A Simple Homotopy Algorithm for Compressive Sensing

Lijun Zhang* **Tianbao Yang†** **Rong Jin‡** **Zhi-Hua Zhou***
*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
†Department of Computer Science, the University of Iowa, Iowa City, USA
‡Department of Computer Science and Engineering, Michigan State University, East Lansing, USA
‡Institute of Data Science and Technologies at Alibaba Group, Seattle, USA
{zhanglj, zhouzh}@lamda.nju.edu.cn tianbao-yang@uiowa.edu rongjin@cse.msu.edu

Abstract

In this paper, we consider the problem of recovering the s largest elements of an arbitrary vector from noisy measurements. Inspired by previous work, we develop an homotopy algorithm which solves the ℓ_1-regularized least square problem for a sequence of decreasing values of the regularization parameter. Compared to the previous method, our algorithm is more *c nt* in the sense it only updates the solution once for each intermediate problem, and more *pract ca* in the sense it has a simple stopping criterion by checking the sparsity of the intermediate solution. Theoretical analysis reveals that our method enjoys a linear convergence rate in reducing the recovery error. Furthermore, our guarantee for recovering the top s elements of the target vector is tighter than previous results, and that for recovering the target vector itself matches the state of the art in compressive sensing.

1 Introduction

Compressive Sensing (CS) is a new paradigm of data acquisition that enables reconstruction of sparse or compressible signals from a relatively small number of linear measurements (Candès and Tao, 2006; Donoho, 2006). The standard assumption is that one has access to linear measurements of the form

$$\mathbf{y} = U^\top \mathbf{x}_* + \mathbf{e}$$

where $\mathbf{x}_* \in \mathbb{R}^d$ is the *un nown* target vector, $U \in \mathbb{R}^{d\times m}$ is the sensing matrix and $\mathbf{e} \in \mathbb{R}^m$ is a vector of noise. Recently, substantial process has been made in designing the encoder U and the associated decoder Δ which recovers $\mathbf{x}_*$ from U and $\mathbf{y}$ (Davenport et al., 2012).

One of the most famous decoders for CS is the ℓ_1-regularized least squares (ℓ_1-LS) formulation, known as Lasso in statistics (Tibshirani, 1996), given by

$$\min \frac{1}{2}\|\mathbf{y} - U^\top \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

The recovery performance of Lasso has been extensively studied and generally speaking, it can recover $\mathbf{x}_*$ up to the noise level under appropriate assumptions (Daubechies et al., 2004; Tropp, 2006; Zhang, 2009).

Under the assumption that λ is given beforehand, Xiao and Zhang (2012) propose a proximal-gradient homotopy method to improve the efficiency of ℓ_1-LS. The key idea is to solve the ℓ_1-LS problem for a sequence of decreasing values of the regularization parameter, and use an approximate solution at the end of each stage to warm start the next stage. In this study, we make three steps further. We show that (i)

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

ℓ_2-norm of the s largest elements of $\mathbf{x}_* - \mathbf{x}_*^s$. In contrast, previous analysis in CS (Davenport et al., 2012) can only upper bound the recovery error for $\mathbf{x}_*^s$ by

$$O\left(\|\mathbf{e}\|_2 + \|\mathbf{x}_* - \mathbf{x}_*^s\|_2\right).$$

Thus, our recovery guarantee could be significantly better than previous results if the ℓ_2-norm of $\mathbf{x}_* - \mathbf{x}_*^s$ is not concentrated on its s largest elements (i.e., $\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2 \ll \|\mathbf{x}_* - \mathbf{x}_*^s\|_2$). Following the triangle inequality, we obtain the following bound for recovering $\mathbf{x}_*$

$$\|\hat{\mathbf{x}} - \mathbf{x}_*\|_2 \leq O\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \sqrt{\frac{s\log d}{m}}\|\mathbf{e}\|_2\right)$$

which matches state of the art (DeVore et al., 2009).

2 Related Work

Existing algorithms in CS could be roughly categorized into convex optimization based approaches and greedy approaches (Davenport et al., 2012; Blumensath et al., 2012). Roughly speaking, convex approaches have better theoretical guarantee, while greedy approaches are more efficient.

In the noise-free setting, Candès and Tao (2005) pose the following ℓ_1-minimization problem, denoted by Δ_1, for decoding

$$\min \|\mathbf{x}\|_1 \quad \text{s.t. } U^\top \mathbf{x} = \mathbf{y}.$$

To analyze the recovery performance, they introduce the Restricted Isometry Property (RIP) for matrices. Define the isometry constant of U as the smallest number δ_s such that the following holds for all s-sparse vectors $\mathbf{x} \in \mathbb{R}^d$

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \leq \|U^\top \mathbf{x}\|_2^2 \leq (1+\delta_s)\|\mathbf{x}\|_2^2.$$

It has been shown that if $\delta_s + \delta_{2s} + \delta_{3s} < 1$ or $\delta_{2s} < \sqrt{2}-1$, the decoder Δ_1 yields perfect recovery for *a* s-sparse vectors $\mathbf{x}_* \in \mathbb{R}^d$ (Candès and Tao, 2005; Candès, 2008). If U is constructed as the random matrix with independent sub-Gaussian columns, a sufficient condition is to take $m = \Omega(s\log d)$ measurements (Mendelson et al., 2008).

In the general case, Candès (2008) propose the following convex formulation, denoted by Δ_1^ϵ, for decoding

$$\min \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{y} - U^\top \mathbf{x}\|_2 \leq \epsilon$$

where ϵ is an upper bound of $\|\mathbf{e}\|_2$. Let $\Delta_1^\epsilon(U^\top \mathbf{x}_* + \mathbf{e})$ be the solution returned by the above decoder. Suppose U satisfies the RIP of order $2s$ with $\delta_{2s} < \sqrt{2}-1$, we have

$$\|\Delta_1^\epsilon(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2 \leq O\left(\frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + \epsilon\right)$$

for *a* $\mathbf{x}_* \in \mathbb{R}^d$. An obvious drawback of this approach is that we must have a good a priori estimate of $\|\mathbf{e}\|_2$. This limitation is soon addressed by Wojtaszczyk (2010), who shows that the decoder Δ_1 performs very well even in the noise setting. In particular, if U is a Gaussian random matrix and $m = \Omega(s\log d)$, with an overwhelming probability, we have

$$\|\Delta_1(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2 \leq O\left(\frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + \|\mathbf{e}\|_2\right)$$

for *a* $\mathbf{x}_* \in \mathbb{R}^d$. Another possible way is to estimate the noise level under the Bayesian framework (Ji et al., 2008).

Notice that in the above inequalities, the ℓ_2-norm of the recovery error is upper bounded by the ℓ_1-norm of the corresponding error of the best s-term approximation. To make the result more consistent, it is natural to ask whether we could upper bound $\|\Delta(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2$ by $\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ for some decoder Δ. Unfortunately, even in the noise-free setting, if we want the following inequality

$$\|\Delta(U^\top \mathbf{x}_*) - \mathbf{x}_*\|_2 \leq O(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2)$$

to hold for all $\mathbf{x}_* \in \mathbb{R}^d$, the number of measurements m needs to be on the order of d (Cohen et al., 2009). This difficulty motivates the study of *nstanc opt a ty n probab ty*, which means we are looking for some performance guarantee that holds with a high probability for an arbitrary but *x d* vector $\mathbf{x}_*$. When U is a Gaussian random matrix and $m = \Omega(s\log d)$, Wojtaszczyk (2010) shows that with an overwhelming probability

$$\|\Delta_1(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2 \leq O\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \|\mathbf{e}\|_2\right)$$

for any *x d* $\mathbf{x}_* \in \mathbb{R}^d$. This result is extended to more general families of matrices by DeVore et al. (2009).

The typical greedy approaches include Matching Pursuit (MP) (Mallat and Zhang, 1993), Iterative Hard Thresholding (IHT) (Blumensath and Davies, 2008; Garg and Khandekar, 2009), and Compressive Sampling Matching Pursuit (CoSaMP) (Needell and Tropp, 2009). Due to space limitation, we only give a brief description of IHT. In each iteration, IHT first performs gradient descent with respect to $\|U^\top \mathbf{x} - \mathbf{y}\|_2^2$ and then applies hard-thresholding to keep the s largest elements, in contrast to the soft-thresholding in our method. Let $\mathbf{x}_t$ denote the solution of IHT in the t-th iteration. The updating rule is given by

$$\mathbf{x}_{t+1} = \left[\mathbf{x}_t - \eta U(U^\top \mathbf{x}_t - \mathbf{y})\right]^s$$

where η is the step size. Under certain RIP condition, it has been shown that the recovery error of IHT can

Algorithm 1 A Simple Homotopy Algorithm

Input: Sensing Matrix $U \in \mathbb{R}^{d\times n}$, Measurements $\mathbf{y} \in \mathbb{R}^m$, Shrinking Parameter γ, Sparsity s, Maximum Number of Iterations T,

Initialize $\mathbf{x}_1 = 0$, $\lambda_1 = \|U\mathbf{y}\|_\infty$
for $t = 1, 2, \ldots, T$ **do**
  $\mathbf{x}_{t+1} = P_{\lambda_t}\left(\mathbf{x}_t - U(U^\top \mathbf{x}_t - \mathbf{y})\right)$
  if $\|\mathbf{x}_{t+1}\|_0 > 2s$ **then**
    Return $\mathbf{x}_t$
  end if
  $\lambda_{t+1} = \gamma\lambda_t$
end for
Return $\mathbf{x}_{T+1}$

be upper bounded by

$$O\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + \|\mathbf{e}\|_2\right)$$

for $a \quad \mathbf{x}_* \in \mathbb{R}^d$ (Blumensath et al., 2012).

3 A Simple Homotopy Algorithm for Compressive Sensing

We first introduce the proposed homotopy algorithm, and then present its theoretical guarantee.

3.1 The Algorithm

In our algorithm, we solve a sequence of ℓ_1-regularized least squares (ℓ_1-LS) with decreasing regularization parameters. In particular, we set

$$\lambda_1 = \|U\mathbf{y}\|_\infty, \quad \lambda_{t+1} = \gamma\lambda_t$$

for some $\gamma < 1$. For each intermediate ℓ_1-LS problem, we use the solution from the previous iteration as the initial point, and perform composite gradient mapping (Nesterov, 2013) to update it *onc* , i.e.,

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \langle \mathbf{x}, U(U^\top \mathbf{x}_t - \mathbf{y})\rangle + \|\mathbf{x} - \mathbf{x}_t\|_2^2 \text{Bi1 .24T2626 i133 cm BT R15 .6264 Tf 1 1 75.6 52.246 Tm b}$$

w r

$$\Lambda = \sqrt{\frac{s(\tau + \log d)}{m}}\|\mathbf{e}\|_2 + C\left(\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2 + \sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{x}_* - \mathbf{x}_*^s\|_2\right) \quad (1)$$

or so un v rsa constant C prov d d

$$C\sqrt{\frac{\tau + s\log(d/s)}{m}} \leq \frac{1}{6}. \quad (2)$$

Remark The constants in the above theorem should not be treated literally, because we have made no effort to optimize them. Generally speaking, a smaller γ will lead to a fast convergence rate, but a larger constant, which is 6 now, in the recovery guarantee.

The above theorem implies that the recovery error reduces *xpon nt a y* until it reaches $O(\Lambda)$. Thus, with a sufficiently large T, the recovery error of $\mathbf{x}_*^s$ can be upper bounded by

$$O\left(\sqrt{\frac{s\log d}{m}}\left(\|\mathbf{e}\|_2 + \|\mathbf{x}_* - \mathbf{x}_*^s\|_2\right) + \|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2\right).$$

In the noise-free setting, our analysis also implies exact recovery of s-sparse vectors, since $\Lambda = 0$ if $\|\mathbf{e}\|_2 = 0$ and $\|\mathbf{x}_*\|_0 \leq s$.

From the literature of CS (Davenport et al., 2012, Theorem 1.14), we find that previous analysis is able to upper bound the recovery error of $\mathbf{x}_*^s$ by $c\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \|\mathbf{e}\|_2\right)$ for some constant $c > 1$. Thus, when m is large enough, our upper bound could be significantly smaller than the existing results. Finally, it is worth to mention that if our goal is to recover $\mathbf{x}_*$, following the triangle inequality, our analysis yields the same upper bound as previous studies (DeVore et al., 2009).

3.3 A Post-processing Step

Since $\hat{\mathbf{x}}$ is $2s$-sparse and $\mathbf{x}_*^s$ is s-sparse, one may ask whether it is possible to find a good s-sparse vector to approximate $\mathbf{x}_*^s$. The following theorem shows that we can simply select the s largest elements of $\hat{\mathbf{x}}$ to approximate $\mathbf{x}_*^s$ and the recovery error is on the same order.

Theorem 2. *L t* $\mathbf{y} \in \mathbb{R}^d$ *b a s spars v ctor n w av*

$$\|\mathbf{x}^s - \mathbf{y}\|_2 \leq \sqrt{3}\|\mathbf{x} - \mathbf{y}\|_2, \ \forall \mathbf{x} \in \mathbb{R}^d.$$

4 Analysis

We here present the proofs of main theorems. The omitted proofs are provided in the supplementary material.

4.1 Proof of Theorem 1

Notice that starting our algorithm with $\lambda_1 = \|U\mathbf{y}\|_\infty$ has the same effect as staring with $\|U\mathbf{y}\|_\infty\gamma^{-k}$, $k \in \mathbb{Z}$, which means we can set λ_1 as large as we need. Thus, without loss of generality, we can assume

$$\lambda_1 \geq \frac{1}{3\sqrt{s}}\max\left(\|\mathbf{x}_*^s\|_2, 6\Lambda\right). \quad (3)$$

We first state two theorems that are central to our analysis. Theorem 3 reveals that the recovery error of our algorithm will reduce by a constant factor until it reaches the optimal level. Then, Theorem 4 shows that the recovery error will remain small, as long as the sparsity of the solution does not exceed $2s$.

We denote by $\mathcal{S}_*$ and $\mathcal{S}_t$ the support set of $\mathbf{x}_*^s$ and $\mathbf{x}_t$, respectively.

Theorem 3. *Assu* $|\mathcal{S}_t \setminus \mathcal{S}_*| \leq s$ $\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq 3\lambda_t\sqrt{s}$ *and* $\Lambda \leq \frac{1}{2}\lambda_t\sqrt{s}$ *w r* Λ *s v n n n w t a probab ty at ast* $1 - 6e^{-\tau}$ *w av*

$$|\mathcal{S}_{t+1} \setminus \mathcal{S}_*| \leq s \text{ and } \|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq 3\lambda_{t+1}\sqrt{s}$$

prov d d t cond t on n s tru

Theorem 4. *Assu* $|\mathcal{S}_t| \leq 2s$ $\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq 6\Lambda$ *and* $\Lambda > \frac{1}{2}\lambda_t\sqrt{s}$ *w r* Λ *s v n n I* $|\mathcal{S}_{t+1}| \leq 2s$ *t n w t a probab ty at ast* $1 - 6e^{-\tau}$ *w av*

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

prov d d t cond t on n s tru

We continue the proof of Theorem 1 in the following. Let

$$k = \min\left\{t : \Lambda > \frac{1}{2}\lambda_t\sqrt{s}\right\} \overset{(3)}{>} 1.$$

In the following, we will show that the recovery error $\|\mathbf{x}_t - \mathbf{x}_*^s\|_2$ will first decrease exponentially as t approaches k, and then keep below 6Λ.

$T < k$ From (3), we have $\|\mathbf{x}_1 - \mathbf{x}_*^s\|_2 = \|\mathbf{x}_*^s\|_2 \leq 3\lambda_1\sqrt{s}$. Since the condition $\Lambda \leq \frac{\lambda_t\sqrt{s}}{2}$ holds for $t = 1, \ldots, T$, we can apply Theorem 3 to bound the recovery error in each iteration. Thus, with a probability at least $1 - 6Te^{-\tau}$, we have

$$\|\hat{\mathbf{x}} - \mathbf{x}_*^s\|_2 = \|\mathbf{x}_{T+1} - \mathbf{x}_*^s\|_2 \leq 3\lambda_{T+1}\sqrt{s} = 3\lambda_1\sqrt{s}\gamma^T.$$

$T \geq k$ From the above analysis, with a probability at least $1-6(k-1)e^{-\tau}$, we have $\|\mathbf{x}_k - \mathbf{x}_*^s\|_2 \leq 3\sqrt{s}\lambda_k$ and $|\mathcal{S}_k \setminus \mathcal{S}_*| \leq s$, which also means our algorithm arrives the k-th iteration. In the k-th iteration, there will be two cases: $|\mathcal{S}_{k+1}| > 2s$ and $|\mathcal{S}_{k+1}| \leq 2s$. For the first case, our algorithm terminates, and return $\mathbf{x}_k$ as the final solution, implying

$$\|\hat{\mathbf{x}} - \mathbf{x}_*^s\|_2 = \|\mathbf{x}_k - \mathbf{x}_*^s\|_2 \leq 3\lambda_k\sqrt{s} \leq 6\Lambda.$$

For the second case, our algorithm keeps running, and we can bound the recovery error of $\mathbf{x}_{k+1}$ by Theorem 4. If $T = k$ or $|\mathcal{S}_{k+2}| > 2s$, our algorithm terminates and return $\mathbf{x}_{k+1}$ as the final solution, which implies

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

Otherwise, our algorithm keeps running. Since $2(1+\sqrt{3}) \leq 6$, the condition in Theorem 4 are satisfied, and thus can be applied repeatedly to bound the recovery error for all the rest iterations.

4.2 Proof of Theorem 3

We need the following theorem to analyze the behavior of the composite gradient descent.

Theorem 5. *uppos* $\mathbf{x}_t - \mathbf{x}_*^s$ *s a* $3s$ *spars v ctor t a probab ty at ast* $1-6e^{-\tau}$ *w av*

$$\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 \leq \Lambda + C\sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2$$

w r Λ *s v n n*

Given a set $\mathcal{S} \subseteq [d]$, $\mathbf{x}_\mathcal{S}$ denotes the vector which coincides with $\mathbf{x}$ on $\mathcal{S}$ and has zero coordinates outside $\mathcal{S}$. We denote the sub-gradient of $\|\cdot\|_1$ by $\partial\|\cdot\|_1$.

Using the fact that

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

we have

$$\begin{aligned}
0 \leq& \left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) + \lambda_t\partial\|\mathbf{x}_{t+1}\|_1, \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle \\
\leq& \left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle + \lambda_t\|\mathbf{x}_*^s\|_1 - \lambda_t\|\mathbf{x}_{t+1}\|_1 \\
\leq& \left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle + \lambda_t\|\mathbf{x}_*^s\|_1 - \lambda_t\|[\mathbf{x}_{t+1}]_{\mathcal{S}_*}\|_1 \\
\leq& \left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle + \lambda_t\|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_*}\|_1 \\
\leq& \left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle + \lambda_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2
\end{aligned}$$

and thus

$$\begin{aligned}
&\lambda_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \\
\geq& \left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_{t+1} - \mathbf{x}_*^s\right\rangle \\
=& \|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2^2 + (\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top\left(U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right).
\end{aligned} \tag{4}$$

According to Theorem 5, with a probability at least $1-6e^{-\tau}$, we have

$$\begin{aligned}
&\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 \\
\leq& \Lambda + C\sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \\
\overset{(2)}{\leq}& \Lambda + \frac{1}{6}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq \lambda_t\sqrt{s}.
\end{aligned}$$

The above inequality implies the magnitude of the smallest $d-s$ elements of $\mathbf{x}_t - U(U^\top\mathbf{x}_t - \mathbf{y}) - \mathbf{x}_*^s$ is smaller than λ_t. Combining with the fact that

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

it is easy to verify that $|\mathcal{S}_{t+1} \setminus \mathcal{S}_*| \leq s$. Furthermore,

$$\begin{aligned}
&\left|(\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top\left(U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right)\right| \\
\leq& \left(\|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_*}\|_2 + \|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_{t+1}\setminus\mathcal{S}_*}\|_2\right)\lambda_t\sqrt{s} \\
\leq& \lambda_t\sqrt{2s}\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2.
\end{aligned}$$

Substituting the above inequality into (4), with a probability at least $1-6e^{-\tau}$, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2^2 \leq \left(\lambda_t\sqrt{s} + \lambda_t\sqrt{2s}\right)\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2$$

and thus

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

4.3 Proof of Theorem 4

We need to reuse (4) that appears in the analysis of Theorem 3. According to Theorem 5, with a probability at least $1-6e^{-\tau}$, we have

$$\begin{aligned}
&\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 \\
\overset{(2)}{\leq}& \Lambda + \frac{1}{6}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq 2\Lambda.
\end{aligned}$$

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

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

Substituting the above inequality into (4), we have

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

and thus

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

4.4 Proof of Theorem 5

In the analysis, we need to bound $\|(UU^\top\mathbf{z})^s\|_2$ for a x d vector $\mathbf{z}\in\mathbb{R}^d$, and $\|[(UU^\top-I)\mathbf{z}]^s\|_2$, for a $3s$-sparse vectors $\mathbf{z}\in\mathbb{R}^d$. Thus, we build the following two theorems.

Theorem 6. *For a x d* $\mathbf{z}\in\mathbb{R}^d$ *w t a probab ty at ast* $1-2e^{-\tau}$ *w av*

$$\left\|\left(UU^\top\mathbf{z}\right)^s\right\|_2 \le C\left(\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{z}\|_2+\|\mathbf{z}^s\|_2\right)$$

or so constant $C>0$

Theorem 7. *t a probab ty at ast* $1-2e^{-\tau}$ *or a* $\mathbf{z}\in\mathbb{R}^d$ *w t* $\|\mathbf{z}\|_0\le 3s$ *w av*

$$\left\|\left[(UU^\top-I)\mathbf{z}\right]^s\right\|_2 \le C\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{z}\|_2$$

or so constant $C>0$

We rewrite $U\left(U^\top\mathbf{x}_t-\mathbf{y}\right)-(\mathbf{x}_t-\mathbf{x}_*^s)$ as

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

Then, we have

$$\begin{aligned}&\left\|\left[U\left(U^\top\mathbf{x}_t-\mathbf{y}\right)-(\mathbf{x}_t-\mathbf{x}_*^s)\right]^s\right\|_2\\ \le&\|\mathbf{w}_a^s\|_2+\|\mathbf{w}_b^s\|_2+\|\mathbf{w}_c^s\|_2.\end{aligned}$$

Bounding $\|\mathbf{w}_a^s\|_2$ According to Theorem 6, with a probability at least $1-2e^{-\tau}$, we have

$$\begin{aligned}&\|\mathbf{w}_a^s\|_2\\ =&\left\|\left[UU^\top(\mathbf{x}_*^s-\mathbf{x}_*)\right]^s\right\|_2\\ \le&C\left(\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{x}_*-\mathbf{x}_*^s\|_2+\|(\mathbf{x}_*-\mathbf{x}_*^s)^s\|_2\right)\end{aligned}$$

for some constant $C>0$.

Bounding $\|\mathbf{w}_b^s\|_2$ Notice that $\mathbf{x}_t-\mathbf{x}_*^s$ is a $3s$-sparse vector. According to Theorem 7, with a probability at least $1-2e^{-\tau}$, we have

$$\begin{aligned}\|\mathbf{w}_b^s\|_2=&\left\|\left[(UU^\top-I)(\mathbf{x}_t-\mathbf{x}_*^s)\right]^s\right\|_2\\ \le&C\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{x}_t-\mathbf{x}_*^s\|_2\end{aligned}$$

for some constant $C>0$.

Bounding $\|\mathbf{w}_c^s\|_2$ Since $U=\frac{1}{\sqrt{m}}[\mathbf{v}_1,\ldots,\mathbf{v}_d]^\top$, and we assume $\mathbf{v}_i$ is a sub-Gaussian vector. We have

$$\|\mathbf{v}_j^\top\mathbf{e}\|_{\psi_2}\le\|\mathbf{e}\|_2,\ j=1,\ldots,d.$$

Using the property of Orliz norm (Koltchinskii, 2009, 2011), with a probability at least $1-2e^{-\tau}$, we have

$$|\mathbf{v}_j^\top\mathbf{e}|\le\|\mathbf{v}_j^\top\mathbf{e}\|_{\psi_2}\sqrt{\tau}\le\|\mathbf{e}\|_2\sqrt{\tau}.$$

By taking the union bound, we have, with a probability at least $1-2e^{-\tau}$,

$$\|U\mathbf{e}\|_\infty=\frac{1}{\sqrt{m}}\max_j|\mathbf{v}_j^\top\mathbf{e}|\le\|\mathbf{e}\|_2\sqrt{\frac{\tau+\log d}{m}}$$

implying

$$\|\mathbf{w}_c^s\|_2=\|(U\mathbf{e})^s\|_2\le\|\mathbf{e}\|_2\sqrt{\frac{s(\tau+\log d)}{m}}.$$

We complete the proof by combining the bounds for $\|\mathbf{w}_a^s\|_2$, $\|\mathbf{w}_b^s\|_2$, and $\|\mathbf{w}_c^s\|_2$.

4.5 Proof of Theorem 6

We define the set of s-sparse vectors with length smaller than 1 as

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

Then, it is easy to check that

$$\mathcal{E}_s(\mathbf{z}):=\max_{\mathbf{w}\in\mathcal{K}_{d,s}}\mathbf{w}^\top UU^\top\mathbf{z}=\|(UU^\top\mathbf{z})^s\|_2.$$

Let $\mathcal{K}_{d,s}(\epsilon)$ be a proper ϵ-net for $\mathcal{K}_{d,s}$ with the smallest cardinality, and $N(\mathcal{K}_{d,s},\epsilon)=|\mathcal{K}_{d,s}(\epsilon)|$ be the covering number for $\mathcal{K}_{d,s}$. We have the following lemma for bounding $N(\mathcal{K}_{d,s},\epsilon)$ (Plan and Vershynin, 2013, Lemma 3.3).

Lemma 1. *For* $\epsilon\in(0,1)$ *and* $s\le d$ *w av*

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

Using the ϵ

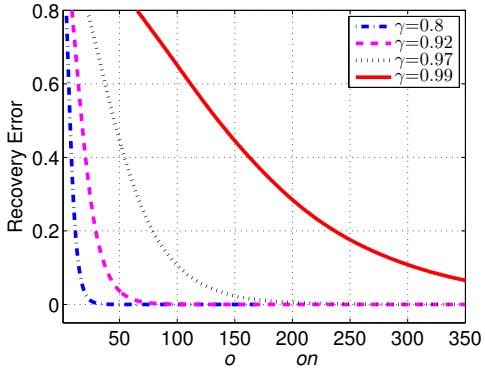


Figure 1: $\mathbf{x}_*$ is sparse.

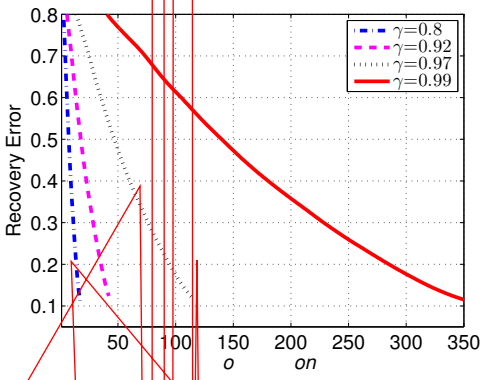
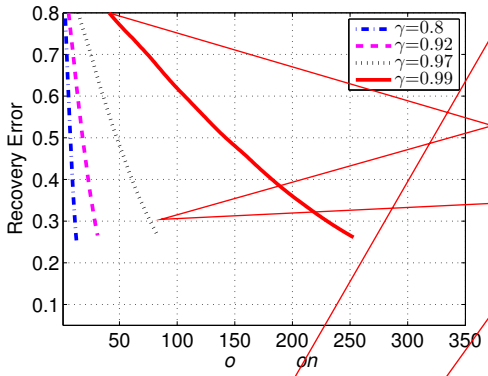


Figure 2: $\mathbf{x}_*$ is dense.



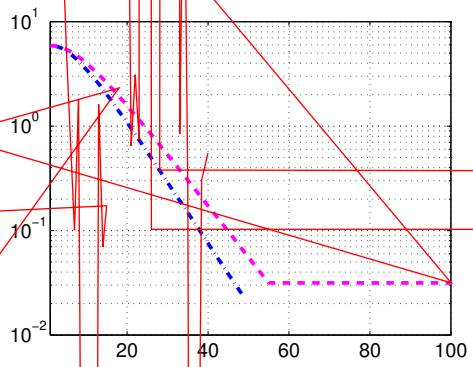


Figure 3: $\mathbf{x}_*$ is dense and $\mathbf{y}$ contains noise.

References

T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journa o Four r Ana ys s and App cat ons*, 14(5-6):629–654, 2008.

T. Blumensath, M. E. Davies, and G. Rilling. Greedy algorithms for compressed sensing. In *Co pr ss d ns n ory and App cat ons*, chapter 8, pages 348–393. Cambridge University Press, 2012.

P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *roc d n s o t nd Annua Con r nc on In or at on c nc s and yst s*, pages 16–21, 2008.

E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Co pt s ndus Mat at qu* , 346(9–10):589–592, 2008.

E. J. Candès and T. Tao. Decoding by linear programming. *In or at on ory IEEE ransact ons on*, 51(12):4203–4215, 2005.

E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE ransact ons on In or at on ory*, 52(12):5406–5425, 2006.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journa o t A r can Mat at ca oc ty*, 22(1):211C–231, 2009.

I. Daubechies, M. Defrise, and C. D. Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Co un cat ons on ur and App d Mat at cs*, 57(11):1413–1457, 2004.

M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing. In *Co pr ss d ns n ory and App cat ons*, chapter 1, pages 1–64. Cambridge University Press, 2012.

R. DeVore, G. Petrova, and P. Wojtaszczyk. Instance-optimality in probability with an ℓ_1-minimization decoder. *App d and Co putat ona Har on c Ana ys s*, 27(3):275–288, 2009.

D. L. Donoho. De-noising by soft-thresholding. *IEEE ransact ons on In or at on ory*, 41(3):613–627, 1995.

D. L. Donoho. Compressed sensing. *IEEE rans act ons on In or at on ory*, 52(4):1289–1306, 2006.

R. Garg and R. Khandekar. Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *roc d n s o t t Annua Int rnat ona Con r nc on Mac n L arn n* , pages 337–344, 2009.

S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE ransact ons on na roc ss n* , 56(6):2346–2356, 2008.

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *B rnou* , 15(3):799–828, 2009.

V. Koltchinskii. *Orac In qua t s n E p r ca s M n zat on and pars cov ry rob s.* Springer, 2011.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE ransact ons on na roc ss n* , 41(12):3397–3415, 1993.

S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Construct v Approx at on*, 28(3):277–289, 2008.

D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *App d and Co putat ona Har on c Ana ys s*, 26 (3):301–321, 2009.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mat at ca ro ra n* , 140 (1), 2013.

Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Co un cat ons on ur and App d Mat at cs*, 66(8):1275–1297, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journa o t oya tat st ca oc ty r s B M t odo o ca* , 58(1):267–288, 1996.

J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE ransact ons on In or at on ory*, 52(3):1030–1051, 2006.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Co pr ss d ns n ory and App cat ons*, chapter 5, pages 210–268. Cambridge University Press, 2012.

P. Wojtaszczyk. Stability and instance optimality for gaussian measurements in compressed sensing. *Foundat ons o Co putat ona Mat at cs*, 10(1), 2010. ISSN 1615-3375.

L. Xiao and T. Zhang. A proximal-gradient homotopy method for the ℓ_1-regularized least-squares problem. In *roc d n s o t t Int rnat ona Con r nc on Mac n L arn n* , pages 839–846, 2012.

L. Zhang, J. Yi, and R. Jin. Efficient algorithms for robust one-bit compressive sensing. In *roc d n s o t st Int rnat ona Con r nc on Mac n L arn n* , 2014.

T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *An na s o tat st cs*, 37(5A):2109–2144, 2009.