kutyniok (2012).

2.5 Optimization algorithms

There has been extensive research on solving the ℓ_1 minimization problems in (1) and (2), and the ℓ_1 regularized minimization problem in (3). Various algorithms have been developed, including greedy algorithms (Davis et al. 2004; Tropp 2006a; Needell and Tropp 2010; Mallat and Zhang 1993; Tropp and Gilbert 2007; Donoho et al. 2012; Needell and Vershynin 2009), interior-point methods (Chen et al. 2001; Turlach et al. 2005; Kim et al. 2008), proximal gradient methods (Nesterov 2007; Tseng 2008; Beck and Teboulle 2009; Becker et al. 2011), homotopy-type path-following methods (Osborne et al. 2000, 1999; Efron et al. 2004), iterative hard-thresholding methods (Garg and Khandekar 2009; Blumensath and Davies 2009; Foucart 2011; Kyrillidis and Cevher 2014), and many other methods (van den Berg and Friedlander 2008; Wright et al. 2009; Lorenz et al. 2014a; Asif and Romberg 2014; Brauer et al. 2018; Wen et al. 2010; Hale et al. 2008). It is notable that this list is by no means

[3] The constant c is just a positive constant that does not depend on the RIP constants of the matrix U like C_0, C_2. Its exact value depends on the parameters of sub-Gaussian distribution. For example, if X follows a sub-Gaussian distribution that satisfies $\Pr(|X| \geq t) \leq C\int_t^\infty e^{-\gamma t^2} dt$, then c is a constant depends on C and γ (Hanson and Wright 1971). When $U_{i,j}$ is a Gaussian variable following $\mathcal{N}(0, 1/\sqrt{n})$, one can easily derive that $c \geq 16$ works by following the analysis. C_0, C_2 are constants depending on δ_{2s}, please refer to Davenport et al. (2012) for the exact expressions.

complete. In Garg and Khandekar (2009) and Lorenz et al. (2014b), the authors gave a nice review of the convergence rates and their computational costs for different optimization algorithms. Below, we focus on two classes of algorithms that are closely related to the proposed work, with one employing the iterative hard-thresholding and the other exploiting the iterative soft-thresholding.

The hard-thresholding amounts to updating the solution based on the exact sparsification, i.e.,

$$\mathbf{x}_{t+1} = H_s\left(\mathbf{x}_t - \frac{1}{\gamma}U^\top(U\mathbf{x}_t - \mathbf{y})\right),$$

where γ is a constant and $H_s(\mathbf{x}) = \mathbf{x}^s$ is the hard-thresholding operator that gives the best s-sparse approximation of a vector x, i.e., setting all elements in $\mathbf{x}$ to zeros except for the s largest elements in magnitude. In Blumensath and Davies (2009), the authors analyzed the iterative hard-thresholding algorithm with $\gamma = 1$. They show that when U satisfies RIP with $\delta_{3s} < 1/\sqrt{32}$, the sequence $\{\mathbf{x}_t\}$ converges linearly to the best attainable solution up to a constant, namely,

$$\|\mathbf{x}_t - \mathbf{x}_*\|_2 \le 2^{-t}\|\mathbf{x}_*\|_2 + 6\left[\|\mathbf{e}\|_2 + \|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \frac{1}{\sqrt{s}}\|\mathbf{x}_* - \mathbf{x}_*^s\|_1\right]. \tag{9}$$

Similarly, Garg and Khandekar (2009) analyzed the iterative hard-thresholding with $\gamma = 1+\delta_{2s}$ under the **Settings I and II**, and showed the sequence $\{\mathbf{x}_t\}$ converges to a solution $\bar{\mathbf{x}}$ that recovers any s-sparse signal $\mathbf{x}_*$ signal up to the noise level, i.e., $\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2 \le 4/(1-\delta_{2s})\|\mathbf{e}\|_2$ with a rate of $((8\delta_{2s})/(1-\delta_{2s}))^t$, under the condition $\delta_{2s} < 1/3$. In contrast, the proposed algorithm in **Settings I and II** only requires $\delta_s + \sqrt{2}\delta_{2s} + \delta_{3s} < 1$, which is less restrictive than the condition $\delta_{3s} < 1/\sqrt{32}$ in Blumensath and Davies (2009) (due to that $\delta_s \le \delta_{2s} \le \delta_{3s}$). In **Setting III**, we proved a recovery for a fixed signal $\mathbf{x}_*$ with high probability for a sub-Gaussian measurement matrix. We can literally compare the non-diminishing terms in (5) and (9)[4] by ignoring the constant factors. First, $\sqrt{s}\|U^\top\mathbf{e}\|_\infty$ in (5) is much smaller than $\|\mathbf{e}\|_2$ in (9) when $n \ge \Omega(s\log d)$ as implied by "Appendix E". Second, $\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|$ in (5) is smaller than $\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ in (9). Third, $\sqrt{(\tau + s\log[d/s])/n}\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ in (5) is smaller than $\|\mathbf{x}_* - \mathbf{x}_*^s\|_1/\sqrt{s}$ when $\|\mathbf{x}_* - \mathbf{x}_*^s\|_2/\|\mathbf{x}_* - \mathbf{x}_*^s\|_1 \le \sqrt{n/(s(\tau + s\log d/s))}$.

The iterative soft-thresholding algorithm (ISTA) is based on the proximal mapping of ℓ_1 regularization for solving the ℓ_1 regularized minimization problem (3), where the updates are given by

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\left\|\mathbf{x} - \frac{1}{\gamma_t}U^\top(U\mathbf{x}_t - \mathbf{y})\right\|_2 + \frac{\lambda}{\gamma_t}\|\mathbf{x}\|_1,$$

where $1/\gamma_t$ is a step size. The proximal mapping springs from Nesterov's first order method for composite optimization (Nesterov 2007). In Bredies and Lorenz (2008) and Hale et al. (2008), the authors studied the soft-thresholding update with a constant step size and established local linear convergence rates as the iterates are close enough to the optimum. Although the update of the proposed algorithms is very similar to that of ISTA, there are several striking differences between ISTA and the proposed algorithms, including Algorithms 1, 2 and 3. First, ISTA solves exactly the ℓ_1 regularized least-squares problem (i.e., the BPDN problem) with a *fixed* regularization parameter. The proposed algorithms are to directly reconstruct a sparse signal from noisy measurements with a target sparsity as an input parameter. Second,

[4] Though the two results are not directly comparable because (9) is a deterministic result.

ISTA for optimizing the BPDN formulation to recover a sparse signal requires a regularization parameter λ such that $\lambda \geq \Omega(\|U^\top \mathbf{e}\|_\infty)$. The proposed Algorithm 3 does not need any knowledge about the order of $\|U^\top \mathbf{e}\|_\infty$ but instead need a target sparsity as an input parameter. It uses the proximal mapping of an ℓ_1 norm regularizer with a gradually decreasing regularization parameter λ_t until the solution exceeds the target sparsity by two times. Third, the proposed algorithms enjoy global linear convergence, while ISTA has only local linear convergence when the solution is close enough to the optimal solution. Last but not least, the presented algorithms and analysis provide a unified framework of optimization and recovery of sparse signals. In contrast, ISTA is only an optimization algorithm which solely provides no guarantee on the recovery of underlying true sparse signal. By a unified framework we mean that the convergence in terms of optimization and recovery of the target signal (please see Theorem 2, Part I and Part II) can be simultaneously achieved. In contrast, ISTA itself is just an optimization algorithm for solving ℓ_1 regularized problem with a fixed regularization parameter, which only has convergence guarantee for optimization. If one wants to make a statement about recovery error, one has to utilize other analysis tools for analyzing the ℓ_1 regularized formulation with an appropriate parameter, which is separate from the analysis for optimization.

Recently, several algorithms were shown to exhibit global linear convergence for the BPDN problem. Agarwal et al. (2010) studied a more general problem than (3) for statistical recovery, where the first quadratic term is replaced by $L_n(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}; \mathbf{z}_i)$ with $\ell(\mathbf{x}; \mathbf{z}_i)$ denoting a loss of the model $\mathbf{x}$ measured on the data $\mathbf{z}_i$. They used a different update

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top \nabla L_n(\mathbf{x}_t) + \frac{\gamma_u}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{10}$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_1 \leq \rho\}$, and γ_u is a parameter related to the restricted smoothness of $L_n(\mathbf{x})$. They proved a global linear convergence of the above update with $\rho = \Theta(\|\mathbf{x}_*\|_1)$ for finding a solution up to the statistical precision, meaning the typical distance between the true unknown parameter $\mathbf{x}_*$ and an optimal solution to the ℓ_1 regularized problem. Xiao and Zhang (2013) studied a proximal-gradient homotopy gradient method for solving (3). They iteratively solve the problem (3) by the proximal gradient descent with a decreasing regularization parameter λ and an increasing accuracy at each stage, and use the solution obtained at each stage to warm start the next stage. Global linear convergence was also established. Several improvements and extensions over Xiao and Zhang's results were made in Lin and Xiao (2015) and Eghbali and Fazel (2017). For example, Lin and Xiao (2015) proposed an adaptive accelerated proximal gradient method and a homotopy variant for solving BPDN problem with improved convergence. Eghbali and Fazel (2017) generalized the linear convergence rate analysis of the homotopy algorithm studied in Xiao and Zhang (2013) to problem (3) with the ℓ_1 norm replaced by a general class of decomposable norms.

Although there are many parallels between this work and Agarwal et al. (2010) and Xiao and Zhang (2013), there are big differences. (i) The proposed work is dedicated to sparse signal recovery, exhibiting the conditions in different settings under which the recovery is optimal. (ii) Different from Agarwal et al. (2010) that updates the solution using the constrained proximal mapping in (10), our algorithms solve a simple proximal mapping of the ℓ_1 norm regularization at each iteration. (iii) Different from Xiao and Zhang (2013) that updates the solution using a stage-wise proximal gradient descent with more parameters that need to be tuned in practice, the proposed homotopy proximal mapping algorithm as well as the analysis are much simpler. (iv) Our algorithm and analysis provide arguably better guarantees for the solutions. First, both the convergence rates and the recovery error of the proposed algorithms imply that more observations lead to not only more accurate recovery but

also faster convergence, which shed more insights about the sparse signal recovery problem and our algorithms, and was also observed by Oymak et al. (2018) on a relevant problem. Second, our algorithm can guarantee that the support sets of the intermediate solutions do not exceed the target support set by s, the target sparsity. In contrast, Agarwal et al. (2010) provides no explicit guarantee of sparsity bound for the intermediate solutions, and in Xiao and Zhang (2013) the support sets of the intermediate solutions beyond the target support set could be much larger than s.

3 Sparse signal recovery

Below, we first give some notations and definitions that will be used in the sequel. Then we present the algorithms, main results and their proofs.

3.1 Notations and definitions

We denote by $\mathcal{S}(\mathbf{x})$ the support set (or support for short) for $\mathbf{x}$ that includes all the indices of the non-zero entries in $\mathbf{x}$, i.e.,

$$\mathcal{S}(\mathbf{x}) = \{i \in \{1, \ldots, d\} : [\mathbf{x}]_i \neq 0\}, \tag{11}$$

where $[\mathbf{x}]_i$ denote the ith element in $\mathbf{x}$. Denote by $\mathcal{S}_1 \backslash \mathcal{S}_2$ a subset of $\mathcal{S}_1$ that contains all elements in $\mathcal{S}_1$ but not in $\mathcal{S}_2$. We also denote by $\overline{\mathcal{S}}(\mathbf{x}) = \{1, \ldots, d\} \backslash \mathcal{S}(\mathbf{x})$ the complementary set (or complement for short) of $\mathcal{S}(\mathbf{x})$. In particular, we use $\mathcal{S}_*, \overline{\mathcal{S}}_*$ to denote the support set and its complementary set of $\mathbf{x}_*$.

Considering a vector $\mathbf{x} \in \mathbb{R}^d$ and a matrix $M \in \mathbb{R}^{n \times d}$, given a set $\mathcal{S} \subseteq \{1, \ldots, d\}$, we denote by $[\mathbf{x}]_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ the vector that only includes the entries of $\mathbf{x}$ in the subset $\mathcal{S}$, and by $M_{\mathcal{S}}$ a sub-matrix that only contains the columns of M indexed by $\mathcal{S}$. Given two subsets $\mathcal{A} \subseteq \{1, \ldots, d\}$ and $\mathcal{B} \subseteq \{1, \ldots, d\}$, we denote by $[M]_{\mathcal{A},\mathcal{B}}$ a sub-matrix that includes all the entries (i, j) in matrix M with $i \in \mathcal{A}$ and $j \in \mathcal{B}$. $\|M\|_2$ denotes the spectral norm of a matrix M.

Let $U \in \mathbb{R}^{n \times d}$ be a measurement matrix and

$$\mathbf{y} = U\mathbf{x}_* + \mathbf{e} \tag{12}$$

be the corresponding n observations of the target signal $\mathbf{x}_*$. Similar to many works in compressive sensing, we assume the measurement matrix U satisfies the following some restricted isometry property (RIP) (with an overwhelming probability). Besides s-restricted isometry constant δ_s, we will also use the following RIP constant.

Definition 2 (s, s*-restricted orthogonality constant*) Let $\theta_{s,s}$ be the smallest constant such that for any two disjoint subsets $\mathcal{T}, \mathcal{T}' \subseteq \{1, \ldots, d\}$ with $|\mathcal{T}| \leq s$, $|\mathcal{T}'| \leq s$, $2s \leq d$, and for any $\mathbf{x} \in \mathbb{R}^{|\mathcal{T}|}$, $\mathbf{x}' \in \mathbb{R}^{|\mathcal{T}'|}$,

$$|\langle U_{\mathcal{T}}\mathbf{x}, U_{\mathcal{T}'}\mathbf{x}'\rangle| \leq \theta_{s,s}\|\mathbf{x}\|_2\|\mathbf{x}'\|_2.$$

Remark 1 Although the results in the following are stated using δ_s and $\theta_{s,s}$, we can easily obtain the results with only restricted isometry constants by noting that $\theta_{s,s} \leq \delta_{2s}$ (Candès and Tao 2005).

Based on the definitions of δ_s and $\theta_{s,s}$, we can derive the following facts.

Fact 1 For any subset $\mathcal{T} \subseteq \{1, \ldots, d\}$ with $|\mathcal{T}| \leq s$, the definition of δ_s implies that $U_{\mathcal{T}}^\top U_{\mathcal{T}}$ has all of its eigen-values in $[1-\delta_s, 1+\delta_s]$. As a result $\|(U_{\mathcal{T}}^\top U_{\mathcal{T}} - I)\mathbf{x}\|_2 \leq \delta_s \|\mathbf{x}\|_2$.

Fact 2 for any two disjoint subsets $\mathcal{T}, \mathcal{T}' \subseteq \{1, \ldots, d\}$ with $|\mathcal{T}| \leq s$, $|\mathcal{T}'| \leq s$, $2s \leq d$, the definition of restricted orthogonality constant implies that $\|U_{\mathcal{T}}^\top U_{\mathcal{T}'}\|_2 \leq \theta_{s,s}$.

The above two constants are standard tools in the analysis of compressive sensing. It has been shown that several random measurement matrices, including sub-Gaussian, partial Fourier and incoherent matrices, satisfy the RIP with small δ_s and $\theta_{s,s}$ (Candès et al. 2006). However, it should be noted that it is NP-hard to evaluate the RIP and compute RIP constants in general (Tillmann and Pfetsch 2014).

3.2 Algorithms and main results

To motivate our approach, we first consider the following optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad \mathcal{L}(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_*\|_2^2. \tag{13}$$

Evidently, the optimal solution to (13) is $\mathbf{x}_*$. We now consider a gradient descent method for optimizing the problem in (13), leading to the following updating equation for $\mathbf{x}_t$

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{x} - (\mathbf{x}_t - \nabla\mathcal{L}(\mathbf{x}_t))\|_2^2, \tag{14}$$

where $\nabla\mathcal{L}(\mathbf{x}_t) = \mathbf{x}_t - \mathbf{x}_*$. Since the problem in (13) is both smooth and strongly convex, the above updating enjoys a linear convergence rate with, in fact, only one step, allowing an efficient reconstruction of $\mathbf{x}_*$.

However, the updating rule in (14) can not be used because it requires knowing $\mathbf{x}_*$, the full information of the sparse signal to be recovered. In compressive sensing, the only available information about the target signal $\mathbf{x}_*$ is through a set of $n < d$ observations given in (12). Using the observations, we construct an approximate gradient as

$$\widehat{\nabla}\mathcal{L}(\mathbf{x}_t) = U^\top(U\mathbf{x}_t - \mathbf{y}) = U^\top U(\mathbf{x}_t - \mathbf{x}_*) - U^\top\mathbf{e}. \tag{15}$$

As can be seen if $U^T U(\mathbf{x}_t - \mathbf{x}_*)$ is close to $\mathbf{x}_t - \mathbf{x}_*$ and $U^\top\mathbf{e}$ is not significantly large in magnitude, $\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)$ would provide a useful estimate of $\nabla\mathcal{L}(\mathbf{x}_t)$. To ensure this, we should assume certain restrictive conditions on U and a small noise $\mathbf{e}$.

Next, we will use $\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)$ as an approximation of $\nabla\mathcal{L}(\mathbf{x}_t)$ and update the solution by performing the following proximal mapping:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \lambda_t\|\mathbf{x}\|_1 + (\mathbf{x} - \mathbf{x}_t)^\top\widehat{\nabla}\mathcal{L}(\mathbf{x}_t) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2, \tag{16}$$

where $\lambda_t > 0$ is a ℓ_1 norm regularization parameter that decreases over iterations. The updating rule given in (16) differs from (14) in that (i) the true gradient $\nabla\mathcal{L}(\mathbf{x}_t)$ is replaced with an approximate gradient $\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)$ and (ii) an ℓ_1 regularization term $\lambda_t\|\mathbf{x}\|_1$ is added. With appropriate choice of λ_t, this regularization term will essentially remove the noise arising from the gradient approximation and consequentially lead to a global linear convergence rate as shown in the following.

To give the solution of $\mathbf{x}_{t+1}$ in a closed form, we write (16) as

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})\right)\right\|_2^2 + \lambda_t\|\mathbf{x}\|_1. \tag{17}$$

Algorithm 1 Homotopy Proximal Mapping (HPM) Algorithm for Recovering a Sparse Signal

1: **Input:** The measurement matrix $U \in \mathbb{R}^{n\times d}$, observations $\mathbf{y} = U\mathbf{x}_* + \mathbf{e}$, a sequence of regularization parameters $\lambda_1, \ldots, \lambda_T$
2: **Initialize** $\mathbf{x}_1 = 0$.
3: **for** $t = 1, \ldots, T$ **do**
4: Compute $\widehat{\mathbf{x}}_t = \mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})$
5: Update the solution $\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\,[|\widehat{\mathbf{x}}_t| - \lambda_t]_+$
6: **end for**
7: **Output** the final solution $\mathbf{x}_{T+1}$

It is commonly known that the value of $\mathbf{x}_{t+1}$ is given by Beck and Teboulle (2009)

$$\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\,[|\widehat{\mathbf{x}}_t| - \lambda_t]_+\,, \tag{18}$$

where $\widehat{\mathbf{x}}_t$ denotes the intermediate solution before soft-thresholding given by

$$\widehat{\mathbf{x}}_t = \mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y}), \tag{19}$$

and $[v]_+ = \max(0, v)$. We present the detailed steps of the proposed algorithm in Algorithm 1 for reconstructing the sparse signal given a set of noiseless/noisy observations. To end this section, we present our main result in the following two theorems regarding the sparse signal recovery with noiseless observations and with noisy observations.

Theorem 1 *Let $\mathbf{x}_* \in \mathbb{R}^d$ be an s-sparse signal and $\mathbf{y} = U\mathbf{x}_*$ be a set of n measurements of $\mathbf{x}_*$. Assume U satisfies RIP such that*

$$\gamma = \delta_s + \sqrt{2}\theta_{s,s} + \delta_{3s} < 1.$$

Let $\{\Delta_1, \ldots, \Delta_t\}$ be a sequence such that $\|\mathbf{x}_1 - \mathbf{x}_*\|_2 \le \Delta_1$, and

$$\Delta_{t+1} = (\delta_s + \sqrt{2}\theta_{s,s} + \delta_{3s})\Delta_t.$$

If we run Algorithm 1 with $\lambda_t = \Delta_t(\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s}$, then for all $t \ge 0$ we have

(1) $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \le s$ *and,*
(2) $\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \le \gamma^t\Delta_1$.

Remark 2 Similar to iterative hard-thresholding algorithms (Garg and Khandekar 2009; Blumensath and Davies 2009), Algorithm 1 also requires knowledge of sparsity s and RIP constants in order to enjoy the guarantee stated in Theorem 1. These requirements (especially the knowledge of RIP constants) make Algorithm 1 quite restrictive. In the next section, we present a more practical algorithm. For the requirement of an upper bound Δ_1 on $\|\mathbf{x}_1 - \mathbf{x}_*\|_2 = \|\mathbf{x}_*\|_2 \le \Delta_1$, in practice one might be able to derive such an upper bound given some prior knowledge on the magnitude of the target signal. For example, if each entry of the target signal is known to be in some range $[a, b]$, then one can easily derive an upper bound of $\|\mathbf{x}_*\|_2$.

Theorem 2 *Let $\mathbf{x}_* \in \mathbb{R}^d$ be an s-sparse signal and $\mathbf{y} = U\mathbf{x}_* + \mathbf{e}$ be a set of n noisy measurements of $\mathbf{x}_*$. Assume U satisfies RIP such that*

$$\gamma = \delta_s + \sqrt{2}\theta_{s,s} + \delta_{3s} < 1.$$

Let $\{\Delta_1, \ldots, \Delta_t\}$ be a sequence such that $\|\mathbf{x}_1 - \mathbf{x}_*\|_2 \le \Delta_1$, and

$$\Delta_{t+1} = \gamma\Delta_t + (1 + \sqrt{2})\sqrt{s}\|U^\top\mathbf{e}\|_\infty, \quad t \ge 1.$$

Part I: If we run Algorithm 1 with

$$\lambda_t = \frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}\Delta_t + \|U^\top \mathbf{e}\|_\infty,$$

then for all $t \geq 0$ *we have*

(1) $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \leq s$ *and,*
(2) $\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq \gamma^t \Delta_1 + \frac{1-\gamma^t}{1-\gamma}(1+\sqrt{2})\sqrt{s}\|U^\top \mathbf{e}\|_\infty$.

Part II: In addition, if $\max(2\delta_{6s}, \gamma + (\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s}) < 1$, *the sequence* $\{\mathbf{x}_t\}$ *converges to a unique fixed-point* $\bar{\mathbf{x}}$ *such that* $|\mathcal{S}(\bar{\mathbf{x}})\backslash\mathcal{S}_*| \leq s$ *and* $\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2 \leq \sqrt{s}\|U^\top \mathbf{e}\|_\infty(1+\sqrt{2})/(1-\gamma)$. *Moreover,* $\bar{\mathbf{x}}$ *is an optimal solution to the following* ℓ_1*-regularized problem:*

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|U\mathbf{x} - \mathbf{y}\|_2^2 + \bar{\lambda}\|\mathbf{x}\|_1,$$

where $\bar{\lambda} = (\sqrt{2}(\delta_s + \sqrt{2}\theta_{s,s}) + 1 - \delta_{3s})\|U^\top \mathbf{e}\|_\infty/(1-\gamma)$.

Remark 3 Similar to solving Dantzig selector (2) and the ℓ_1 regularized problem (3) for sparse signal recovery that requires $\lambda \geq c\|U^\top \mathbf{e}\|_\infty$, the regularization parameters in our algorithm are also larger than $\|U^\top \mathbf{e}\|_\infty$ and eventually $\lambda_t \geq c\|U^\top \mathbf{e}\|_\infty$, where c depends on RIP constants.

Remark 4 While Theorems 1 and 2 are theoretically interesting, the value of λ_t depends on the RIP constants. In Sect. 4, we present more practical algorithms for (nearly) sparse signal recovery with a sub-Gaussian measurement matrix.

3.3 Proof of Theorem 1

To pave the path for proving Theorem 1, we will present and prove a series of propositions and lemmas. We first prove the following proposition regarding the magnitude of elements in $[\widehat{\mathbf{x}}_t]_{\overline{\mathcal{S}}_*}$.

Proposition 1 *Let* $\mathcal{S}_t$ *be the support set of* $\mathbf{x}_t$ *(the* t*th iterate of Algorithm 1) and* $\mathcal{S}_*$ *be the support set of* $\mathbf{x}_*$. *Define* $\mathcal{S}_t^c = \mathcal{S}_t \cup \mathcal{S}_*$, $\mathcal{S}_t^a = \mathcal{S}_t\backslash\mathcal{S}_*$ *and* $\widetilde{\mathbf{x}}_t = \mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)$. *If we assume* $|\mathcal{S}_t\backslash\mathcal{S}_*| \leq s$, *then there are at most* s *entries of* $[\widetilde{\mathbf{x}}_t]_{\overline{\mathcal{S}}_*}$ *with magnitude larger than* $(\delta_s + \sqrt{2}\theta_{s,s})\|\mathbf{x}_t - \mathbf{x}_*\|_2/\sqrt{s}$.

Proof For any subset $\mathcal{S}' \subset \overline{\mathcal{S}}_*$ of size s, let $\mathcal{S}_1' = \mathcal{S}' \cap \mathcal{S}_t^a$ and $\mathcal{S}_2' = \mathcal{S}'\backslash\mathcal{S}_t^a$. First, we have

$$\begin{aligned}\|[\widetilde{\mathbf{x}}_t]_{\mathcal{S}'}\|_2 &= \left\|[U^\top U(\mathbf{x}_t - \mathbf{x}_*)]_{\mathcal{S}'} - [\mathbf{x}_t]_{\mathcal{S}'}\right\|_2 \\ &= \left\|U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*} + U_{\mathcal{S}'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - [\mathbf{x}_t]_{\mathcal{S}'}\right\|_2,\end{aligned}$$

where the second equality is due to that the support of $\mathbf{x}_t - \mathbf{x}_*$ is $\mathcal{S}_t^c$ and we split that into two disjoint subsets $\mathcal{S}_t^a$ and $\mathcal{S}_*$. By noting that $\mathcal{S}'$ can be split into two subsets $\mathcal{S}_1'$ and $\mathcal{S}_2'$ that do not intersect with each other and that $\|[\mathbf{v}]_{\mathcal{S}'}\|_2 \leq \|[\mathbf{v}]_{\mathcal{S}_1'}\|_2 + \|[\mathbf{v}]_{\mathcal{S}_2'}\|_2$ with

$\mathbf{v} = U^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - \mathbf{x}_t$, we have

$$
\begin{aligned}
&\left\| U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*} + U_{\mathcal{S}'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - [\mathbf{x}_t]_{\mathcal{S}'} \right\|_2 \\
&\quad \le \left\| U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*} \right\|_2 + \left\| U_{\mathcal{S}'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - [\mathbf{x}_t]_{\mathcal{S}'} \right\|_2 \\
&\quad \le \left\| U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*} \right\|_2 + \left\| U_{\mathcal{S}_2'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - [\mathbf{x}_t]_{\mathcal{S}_2'} \right\|_2 + \left\| U_{\mathcal{S}_1'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - [\mathbf{x}_t]_{\mathcal{S}_1'} \right\|_2 \\
&\quad = \left\| U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*} \right\|_2 + \left\| U_{\mathcal{S}_2'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} \right\|_2 + \left\| U_{\mathcal{S}_1'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - [\mathbf{x}_t]_{\mathcal{S}_1'} \right\|_2 \\
&\quad \le \left\| U_{\mathcal{S}'}^\top U_{\mathcal{S}_*} \right\|_2 \|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\|_2 + \left\| U_{\mathcal{S}_2'}^\top U_{\mathcal{S}_t^a} \right\|_2 \|[\mathbf{x}_t]_{\mathcal{S}_t^a}\|_2 + \left\| U_{\mathcal{S}_t^a}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - [\mathbf{x}_t]_{\mathcal{S}_t^a} \right\|_2 \\
&\quad = \left\| U_{\mathcal{S}'}^\top U_{\mathcal{S}_*} \right\|_2 \|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\|_2 + \left\| U_{\mathcal{S}_2'}^\top U_{\mathcal{S}_t^a} \right\|_2 \|[\mathbf{x}_t]_{\mathcal{S}_t^a}\|_2 + \left\| (U_{\mathcal{S}_t^a}^\top U_{\mathcal{S}_t^a} - I)[\mathbf{x}_t]_{\mathcal{S}_t^a} \right\|_2 \\
&\quad \le \theta_{s,s}\|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\|_2 + \theta_{s,s}\|[\mathbf{x}_t]_{\mathcal{S}_t^a}\|_2 + \delta_s\|[\mathbf{x}_t]_{\mathcal{S}_t^a}\|_2 \\
&\quad = \theta_{s,s}\|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\|_2 + \theta_{s,s}\|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_t^a}\|_2 + \delta_s\|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_t^a}\|_2 \\
&\quad \le (\delta_s + \sqrt{2}\theta_{s,s})\|\mathbf{x}_t - \mathbf{x}_*\|_2.
\end{aligned}
$$

where the first equality uses the fact that $[\mathbf{x}_t]_{\mathcal{S}_2'} = 0$, the third inequality uses the fact that $\mathcal{S}_1' \subseteq \mathcal{S}_t^a$, the fourth inequality uses the RIP conditions (see Fact 1 and Fact 2) by noting that $|\mathcal{S}_t^a| \le s$, $|\mathcal{S}_2'| \le s$, $|\mathcal{S}'| \le s$ and $|\mathcal{S}_*| \le s$, and the last inequality uses the fact that $a+b \le \sqrt{2(a^2+b^2)}$ for $a = \|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\|_2$ and $b = \|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_t^a}\|_2$. Combining the above inequalities we have

$$\|[\widetilde{\mathbf{x}}_t]_{\mathcal{S}'}\|_2 \le (\delta_s + \sqrt{2}\theta_{s,s})\|\mathbf{x}_t - \mathbf{x}_*\|_2. \tag{20}$$

Since the above inequality holds for any subset $\mathcal{S}' \subseteq \overline{\mathcal{S}}_*$ of size s, we form a particular set $\mathcal{S}'$ by including the largest s entries in absolute value of $[\widetilde{\mathbf{x}}_t]_{\overline{\mathcal{S}}_*}$. Then the smallest absolute value in $[\widetilde{\mathbf{x}}_t]_{\mathcal{S}'}$ is less than $((\delta_s+\sqrt{2}\theta_{s,s})/\sqrt{s})\|\mathbf{x}_t-\mathbf{x}_*\|_2$. If not, then $\|[\widetilde{\mathbf{x}}_t]_{\mathcal{S}'}\|_2 \ge \sqrt{s}\frac{\delta_s+\sqrt{2}\theta_{s,s}}{\sqrt{s}}\|\mathbf{x}_t - \mathbf{x}_*\|_2 = (\delta_s + \sqrt{2}\theta_{s,s})\|\mathbf{x}_t - \mathbf{x}_*\|_2$, which contradicts the result in (20). By the construction of $\mathcal{S}'$, the smallest entry (in magnitude) in $\mathcal{S}'$ is the sth largest entry (in magnitude) in $[\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)]_{\overline{\mathcal{S}}_*}$, we conclude that at most s entries in $[\widetilde{\mathbf{x}}_t]_{\overline{\mathcal{S}}_*} = [\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)]_{\overline{\mathcal{S}}_*}$ are larger than $((\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s})\|\mathbf{x}_t - \mathbf{x}_*\|_2$ in magnitude. □

As an immediate result of Proposition 1, we prove the following Corollary.

Corollary 1 *Assume the noiseless setting* $\mathbf{e} = 0$. *Let* $\mathcal{S}_t$ *be the support set of* $\mathbf{x}_t$ *and* $\mathcal{S}_*$ *be the support set of* $\mathbf{x}_*$. *If* $|\mathcal{S}_t \backslash \mathcal{S}_*| \le s$ *and* $\lambda_t \ge ((\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s})\|\mathbf{x}_t - \mathbf{x}_*\|_2$, *then* $|\mathcal{S}_{t+1} \backslash \mathcal{S}_*| \le s$ *and* $|\mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}| \le 3s$.

Proof Note that in the noiseless setting when $\mathbf{e} = 0$, the intermediate solution $\widehat{\mathbf{x}}_t$ defined in (19) is equal to

$$\widehat{\mathbf{x}}_t = \mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y}) = \mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*). \tag{21}$$

As shown in (18), $\mathbf{x}_{t+1}$ is given by

$$\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\left[\left|\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)\right| - \lambda_t\right]_+.$$

By Proposition 1, we know that there are at most s entries in $\left[\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)\right]_{\overline{\mathcal{S}}_*}$ whose absolute values are larger than $(\delta_s + \sqrt{2}\theta_{s,s})\|\mathbf{x}_t - \mathbf{x}_*\|_2/\sqrt{s}$. Therefore, $[\mathbf{x}_{t+1}]_{\overline{\mathcal{S}}_*}$ has at most

s non-zero entries by setting the value of $\lambda_t \geq ((\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s})\|\mathbf{x}_t - \mathbf{x}_*\|_2$. We conclude that $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \leq s$ and $|\mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}| \leq 3s$. □

Proposition 2 *Assume the noiseless setting* $\mathbf{e} = 0$. *Let* $\mathcal{S}_t$ *be the support set of* $\mathbf{x}_t$ *and* $\mathcal{S}_*$ *be the support set of* $\mathbf{x}_*$. *If* $|\mathcal{S}_t\backslash\mathcal{S}_*| \leq s$, $\|\mathbf{x}_t - \mathbf{x}_*\|_2 \leq \Delta_t$, *and* $\lambda_t = ((\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s})\Delta_t$, *then we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq (\delta_s + \sqrt{2}\theta_{s,s} + \delta_{3s})\Delta_t.$$

To prove the above proposition, we need the following lemma, whose proof is deferred to the "Appendix".

Lemma 1 *Let* $\mathbf{x}$ *by any* s*-sparse vector and* $\mathbf{x}_{t+1}$ *given by (18), we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}\|_2^2 \leq \lambda_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}\|_2 + |(\mathbf{x}_{t+1} - \mathbf{x})^\top(U^\top(U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}))|.$$

Proof of Proposition 2 Let $\mathcal{T} = \mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}$; by Corollary 1, we have $|\mathcal{T}| \leq 3s$. By the definition of δ_s, $\|U_{\mathcal{T}}^\top U_{\mathcal{T}} - I\|_2 \leq \delta_{3s}$ (see Fact 1). First, since $\mathbf{y} = U\mathbf{x}_*$ we have

$$(\mathbf{x}_{t+1} - \mathbf{x}_*)^\top\left(U^\top(U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*)\right) = (\mathbf{x}_{t+1} - \mathbf{x}_*)^\top(U^\top U - I)(\mathbf{x}_t - \mathbf{x}_*).$$

Due to RIP of U and $|\mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}| \leq 3s$, we have

$$|(\mathbf{x}_{t+1} - \mathbf{x}_*)^\top(U^\top U - I)(\mathbf{x}_t - \mathbf{x}_*)| \leq \delta_{3s}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2\|\mathbf{x}_t - \mathbf{x}_*\|_2.$$

Thus, by applying Lemma 1 with $\mathbf{x} = \mathbf{x}_*$, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \leq \lambda_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 + \delta_{3s}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2\|\mathbf{x}_t - \mathbf{x}_*\|_2.$$

If $\mathbf{x}_{t+1} = \mathbf{x}_*$, we are done. Otherwise, dividing by $\|\mathbf{x}_{t+1} - \mathbf{x}_*\|$ we get

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq \lambda_t\sqrt{s} + \delta_{3s}\|\mathbf{x}_t - \mathbf{x}_*\|_2.$$

Assuming $\|\mathbf{x}_t - \mathbf{x}_*\|_2 \leq \Delta_t$ and plugging in the value of λ_t, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq (\delta_s + \sqrt{2}\theta_{s,s} + \delta_{3s})\Delta_t.$$

□

Proof of Theorem 1 We aim to prove $\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq \gamma^t\Delta_1$ and $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \leq s$ by induction. This is true when $t = 0$ due to the initialization and the assumption $\|\mathbf{x}_1 - \mathbf{x}_*\|_2 \leq \Delta_1$. Next, assume we have $\|\mathbf{x}_t - \mathbf{x}_*\|_2 \leq \gamma^{t-1}\Delta_1$ and $|\mathcal{S}_t\backslash\mathcal{S}_*| \leq s$ for any $t \geq 1$. We prove that it also holds for $t + 1$. By the definition of Δ_t, we have $\Delta_t = \gamma^{t-1}\Delta_1$. Thus $\|\mathbf{x}_t - \mathbf{x}_*\|_2 \leq \Delta_t$. By the value of λ_t, we have $\lambda_t = ((\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s})\Delta_t \geq ((\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s})\|\mathbf{x}_t - \mathbf{x}_*\|_2$. Hence, the condition in Corollary 1 hold, and as a result $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \leq s$. From Proposition 2, we also have $\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \leq (\delta_s + \theta_{s,s} + \delta_{3s})\Delta_t = \gamma\Delta_t = \gamma^t\Delta_1$. □

3.4 Proof of Theorem 2

The logic for proving Theorem 2 is similar to proving Theorem 1.

Corollary 2 *Let* S_t *be the support set of* $\mathbf{x}_t$ *and* $\mathcal{S}_*$ *be the support set of* $\mathbf{x}_*$. *If* $|\mathcal{S}_t\backslash\mathcal{S}_*| \leq s$ *and* $\lambda_t \geq \|U^\top\mathbf{e}\|_\infty + ((\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s})\|\mathbf{x}_t - \mathbf{x}_*\|_2$, *then* $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \leq s$ *and* $|\mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}| \leq 3s$.

Proof $\mathbf{x}_{t+1}$ is given by

$$\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\left[\left|\mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})\right| - \lambda_t\right]_+.$$

Due to $\mathbf{y} = U\mathbf{x}_* + \mathbf{e}$, we have

$$\mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y}) = \mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*) + U^\top\mathbf{e}.$$

By Proposition 1, there are at most s entries in $\left[\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)\right]_{\overline{\mathcal{S}}_*}$ with magnitude larger than $\frac{\delta_s+\sqrt{2}\theta_{s,s}}{\sqrt{s}}\|\mathbf{x}_t - \mathbf{x}_*\|_2$. As a result, $[\mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})]_{\overline{\mathcal{S}}_*}$ has at most s entries whose magnitudes larger than $\|U^\top\mathbf{e}\|_\infty + (\delta_s + \sqrt{2}\theta_{s,s})\|\mathbf{x}_t - \mathbf{x}_*\|_2/\sqrt{s}$. Therefore, given the assumed bound on λ_t, $[\mathbf{x}_{t+1}]_{\overline{\mathcal{S}}_*}$ has at most s entries larger than zero. We conclude that $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \le s$ and $|\mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}| \le 3s$. □

Proposition 3 *Let S_t be the support set of $\mathbf{x}_t$ and $\mathcal{S}_*$ be the support set of $\mathbf{x}_*$. If $|\mathcal{S}_t\backslash S_*| \le s$, $\|\mathbf{x}_t - \mathbf{x}_*\|_2 \le \Delta_t$ and $\lambda_t = \|U^\top\mathbf{e}\|_\infty + ((\delta_s + \sqrt{2}\theta_{s,s})/(\sqrt{s}))\Delta_t$, then we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \le (\delta_s + \sqrt{2}\theta_{s,s} + \delta_{3s})\Delta_t + (1+\sqrt{2})\sqrt{s}\|U^\top\mathbf{e}\|_\infty.$$

Proof Since $\mathbf{y} = U\mathbf{x}_* + \mathbf{e}$, we have

$$\begin{aligned}(\mathbf{x}_{t+1} - \mathbf{x}_*)^\top U^\top(U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*) =& (\mathbf{x}_{t+1} - \mathbf{x}_*)^\top(U^\top U - I)(\mathbf{x}_t - \mathbf{x}_*)\\ &- (\mathbf{x}_{t+1} - \mathbf{x}_*)^\top U^\top\mathbf{e}.\end{aligned}$$

Due to the restricted isometry property, we have

$$|(\mathbf{x}_{t+1} - \mathbf{x}_*)^\top(U^\top U - I)(\mathbf{x}_t - \mathbf{x}_*)| \le \delta_{3s}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2\|\mathbf{x}_t - \mathbf{x}_*\|_2,$$

and by the Cauchy-Schwartz inequality, we have

$$|(\mathbf{x}_{t+1} - \mathbf{x}_*)^\top U^\top\mathbf{e}| \le \sqrt{2s}\|U^\top\mathbf{e}\|_\infty\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2,$$

where we use the fact that $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \le s$ due to Corollary 2. Thus, by combining the two inequalities with Lemma 1 applied to $\mathbf{x} = \mathbf{x}_*$, we have

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \le& \lambda_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 + \delta_{3s}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2\|\mathbf{x}_t - \mathbf{x}_*\|_2\\ &+ \sqrt{2s}\|U^\top\mathbf{e}\|_\infty\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2.\end{aligned}$$

Then we get

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \le \lambda_t\sqrt{s} + \delta_{3s}\|\mathbf{x}_t - \mathbf{x}_*\|_2 + \sqrt{2s}\|U^\top\mathbf{e}\|_\infty.$$

Plugging in the value of λ_t, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \le (\delta_s + \sqrt{2}\theta_{s,s} + \delta_{3s})\|\mathbf{x}_t - \mathbf{x}_*\|_2 + (1+\sqrt{2})\sqrt{s}\|U^\top\mathbf{e}\|_\infty.$$

□

Proof of Theorem 2 First, we assume $\|\mathbf{x}_t - \mathbf{x}_*\|_2 \le \Delta_t$. Then by Proposition 3, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 \le \gamma\Delta_t + (1+\sqrt{2})\sqrt{s}\|U^\top\mathbf{e}\|_\infty \triangleq \Delta_{t+1}.$$

Similarly, we can use Corollary 2 to show that $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \le s$ given $|\mathcal{S}_t\backslash\mathcal{S}_*| \le s$. Since $\mathcal{S}_1 = \emptyset$ and $\|\mathbf{x}_1 - \mathbf{x}_*\| \le \Delta_1$, the claim of Part I follows by induction.

To prove Part II, let us define $\bar{\mathbf{x}}_{t+1} = (\mathbf{x}_{t+1}, \Delta_{t+1})^\top$ and $\mathcal{M}(\bar{\mathbf{x}}_t) \triangleq \bar{\mathbf{x}}_{t+1}$. Define a metric space $(\mathcal{X}, \mathcal{D})$ such that $\mathcal{X} = \left\{(\mathbf{x}, \Delta)^\top ; |\mathcal{S}(\mathbf{x})\backslash\mathcal{S}_*| \leq s, \Delta \geq \|\mathbf{x} - \mathbf{x}_*\|_2\right\}$ and $\mathcal{D}(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2 + |\Delta_1 - \Delta_2|$. The first part of this theorem implies that $\bar{\mathbf{x}}_t \in \mathcal{X}$.

We will show that $\mathcal{M}$ is a contraction mapping. Let $\mathcal{M}^1(\mathbf{x}, \Delta)$ denote the component of $\mathcal{M}$ corresponding to $\mathbf{x}$ and $\mathcal{M}^2(\mathbf{x}, \Delta)$ denote the component of $\mathcal{M}$ corresponding to Δ. Then

$$\mathcal{M}^1(\mathbf{x}, \Delta) = \arg\min_{\mathbf{u}\in\mathbb{R}^d} \underbrace{\frac{1}{2}\left\|\mathbf{u} - \left(\mathbf{x} - U^\top(U\mathbf{x} - \mathbf{y})\right)\right\|_2^2 + \left(\frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}\Delta + \|U^\top\mathbf{e}\|_\infty\right)\|\mathbf{u}\|_1}_{H_{\mathbf{x}}(\mathbf{u})}, \quad (22)$$

and $\mathcal{M}^2(\mathbf{x}, \Delta) = \gamma\Delta + (1 + \sqrt{2})\sqrt{s}\|U^\top\mathbf{e}\|_\infty$. Let $\bar{\mathbf{x}} = (\mathbf{x}, \Delta)^\top$ and $\bar{\mathbf{x}}' = \left(\mathbf{x}', \Delta'\right)^\top$ be two points from $\mathcal{X}$. Without loss of generality, let us assume $\Delta \geq \Delta'$. According to Corollary 2 we have $|\mathcal{S}(\mathcal{M}^1(\mathbf{x}, \Delta))\backslash\mathcal{S}_*| \leq s$. Similarly, since $\Delta \geq \Delta' \geq \|\mathbf{x}' - \mathbf{x}_*\|_2$, we also have $|\mathcal{S}(\mathcal{M}^1(\mathbf{x}', \Delta))\backslash\mathcal{S}_*| \leq s$. As a result, the cardinality of the support set of $\mathcal{M}^1(\mathbf{x}, \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta)$ is at most $3s$. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ denote the support set of $\mathbf{x} - \mathbf{x}'$ and $\mathcal{M}^1(\mathbf{x}, \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta)$, respectively. By the strong convexity of the problem (22), we can prove that

$$\begin{aligned}
\frac{1}{2}\|\mathcal{M}^1(\mathbf{x}, \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta)\|_2^2 &\leq H_{\mathbf{x}}(\mathcal{M}^1(\mathbf{x}', \Delta)) - H_{\mathbf{x}}(\mathcal{M}^1(\mathbf{x}, \Delta)) \\
&= H_{\mathbf{x}'}(\mathcal{M}^1(\mathbf{x}', \Delta)) - H_{\mathbf{x}'}(\mathcal{M}^1(\mathbf{x}, \Delta)) \\
&\quad - (\mathcal{M}^1(\mathbf{x}, \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta))^\top\left[\left(\mathbf{x} - U^\top(U\mathbf{x} - \mathbf{y})\right) - \left(\mathbf{x}' - U^\top(U\mathbf{x}' - \mathbf{y})\right)\right] \\
&\leq \|(\mathcal{M}^1(\mathbf{x}, \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta))_{\mathcal{T}_2}\|_2 \\
&\quad \times \left\|\left[\left(\mathbf{x} - U^\top(U\mathbf{x} - \mathbf{y})\right) - \left(\mathbf{x}' - U^\top(U\mathbf{x}' - \mathbf{y})\right)\right]_{\mathcal{T}_2}\right\|_2.
\end{aligned}$$

By the closed form of $\mathcal{M}^1(\mathbf{x}, \Delta)$ according to that similar to (18), it is not difficult to prove that $\|\mathcal{M}^1(\mathbf{x}', \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta')\|_2 \leq \frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}|\Delta - \Delta'|$. Then

$$\begin{aligned}
\|\mathcal{M}^1(\mathbf{x}, \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta')\|_2 &\leq \|\mathcal{M}^1(\mathbf{x}, \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta)\|_2 \\
&\quad + \|\mathcal{M}^1(\mathbf{x}', \Delta) - \mathcal{M}^1(\mathbf{x}', \Delta')\|_2 \\
&\leq 2\left\|\left[\left(\mathbf{x} - U^\top(U\mathbf{x} - \mathbf{y})\right) - \left(\mathbf{x}' - U^\top(U\mathbf{x}' - \mathbf{y})\right)\right]_{\mathcal{T}_2}\right\|_2 + \frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}|\Delta - \Delta'| \\
&\leq 2\|(I - U_{\mathcal{T}_1\cup\mathcal{T}_2}^\top U_{\mathcal{T}_1\cup\mathcal{T}_2})(\mathbf{x} - \mathbf{x}')_{\mathcal{T}_1\cup\mathcal{T}_2}\|_2 + \frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}|\Delta - \Delta'| \\
&\leq 2\delta_{6s}\|\mathbf{x} - \mathbf{x}'\|_2 + \frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}|\Delta - \Delta'|.
\end{aligned}$$

In addition, $|\mathcal{M}^2(\mathbf{x}, \Delta) - \mathcal{M}^2(\mathbf{x}', \Delta')| \leq \gamma|\Delta - \Delta'|$. Thus,

$$\begin{aligned}
\mathcal{D}(\mathcal{M}(\bar{\mathbf{x}}), \mathcal{M}(\bar{\mathbf{x}}')) &\leq 2\delta_{6s}\|\mathbf{x} - \mathbf{x}'\|_2 + \left(\gamma + \frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}\right)|\Delta - \Delta'| \\
&\leq \max\left(2\delta_{6s}, \gamma + \frac{\delta_s + \sqrt{2}\theta_{s,s}}{\sqrt{s}}\right)\mathcal{D}(\bar{\mathbf{x}}, \bar{\mathbf{x}}').
\end{aligned}$$

Thus, under the condition that $\max(2\delta_{6s}, \gamma + (\delta_s + \sqrt{2}\theta_{s,s})/\sqrt{s}) < 1$, $\mathcal{M}(\bar{\mathbf{x}})$ is indeed a contraction mapping. Since $(\mathbf{x}_{t+1}, \Delta_{t+1})^\top = \mathcal{M}\left((\mathbf{x}_t, \Delta_t)^\top\right)$, by Banach fixed-point theorem (Amster 2014) we have that $\mathbf{x}_{t+1}$ converges to a unique fixed point $\bar{\mathbf{x}}$.

Finally, we prove that $\bar{\mathbf{x}}$ is the solution to an ℓ_1-regularized problem with a certain regularization parameter $\bar{\lambda} = (\sqrt{2}(\delta_s + \sqrt{2}\theta_{s,s}) + 1 - \delta_{3s})\|U^\top \mathbf{e}\|_\infty/(1-\gamma)$. By the optimality condition of (17), we have

$$\mathbf{x}_{t+1} - \mathbf{x}_t + U^\top(U\mathbf{x}_t - \mathbf{y}) + \lambda_t \mathbf{p}_t = 0, \quad \text{where } \mathbf{p}_t \in \partial\|\mathbf{x}_{t+1}\|_1,$$

where $\partial\|\cdot\|_1$ denotes the subdifferential of the ℓ_1 norm function. Since $\lim_{t\to\infty} \mathbf{x}_t = \bar{\mathbf{x}}$ and $\lim_{t\to\infty} \lambda_t = \bar{\lambda}$, then there exists $\bar{\mathbf{p}}$ such that $\lim_{t\to\infty} \mathbf{p}_t = \bar{\mathbf{p}}$. Since $\|\mathbf{x}\|_1$ is a closed proper convex function, the sequence $\mathbf{x}_{t+1}$ converges to $\bar{\mathbf{x}}$ and the sequence $\mathbf{p}_t \in \partial\|\mathbf{x}_{t+1}\|_1$ converges to $\bar{\mathbf{p}}$, according to Rockafellar (1970, Theorem 24.4) we have $\bar{\mathbf{p}} \in \partial\|\bar{\mathbf{x}}\|_1$. As a result,

$$U^\top(U\bar{\mathbf{x}} - \mathbf{y}) + \bar{\lambda}\bar{\mathbf{p}} = 0, \quad \text{where } \bar{\mathbf{p}} \in \partial\|\bar{\mathbf{x}}\|_1,$$

which implies that $\bar{\mathbf{x}}$ is the optimal solution to the following problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|U\mathbf{x} - \mathbf{y}\|_2^2 + \bar{\lambda}\|\mathbf{x}\|_1.$$

□

4 Nearly-sparse signal recovery

In this section, we present algorithms and analysis for finding a sparse solution that approximates a nearly-sparse signal $\mathbf{x}_*$ with a small error.

4.1 Algorithms and main results

In order to derive a practical algorithm and a better recovery result, we assume that the random measurement matrix $U \in \mathbb{R}^{n\times d}$ contains sub-Gaussian measurements, i.e., each element U_{ij} is a sub-Gaussian random variable and has mean zero and variance $1/n$. The details of the algorithm are presented in Algorithm 2. The values of Δ_1 and Λ can be set according to our analysis. In the section, we abuse the notation $\mathcal{S}_*$ to denote the support set of $\mathbf{x}_*^s$. We first state the main theorem regarding the nearly-sparse signal recovery of Algorithm 2.

Theorem 3 *Assume that the random measurement matrix $U \in \mathbb{R}^{n\times d}$ contains sub-Gaussian measurements, i.e., each element U_{ij} is a sub-Gaussian random variable and has mean zero and variance $1/n$. For any $\tau > 0$ and some universal constant $c > 0$ define:*

$$\Lambda \triangleq \sqrt{s}\|U^\top \mathbf{e}\|_\infty + cD(\mathbf{x}_*, \mathbf{x}_*^s), \tag{23}$$

$$D(\mathbf{x}_*, \mathbf{x}_*^s) \triangleq \|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2 + \sqrt{\frac{\tau + s\ln[d/s]}{n}}\|\mathbf{x}_* - \mathbf{x}_*^s\|_2. \tag{24}$$

Assume that n is large enough such that for $n \geq (c^2(\tau + s\log[d/s]))/\eta^2$ holds for the given input parameter $\eta < \sqrt{2} - 1$ of Algorithm 2 and some $\tau > 0$ and some universal constant $c > 0$. Let $\gamma = (1+\sqrt{2})\eta$, $\{\Delta_t, t = 1, \ldots, T\}$ be a sequence such that $\Delta_1 \geq \max\left(\|\mathbf{x}_*^s\|_2, \Lambda\right)$, and

$$\Delta_{t+1} = \gamma\Delta_t + (1+\sqrt{2})\Lambda.$$

Algorithm 2 Homotopy Proximal Mapping Algorithm for Recovering a Sparse Signal (HPM1)

Input: initial size $\Delta_1 \geq \max(\|\mathbf{x}_*^s\|_2, \Lambda)$, the target sparsity s, a random measurement matrix $U \in \mathbb{R}^{d\times n}$ and measurements $\mathbf{y} \in \mathbb{R}^n$, and $0 < \eta < \sqrt{2} - 1$, where Λ is defined in (23).

1: Initialize $\mathbf{x}_1 = 0$, $\gamma = (1 + \sqrt{2})\eta$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad \lambda_t = (\Lambda + \eta\Delta_t)/\sqrt{s}$
4: $\quad \widehat{\mathbf{x}}_{t+1} = \mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})$
5: $\quad \mathbf{x}_{t+1} = \text{sign}(\widehat{\mathbf{x}}_{t+1})\left[\widehat{\mathbf{x}}_{t+1} - \lambda_t\right]_+$
6: $\quad \Delta_{t+1} = \gamma\Delta_t + (1 + \sqrt{2})\Lambda$
7: **end for**

Return $\mathbf{x}_{T+1}$

For any $t \geq 0$, with probability $1-2te^{-\tau}$, the iterates $\{\mathbf{x}_1, \ldots, \mathbf{x}_{t+1}\}$ generated by Algorithm 2 satisfy

$$|\mathcal{S}_{i+1}\backslash\mathcal{S}_*| \leq s, \quad \|\mathbf{x}_{i+1} - \mathbf{x}_*^s\|_2 \leq \Delta_{i+1}, \forall i \in \{0, \ldots, t\}.$$

In particular, let T_0 be the smallest value such that

$$\gamma^{T_0}\Delta_1 \leq \frac{\Lambda}{1-\gamma}.$$

We run Algorithm 2 with T_0 iterations and denote by $\bar{\mathbf{x}}$ the output solution. with probability $1 - 2T_0e^{-\tau}$, we have

$$\|\bar{\mathbf{x}} - \mathbf{x}_*^s\|_2 \leq \frac{\sqrt{2}(1+\sqrt{2})}{1-\gamma}\Lambda. \tag{25}$$

Remark 5 First, we note that the above result is meaningful when $\Lambda \leq \|\mathbf{x}_*^s\|_2$, otherwise a zero vector would recover $\mathbf{x}_*^s$ with an error less than Λ. Second, we note that the final solution returned by Algorithm 2 is at most $2s$-sparse. We can also take the s-largest element in $\bar{\mathbf{x}}$ to form an s-sparse approximation. Proposition 5 in the "Appendix" guarantees that the error $\|\bar{\mathbf{x}}^s - \mathbf{x}_*^s\|_2$ is only amplified by a constant factor of $\sqrt{3}$.

Remark 6 It can be seen that when $\mathbf{x}_* = \mathbf{x}_*^s$, i.e., the signal is sparse, the problem boils down to sparse signal recovery with noisy observations and the result in Theorem 3 is similar to Theorem 2 except that the RIP constants are replaced with a quantity dependent on n since we directly bound RIP constants of a sub-Gaussian matrix. Further, when $\mathbf{e} = 0$, then we can set $\Lambda = 0$ in Algorithm 2 and the result in Theorem 3 is similar to that in Theorem 1 for sparse signal recovery under noiseless observations.

Remark 7 The result in Theorem 3 also implies that more observations (i.e., larger n) may lead to more accurate recovery and fast convergence. Also, we note that the key property of the measurement matrix U is that it satisfies the JL lemma with a high probability. Therefore, any JL transforms can be used, including sparse JL transforms based on random hashing (Dasgupta et al. 2010; Kane and Nelson 2014), which can speed up the computation.

One issue of Algorithm 2 is that it needs to estimate $\|U^\top\mathbf{e}\|_\infty$ and $cD(\mathbf{x}_*, \mathbf{x}_*^s)$ for setting λ_t and for stopping the algorithm, which could be difficult in many circumstances. In addition, an overestimated Λ could increase the number of iterations and the recovery error. To alleviate this issue, below we present a more practical algorithm for nearly sparse signal recovery which

Algorithm 3 Homotopy Proximal Mapping Algorithm for Recovering a Sparse Signal (HPM2)

Input: the target sparsity s, a random measurement matrix $U \in \mathbb{R}^{d\times n}$ and measurements $\mathbf{y} \in \mathbb{R}^n$ and $\eta > 0$, and the total number of iterations T.

Initialize $\mathbf{x}_1 = 0$, $\gamma = 2(1+\sqrt{2})\eta$, and $\lambda_1 = 2\eta\Delta_1/\sqrt{s}$
for $t = 1, 2, \ldots, T$ **do**
  $\widehat{\mathbf{x}}_{t+1} = \mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})$
  $\mathbf{x}_{t+1} = \text{sign}(\widehat{\mathbf{x}}_{t+1})\left[\widehat{\mathbf{x}}_{t+1} - \lambda_t\right]_+$
  $\lambda_{t+1} = \gamma\lambda_t$
  if $\|\mathbf{x}_{t+1}\|_0 > 2s$ **then**
    Set $\widehat{\mathbf{x}} = \mathbf{x}_t$
    Break
  end if
end for
Return $\widehat{\mathbf{x}}$

performs better in absence of prior knowledge. The key idea is motivated by Theorem 3. At earlier stages of Algorithm 2, we would expect that $\Lambda \le O(\Delta_t)$ and therefore we can absorb Λ into Δ_t for setting λ_t. For stopping the algorithm, we note that as long as $|\mathcal{S}_{t+1}| \le 2s$, we can have the recovery error bounded by Δ_{t+1} (Theorem 5) or $O(\Lambda)$ (Theorem 6); therefore we stop the algorithm when $|\mathcal{S}_{t+1}| > 2s$. The detailed steps of the practical algorithm are presented in Algorithm 3. The recovery error of Algorithm 3 is provided by the following theorem.

Theorem 4 *Suppose the same random assumption on U holds as in Theorem 3. Let $\Delta_1 \ge \|\mathbf{x}_*^s\|_2$ be a constant. Let $\widehat{\mathbf{x}}$ be the solution output from Algorithm 3 and T is the maximum number of iteration allowed. Assume*

$$c\sqrt{\frac{\tau + s\log(d/s)}{n}} \le \eta \le \frac{1}{2(1+\sqrt{3})}$$

Then, with probability at least $1 - 6Te^{-\tau}$, we have

$$\|\widehat{\mathbf{x}} - \mathbf{x}_*^s\|_2 \le \max\left(\frac{\Lambda}{\eta}, \gamma^T\Delta_1\right)$$

where $\gamma = 2(1+\sqrt{2})\eta < 1$, $\Lambda = \sqrt{s}\|U^\top\mathbf{e}\|_\infty + cD(\mathbf{x}_*, \mathbf{x}_*^s)$, $D(\mathbf{x}_*, \mathbf{x}_*^s)$ is defined in Theorem 3 and c is some universal constant.

Remark 8 Although in Algorithm 3 we still use an estimate $\Delta_1 \ge \|\mathbf{x}_*^s\|_2$ for setting the initial value of λ, in practice we can set it to a sufficiently large value (e.g., $\|U^\top\mathbf{y}\|_\infty$) such that $\mathbf{x}_2 = 0$.

Remark 9 The universal constant c in Theorem 4 should not be treated literally the same as in Theorem 3. In numerical simulations, we observe that Algorithm 3 is more robust to smaller values of η than Algorithm 2.

Remark 10 Theorem 4 reveals a tradeoff in setting the value of η. A smaller value of η will lead to faster convergence but larger recovery error.

4.2 Proof of Theorem 3

The proof of Theorem 3 will be presented at the end of this subsection after a series of results. We first give the following lemma.

Lemma 2 *Assume $U \in \mathbb{R}^{n\times d}$ is a sub-Gaussian measurement matrix, where each element in U has zero mean and variance $1/n$. If $|\mathcal{S}_t \backslash \mathcal{S}_*| \leq s$, then with probability $1-2e^{-\tau}$, we have*

$$\left\| (U^\top (U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*^s))^s \right\|_2 \leq \sqrt{s}\|U^\top \mathbf{e}\|_\infty + cD(\mathbf{x}_*, \mathbf{x}_*^s) + c\sqrt{\frac{\tau + s\log[d/s]}{n}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2,$$

where $D(\mathbf{x}_*, \mathbf{x}_*^s)$ is defined in Theorem 3 and c is some universal constant.

Lemma 2 is proved in the "Appendix". Following Lemma 2, we prove the following corollary.

Corollary 3 *Let $\mathcal{S}_t$ and $\mathcal{S}_{t+1}$ be the support sets of $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$, respectively. If $|\mathcal{S}_t \backslash \mathcal{S}_*| \leq s$, then with probability $1-2e^{-\tau}$, we have*

$$|\mathcal{S}_{t+1} \backslash \mathcal{S}_*| \leq s,$$

provided that

$$\lambda_t \geq \|U^\top \mathbf{e}\|_\infty + \frac{cD(\mathbf{x}_*, \mathbf{x}_*^s)}{\sqrt{s}} + \frac{c}{\sqrt{s}}\sqrt{\frac{\tau + s\log[d/s]}{n}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2. \tag{26}$$

Proof The proof is similar to that of Corollary 2. Recall that $\mathbf{x}^s$ denotes the vector $\mathbf{x}$ with all but the s largest entries (in magnitude) set to zero, and $\mathcal{S}_*$ denotes the support of the s-largest entries in $\mathbf{x}_*$. From Lemma 2, we can conclude that $[\mathbf{x}_t - U^\top (U\mathbf{x}_t - \mathbf{y})]_{\overline{\mathcal{S}}_*}$ has at most s entries with magnitude larger than the quantity in the right hand side of (26). This can be verified by contradiction. Suppose there exists $\mathcal{A} \subseteq \overline{\mathcal{S}}_*$ such that $|\mathcal{A}| > s$ and for all $i \in \mathcal{A}$,

$$[\mathbf{x}_t - U^\top (U\mathbf{x}_t - \mathbf{y})]_i \geq \|U^\top \mathbf{e}\|_\infty + \frac{cD(\mathbf{x}_*, \mathbf{x}_*^s)}{\sqrt{s}} + \frac{c}{\sqrt{s}}\sqrt{\frac{\tau + s\log[d/s]}{n}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2.$$

Let $\mathcal{A}_s \subseteq \mathcal{A}$ such that $|\mathcal{A}_s| = s$. Then

$$\|[\mathbf{x}_t - \mathbf{x}_*^s - U^\top (U\mathbf{x}_t - \mathbf{y})]_{\mathcal{A}_s}\|_2 \geq \sqrt{s}\|U^\top \mathbf{e}\|_\infty + cD(\mathbf{x}_*, \mathbf{x}_*^s) + c\sqrt{\frac{\tau + s\log[d/s]}{n}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2,$$

where we use the fact that $[\mathbf{x}_*^s]_{\mathcal{A}_s} = 0$. However, the above inequality contradicts Lemma 2. Since $\mathbf{x}_{t+1}$ is given by

$$\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\left[\left|\mathbf{x}_t - U^\top (U\mathbf{x}_t - \mathbf{y})\right| - \lambda_t\right]_+,$$

therefore, given the assumed bound on λ_t, $[\mathbf{x}_{t+1}]_{\overline{\mathcal{S}}_*}$ has at most s entries larger than zero. We conclude that $|\mathcal{S}_{t+1} \backslash \mathcal{S}_*| \leq s$. □

Based on the above corollary, we can prove the following proposition that serves as the key to proving the main theorem.

Proposition 4 *Assume* $|\mathcal{S}_t \backslash \mathcal{S}_*| \leq s$, $\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq \Delta_t$, *and define*

$$\Lambda \triangleq \sqrt{s}\|U^\top \mathbf{e}\|_\infty + cD(\mathbf{x}_*, \mathbf{x}_*^s). \tag{27}$$

Let $\lambda_t = \dfrac{\Lambda + \eta\Delta_t}{\sqrt{s}}$. *Then, with probability* $1 - 2e^{-\tau}$, *we have*

$$|\mathcal{S}_{t+1} \backslash \mathcal{S}_*| \leq s, \quad \|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq \Delta_{t+1} \triangleq (1+\sqrt{2})\eta\Delta_t + (1+\sqrt{2})\Lambda,$$

provided

$$c\sqrt{\frac{\tau + s\log[d/s]}{n}} \leq \eta.$$

Proof It is easy to verify that the condition for λ_t in Corollary 3 is satisfied. Combining with the fact that $\mathbf{x}_t$ is a $2s$-sparse vector, we have $|\mathcal{S}_{t+1} \backslash \mathcal{S}_*| \leq s$ due to Corollary 3. Applying Lemma 1 with $\mathbf{x} = \mathbf{x}_*^s$, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2^2 \leq \lambda_t\sqrt{s}\left\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\right\|_2 + \left|(\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top (U^\top (U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*^s))\right|. \tag{28}$$

According to Lemma 2, with probability $1 - 2e^{-\tau}$, we have

$$\left\|\left(U^\top (U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*^s)\right)^s\right\|_2 \leq \Lambda + \eta\Delta_t. \tag{29}$$

Thus,

$$\begin{aligned}
&\left|(\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top \left(U\left(U^\top \mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right)\right| \\
&\leq \left|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_*}^\top \left[U\left(U^\top \mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]_{\mathcal{S}_*}\right| \\
&\quad + \left|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_{t+1}\backslash\mathcal{S}_*}^\top \left[U\left(U^\top \mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]_{\mathcal{S}_{t+1}\backslash\mathcal{S}_*}\right| \\
&\leq \left(\left\|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_*}\right\|_2 + \left\|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_{t+1}\backslash\mathcal{S}_*}\right\|_2\right)(\Lambda + \eta\Delta_t) \\
&\leq \sqrt{2}(\Lambda + \eta\Delta_t)\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2,
\end{aligned}$$

where we use the fact that $|\mathcal{S}_*| \leq s$ and $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \leq s$, and inequality (29), and the last inequality uses the fact that $a+b \leq \sqrt{2(a^2+b^2)}$. Combining the above inequality with (28), with probability $1 - 2e^{-\tau}$, we have

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2^2 &\leq \left(\lambda_t\sqrt{s} + \sqrt{2}\eta\Delta_t + \sqrt{2}\Lambda\right)\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \\
&\leq [(1+\sqrt{2})\eta\Delta_t + (1+\sqrt{2})\Lambda]\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2.
\end{aligned}$$

Therefore,

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq (1+\sqrt{2})\eta\Delta_t + (1+\sqrt{2})\Lambda. \tag{30}$$

□

Proof of Theorem 3 Following Proposition 4 and by induction, we can prove that for any t, $|\mathcal{S}_{t+1}\backslash\mathcal{S}_*| \leq s$ and $\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq \Delta_{t+1}$ hold with probability $1 - 2te^{-\tau}$. Since $\Delta_{t+1} = \gamma\Delta_t + (1+\sqrt{2})\Lambda$, we have

$$\Delta_{t+1} \leq \gamma^t\Delta_1 + \frac{1-\gamma^t}{1-\gamma}(1+\sqrt{2})\Lambda \leq \gamma^t\Delta_1 + \frac{1}{1-\gamma}(1+\sqrt{2})\Lambda.$$

Letting $t = T_0$ such that $\gamma^{T_0}\Delta_1 \leq \Lambda/(1-\gamma)$, we then have

$$\|\mathbf{x}_{T_0+1} - \mathbf{x}_*^s\|_2 \leq \frac{\sqrt{2}(1+\sqrt{2})\Lambda}{1-\gamma}$$

with probability $1 - 2T_0 e^{-\tau}$, which completes the proof of Theorem 3. □

4.3 Proof of Theorem 4

We first state two theorems that are central to our analysis. Theorem 5 reveals that the recovery error of Algorithm 3 will decrease by a constant factor at the beginning, and Theorem 6 shows that the recovery error will stay small in the later stages.

Theorem 5 *Let $\Delta_1 \geq \|\mathbf{x}_*^s\|_2$ be a constant, $\gamma = 2(1+\sqrt{2})\eta$, and $\{\mathbf{x}_t, t = 1, \ldots\}$ be iterates generated by Algorithm 3. Assume $|\mathcal{S}_t \backslash \mathcal{S}_*| \leq s$, $\|\mathbf{x}_t - \mathbf{x}_*^s\| \leq \Delta_t$, and $\Lambda \leq \eta\Delta_t$. Then, with probability at least $1 - 2e^{-\tau}$, we have*

$$|\mathcal{S}_{t+1} \backslash \mathcal{S}_*| \leq s \text{ and } \|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq \Delta_{t+1} \triangleq \gamma\Delta_t,$$

provided the condition in Theorem 4 is true.

Proof The proof is very similar to that of Proposition 4 by noting that $\lambda_t = 2\eta\Delta_t/\sqrt{s} > (\Lambda + \eta\Delta_t)/\sqrt{s}$ satisfies the condition in Corollary 3. Then we can bound (30) by $2(1+\sqrt{2})\eta\Delta_t = \gamma\Delta_t$, which finishes the proof. □

Theorem 6 *Let $\{\mathbf{x}_t, t = 1, \ldots\}$ be iterates generated by Algorithm 3. Assume $|\mathcal{S}_t| \leq 2s$, $\|\mathbf{x}_t - \mathbf{x}_*^s\| \leq \Lambda/\eta$, and $\Lambda > \eta\Delta_t$. If $|\mathcal{S}_{t+1}| \leq 2s$, then with probability at least $1 - 2e^{-\tau}$, we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq 2(1+\sqrt{3})\Lambda \leq \Lambda/\eta,$$

provided the condition in Theorem 4 is true.

Proof First we note that $\mathbf{x}_t - \mathbf{x}_*^s$ is at most $3s$-sparse. With a slight change of the universal constant, we still have Lemma 2 (cf. the proof of Lemma 2 in the "Appendix"). Then, with probability at least $1 - 2e^{-\tau}$, we have

$$\begin{aligned}\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 &\leq \Lambda + c\sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \\ &\leq \Lambda + \eta\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq 2\Lambda.\end{aligned}$$

Notice that $\mathbf{x}_{t+1} - \mathbf{x}_*^s$ is $3s$-sparse in this case, and we can verify that

$$\left|(\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top\left(U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right)\right| \leq 2\sqrt{3}\Lambda\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2.$$

To see this, we can split $\mathbf{x}_{t+1} - \mathbf{x}_*^s = \mathbf{a} + \mathbf{b} + \mathbf{c}$ into three components, each with at most s non-zero entries and non-overlapping support. Then

$$\begin{aligned}&\left|(\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top\left(U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right)\right| \\ &\leq (\|\mathbf{a}\|_2 + \|\mathbf{b}\|_2 + \|\mathbf{c}\|_2)\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 \\ &\leq (\|\mathbf{a}\|_2 + \|\mathbf{b}\|_2 + \|\mathbf{c}\|_2)2\Lambda \leq 2\sqrt{3}\Lambda\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2.\end{aligned}$$

where we use the fact that $a+b+c \leq \sqrt{3(a^2+b^2+c^2)}$. Applying Lemma 1 with $\mathbf{x} = \mathbf{x}_*^s$, we have, with probability at least $1-2e^{-\tau}$,

$$\begin{aligned}\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2^2 &\leq \lambda_t\sqrt{s}\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2 + |(\mathbf{x}_{t+1}-\mathbf{x}_*^s)^\top(U^\top(U\mathbf{x}_t-\mathbf{y})-(\mathbf{x}_t-\mathbf{x}_*^s))| \\ &\leq \left(\lambda_t\sqrt{s}+2\sqrt{3}\Lambda\right)\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2 \leq 2(1+\sqrt{3})\Lambda\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2,\end{aligned}$$

where we use the fact that $\eta = \frac{2\eta\Delta_t}{\sqrt{s}} \leq \frac{2\Lambda}{\sqrt{t}}$ due to the assumption $\Lambda > \eta\Delta_t$. Thus,

$$\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2 \leq 2(1+\sqrt{3})\Lambda.$$

□

Proof of Theorem 4 Let $k = \min\{t : \Lambda > \eta\Delta_t\}$. Assume $k \geq 1$, otherwise Theorem 4 holds with $T = 1$. In the following, we consider the two cases $T < k$ and $T \geq k$, where T is the input to the algorithm.

$T < k$: Since the condition $\Lambda \leq \eta\Delta_t$ holds for $t = 1, \ldots, T$, we can apply Theorem 5 to bound the recovery error in each iteration. Thus, with probability at least $1-2Te^{-\tau}$, we have

$$\|\widehat{\mathbf{x}}-\mathbf{x}_*^s\|_2 = \|\mathbf{x}_{T+1}-\mathbf{x}_*^s\|_2 \leq \Delta_{T+1} = \gamma^T\Delta_1.$$

$T \geq k$: From the above analysis, with probability at least $1-2(k-1)e^{-\tau}$, we have $\|\mathbf{x}_k - \mathbf{x}_*^s\|_2 \leq \Delta_k$ and $|\mathcal{S}_k \backslash \mathcal{S}_*| \leq s$, which also means our algorithm arrives at the kth iteration (i.e., it does not terminate before kth iteration). In the kth iteration, one of the two cases holds: $|\mathcal{S}_{k+1}| > 2s$ and $|\mathcal{S}_{k+1}| \leq 2s$. In the first case, our algorithm terminates and returns $\mathbf{x}_k$ as the final solution, implying

$$\|\widehat{\mathbf{x}}-\mathbf{x}_*^s\|_2 = \|\mathbf{x}_k-\mathbf{x}_*^s\|_2 \leq \Delta_k \leq \Lambda/\eta.$$

In the second case, Algorithm 3 keeps running, and we can bound the recovery error by Theorem 6. In particular, if at $T' \geq k$, $|\mathcal{S}_{T'+1}| > 2s$, our algorithm terminates and returns $\mathbf{x}_{T'}$ as the final solution, which implies $|\mathcal{S}_t| < 2s$ and $t \leq T'$. Thus, by applying induction of Theorem 6 from $t = k$, we have

$$\|\widehat{\mathbf{x}}-\mathbf{x}_*^s\|_2 = \|\mathbf{x}_{T'}-\mathbf{x}_*^s\|_2 \leq 2(1+\sqrt{3})\Lambda \leq \Lambda/\eta.$$

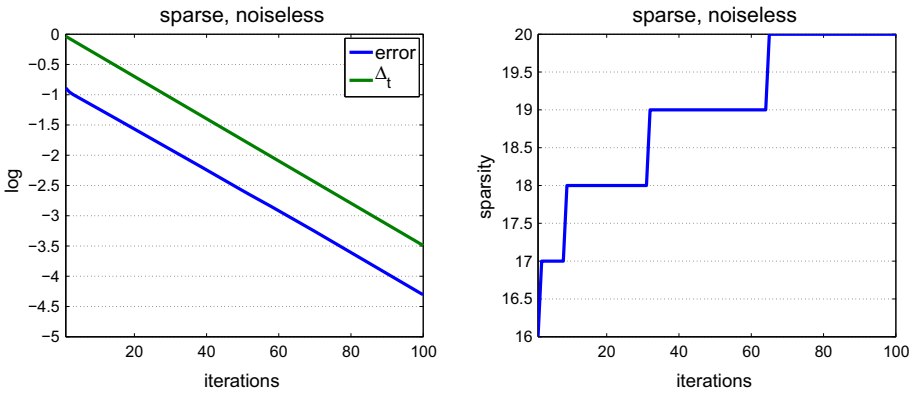
□

5 Numerical simulations

In this section, we conduct numerical simulations to verify the proposed algorithms and the developed analysis, and also compare with previous algorithms. We implemented all involved algorithms using Matlab R2017a and evaluate them in the same computing environment—a linux machine with 3.10 GHz CPU and 264 G memory. All algorithms were run using single computational thread option on Matlab.

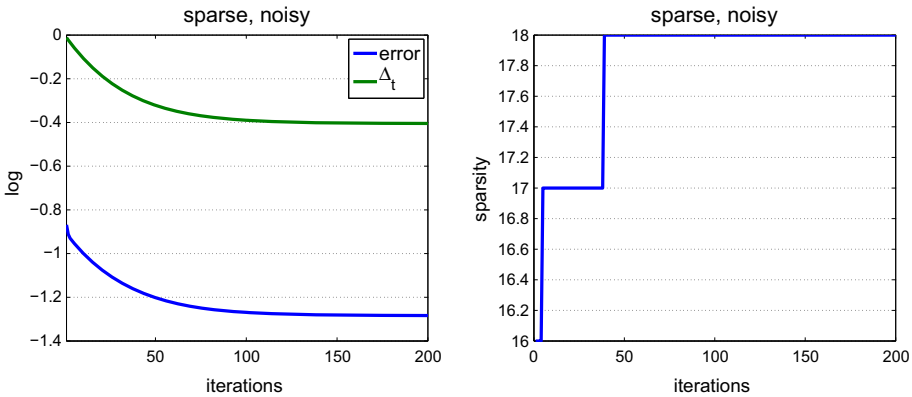
5.1 Verifying Theorem 3

We first conduct some empirical studies to verify the results in Theorem 3. We generate a measurement matrix $U \in \mathbb{R}^{n\times d}$ such that each element follows an i.i.d distribution $\mathcal{N}(0, 1/n)$. To generate an s-sparse target signal, we sample the non-zeros elements from the standard



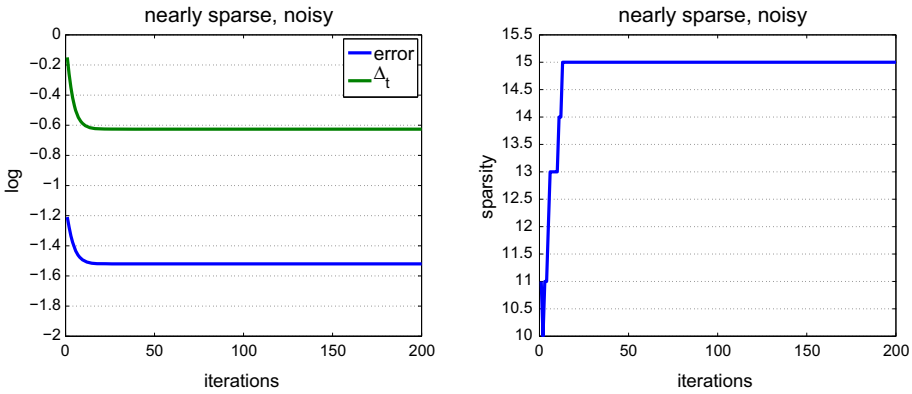
(a) recovery error **(b)** sparsity

Fig. 1 Algorithm 2: recovery error and sparsity versus iterations in setting I



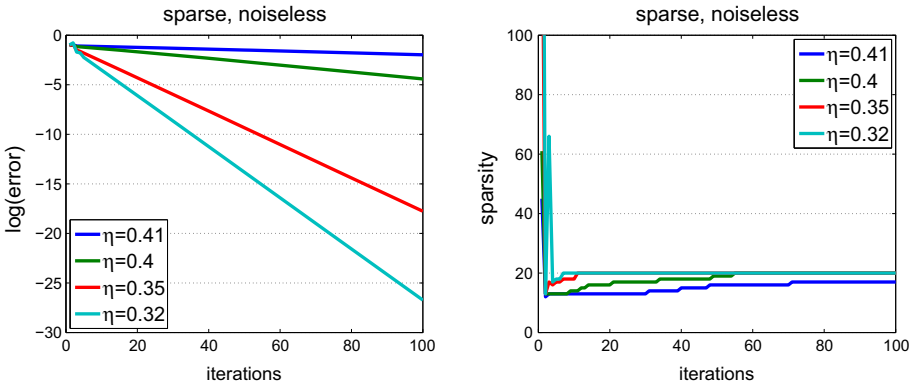
(a) recovery error **(b)** sparsity

Fig. 2 Algorithm 2: recovery error and sparsity versus iterations in setting II



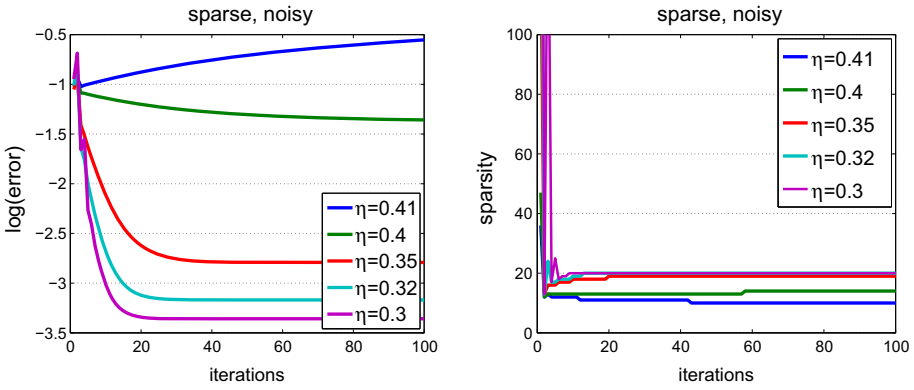
(a) recovery error **(b)** sparsity

Fig. 3 Algorithm 2: recovery error and sparsity versus iterations in setting III



(a) error **(b)** sparsity

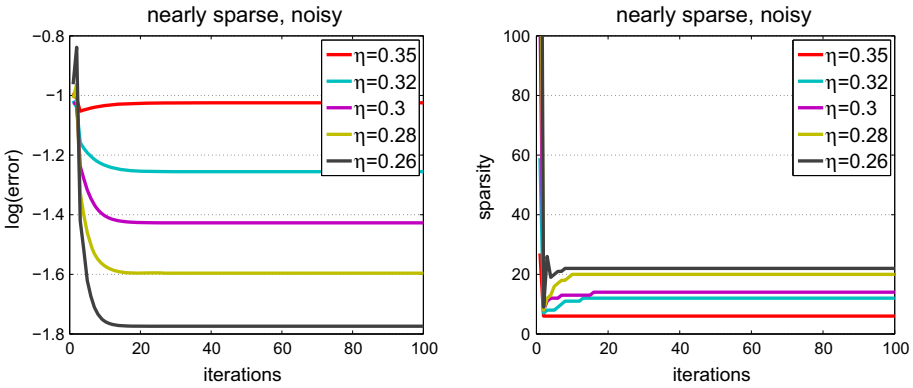
Fig. 4 Algorithm 2: recovery error and sparsity versus iterations in setting I for different values of η



(a) error **(b)** sparsity

Fig. 5 Algorithm 2: recovery error and sparsity versus iterations in setting II for different values of η

normal distribution followed by ℓ_2 norm normalization. To generate a nearly sparse target signal, we set the ith element of $\mathbf{x}_*$ to i^{-1} followed by ℓ_2 norm normalization. The noise vector $\mathbf{e}$ is drawn from uniform distribution $[-\sigma, \sigma]$. We run Algorithm 2 with hundreds of iterations and plot recovery error ($\|\mathbf{x}_t - \mathbf{x}_*^s\|_2$) and sparsity versus the number of iterations in Figs. 1, 2 and 3 for $n = 2000, d = 10000, s = 20, \sigma = 0.001$ under three different settings, respectively. The value of Δ_1 is set to $\|\mathbf{x}_*^s\|_2$, the value of η is set to 0.4, 0.4, 0.3 for three different settings, respectively, and the value of Λ is set to 0, $\sqrt{s}\|U^\top \mathbf{e}\|_\infty$ and $\sqrt{s}\|U^\top \mathbf{e}\|_\infty + \eta\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$, respectively. The curves of recovery error in Figs. 1a, 2a and 3a demonstrate that the recovery error is upper bounded by Δ_t, which clearly validates the recovery error bounds in Theorem 3. The curves of sparsity level in Figs. 1b, 2b and 3b demonstrate that the number of non-zero elements of all iterates $\mathbf{x}_t$ does not exceed $s = 20$, which are consistent with the result about support sizes of $\mathbf{x}_t$ in Theorem 3.



(a) error **(b)** sparsity

Fig. 6 Algorithm 2: recovery error and sparsity versus iterations in setting III for different values of η

5.2 Varying η

We conduct more experiments to demonstrate that the robustness of the proposed HPM algorithm (Algorithm 2) with respect to the value of η. The data is generated similarly as before for the three settings with $n = 1000, d = 10000, s = 20, \sigma = 0.001$. The results are shown in Figs. 4, 5 and 6 for different values of η, not exceeding its upper limit $\sqrt{2} - 1 \approx 0.414$. The smallest value of η in each figure is the smallest one[5] that guarantees convergence. From these results, we have several interesting observations: (i) from noisy to noiseless observations and from sparse signal to nearly sparse signal, the algorithm becomes more robust for smaller values of η and less robust for larger values of η. For example in setting I, the smallest value of η that guarantees convergence is 0.32, but when adding some noise to the observations, the smallest value of η reduces to 0.3. However, the value $\eta = 0.41$, which originally works for noiseless observations, will cause the algorithm to diverge in setting II. (ii) As long as convergence is observed, a smaller value of η yields faster convergence in all cases and more accurate recovery in settings II and III. (iii) Even though the sparsity of intermediate solutions exceeds $2s$, the algorithm still converges.

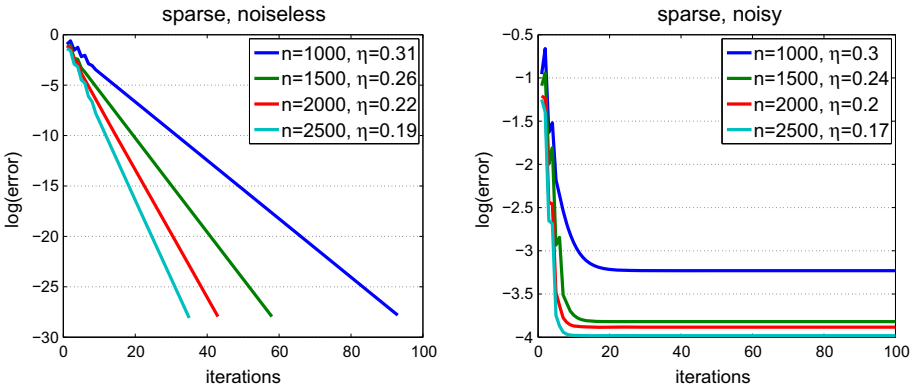
5.3 Varying n

We also verify that more observations lead to faster convergence and more accurate recovery. To this end, we generate data similarly as before with different values of $n = 1000, 1500, 2000, 2500$. For each value of n, we choose the smallest η that can guarantee the convergence. The results for the first two settings are shown in Fig. 7, which clearly demonstrate that the with more observations, we can use smaller η to get a faster convergence and a more accurate recovery in setting II. Similar result has been observed for setting III.

5.4 HPM1 versus HPM2

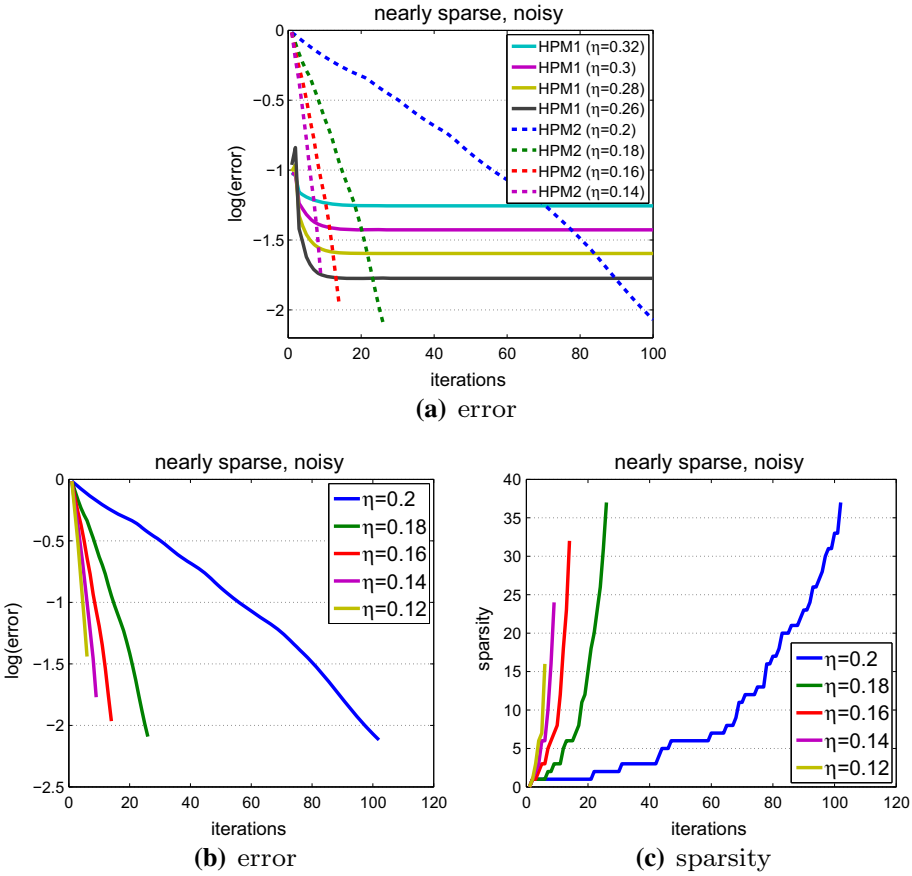
We also compare HPM2 with HPM1 in setting III to demonstrate the benefit of HPM2. The data is generated similarly as before with $n = 1000, d = 10{,}000, s = 20, \sigma = 0.001$. The

[5] We start a value of $\eta = 0.41$ and decrease by 0.01 until we observe divergence.



(a) setting I: different n **(b)** setting II: different n

Fig. 7 Algorithm 2: recovery error versus different n. The value of η is chosen as the best one for each value of n



(a) error **(b)** error **(c)** sparsity

Fig. 8 **a** HPM1 versus HPM2 with different values of η in setting III. **b**, **c** Recovery error and sparsity of solutions in HPM2 versus iterations in setting III for different values of η resp

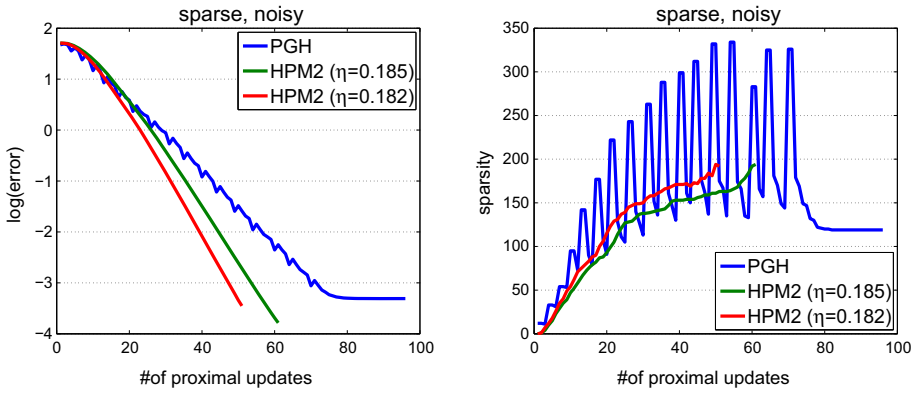


Fig. 9 HPM2 versus PGH. Two different values of η are used in HPM2. PGH requires 96 proximal updates and HPM2 requires 51 and 61 proximal updates with $\eta = 0.182$ and $\eta = 0.185$, respectively. The error of the final solution returned by PGH is 0.0365, and the error of the final solution returned by HPM2 with the two different values of η is 0.0317, and 0.0227, respectively. The recovery error of the 100-sparse solution formed by taking the top 100 elements in the returned solution by HPM2 is 0.0312 and 0.0223, and that by PGH is 0.0362

result is shown in Fig. 8a. The initial value of λ in HPM2 is set to $\|U^\top \mathbf{y}\|_\infty$. It shows that HPM2 with an appropriate value of η can achieve similar convergence speed and even more accurate recovery than HPM1. We also plot the recovery error and sparsity of intermediate solutions for HPM2 in Fig. 8b, c. The curves exhibit a tradeoff in setting the value of η, namely a smaller value of η leads to a faster convergence but a worse recovery, which is consistent with Theorem 4.

5.5 Comparing with proximal-gradient homotopy method (PGH)

We compare HPM2 with the PGH method that solves the BPDN problem for sparse signal recovery (Xiao and Zhang 2013). The data is generated exactly the same as in Xiao and Zhang (2013). In particular, we generate a random measurement matrix $U \in \mathbb{R}^{n\times d}$ with $n = 1000$ and $d = 5000$. The entries of the matrix U are generated independently with the uniform distribution over the interval $[-1, +1]$ and are scaled to have a variance $1/n$. The vector $\mathbf{x}_* \in \mathbb{R}^d$ is generated with the same distribution at 100 randomly chosen coordinates (i.e., $\mathcal{S}_* = 100$). The noise $\mathbf{e} \in \mathbb{R}^n$ is a dense vector with independent random entries with the uniform distribution over the interval $[-\sigma, \sigma]$, where σ is the noise magnitude and is set to 0.01. Finally the vector $\mathbf{y}$ was obtained as $\mathbf{y} = U\mathbf{x}_* + \mathbf{e}$. The target value of λ in PGH is chosen to be $\lambda_{\text{target}} = 1$ according to Xiao and Zhang (2013). The parameters in PGH (e.g., γ_{inc}, γ_{dec}, η, δ) are exactly the same as those used in Xiao and Zhang (2013). The initial value of λ for both PGH and HPM2 is set to $\|U^\top \mathbf{y}\|_\infty$. We plot the recovery error and sparsity of generated solutions versus the number of proximal updates in Fig. 9. We can see that HPM2 achieves faster convergence and better recovery than PGH for sparse signal recovery.

5.6 Comparing with iterative soft thresholding algorithm (ISTA) and iterative hard thresholding (IHT)

Finally, we compare HPM2 with two other algorithms, namely ISTA and IHT (Garg and Khandekar 2009). The measurement matrix U and the noise vector $\mathbf{e}$ are generated the same

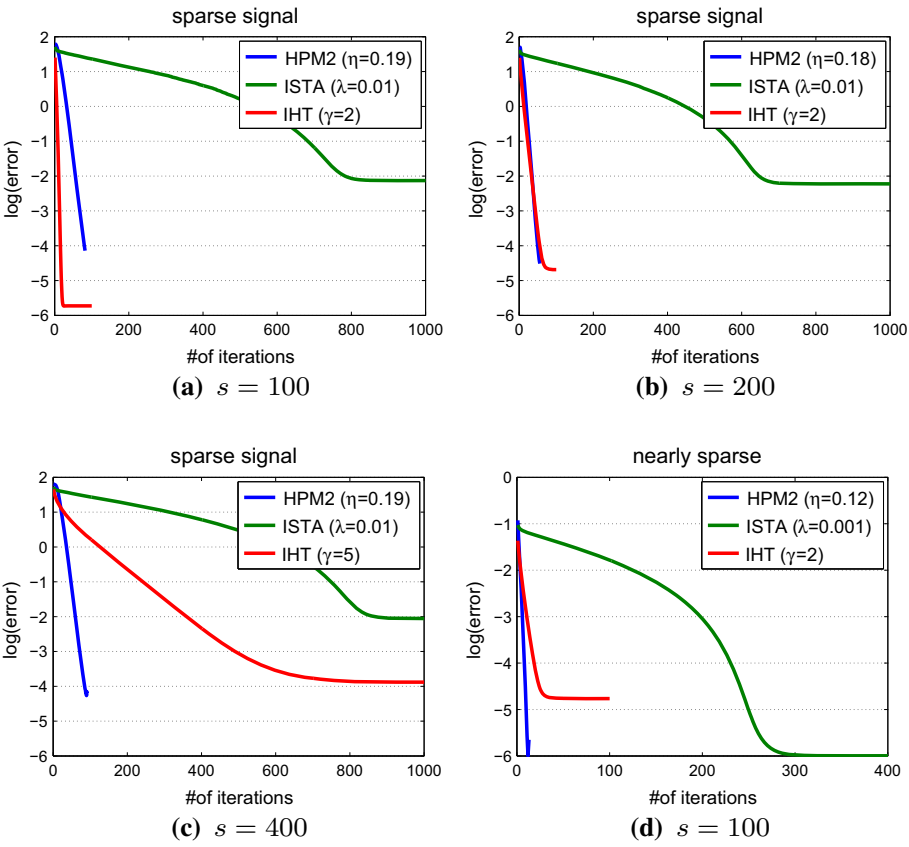


Fig. 10 HPM2 versus ISTA and IHT. From left to right: **a** $\mathbf{x}_*$ is sparse with only 100 non-zero entries and the parameter s in HPM2 and IHT is set to 100; **b** $\mathbf{x}_*$ is sparse with only 100 non-zero entries and the parameter s in HPM2 and IHT is set to 200; **c** $\mathbf{x}_*$ is sparse with only 100 non-zero entries and the parameter s in HPM2 and IHT is set to 400; **d** $\mathbf{x}_*$ is nearly sparse and the parameter s in HPM2 and IHT is set to 100

as above, i.e., $U \in \mathbb{R}^{1000\times 5000}$ and each entry is sampled from a uniform distribution over $[-1, +1]$ and is scaled to have a variance of $1/n$. For the ground-truth signal $\mathbf{x}_*$, we consider two scenarios: (i) a sparse signal with 100 randomly chosen coordinates sampled from the uniform distribution over $[-1, +1]$; (ii) a nearly sparse signal such that the entries follow an exponential decay, i.e., $[\mathbf{x}_*]_i = e^{-i}$. Since the proposed HPM2 and IHT require a parameter s that estimates the sparsity of the target signal, in the first scenario we vary s among three values $s = 100, s = 200$ and $s = 400$. In the second scenario, we fix $s = 100$. For other parameters that each algorithm relies on (e.g., η in HPM2, the step size parameter $1/\gamma$ in IHT and the regularization parameter λ in ISTA), we tune them among numerous values and report the performance of the best one. We vary the value of η in $[0.1, 0.2]$, the value of γ in $[1, 10]$ and the value of λ in $[0.001, 1]$. The recovery error measured by the difference between the top s components of the returned solution and the top s components of the ground-truth signal is plotted in Fig. 10. From the results, we observe that (i) IHT and HPM2 converge much faster than ISTA; (ii) when the ground-truth signal is sparse and the parameter s is set to the exact number of non-zeros in the ground-truth signal, IHT performs better than

HPM2; (iii) however, when the parameter s is overestimated and the ground-truth signal is not exactly sparse, the proposed algorithm HPM2 performs better than IHT, where the latter case is consistent with our comparison in Sect. 2.

6 Conclusions

In this paper, we have presented simple homotopy proximal mapping algorithms for reconstructing a sparse signal from (noisy) linear measurements of the signal. We proved a global linear convergence for the proposed homotopy proximal mapping algorithms under three different settings. For sparse signal recovery, one of the proposed algorithms with an appropriate setting of a parameter based on the RIP constants converges linearly to the optimal solution up to the noise level. For nearly sparse signal recovery with a sub-Gaussian measurement matrix, our high probability result is better than previous results for instance-level recovery in terms of the order of recovery error. In addition, we develop a practical algorithm that runs without any knowledge of noise level but requires the target sparsity level and an upper bound of the target signal's Euclidean norm. Numerical simulations verify the proposed algorithms and the established theorems. As future work, we will consider how to incorporate hard thresholding into the proposed HPM schemes and prove the convergence and recovery bounds. We will also consider how to develop more practical algorithms without relying on prior knowledge of the target sparsity and the upper bound of the target signal's Euclidean norm.

Acknowledgements We thank all reviewers for their constructive comments. Z.-H. Zhou is partially supported by National Key R&D Program of China (2018YFB1004300), and NSFC (61751306). L. Zhang is partially supported by JiangsuSF (BK20160658), and YESS (2017QNRC001). T. Yang is partially supported by NSF (1545995).

A Proof of Lemma 1

Define $L_t(\mathbf{x})$ as

$$L_t(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - U^\top(U\mathbf{x}_t - \mathbf{y})\right)\right\|_2^2 + \lambda_t\|\mathbf{x}\|_1.$$

Since $\mathbf{x}_{t+1}$ is the optimal solution to $\min_{\mathbf{x}} L_t(\mathbf{x})$, we have for any $\mathbf{x}$

$$(\mathbf{x}_{t+1} - \mathbf{x})^\top \partial L_t(\mathbf{x}_{t+1}) \leq 0,$$

where $\partial L_t(\cdot)$ denotes the subdifferential of the function $L_t(\cdot)$, i.e., there exists a $g_{t+1} \in \partial\|\mathbf{x}_{t+1}\|_1$ such that

$$(\mathbf{x}_{t+1} - \mathbf{x})^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + (\mathbf{x}_{t+1} - \mathbf{x})^\top U^\top(U\mathbf{x}_t - \mathbf{y}) + \lambda_t(\mathbf{x}_{t+1} - \mathbf{x})^\top g_{t+1} \leq 0.$$

Let $\mathbf{x}$ be an s-sparse vector with support set $\mathcal{S}$. Then we have

$$(\mathbf{x}_{t+1} - \mathbf{x})^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + (\mathbf{x}_{t+1} - \mathbf{x})^\top U^\top(U\mathbf{x}_t - \mathbf{y}) + \lambda_t\|[\mathbf{x}_{t+1}]_{\mathcal{S}_{t+1}\setminus\mathcal{S}}\|_1 \leq \lambda_t\|[\mathbf{x}_{t+1} - \mathbf{x}]_{\mathcal{S}}\|_1,$$

where we use $([\mathbf{x}_{t+1}]_{\mathcal{S}_{t+1}\setminus\mathcal{S}})^\top g_{t+1} = \|\mathbf{x}_{t+1}\|_1$ and $\|g_{t+1}\|_\infty \leq 1$. Note that

$$\begin{aligned}&(\mathbf{x}_{t+1} - \mathbf{x})^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + (\mathbf{x}_{t+1} - \mathbf{x})^\top U^\top(U\mathbf{x}_t - \mathbf{y})\\ &\quad = \|\mathbf{x}_{t+1} - \mathbf{x}\|_2^2 + (\mathbf{x}_{t+1} - \mathbf{x})^\top\left(U^\top(U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x})\right).\end{aligned}$$

We complete the proof by noting that $\lambda_t \|[\mathbf{x}_{t+1}]_{\mathcal{S}_{t+1}\setminus\mathcal{S}}\|_1 \geq 0$ and

$$\|[\mathbf{x}_{t+1} - \mathbf{x}]_{\mathcal{S}}\|_1 \leq \sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}\|_2.$$

□

B Proof of Lemma 2

We first decompose $U^\top (U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*^s)$ into 3 components:

$$\begin{aligned} & U^\top (U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*^s) \\ &= U^\top (U\mathbf{x}_t - U\mathbf{x}_* - \mathbf{e}) - (\mathbf{x}_t - \mathbf{x}_*^s) \\ &= \underbrace{U^\top U(\mathbf{x}_*^s - \mathbf{x}_*)}_{:=\mathbf{w}_a} + \underbrace{(U^\top U - I)(\mathbf{x}_t - \mathbf{x}_*^s)}_{:=\mathbf{w}_b} - \underbrace{U^\top \mathbf{e}}_{:=\mathbf{w}_c}. \end{aligned}$$

Then, we have

$$\left\| \left[U^\top (U\mathbf{x}_t - \mathbf{y}) - (\mathbf{x}_t - \mathbf{x}_*^s) \right]^s \right\|_2 \leq \|\mathbf{w}_a^s\|_2 + \|\mathbf{w}_b^s\|_2 + \|\mathbf{w}_c^s\|_2. \tag{31}$$

The last term can be bounded by $\|\mathbf{w}_c^s\|_2 \leq \sqrt{s}\|U^\top \mathbf{e}\|_\infty$. In the following analysis, we intend to bound $\|(U^\top U\mathbf{z})^s\|_2$ for a fixed vector $\mathbf{z}$, and $\left\|((UU^\top - I)\mathbf{z})^s\right\|_2$ for any sparse vector $\mathbf{z}$. We will address these two bounds in the following two subsections.

B.1 Bounding $\|(U^\top U\mathbf{z})^s\|_2$ for a fixed $\mathbf{z}$

First, we define

$$\mathcal{K}_{d,s} = \left\{ \mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq 1, \|\mathbf{w}\|_0 \leq s \right\},$$

and

$$\mathcal{E}_s(\mathbf{z}) = \max_{\mathbf{w} \in \mathcal{K}_{d,s}} \mathbf{w}^\top U^\top U\mathbf{z}.$$

It is easy to verify that

$$\|(U^\top U\mathbf{z})^s\|_2 = \mathcal{E}_s(\mathbf{z}).$$

This can be seen that the maximum in the definition of $\mathcal{E}_s(\mathbf{z})$ is achieved when the support set of $\mathbf{w}$ is the support set of the top s-elements (in magnitude) in $U^\top U\mathbf{z}$. Hence, to bound $\|(U^\top U\mathbf{z})^s\|_2$, it suffices to bound $\mathcal{E}_s(\mathbf{z})$.

Theorem 7 *For a fixed* $\mathbf{z}$*, with probability* $1 - e^{-\tau}$ *with some arbitrary* $\tau > 0$*, we have*

$$\mathcal{E}_s(\mathbf{z}) \leq c\left(\sqrt{\frac{\tau + s\log(d/s)}{n}} \|\mathbf{z}\|_2 + \|\mathbf{z}^s\|_2 \right),$$

where c *is some universal constant.*

Proof Let $\mathcal{K}_{d,s}(\epsilon)$ be the proper ϵ-net for $\mathcal{K}_{d,s}$ with the smallest cardinality (i.e., covering number) [for definitions of ϵ-net and covering number, please refer to Plan and Vershynin (2011)], and let $N(\mathcal{K}_{d,s}, \epsilon)$ be the covering number for $\mathcal{K}_{d,s}$. Lemma 3 in "Appendix C"

bounds the covering number $N(\mathcal{K}_{d,s}, \epsilon)$. By taking the union bound [aka Boole's inequality Galambos (1977)] of the result in Lemma 5 over all $\mathbf{w} \in \mathcal{K}_{d,s}(\epsilon)$, we have, with probability $1 - e^{-\tau}$,

$$\max_{\mathbf{w} \in \mathcal{K}_{d,s}(\epsilon)} \left| \mathbf{w}^\top U^\top U \mathbf{z} - \mathbf{w}^\top \mathbf{z} \right| \leq c \sqrt{\frac{\tau + s \log(9d/[\epsilon s])}{n}} |\mathbf{z}|_2,$$

if $\epsilon \in (0, 1)$, and therefore

$$\mathcal{E}_s(\mathbf{z}, \epsilon) \leq c \sqrt{\frac{\tau + s \log(9d/[\epsilon s])}{n}} \|\mathbf{z}\|_2 + \|\mathbf{z}^s\|_2.$$

We complete the proof by using Lemma 4 in "Appendix C" with $\epsilon = \frac{1}{2\sqrt{2}}$ and assuming d is sufficiently large. □

B.2 Bound $\left\|((U^\top U - I)\mathbf{z})^s\right\|_2$ for any sparse $\mathbf{z}$

Theorem 8 *For any* $\mathbf{z}$ *with* $\|\mathbf{z}\|_0 \leq s$, *and any* $\tau > 0$, *with probability* $1 - e^{-\tau}$, *we have*

$$\left\|((U^\top U - I)\mathbf{z})^s\right\|_2 \leq c \sqrt{\frac{\tau + s \log[d/s]}{n}} \|\mathbf{z}\|_2,$$

where c *is some universal constant.*

Proof We define $\Sigma_s(\mathbf{z})$ as

$$\Sigma_s(\mathbf{z}) = \max_{\mathbf{w} \in \mathcal{K}_{d,s}} \mathbf{w}^\top (U^\top U - I)\mathbf{z}.$$

It is easy to see $\left\|((U^\top U - I)\mathbf{z})^s\right\|_2 = \Sigma_s(\mathbf{z})$. Following the analysis of Theorem 7, it is easy to verify that, with probability $1 - e^{-\tau}$, for a fixed $\mathbf{z}$, we have

$$\Sigma_s(\mathbf{z}) \leq c \sqrt{\frac{\tau + s \log[d/s]}{n}} \|\mathbf{z}\|_2,$$

for some universal constant c. To extend this result to any s-sparse $\mathbf{z}$, we define

$$\mu_s = \max_{\mathbf{z} \in \mathcal{K}_{d,s}} \Sigma_s(\mathbf{z}).$$

Evidently, for any $\mathbf{z}$ with $\|\mathbf{z}\|_0 \leq s$, we have

$$\Sigma_s(\mathbf{z}) \leq \mu_s \|\mathbf{z}\|_2.$$

Using the same idea as Theorem 7, we define a discrete version of μ_s as

$$\mu_s(\epsilon) = \max_{\mathbf{z} \in \mathcal{K}_{d,s}(\epsilon)} \Sigma_s(\mathbf{z}),$$

and following the same argument as Lemma 4, we have

$$\mu_s \leq \frac{\mu_s(\epsilon)}{1 - \sqrt{2}\epsilon}.$$

Since for any fixed $\mathbf{z} \in \mathcal{K}_{d,s}$, with probability $1 - e^{-\tau}$, we have

$$\Sigma_s(\mathbf{z}) \leq c \sqrt{\frac{\tau + s \log[d/s]}{n}}.$$

By taking the union bound and using the relationship between μ_s and $\mu_s(\epsilon)$, with probability $1 - e^{-\tau}$, we have

$$\mu_s \leq c\sqrt{\frac{\tau + s\log[d/s]}{n}}.$$

We complete the proof by using $\Sigma_s(\mathbf{z}) \leq \mu_s \|\mathbf{z}\|_2$. □

Proof of Lemma 2 Combining the above results, we can complete the proof of Lemma 2. In particular, we apply Theorem 7 to bound $\|\mathbf{w}_a^s\|_2 = \|(U^\top U(\mathbf{x}_*^s - \mathbf{x}_*))^s\|_2$ in (31), and apply Theorem 8 to bound $\|\mathbf{w}_b^s\|_2 = \|(U^\top U - I)(\mathbf{x}_t - \mathbf{x}_*^s)\|_2$ for any $2s$-sparse $\mathbf{x}_t - \mathbf{x}_*^s$. □

C Other Lemmas and Proofs

Lemma 3 (Lemma 3.3 from Plan and Vershynin (2011)) *For $\epsilon \in (0, 1)$ and $s \leq d$, we have*

$$\log N(\mathcal{K}_{d,s}, \epsilon) \leq s \log\left(\frac{9d}{\epsilon s}\right).$$

Using the ϵ-net $\mathcal{K}_{d,s}(\epsilon)$, we define a discretized version of $\mathcal{E}_s(\mathbf{z})$ as

$$\mathcal{E}_s(\mathbf{z}, \epsilon) = \max_{\mathbf{w} \in \mathcal{K}_{d,s}(\epsilon)} \mathbf{w}^\top U^\top U \mathbf{z}.$$

The following lemma relates $\mathcal{E}_s(\mathbf{z}, \epsilon)$ with $\mathcal{E}_s(\mathbf{z})$.

Lemma 4 *For $\epsilon \in (0, 1/\sqrt{2})$, we have*

$$\mathcal{E}_s(\mathbf{z}) \leq \frac{\mathcal{E}_s(\mathbf{z}, \epsilon)}{1 - \sqrt{2}\epsilon}.$$

Based on the conclusion from Lemma 4, it is sufficient to bound $\mathcal{E}_s(\mathbf{z}, \epsilon)$. The lemma below is useful for bounding $\mathcal{E}_s(\mathbf{z}, \epsilon)$ that follows from the JL lemma for a sub-Gaussian matrix.

Lemma 5 *For fixed $\mathbf{w}$ and $\mathbf{z}$ such that $\|\mathbf{w}\|_2 \leq 1$, and any $\tau > 0$, with probability $1 - e^{-\tau}$, we have*

$$\mathbf{w}^\top U^\top U \mathbf{z} - \mathbf{w}^\top \mathbf{z} \leq c\sqrt{\frac{\tau}{n}} \|\mathbf{z}\|_2,$$

where c is some universal constant.

C.1 Proof of Lemma 4

The analysis is the same as that for Lemma 9.2 of Koltchinskii (2011), we include it for completeness. For any $\mathbf{x}, \mathbf{x}' \in \mathcal{K}_{d,s}$, we can always find two vectors $\mathbf{y}, \mathbf{y}'$ such that

$$\mathbf{x} - \mathbf{x}' = \mathbf{y} - \mathbf{y}', \ \|\mathbf{y}\|_0 \leq s, \ \|\mathbf{y}'\|_0 \leq s, \ \mathbf{y}^\top \mathbf{y}' = 0.$$

Thus

$$\begin{aligned}
\langle \mathbf{x} - \mathbf{x}', UU^\top \mathbf{z} \rangle &= \langle \mathbf{y}, UU^\top \mathbf{z} \rangle + \langle -\mathbf{y}', UU^\top \mathbf{z} \rangle \\
&= \|\mathbf{y}\|_2 \left\langle \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, UU^\top \mathbf{z} \right\rangle + \|\mathbf{y}'\|_2 \left\langle \frac{-\mathbf{y}'}{\|\mathbf{y}'\|_2}, UU^\top \mathbf{z} \right\rangle \\
&\leq (\|\mathbf{y}\|_2 + \|\mathbf{y}'\|_2)\mathcal{E}_s(\mathbf{z}) \leq \mathcal{E}_s(\mathbf{z})\sqrt{2}\sqrt{\|\mathbf{y}\|_2^2 + \|\mathbf{y}'\|_2^2} \\
&= \mathcal{E}_s(\mathbf{z})\sqrt{2}\|\mathbf{y} - \mathbf{y}'\|_2 = \mathcal{E}_s(\mathbf{z})\sqrt{2}\|\mathbf{x} - \mathbf{x}'\|_2.
\end{aligned}$$

Then, we have

$$\begin{aligned}\mathcal{E}_s(\mathbf{z}) &= \max_{\mathbf{w}\in\mathcal{K}_{d,s}} \mathbf{w}^\top U U^\top \mathbf{z} \\ &\le \mathcal{E}_s(\mathbf{z},\epsilon) + \sup_{\mathbf{x}\in\mathcal{K}_{d,s},\mathbf{x}'\in\mathcal{K}_{d,s}(\epsilon),\|\mathbf{x}-\mathbf{x}'\|_2\le\epsilon} \langle \mathbf{x}-\mathbf{x}', UU^\top\mathbf{z}\rangle \\ &\le \mathcal{E}_s(\mathbf{z},\epsilon) + \sqrt{2}\epsilon\mathcal{E}_s(\mathbf{z}),\end{aligned}$$

which implies

$$\mathcal{E}_s(\mathbf{z}) \le \frac{\mathcal{E}_s(\mathbf{z},\epsilon)}{1-\sqrt{2}\epsilon}.$$

C.2 Proof of Lemma 5

Let us first assume $\|\mathbf{z}\|_2 = 1$, otherwise

$$\mathbf{w}^\top U^\top U\mathbf{z} - \mathbf{w}^\top\mathbf{z} \le (\mathbf{w}^\top U^\top U\mathbf{z}' - \mathbf{w}^\top\mathbf{z}')\|\mathbf{z}\|_2,$$

where $\mathbf{z}' = \mathbf{z}/\|\mathbf{z}\|_2$. Following the JL lemma for a sub-Gaussian matrix (Nelson 2013), we know that with probability $1-\exp(-c'\epsilon^2 n)$, where c' is some constant (indeed, $c' < 1/8$ works for a Gaussian matrix $U_{ij} \sim \mathcal{N}(0, 1/\sqrt{n})$),

$$(1-\epsilon)\|\mathbf{z}\|_2^2 \le \|U\mathbf{z}\|_2^2 \le (1+\epsilon)\|\mathbf{z}\|_2^2.$$

Therefore,

$$\begin{aligned}\mathbf{w}^\top U^\top U\mathbf{z} - \mathbf{w}^\top\mathbf{z} &= \frac{\|U(\mathbf{w}+\mathbf{z})\|_2^2 - \|U(\mathbf{w}-\mathbf{z})\|_2^2}{4} - \mathbf{w}^\top\mathbf{z} \\ &\le \frac{\epsilon}{2}(\|\mathbf{w}\|_2^2 + \|\mathbf{z}\|_2^2) \le \epsilon.\end{aligned}$$

Therefore, with probability $1-e^{-\tau}$, we have

$$\mathbf{w}^\top U^\top U\mathbf{z} - \mathbf{w}^\top\mathbf{z} \le c\sqrt{\frac{\tau}{n}}\|\mathbf{z}\|_2,$$

where $c = 1/\sqrt{c'}$.

D Top-*s* recovery error

Proposition 5 *Let $\mathbf{y} \in \mathbb{R}^{2s}$ be an arbitrary s-sparse vector. Then we have*

$$\|\mathbf{x}^s - \mathbf{y}\|_2 \le \sqrt{3}\|\mathbf{x}-\mathbf{y}\|_2 \quad \forall \mathbf{x} \in \mathbb{R}^{2s}.$$

Proof Let $\mathcal{X}$ and $\mathcal{Y}$ be the support set of $\mathbf{x}$ and $\mathbf{y}$, respectively. If $|\mathcal{X}| \le s$, we have

$$\|\mathbf{x}^s - \mathbf{y}\|_2 = \|\mathbf{x}-\mathbf{y}\|_2.$$

Thus, in the following, we only need to consider the case $|\mathcal{X}| > s$. Let $\mathcal{A}$ be the indices of the s largest elements in $\mathbf{x}$, and $\mathcal{B} = \mathcal{X}\backslash\mathcal{A}$. Then, we have

$$\|\mathbf{x} - \mathbf{y}\|_2^2 = \sum_{i\in\mathcal{A}\backslash\mathcal{Y}} x_i^2 + \sum_{i\in\mathcal{A}\cap\mathcal{Y}} (x_i - y_i)^2 + \sum_{i\in\mathcal{B}\cap\mathcal{Y}} (x_i - y_i)^2 + \sum_{i\in\mathcal{B}\backslash\mathcal{Y}} x_i^2,$$
$$\|\mathbf{x}^s - \mathbf{y}\|_2^2 = \sum_{i\in\mathcal{A}\backslash\mathcal{Y}} x_i^2 + \sum_{i\in\mathcal{A}\cap\mathcal{Y}} (x_i - y_i)^2 + \sum_{i\in\mathcal{B}\cap\mathcal{Y}} y_i^2.$$

Since

$$|\mathcal{A}\backslash\mathcal{Y}| + |\mathcal{A}\cap\mathcal{Y}| = |\mathcal{A}| = s \geq |\mathcal{Y}| = |\mathcal{A}\cap\mathcal{Y}| + |\mathcal{B}\cap\mathcal{Y}|,$$

we have $|\mathcal{A}\backslash\mathcal{Y}| \geq |\mathcal{B}\cap\mathcal{Y}|$. As a result, we must have

$$\sum_{i\in\mathcal{B}\cap\mathcal{Y}} x_i^2 \leq \sum_{i\in\mathcal{A}\backslash\mathcal{Y}} x_i^2. \tag{32}$$

Since

$$\sum_{i\in\mathcal{B}\cap\mathcal{Y}} y_i^2 \leq 2\sum_{i\in\mathcal{B}\cap\mathcal{Y}} (x_i - y_i)^2 + 2\sum_{i\in\mathcal{B}\cap\mathcal{Y}} x_i^2 \overset{(32)}{\leq} 2\sum_{i\in\mathcal{B}\cap\mathcal{Y}} (x_i - y_i)^2 + 2\sum_{i\in\mathcal{A}\backslash\mathcal{Y}} x_i^2,$$

we have

$$\|\mathbf{x}^s - \mathbf{y}\|_2^2 \leq 3\sum_{i\in\mathcal{A}\backslash\mathcal{Y}} x_i^2 + \sum_{i\in\mathcal{A}\cap\mathcal{Y}} (x_i - y_i)^2 + 2\sum_{i\in\mathcal{B}\cap\mathcal{Y}} (x_i - y_i)^2 \leq 3\|\mathbf{x} - \mathbf{y}\|_2^2.$$

□

E Upper bound of $\|U^\top \mathbf{e}\|_\infty$

Proposition 6 *Let $U \in \mathbb{R}^{n\times d}$ be a random matrix with sub-Gaussian entries of mean 0 and variance $1/n$. For any $\tau > 0$, with probability $1 - 2e^{-\tau}$, we have*

$$\|U^\top \mathbf{e}\|_\infty \leq \theta\|\mathbf{e}\|_2\sqrt{\frac{\tau + \log d}{n}}, \tag{33}$$

where $\theta > 0$ is a constant.

Proof Let $\mathbf{u}_i$ denote the ith column vector of U. Since $[\mathbf{u}_i]_j$, $j = 1, \ldots, n$, are independent $(1/\sqrt{n})$- sub-Gaussian variables, $\mathbf{u}_i^T\mathbf{e}$ is a $(\|\mathbf{e}\|_2/\sqrt{n})$- sub-Gaussian variable. According to the property of a sub-Gaussian vector, there exists $\theta > 0$ such that

$$\|\mathbf{u}_i^\top \mathbf{e}\|_{\psi_2} \leq \theta\frac{\|\mathbf{e}\|_2}{\sqrt{n}}, i = 1, \ldots, d,$$

where $\|\cdot\|_{\psi_2}$ is the Orlicz norm (Rao and Ren 1991). It is known that the following property of the Orlicz norm $|\mathbf{u}_i^\top \mathbf{e}| \leq \|\mathbf{u}_i^\top \mathbf{e}\|_{\psi_2}\sqrt{\tau}$ holds. [cf. Koltchinskii (2011)]. Then, with probability $1 - 2e^{-\tau}$, we have

$$|\mathbf{u}_i^\top \mathbf{e}| \leq \|\mathbf{u}_i^\top \mathbf{e}\|_{\psi_2}\sqrt{\tau} \leq \theta\|\mathbf{e}\|_2\sqrt{\frac{\tau}{n}}.$$

Taking the union bound, we can complete the proof. □

References

Agarwal, A., Negahban, S., & Wainwright, M. J. (2010). Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, *23*, 37–45.

Amster, P. (2014). *The Banach fixed point theorem* (pp. 29–51). Boston, MA: Springer.

Asif, M. S., & Romberg, J. K. (2014). Sparse recovery of streaming signals using ℓ_1-homotopy. *IEEE Transactions on Signal Processing*, *62*(16), 4209–4223.

Becker, S., Bobin, J., & Candès, E. J. (2011). Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, *4*, 1–39.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*(1), 183–202.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, *37*(4), 1705–1732.

Blumensath, T., & Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, *27*, 265–274.

Brauer, C., Lorenz, D. A., & Tillmann, A. M. (2018). A primal-dual homotopy algorithm for ℓ_1-minimization with ℓ_∞-constraints. *Computational Optimization and Applications*, *70*(2), 443–478.

Bredies, K., & Lorenz, D. A. (2008). Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, *14*(5–6), 813–837.

Cai, T. T., & Zhang, A. (2014). Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Transactions on Information Theory*, *60*(1), 122–132.

Candès, E. (2008). The restricted isometry property and its implications for compressed sensing. *Comptes rendus de l'Académie des Sciences Serie*, *I*, 589–592.

Candès, E. J., Romberg, J. K., & Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, *59*, 1207–1223. https://doi.org/10.1002/cpa.20124.

Candès, E. J., & Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, *51*, 4203–4215.

Candès, E., & Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*(6), 2313–2351.

Candès, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, *25*, 21–30.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*(1), 33–61.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, *43*, 129–159.

Dasgupta, A., Kumar, R., & Sarlós, T. (2010). A sparse Johnson–Lindenstrauss transform. In *Proceedings of the 42nd ACM symposium on theory of computing, STOC '10* (pp. 341–350).

Davenport, M. A., Duarte, M. F., Eldar, Y. C., & Kutyniok, G. (2012). Introduction to compressed sensing. In *Compressed sensing: Theory and applications*. Cambridge University Press

Davis, G., Mallat, S., & Avellaneda, M. (2004). Adaptive greedy approximations. *Constructive Approximation*, *13*, 57–98.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*, 1289–1306.

Donoho, D. L., & Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. CoRR abs/0906.2530.

Donoho, D. L., Johnstone, I., & Montanari, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Transactions on Information Theory*, *59*(6), 3396–3433.

Donoho, D. L., Maleki, A., & Montanari, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, *57*(10), 6920–6941.

Donoho, D. L., & Tsaig, Y. (2008). Fast solution of l1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, *54*, 4789–4812.

Donoho, D. L., Tsaig, Y., Drori, I., & Starck, J. L. (2012). Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Transactions on Information Theory*, *58*, 1094–1121.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Eghbali, R., & Fazel, M. (2017). Decomposable norm minimization with proximal-gradient homotopy algorithm. *Computational Optimization and Applications*, *66*(2), 345–381. https://doi.org/10.1007/s10589-016-9871-8.

Eldar, Y., & Kutyniok, G. (2012). *Compressed sensing: Theory and applications*. Cambridge: Cambridge University Press.
Foucart, S. (2011). Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, *49*(6), 2543–2563.
Galambos, J. (1977). Bonferroni inequalities. *Annals of Probability*, *5*(4), 577–581. https://doi.org/10.1214/aop/1176995765.
Garg, R., & Khandekar, R. (2009). Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th annual international conference on machine learning* (pp. 337–344). ACM.
Hale, E. T., Wotao, Y., & Zhang, Y. (2008). Fixed-point continuation for l1-minimization: methodology and convergence. *SIAM Journal on Optimization*, *19*(3), 1107–1130.
Hanson, D. L., & Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Annals of Mathematical Statistics*, *42*(3), 1079–1083. https://doi.org/10.1214/aoms/1177693335.
Johnson, W., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability* (New Haven, CT, 1982) (Vol. 26, pp. 189–206).
Kane, D. M., & Nelson, J. (2014). Sparser Johnson–Lindenstrauss transforms. *Journal of the ACM*, *61, 4:1–4*, 23.
Kim, S., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2008). An interior-point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, *1*, 606–617.
Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems: École DÉté de Probabilités de Saint-Flour XXXVIII-2008. Ecole d'été de probabilités de Saint-Flour*. New York: Springer.
Kyrillidis, A. T., & Cevher, V. (2012). Combinatorial selection and least absolute shrinkage via the clash algorithm. In *ISIT* (pp. 2216–2220).
Kyrillidis, A. T., & Cevher, V. (2014). Matrix recipes for hard thresholding methods. *Journal of Mathematical Imaging and Vision*, *48*(2), 235–265.
Lin, Q., & Xiao, L. (2015). An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. *Computational Optimization and Applications*, *60*(3), 633–674. https://doi.org/10.1007/s10589-014-9694-4.
Lorenz, D. A., Pfetsch, M. E., & Tillmann, A. M. (2014a). An infeasible-point subgradient method using adaptive approximate projections. *Computational Optimization and Applications*, *57*(2), 271–306. https://doi.org/10.1007/s10589-013-9602-3.
Lorenz, D. A., Pfetsch, M. E., & Tillmann, A. M. (2014b). Solving basis pursuit: Heuristic optimality check and solver comparison. *ACM Transactions on Mathematical Software*, *41*, 1–29.
Maleki, A., & Donoho, D. L. (2010). Optimally tuned iterative reconstruction algorithms for compressed sensing. *The Journal of Selected Topics in Signal Processing*, *4*(2), 330–341.
Mallat, S., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, *41*, 3397–3415.
Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.
Needell, D., & Tropp, J. A. (2010). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Communications of the ACM*, *53*, 93–100.
Needell, D., & Vershynin, R. (2009). Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, *9*, 317–334.
Nelson, J. (2013). Johnson–Lindenstrauss notes. Technical report.
Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Core discussion papers, Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE).
Osborne, M. R., Presnell, B., & Turlach, B. A. (1999). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, *9*, 319–337.
Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, *20*, 389–403.
Oymak, S., Recht, B., & Soltanolkotabi, M. (2018). Sharp time?data tradeoffs for linear inverse problems. *IEEE Transactions on Information Theory*, *64*(6), 4129–4158. https://doi.org/10.1109/TIT.2017.2773497.
Plan, Y., & Vershynin, R. (2011). One-bit compressed sensing by linear programming. CoRR abs/1109.4299.
Rao, M., & Ren, Z. (1991). *Theory of orlicz spaces. Chapman and Hall pure and applied mathematics*. Boca Raton: CRC Press.
Rockafellar, R. T. (1970). *Convex analysis. Princeton mathematical series*. Princeton, NJ: Princeton University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, *58*, 267–288.
Tillmann, A. M., & Pfetsch, M. E. (2014). The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, *60*(2), 1248–1259. https://doi.org/10.1109/TIT.2013.2290112.
Tropp, J. A. (2006a). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, *50*, 2231–2242.
Tropp, J. A. (2006b). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, *52*, 1030–1051.
Tropp, J. A., & Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, *53*, 4655–4666.
Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization* **(submitted)**.
Turlach, B. A., Venables, W. N., & Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, *47*, 349–363.
van de Geer, S. A., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *The Electronic Journal of Statistics*, *3*, 1360–1392.
van den Berg, E., & Friedlander, M. P. (2008). Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, *31*(2), 890–912. https://doi.org/10.1137/080714488.
Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, *55*, 2183–2202.
Wen, Z., Yin, W., Goldfarb, D., & Zhang, Y. (2010). A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM Journal on Scientific Computing*, *32*(4), 1832–1857. https://doi.org/10.1137/090747695.
Wright, S., Nowak, R., & Figueiredo, M. A. T. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*. https://doi.org/10.1109/TSP.2009.2016892.
Xiao, L., & Zhang, T. (2013). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, *23*(2), 1062–1091.
Zhang, T. (2009). Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, *37*, 2109–2144.
Zhang, C. H., & Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, *36*, 1567–1594.
Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Affiliations

Tianbao Yang[1] · Lijun Zhang[2] · Rong Jin[3] · Shenghuo Zhu[3] · Zhi-Hua Zhou[2]

Lijun Zhang
zhanglj@lamda.nju.edu.cn

Rong Jin
jinrong.jr@alibaba-inc.com

Shenghuo Zhu
shenghuo@gmail.com

Zhi-Hua Zhou
zhouzh@lamda.nju.edu.cn

1 Department of Computer Science, University of Iowa, Iowa City, IA 52242, USA

2 National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

3 Alibaba Group, Seattle, WA, USA