Cross Validation in Compressive Sensing and its Application of OMP-CV Algorithm

Jinye Zhang, Laming Chen, Petros T. Boufounos, and Yuantao Gu

Abstract

Compressive sensing (CS) is a data acquisition technique that measures sparse or compressible signals at a sampling rate lower than their Nyquist rate. Results show that sparse signals can be reconstructed using greedy algorithms, often requiring prior knowledge such as the signal sparsity or the noise level. As a substitute to prior knowledge, cross validation (CV), a statistical method that examines whether a model overfits its data, has been proposed to determine the stopping condition of greedy algorithms. This paper first analyzes cross validation in a general compressive sensing framework and developed general cross validation techniques which could be used to understand CV-based sparse recovery algorithms. Furthermore, we provide theoretical analysis for OMP-CV, a cross validation modification of orthogonal matching pursuit, which has very good sparse recovery performance. Finally, numerical experiments are given to validate our theoretical results and investigate the behaviors of OMP-CV.

Keywords: Compressive sensing, signal reconstruction, cross validation, orthogonal matching pursuit

1 Introduction

Compressive sensing (CS) is a new data acquisition technique that aims to measure sparse and compressible signals at sampling rate smaller than the Nyquist rate [1] [2]. Its fundamental promise is that some certain classes of signals, such as natural images, have a sparse representation where most of the coefficients are approximately zero. Generally speaking, compressive sensing consists of two main building blocks: sampling an N-dimensional k-sparse signal by computing its M (which is much smaller than N) linear projections and reconstructing the signal using various recovery methods.

Although the reconstruction of the original signal $\mathbf{x}$ from its M measurements is an ill-posed problem, it can be achieved by using the prior knowledge that $\mathbf{x}$ is sparse, i.e. $k \ll N$. One important result in CS theory is that $\mathbf{x}$ can be reconstructed using optimization strategies aiming to find the sparsest signal matching with the measurements, which can be viewed as an l_0 norm minimization problem [3]. Although the l_0 minimization is NP-hard, it was demonstrated [4] that it is equivalent to an l_1 optimization problem as long as the sensing matrix satisfies the so-called Restricted Isometry Property (RIP) with a constant parameter. In addition, such l_1 optimization problem could be solved efficiently via linear programming (LP) techniques.

Apart from the LP techniques, a family of iterative greedy algorithms also received significant attention due to their low computational complexity. They include OMP, ROMP, StOMP, SP, and CoSaMP [5, 6, 7, 8, 9]. The basic idea behind these algorithms is to find the support set of the signal iteratively. At each iteration, one or several coordinates of the vector $\mathbf{x}$ are selected into the current support set based on the correlation between the columns of sensing matrix and the measurement residual. With k iterations needed, the computational complexity of OMP, for example, is roughly $O(kmN)$ [8].

Most greedy algorithms require prior knowledge such as sparsity or noise level to properly stop the iteration, which however could not be satisfied in most practical cases. Without such information the termination of the algorithm may be too early or too late. In the former case, the signal will not be completely recovered (underfitting), while in the latter case some portion of the noise will be treated as signal (overfitting). In both cases, therefore, the reconstruction quality may greatly deteriorate.

As a substitute to prior knowledge, cross validation (CV) was proposed [10] to determine the stopping condition of greedy algorithms. Cross validation [11, 12, 13] is a statistical technique that separates a data set into a training (estimation) set and a testing (cross validation) set. The training set is used to construct the model and the testing set is used to adjust the model order so that the noise is not overfitted. When CV is utilized in compressive sensing, the measurement vector is split into a reconstruction measurement and a cross validation measurement. The former is used to reconstruct the sparse signal using a greedy algorithm, while the latter is to decide the stopping condition. The basic idea behind this technique is to sacrifice a small amount of measurements in exchange of prior knowledge. In a nutshell, this technique makes it possible for greedy algorithms to reconstruct the signal without prior knowledge like sparsity or noise level.

1.1 Related Works

The idea of applying cross validation in compressive sensing is first proposed by Boufounos, Duarte, and Baraniuk in [10], where the general framework of CS-CV modification is founded. In a CS-CV modified algorithm, both the sensing matrix and the measurement vector are separated into the reconstruction part and the cross validation part. When the former part is utilized iteratively to construct the support set, the latter is adopted to calculate the CV residual and to determine the stopping condition. As soon as the CV residual is smaller than a given constant, its corresponding recovered signal will therefore be outputted as the reconstructed signal. The above work is the first step that introduced cross validation into the field of compressive sensing.

Another important work in CS literature related to cross validation is made by Ward [14], who cleverly used the Johnson-Lindenstrauss (JL) Lemma to evaluate the recovery status. In this work, the reconstruction matrix is used for recovering the sparse signal and the cross validation matrix is used for estimating the reconstruction error. The dependence of the desired estimation accuracy and the confidence level in the prediction on the number of CV measurements is also studied. The above work offers us a tractable way for parameter selection in OMP-CV algorithm, which is theoretically analyzed in this paper.

1.2 Our Contribution

The main contribution of this work is two-fold. We first develop some general cross validation techniques for compressive sensing. Then the theoretical analysis on OMP-CV algorithm is conducted comprehensively.

The general cross validation techniques we provide could basically answer the following two questions, referred to as *general CV problems* in the reminder of this paper.

1. (Recovery error estimation) Given a reconstructed signal, with what accuracy and what probability could its CV residual provide bounds on its recovery error?

2. (Recovery error comparison) Given a pair of reconstructed signals, with what probability could the comparison between their CV residuals correctly evaluate their recovery errors?

To solve these problems, we first calculate the probability distribution of CV residuals. Consequently, by transforming the distribution into inequalities that hold with certain probability, we directly answer the above two questions.

Equipped with the general cross validation techniques, we then analyze the OMP-CV algorithm. We refer to the algorithm output, which is the reconstructed signal with the smallest CV residual, as the *OMP-CV output*. The reconstructed signal with the smallest recovery error is referred to as the *oracle output*.

Our analysis result shows that the recovery error of the OMP-CV output is very close to that of the oracle output with high probability, given that the oracle output recovers all indices in the support set of the original signal. In order to achieve the above result, we first analyze the internal structure of two recovered signals in different OMP iterations. We then study how their CV residuals affect the recovery errors using the techniques developed for the general CV problems. Finally we generalize the recovery error comparison between two recovered signals, which are generated in different OMP iterations, to that of all recovered signals. Therefore we could estimate how close the OMP-CV output is to the oracle output.

The reminder of the paper is organized as follows. Section 2 contains the problem formulation, OMP-CV description, and some mathematical tools required in the following analysis. Section 3 analyzes the cross validation techniques for compressive sensing while

Section 4 provides a comprehensive discussion on OMP-CV algorithm. Numerical simulations are given in Section 5 to verify the theoretical content. Concluding remarks are drawn in Section 6, while proofs of some theorems are presented in the Appendix.

2 Preliminaries

2.1 Notations

We consider an unknown k-sparse signal $\mathbf{x} \in \mathbb{R}^N$ observed using M linear measurements corrupted by additive noise. Let T be the support set of $\mathbf{x}$ and $|T| = k$ to denote the cardinality of T. The vector $\mathbf{x}_T$ contains the entries of $\mathbf{x}$ indexed by T. To implement the CV-based modification, we separate the original M by N sensing matrix to a reconstruction matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ and a CV matrix $\mathbf{A}_{\mathrm{cv}} \in \mathbb{R}^{m_{\mathrm{cv}} \times N}$. The measurement vector is also separated accordingly, to a reconstruction measurement $\mathbf{y} \in \mathbb{R}^m$ and a CV measurement $\mathbf{y}_{\mathrm{cv}} \in \mathbb{R}^{m_{\mathrm{cv}}}$. In this paper we only consider Gaussian sensing matrices and additive Gaussian noises. The reconstruction matrix $\mathbf{A}$ is properly normalized to have unit column norm. Because the same data acquisition system is assumed to be used to obtain both the reconstruction and CV measurements, the CV matrix $\mathbf{A}_{\mathrm{cv}}$ is normalized to have column norm equal to $\sqrt{m_{\mathrm{cv}}/m}$ and the CV noise has the same per measurement variance as the measurement noise. In other words, the notations can be formulated as

$$\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{n}, & \mathbf{n} &= \sigma_{\mathrm{n}}\mathbf{a}_{\mathrm{n}}, \\ \mathbf{y}_{\mathrm{cv}} &= \mathbf{A}_{\mathrm{cv}}\mathbf{x} + \mathbf{n}_{\mathrm{cv}}, & \mathbf{n}_{\mathrm{cv}} &= \sigma_{\mathrm{n}}\mathbf{a}_{\mathrm{cv,n}}, \end{aligned}$$

where the entries of $\mathbf{A}$, $\mathbf{A}_{\mathrm{cv}}$, $\mathbf{a}_{\mathrm{n}}$, and $\mathbf{a}_{\mathrm{cv,n}}$ are i.i.d normally distributed with mean zero and variance $1/m$.

2.2 Orthogonal Matching Pursuit with Cross Validation

The OMP-CV algorithm proposed in [10] is a noise- and sparsity-robust greedy recovery algorithm that adopts CV in OMP. In this algorithm, every iteration can be viewed as two separate parts: reconstructing the signal by OMP and evaluating the recovered signal by cross validation techniques, which is utilized to properly terminate the iteration before the recovery starts to overfit the noise. The OMP-CV algorithm that studied in this work is slightly different from its original version. One may refer to Table 1, where the iteratively reconstructed signal is chosen as output based on the criteria that its CV residual is the smallest rather than less than a certain constant.

In Table 1, we use $\hat{\mathbf{x}}^p$ and T^p to denote the recovered signal and its support, respectively, in the p-th iteration. The difference between the recovered signal $\hat{\mathbf{x}}^p$ and the input signal $\mathbf{x}$ is denoted using $\Delta\mathbf{x}^p$. The recovery error and the CV residual corresponding to $\hat{\mathbf{x}}^p$ are denoted by $\varepsilon_{\mathrm{x}}^p$ and ϵ_{cv}^p, respectively.

$$\Delta\mathbf{x}^p \triangleq \mathbf{x} - \hat{\mathbf{x}}^p, \qquad \varepsilon_{\mathrm{x}}^p \triangleq \|\Delta\mathbf{x}^p\|_2^2, \qquad \epsilon_{\mathrm{cv}}^p \triangleq \|\mathbf{y}_{\mathrm{cv}} - \mathbf{A}_{\mathrm{cv}}\hat{\mathbf{x}}^p\|_2^2.$$

Table 1: OMP-CV Algorithm

Input:	$\mathbf{A}, \mathbf{A}_{\text{cv}}, \mathbf{y}, \mathbf{y}_{\text{cv}}, d$;
Output:	$\hat{\mathbf{x}}$.
Initialization:	Set $p = 1$, $\epsilon_{\text{cv}}^0 = \|\mathbf{y}_{\text{cv}}\|_2^2$;
Repeat:	
	Compute $\hat{\mathbf{x}}^p$ using an OMP iteration;
	Compute $\epsilon_{\text{cv}}^p = \|\mathbf{A}_{\text{cv}}\hat{\mathbf{x}}^p - \mathbf{y}_{\text{cv}}\|_2^2$;
	Increment p by 1;
Until:	$p \geq d$
Compute $o = \underset{p}{\text{argmin}}\, \epsilon_{\text{cv}}^p$;	
Return: $\hat{\mathbf{x}} = \hat{\mathbf{x}}^o$.	

Here we are trying to present some intuition about how OMP-CV works. Please refer to Figure 1, which demonstrates the evolution by iteration of residual, CV residual, and recovery error. One may notice that the trend of residual in iteration behaves abruptly different comparing to that of recovery error, as soon as the reconstructed signal starts to overfit the noise. Therefore, residual fails to serve as an indicator for correctly terminating the algorithm. However, the CV residual evolves similarly as that of the recovery error. This is the reason that CV modification could improve the performance of OMP and other related greedy algorithms.

We would like to emphasize that OMP-CV is a highly practical algorithm. OMP-CV does not require prior information such as noise level or sparsity. Instead, only the maximum number of iterations is required as input[1]. Furthermore, OMP-CV provides an estimate of the recovery error in its CV residual. This is very helpful because it could be immediately detected if the algorithm did not recover the signal well. Finally, it is worthwhile to mention that by properly setting m_{cv}, the recovery performance of OMP-CV competes with that of OMP even when the accurate information of noise level is given for the latter. These advantages will be supported both theoretically and empirically in following part of this work.

2.3 Restricted Isometry Property

In CS literature, a large body of works focuses on the theoretical analysis of sparse recovery algorithm performance. Among these theoretical analysis, the RIP [15] becomes one of

[1] One may notice that the number of maximum iteration cannot be greater than that of the measurements, because regular OMP will produce zero residual after that and conclude.

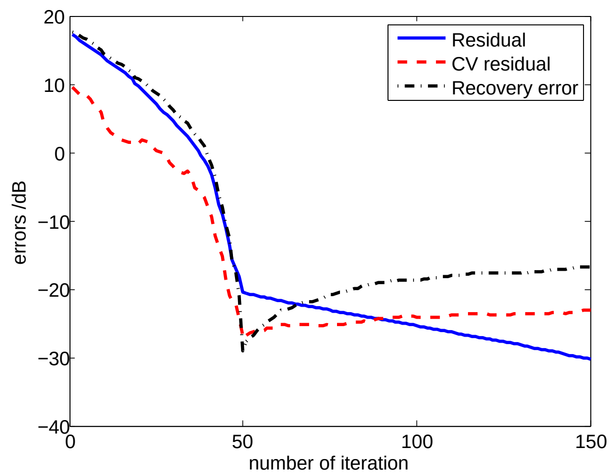


Figure 1: The evolution of residual, CV residual, and recovery error of OMP-CV.

the most helpful and widely used tools. It quantifies the idea that the geometry of sparse signals should be preserved under the mapping of sensing matrix. In this paper, the RIP is frequently used to study the internal structure of reconstructed signals acquired in different iterations of OMP.

Definition 1 *[8] (RIP): A sensing matrix* $\mathbf{A} \in \mathbb{R}^{m\times N}$ *is said to satisfy the Restricted Isometry Property with parameters* (k, δ) *for* $k \leq m$, $0 \leq \delta \leq 1$, *if for all index sets* $T \subset \{1, \cdots, N\}$ *such that* $|T| \leq k$ *and for all* $\mathbf{x} \in \mathbb{R}^N$, *one has*

$$(1-\delta)\|\mathbf{x}_T\|_2^2 \leq \|\mathbf{A}_T\mathbf{x}_T\|_2^2 \leq (1+\delta)\|\mathbf{x}_T\|_2^2. \quad (1)$$

We define δ_k, *the Restricted Isometry Constant (RIC), as the infimum of all parameters* δ *for which the RIP holds.*

Remark 1 *[8] (RIP and eigenvalues):If a sensing matrix* $\mathbf{A} \in \mathbb{R}^{m\times N}$ *satisfies the RIP with parameters* (k, δ_k), *then for all* $T \subset \{1, \cdots, N\}$ *such that* $|T| \leq k$, *it holds that*

$$1-\delta_k \leq \lambda_{\min}(\mathbf{A}_T'\mathbf{A}_T) \leq \lambda_{\max}(\mathbf{A}_T'\mathbf{A}_T) \leq 1+\delta_k, \quad (2)$$

where $(\cdot)', \lambda_{\min}(\cdot)$, *and* $\lambda_{\max}(\cdot)$ *denote the transpose, the minimal and maximal eigenvalues of a matrix, respectively.*

Most known families of matrices satisfying the RIP with optimal or near-optimal performance guarantees are random. Specifically, in this paper we will deal with Gaussian random matrix, which has the following property.

Remark 2 *[9] (Matrices satisfying the RIP): If the entries of* $\sqrt{m}\mathbf{A}$ *are independent and identically distributed standard normal variables and if*

$$m \geq Ck\log(N/k)/\tau^2, \quad (3)$$

where C is a constant, then, one has $\delta_k \le \tau$ except with probability N^{-1}.

If a sensing matrix satisfies the RIP, it has some other properties that are required in our analysis. The first one is a simple translation of Definition 1.

Lemma 1 *Suppose $\mathbf{A}$ has an RIP of δ_k and T is a set of k indices or fewer. For all $\mathbf{x} \in \mathbb{R}^N$, one has*

$$\sqrt{(1-\delta_k)}\|\mathbf{x}_T\|_2 \le \|\mathbf{A}_T\mathbf{x}_T\|_2 \le \sqrt{(1+\delta_k)}\|\mathbf{x}_T\|_2, \tag{4}$$

$$(1-\delta_k)\|\mathbf{x}_T\|_2 \le \|\mathbf{A}'_T\mathbf{A}_T\mathbf{x}_T\|_2 \le (1+\delta_k)\|\mathbf{x}_T\|_2, \tag{5}$$

$$\frac{1}{(1+\delta_k)}\|\mathbf{x}_T\|_2 \le \|(\mathbf{A}'_T\mathbf{A}_T)^{-1}\mathbf{x}_T\|_2 \le \frac{1}{(1-\delta_k)}\|\mathbf{x}_T\|_2. \tag{6}$$

A second consequence is that the disjoint sets of columns from the sensing matrix span nearly orthogonal subspaces. To quantify this observation, we have the following results.

Lemma 2 *[9] (Approximate orthogonality): Suppose $\mathbf{A}$ has a RIC of $\delta_{|S|+|T|}$, where $S, T \subset \{1, 2, \cdots, N\}$ are disjoint sets. One has*

$$\|\mathbf{A}'_S\mathbf{A}_T\|_2 \le \delta_{|S|+|T|}. \tag{7}$$

Lemma 3 *[9] Let $S, T \subset \{1, 2, \cdots, N\}$ be two disjoint sets and suppose that $\delta_{|S|+|T|} < 1$. For all $\mathbf{x} \in \mathbb{R}^N$, one has*

$$\|\mathbf{A}'_S\mathbf{A}_T\mathbf{x}_T\|_2 \le \delta_{|S|+|T|}\|\mathbf{x}_T\|_2. \tag{8}$$

Based on the Lemma 1 and Lemma 3, we derived the following lemma.

Lemma 4 *Let $S, T \subset \{1, 2, \cdots, N\}$ be two disjoint sets and suppose that $\delta_{|S|+|T|} < 1$. For all $\mathbf{x} \in \mathbb{R}^N$, one has*

$$\|\mathbf{A}^\dagger_S\mathbf{A}_T\mathbf{x}_T\|_2 \le \frac{\delta_{|S|+|T|}}{1-\delta_{|S|}}\|\mathbf{x}_T\|_2. \tag{9}$$

Proof The proof of Lemma 4 is postponed to Appendix A. ∎

Before the statement of the last property, we make a definition of an orthogonal projection operator P_T.

Definition 2 *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ and $T \subset \{1, \cdots, N\}$. Define an operation P_T as*

$$P_T \triangleq \mathbf{I} - \mathbf{A}_T\mathbf{A}^\dagger_T, \tag{10}$$

where $(\cdot)^\dagger$ denotes the pseudo-inverse of a matrix.

Remark 3 *P_T is an orthogonal projection operator whose function is to remove the component of a vector that is in the space spanned by $\mathbf{A}_T$.*

Lemma 5 *[16] Let $S, T \subset \{1, 2, \cdots, N\}$ be two disjoint sets and suppose that $\delta_{|S|+|T|} < 1$. For all $\mathbf{x} \in \mathbb{R}^N$, one has*

$$\sqrt{1-\left(\frac{\delta_{|S|+|T|}}{1-\delta_{|S|+|T|}}\right)^2}\|\mathbf{A}_T\mathbf{x}_T\|_2 \le \|P_S\mathbf{A}_T\mathbf{x}_T\|_2 \le \|\mathbf{A}_T\mathbf{x}_T\|_2. \tag{11}$$

3 Cross Validation in Compressive Sensing

The results of general cross validation techniques in compressive sensing are presented in this section. These techniques will be utilized in the later analysis of OMP-CV algorithm. We would like to emphasize that besides the application of analyzing OMP-CV algorithm, the content in this section is fundamentally general and can be used to understand other CV-based sparse recovery algorithms as well.

3.1 Recovery Error Estimation

First, we start with calculating the probability distribution of ϵ_{cv}, which is described by the following lemma.

Lemma 6 *Let* $\hat{\mathbf{x}}$ *be the recovered signal. Assuming there are enough measurements for cross validation, one has*

$$\epsilon_{\text{cv}} = \|\mathbf{y}_{\text{cv}} - \mathbf{A}_{\text{cv}}\hat{\mathbf{x}}\|_2^2 \sim \mathcal{N}(\mu, \sigma^2), \tag{12}$$

where $\mu = \frac{m_{\text{cv}}}{m}(\varepsilon_{\text{x}} + \sigma_{\text{n}}^2), \sigma^2 = \frac{2m_{\text{cv}}}{m^2}(\varepsilon_{\text{x}} + \sigma_{\text{n}}^2)^2$*, and* $\varepsilon_{\text{x}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$*.* $\mathbf{y}_{\text{cv}}, \mathbf{A}_{\text{cv}}, m, m_{\text{cv}}$*, and* σ_{n}^2 *are defined in section 2.1.*

Proof The proof is postponed to Appendix B. ∎

The condition "there are enough measurements for cross validation" is required in one step of the proof of Lemma 6, which is the probability distribution approximation via Central Limit Theorem (CLT). According to Central Limit Theorem, the real probability distribution of ϵ_{cv} converges absolutely to the above approximated result with the increase of m_{cv}. In fact, the approximation is rather good when m_{cv} is greater than tens. Considering that compressive sensing always deals with large scale problems, such condition could be readily satisfied. This lemma will be verified by simulation result in section 5.1.

One immediate consequence of Lemma 6 is that we can use ϵ_{cv} to provide estimation for ε_{x} in term of an inequality that holds with certain probability.

Theorem 1 *(CV estimation): Assuming there are enough measurements for cross validation, the following inequality holds with probability* $\text{erf}(\lambda/\sqrt{2})$

$$h(\lambda, +)\epsilon_{\text{cv}} - \sigma_{\text{n}}^2 \le \varepsilon_{\text{x}} \le h(\lambda, -)\epsilon_{\text{cv}} - \sigma_{\text{n}}^2, \tag{13}$$

where

$$h(\lambda, \pm) \triangleq \frac{m}{m_{\text{cv}}} \frac{1}{1 \pm \lambda\sqrt{2/m_{\text{cv}}}}, \tag{14}$$

λ *is a parameter, and* $\text{erf}(u) \triangleq \frac{1}{\sqrt{\pi}}\int_{-u}^{u} \mathrm{e}^{-t^2}\mathrm{d}t$ *denotes the error function of standard Gaussian distribution.*

Proof According to Lemma 6 and the properties of Gaussian distribution, the result can by derived after some basic algebra. ■

Theorem 1 basically solves the first general CV problem that one can estimate the interval of recovery error ε_{x} by the observed CV residual ϵ_{cv} with probability $\mathrm{erf}(\lambda/\sqrt{2})$. One may notice that the width of the interval is

$$\frac{m}{m_{\mathrm{cv}}}\frac{2\lambda\sqrt{2}}{\sqrt{m_{\mathrm{cv}}}-2\lambda^2/\sqrt{m_{\mathrm{cv}}}}\epsilon_{\mathrm{cv}}, \tag{15}$$

which means that the bounds become tighter with the increase of the number of measurements used for cross validation. In particular, if m_{cv} is far larger than $2\lambda^2$, one could accurately estimate ε_{x} by the estimator $\frac{m}{m_{\mathrm{cv}}}\epsilon_{\mathrm{cv}} - \sigma_{\mathrm{n}}^2$. For example, with $\lambda = 3$, $m = 400$, and $m_{\mathrm{cv}} = 80$, it holds with probability 99.4% that

$$3.4\epsilon_{\mathrm{cv}} - \sigma_{\mathrm{n}}^2 \le \varepsilon_{\mathrm{x}} \le 9.5\epsilon_{\mathrm{cv}} - \sigma_{\mathrm{n}}^2. \tag{16}$$

3.2 Recovery Error Comparison

In the second general CV problem, we try to compare two recovered signals, $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^q$. According to Theorem 1, if $\varepsilon_{\mathrm{x}}^p$ is larger than $\varepsilon_{\mathrm{x}}^q$, the CV residuals should also have $\epsilon_{\mathrm{cv}}^p > \epsilon_{\mathrm{cv}}^q$ with high probability. Therefore, we could be able to compare the recovery errors by simply comparing their CV residuals. This section presents the mathematical formulation of this probability.

For simplicity, we have the following definition.

Definition 3 *(Generalized reconstrction matrix and input signal): Let* $\mathbf{A}_{\mathrm{g}} \triangleq [\mathbf{A}, \mathbf{a}_{\mathrm{n}}]$ *and* $\mathbf{x}_{\mathrm{g}} \triangleq [\mathbf{x}', \sigma_{\mathrm{n}}]'$*, then*

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} = \mathbf{A}_{\mathrm{g}}\mathbf{x}_{\mathrm{g}}, \tag{17}$$

where $\mathbf{A}_{\mathrm{g}}$ *is called the generalized reconstruction matrix, and* $\mathbf{x}_{\mathrm{g}}$ *is called the generalized input signal.*

The purpose of making this definition is to simplify our analysis. Compared to the original input signal $\mathbf{x}$, $\mathbf{x}_{\mathrm{g}}$ has an extra term which represents the noise and can never be recovered. It can be understood as a generalized version of input signal with a part that reflects the measurement noise. The generalized versions of $\Delta\mathbf{x}^p$ and $\varepsilon_{\mathrm{x}}^p$ are $\Delta\mathbf{x}_{\mathrm{g}}^p$ and $\varepsilon_{\mathrm{g}}^p$, respectively.

$$\Delta\mathbf{x}_{\mathrm{g}}^p \triangleq [(\Delta\mathbf{x}^p)', \sigma_{\mathrm{n}}]', \qquad \varepsilon_{\mathrm{g}}^p \triangleq \|\Delta\mathbf{x}_{\mathrm{g}}^p\|_2^2.$$

We start with calculating the probability distribution of $\Delta\epsilon_{\mathrm{cv}} = \epsilon_{\mathrm{cv}}^p - \epsilon_{\mathrm{cv}}^q$.

Lemma 7 *Let* $\hat{\mathbf{x}}^p$ *and* $\hat{\mathbf{x}}^q$ *be two recovered signals. Assuming there are enough measurements for cross validation, one has*

$$\Delta\epsilon_{\mathrm{cv}} = \epsilon_{\mathrm{cv}}^p - \epsilon_{\mathrm{cv}}^q \sim \mathcal{N}(\mu, \sigma^2), \tag{18}$$

where $\mu = \frac{m_{\text{cv}}}{m}(\varepsilon_{\text{g}}^p - \varepsilon_{\text{g}}^q)$ *and* $\sigma^2 = \frac{2m_{\text{cv}}}{m^2}[(\varepsilon_{\text{g}}^p)^2 + (\varepsilon_{\text{g}}^q)^2 - 2\rho_{\text{g}}^2\varepsilon_{\text{g}}^p\varepsilon_{\text{g}}^q]$, *where*

$$\rho_{\text{g}} \triangleq \frac{\langle \Delta\mathbf{x}_{\text{g}}^p, \Delta\mathbf{x}_{\text{g}}^q \rangle}{\|\Delta\mathbf{x}_{\text{g}}^p\|_2 \|\Delta\mathbf{x}_{\text{g}}^q\|_2} \tag{19}$$

denotes the correlation coefficient of the two generalized recovery error signals.

Proof The proof of Lemma 7 is deferred to Appendix C. ■

With Lemma 7, we are much closer to the answer to the key question, which is presented in the following theorem.

Theorem 2 *(CV comparison) Let* $\hat{\mathbf{x}}^p$ *and* $\hat{\mathbf{x}}^q$ *be two recovered signals and assume there are enough measurements for cross validation. If* $\varepsilon_{\text{x}}^p \geq \varepsilon_{\text{x}}^q$, *it holds with probability* $\Phi(\lambda)$ *that* $\epsilon_{\text{cv}}^p \geq \epsilon_{\text{cv}}^q$, *where* λ *is determined by*

$$\frac{1}{\lambda^2} = \frac{2}{m_{\text{cv}}}\left[1 + 2(1-\rho_{\text{g}}^2)\frac{\varepsilon_{\text{g}}^p\varepsilon_{\text{g}}^q}{(\varepsilon_{\text{g}}^p - \varepsilon_{\text{g}}^q)^2}\right], \tag{20}$$

and $\Phi(u) \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} \mathrm{e}^{\frac{-t^2}{2}}\mathrm{d}t$ *is the cumulative distribution function of standard Gaussian distribution.*

Proof According to Lemma 7 and the properties of Gaussian distribution, the result can by derived after some basic algebra. ■

$\Phi(\lambda)$ is called *CV comparison success probability*, because it is the probability that the size order of CV residuals correctly evaluates that of the recovery errors. Next we discuss the parameters that determines λ.

- Parameter m_{cv}:
  When m_{cv} increases, $\frac{1}{\lambda^2}$ decreases, thus λ increases. Hence, a larger m_{cv} is required to obtain a larger λ. This corresponds to our intuition: m_{cv} is the number of CV measurements and the more CV measurements we have, the better CV performance, which is shown as a higher CV comparison success probability, we should be able to obtain.

- Parameter ρ_{g}:
  The value of λ is in positive relation with ρ_{g}^2. We would like to note that ρ_{g} describes the correlation of $\Delta\mathbf{x}_{\text{g}}^p$ and $\Delta\mathbf{x}_{\text{g}}^q$. As an interpretation, CV comparison success probability is often higher if the two recovered signals need to be compared are highly correlated. One possible explanation for this fact is that the similar part of the signal yields similar part in CV measurements and the randomness of the comparison problem comes from the dissimilar part of the recovered signal. From this observation we would like to emphasize that the CV comparison for two highly correlated signals is often of a greater success probability.

- Parameter $\varepsilon_{\rm g}^p$ and $\varepsilon_{\rm g}^q$:
  The influence of $\varepsilon_{\rm g}^p$ and $\varepsilon_{\rm g}^q$ is depicted by the following theorem:

Theorem 3 *Let the CV measurement number $m_{\rm cv}$ and the similarity level $\rho_{\rm g}$ be fixed and assume there are enough measurements for cross validation, CV comparison success probability is equal to or higher than $\Phi(\lambda_0)$ if and only if the ratio of generalized signal error $\varepsilon_{\rm g}^p/\varepsilon_{\rm g}^q$ satisfies*

$$\frac{\varepsilon_{\rm g}^p}{\varepsilon_{\rm g}^q} \geq 2C_0 + 1 + 2\sqrt{C_0^2 + C_0}, \tag{21}$$

where $C_0 \triangleq \frac{\lambda_0^2(1-\rho_{\rm g}^2)}{m_{\rm cv}-2\lambda_0^2}$ is a constant related to $m_{\rm cv}$, λ_0 and $\rho_{\rm g}$.

This Theorem can be derived from Lemma 7 and Theorem 2.

Proof In this proof we just prove that if $\Phi(\lambda) \geq \Phi(\lambda_0)$, it holds that

$$\frac{\varepsilon_{\rm g}^p}{\varepsilon_{\rm g}^q} \geq 2C_0 + 1 + 2\sqrt{C_0^2 + C_0}. \tag{22}$$

The counterpart of the proof is extremely similar so we do not present it here. Since the CV comparison success probability is higher than $\Phi(\lambda_0)$ and the function $\Phi(u)$ is a monotonically increasing function, it holds that $\lambda > \lambda_0$. In addition, because λ and λ_0 are both positive, then

$$\frac{1}{\lambda_0^2} \geq \frac{1}{\lambda^2} = \frac{2}{m_{\rm cv}}\left[1 + 2(1-\rho_{\rm g}^2)\frac{\varepsilon_{\rm g}^p\varepsilon_{\rm g}^q}{(\varepsilon_{\rm g}^p - \varepsilon_{\rm g}^q)^2}\right]. \tag{23}$$

Further

$$\frac{\varepsilon_{\rm g}^p}{\varepsilon_{\rm g}^q} + \frac{\varepsilon_{\rm g}^q}{\varepsilon_{\rm g}^p} \geq \frac{4\lambda_0^2}{m_{\rm cv} - 2\lambda_0^2}(1-\rho_{\rm g}^2) + 2. \tag{24}$$

Because $\varepsilon_{\rm g}^p \geq \varepsilon_{\rm g}^q$, the left part of the above inequality monotonically increases with $\varepsilon_{\rm g}^p/\varepsilon_{\rm g}^q$. Thus

$$\frac{\varepsilon_{\rm g}^p}{\varepsilon_{\rm g}^q} \geq \frac{2\lambda_0^2(1-\rho_{\rm g}^2)}{m_{\rm cv} - 2\lambda_0^2} + 1 + \sqrt{\frac{4\lambda_0^4(1-\rho_{\rm g}^2)^2}{(m_{\rm cv} - 2\lambda_0^2)^2} + \frac{4\lambda_0^2(1-\rho_{\rm g}^2)}{m_{\rm cv} - 2\lambda_0^2}}. \tag{25}$$

Using the notation of C_0 we may complete the proof:

$$\frac{\varepsilon_{\rm g}^p}{\varepsilon_{\rm g}^q} \geq 2C_0 + 1 + 2\sqrt{C_0^2 + C_0}. \tag{26}$$

where

$$C_0 = \frac{\lambda_0^2(1-\rho_{\rm g}^2)}{m_{\rm cv} - 2\lambda_0^2}. \tag{27}$$

■

Remark 4 *The reason that $\rho_{\rm g}$ should be discussed can be explained as follows. For two recovered signals with a high correlation coefficient, their CV comparison success probability is often high. Hence, if we know the lower bound of $\rho_{\rm g}$ in advance, a much better CV recovery error comparison performance can be guaranteed. This is also the case of the OMP-CV* *algorithm. In OMP algorithm, a new index is incorporated into the current support set in each iteration and the recovered signal is determined by the current support set. Thus, the recovered signals of neighboring iterations have an extremely high correlation coefficient. As a result, the CV comparison success probability is very high when we compare the recovery errors of these recovered signals.*

The above analysis solves the recovery error estimation problem and the recovery error comparison problem. These results provide powerful tools for CV-based compressive sensing algorithms. Particularly, we studied the application of cross validation in OMP algorithm. The results are presented in following section.

4 Sparsity and Noise Robust OMP

This section presents our theoretical analysis of the behavior of OMP-CV. First make the definitions:

Definition 4 *(Ratio of unrecovered signal and noise): The ratio $\alpha^p \in \mathbb{R}$, defined as*

$$\alpha^p \triangleq \frac{\|\mathbf{x}_{T\setminus T^p}\|_2}{\sigma_{\mathrm{n}}}, \tag{28}$$

measures to what extent the signal $\hat{\mathbf{x}}$ has not been recovered by $\hat{\mathbf{x}}^p$.

Definition 5 *(Oracle output and OMP-CV output): In OPM-CV, by* oracle output *we mean the recovered signal that has the lowest recovery error. By* OMP-CV output *we mean the recovered signal with the lowest CV residual.*

Intuitively, among recovered signals generated in different OMP iterations, the oracle output is most likely to be the OMP-CV output, which does not held all the time due to the randomness of the CV matrix. This section gives result describing how close the OMP-CV output is to the oracle output, i.e.

Theorem 4 *In OMP-CV, assume that the oracle output is $\hat{\mathbf{x}}^o$ and $T \subset T^o$. For any recovered signal $\hat{\mathbf{x}}^p$ other than $\hat{\mathbf{x}}^o$:*

- *if $T\setminus T^p \neq \emptyset$, then $\epsilon_{\mathrm{cv}}^o < \epsilon_{\mathrm{cv}}^p$ with probability $\Phi(\lambda)$, where*

$$\lambda \geq \sqrt{\frac{m_{\mathrm{cv}}}{2}}\sqrt{1-\mathrm{g}(\alpha^{\mathrm{p}})}; \tag{29}$$

- *if $T\setminus T^p = \emptyset$ and if $\hat{\mathbf{x}}^p$ is the OMP-CV output, then with probability greater than $\{1-(d-k)[1-\Phi(\lambda_0)]\}$ we have*

$$\varepsilon_{\mathrm{g}}^p \leq C_1 \varepsilon_{\mathrm{g}}^o, \tag{30}$$

where $\mathrm{g}(\alpha^{\mathrm{p}})$ is roughly proportional to $1/(\alpha^p)^2$ and C_1 is related to λ_0 and m_{cv}.

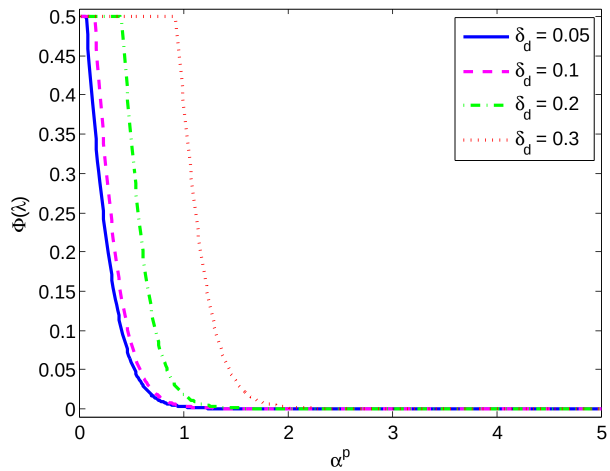


Figure 2: This figure plots $[1 - \Phi(\lambda)]$, the upper bound of the probability that $\epsilon_{\text{cv}}^p < \epsilon_{\text{cv}}^o$, with variation of α^p and RIP constant δ_d. $[1 - \Phi(\lambda)]$ decays sharply as α^p increases.

PROOF One may refer to Appendix 11 for the proof of Theorem 4. ■

Theorem 4 supports the recovery performance of OMP-CV. It divides recovered signals into two categories. For $\hat{\mathbf{x}}^p$ with $T \backslash T^p \neq \emptyset$, the probability $[1 - \Phi(\lambda)]$ decays sharply as α^p increases. Since this is the upper bound of the probability that $\epsilon_{\text{cv}}^p < \epsilon_{\text{cv}}^o$, it is nearly impossible for such $\hat{\mathbf{x}}^p$ to be the OMP-CV output. If $\delta_d < 0.1$, e.g., $[1 - \Phi(\lambda)]$ would be less than 0.5% with $\alpha^p = 1$, and further drop to 0.0063% with $\alpha^p = 2$. The probability $[1 - \Phi(\lambda)]$ is shown in Figure 2 with $m_{\text{cv}} = 48$ and variation of α^p and δ_d. As a result, the OMP-CV output recovers all indices of the support set with overwhelming probability. Further, if the OMP-CV output recovers all indices of the support set, its recovery error can be bounded by ε_{g}^o with high probability showing that the recovery error of OMP-CV output is very close to that of the oracle output.

Remark 5 *(The rationality of assuming the oracle output recovers all indices in the support set, i.e. $T \subset T^o$) Assume otherwise $T \backslash T^o \neq \emptyset$. Since in OMP-CV, the support of the current iteration contains all indices of the support of its previous iterations,*

- *if there exists a recovered signal $\hat{\mathbf{x}}^p$ such that $(T^p \backslash T^o) \cap T \neq \emptyset$, it is nearly impossible that $\epsilon_{\text{cv}}^o < \epsilon_{\text{cv}}^p$ due to inequality (29). Therefore $\hat{\mathbf{x}}^o$ is nearly impossible to be the oracle output, which contradicts with statements of Theorem 4.*

- *if $\hat{\mathbf{x}}^o$ has maximum number of indices of the support T among all recovered signals, then other indices of T is not incorporated during d times of iteration. In this case, $\mathbf{A}_{T \backslash T^o} \hat{\mathbf{x}}_{T \backslash T^o}$ acts as the same role as noise $\mathbf{a}_{\text{n}} \sigma_{\text{n}}$. They can by treated as noise and similar analysis can be conducted by changing σ_{n}^2 to $(\sigma_{\text{n}}^2 + \|\hat{\mathbf{x}}_{T \backslash T^o}\|_2^2)$ and properly modifying the footnote of the RIPs.*

Remark 6 *(Parameter details and setting) Parameter details of Theorem 4 are:*

$$\begin{aligned} \mathrm{g}(\alpha^{\mathrm{p}}) =&[\beta_1(\alpha^p)^2+\beta_2]/[\beta_1(\alpha^p)^2+\beta_2+\max((\alpha^p)^2-\beta_3\alpha^p-\beta_4,0)^2] \approx \beta_1/\left[(\alpha^p)^2+\beta_1\right],\\ C_1 =&2C_0+1+2\sqrt{C_0^2+C_0},\\ C_0 \leq&\beta_5\lambda_0^2/(m_{\mathrm{cv}}-2\lambda_0^2), \end{aligned}$$

where betas are decided by RIP constant [4] of the sensing matrix. β_1 *is far larger than* β_2*,* β_3*, and* β_4*. E.g., if* $\delta_d < 0.1$*, the values of betas are:* $\beta_1 = 2.08$*,* $\beta_2 = \beta_3 = 0.03$*,* $\beta_4 = 0.02$*, and* $\beta_5 = 0.0376$*.*

For parameter setting, m_{cv} *does not need to be a very large number to attain a promising performance. As* C_0 *is proportional to* $1/(m_{\mathrm{cv}} - 2\lambda_0^2)$*, the value of* C_0 *becomes extremely small when* m_{cv} *becomes a little larger than* $2\lambda_0^2$*; meanwhile,* C_0 *does not decay much when* m_{cv} *largely exceeds* $2\lambda_0^2$*. Therefore* m_{cv} *should be properly chosen to be a little larger than* $2\lambda_0^2$*. As for* λ_0*, the probability* $\{1-(d-k)[1-\Phi(\lambda_0)]\}$ *increase significantly with the increase of* λ_0*. For example, it decays by approximately 100 times with* λ_0 *increases by 1. For example, when* $d-k = 100$*, by setting* $\lambda_0 = 4$*, we attain the probability* $1-(d-k)(1-\Phi(\lambda_0)) = 99.7\%$*. When* $d - k$ *increase to 10000, we can attain the similar probability of* 99.4% *by merely increase* λ *by* 1 *to be* 5*.*

If, e.g., $(d-k) = 100$*, setting* $m_{\mathrm{cv}} = 48$ *and* $\lambda_0 = 4$*, produces a numerical form of* (30)*: with probability* 99.7% *we have*

$$\varepsilon_{\mathrm{g}}^p \leq 1.47\varepsilon_{\mathrm{g}}^o. \tag{31}$$

Apart from Theorem 4, we also note that the recovery error of OMP-CV output can be estimated in its CV residual in practical cases. This is achieved via a direct application of Theorem 1:

Remark 7 *(OMP-CV output estimation) Let* $\hat{\mathbf{x}}^p$ *be the OMP-CV output with CV residual* ϵ_{cv}^p *and assume there are enough measurements for cross validation. It holds with probability* $\mathrm{erf}(\frac{\lambda}{\sqrt{2}})$ *that,*

$$h(\lambda,+)\epsilon_{\mathrm{cv}}^p - \sigma_{\mathrm{n}}^2 \leq \varepsilon_{\mathrm{x}}^p \leq h(\lambda,-)\epsilon_{\mathrm{cv}}^p - \sigma_{\mathrm{n}}^2, \tag{32}$$

where $h(\lambda,\pm)$ *is defined in Theorem 1.*

In OMP-CV, therefore, the estimation of the recovery error of the OMP-CV output is available to us in practical cases. If the recovery performance is not as good as expected, some measures could be conducted to improve the recovery performance, e.g., adding more measurements.

5 Numerical Simulation

This section gives simulation results concerning our theoretical analyses. Subsection 5.1 validates Lemma 6 and Lemma 7 respectively, upon which our general CV techniques were developed. Subsection 5.2 simulates the numerical example mentioned in Section 4, providing evidence supporting the rationality of Theorem 4. In Subsection 5.3, the behaviors of OMP-CV are studied with variation of different parameters.

5.1 Validation for Lemma 6 and Lemma 7

As previously noted, CLT is used to approximate the probability distribution of ϵ_{cv} in Lemma 6 and $\Delta\epsilon_{\mathrm{cv}}$ in Lemma 7. The simulations in this section attempt to validate this approximation. In both simulations, parameters are set as $N = 512$, $m = 96$, $m_{\mathrm{cv}} = 48$ and $k = 50$. With recovered signals ($\hat{\mathbf{x}}$ for Lemma 6; $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^q$ for Lemma 7) fixed, the random CV matrix along with its noise is realized 1E5 times and the probability distributions of random variables (ϵ_{cv} for Lemma 6; $\Delta\epsilon_{\mathrm{cv}}$ for Lemma 7) are calculated. The experiment results, shown in Figure3, indicate that the simulation results agree well with the theoretical prediction. This validates our approximation and supports both lemmas.

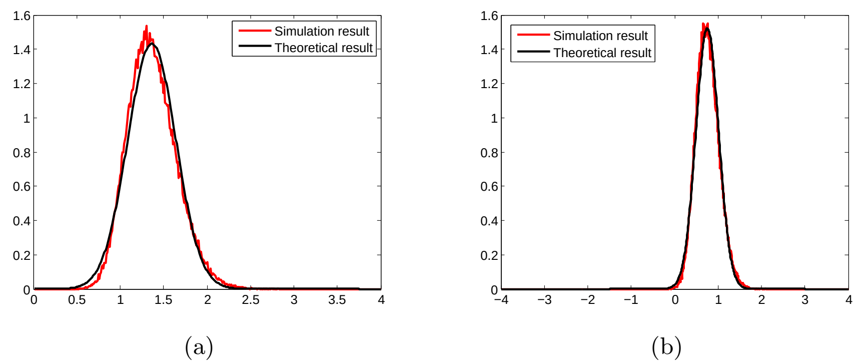
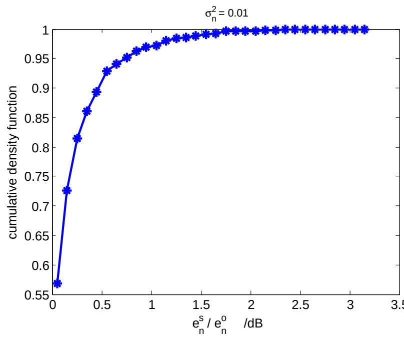


Figure 3: Validation for Lemma 6 and Lemma 7. Figure3(a) validates Lemma 6 by plotting the simulation result of probability distribution of ϵ_{cv}, while Figure3(b) validates Lemma 7 by plotting that of $\Delta\epsilon_{\mathrm{cv}}$ (red curve). The simulation results agree well with the theoretical results, which are plotted in black as reference.

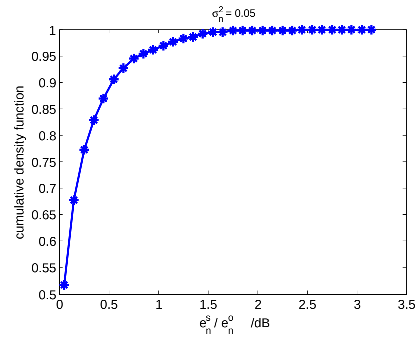
5.2 Validation for Theorem 4

This experiment simulates the numerical example mentioned in Section 4. Parameters are set as $N = 1000$, $m = 400$, $m_{\mathrm{cv}} = 48$, and $k = 50$. The iteration times is set as $d = 150$ such that $d - k = 100$, which corresponds to the example in Section 4. In this situation, as illustrated in Section 4, with Theorem 4 one has that with high probability the OMP-CV output recovers all indices of the support set and if the OMP-CV output recovers all indices of the support set, it holds with probability more than 99.7% that

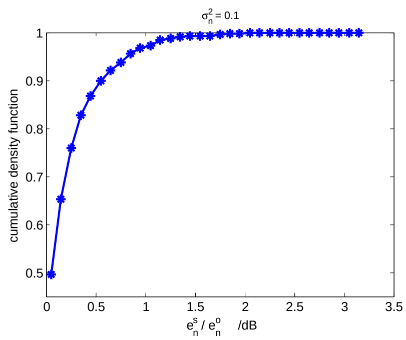
$$\varepsilon_{\mathrm{g}}^{p} \leq 1.47\varepsilon_{\mathrm{g}}^{o} \tag{33}$$



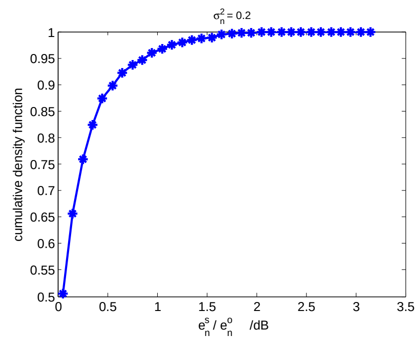
(a) $\sigma_{\rm n}^2 = 0.01$



(b) $\sigma_{\rm n}^2 = 0.05$



(c) $\sigma_{\rm n}^2 = 0.1$



(d) $\sigma_{\rm n}^2 = 0.2$

Figure 4: Cumulative distribution of $\varepsilon_{\rm g}^p/\varepsilon_{\rm g}^o$. Parameters are $N = 1000$, $m = 400$, $m_{\rm cv} = 48$, $k = 50$, and $d - k = 100$, where we have with high probability it holds that $\varepsilon_{\rm g}^p \leq 1.47\varepsilon_{\rm g}^o$ according to Theorem 4. The percentage that the error ratio does not exceeds the bound 1.47 (1.67 dB) are 99.1%, 99.5%, 99.3%, 99.5% for noise level 0.01, 0.05, 0.1, 0.2 respectively, which validates Theorem 4.

where $\hat{\mathbf{x}}^p$ is the OMP-CV output and $\hat{\mathbf{x}}^o$ is the oracle output. The simulation result is shown in Figure 4, where dB is used as our x axis unit and 1.47 is $1.67dB$.

From the result we can see that with overwhelming probability the error ratio $\varepsilon_{\rm g}^p/\varepsilon_{\rm g}^o$ is smaller than $1.67dB$ and in most cases it remains pretty low. More precisely, the percentage that the error ratio does not exceeds the bound 1.67 are 99.1%, 99.5%, 99.3%, 99.5% for noise level 0.01, 0.05, 0.1, 0.2 respectively. Those are approximately the theoretical probability in our theorem, 99.7%. The actual proportion is a little bit smaller than the theoretical probability because there are a few times that OMP-CV output fails to recover all indices of the support set, making the error ratio a little larger.

In addition, although the recovery performance deteriorates with the increase of noise level, the frequency distribution of error ratio $\varepsilon_{\rm g}^p/\varepsilon_{\rm g}^o$ does not vary. This validates our analysis, which states that the upper bound of error ratio stay unchanged when $m_{\rm cv}$ and λ_0 are fixed, where λ_0 is related to the value of $(d - k)$.

5.3 Simulation for OMP-CV Algorithm

This subsection investigates the behaviors of OMP-CV with variation of different parameters. The performance of OMP, referred to in this section as *OMP-residual*, is also given accordingly as reference. The stopping criteria of OMP-residual is based on its residual, i.e. terminating the iteration as long as

$$\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_2^2 < \sigma_n^2. \tag{34}$$

We would like to note that the prior knowledge of noise level is required in OMP-residual but not in OMP-CV. For each parameter setting, algorithms are repeated 1000 times.

5.3.1 Number of CV measurements

This experiment investigates the effect of additional cross validation measurements. Parameters are set as $N = 1000$, $k = 50$, and $\sigma_n^2 = 0.1$. The number of reconstruction measurement m is fixed to be 360 and the number of CV measurements m_{cv} varies from 10 to 80 with step 10. The experiment result is shown in Figure 5, from which we can see with the increase of m_{cv} improves the recovery performance significantly. We can also see the improvement of the recovery performance becomes slow when m_{cv} exceeds 60. Thus, it is not necessary for m_{cv} to be very large to obtain a satisfying recovery performance.

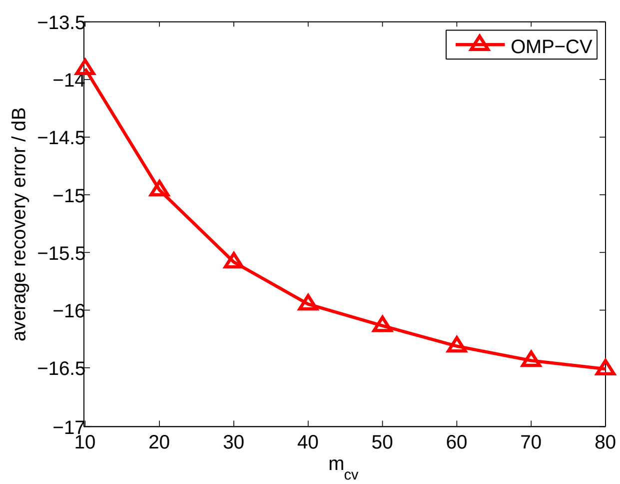


Figure 5: The effect of additional CV measurements. Parameters are $N = 1000$, $k = 50$, $m = 360$, and $\sigma_n^2 = 0.1$. m_{cv} varies from 10 to 80 with step 10. The experiment result shows that the increase of m_{cv} improves the recovery performance significantly and the performance improvement becomes slow when m_{cv} exceeds 60. Thus, it is not necessary for m_{cv} to be very large to obtain a satisfying recovery performance.

5.3.2 Trade Off between m and m_{cv}

Given a fixed total number of measurements M, there is a tradeoff between m and m_{cv} [10]. On one hand, increasing m will reduce the reconstruction error. On the other hand, increasing m_{cv} will improve the CV estimate, and thus make the OMP-CV output closer to the oracle output. This simulation empirically investigates the recovery performance of OMP-CV as m_{cv} varies.

In this experiment, we set $N = 1000$, $k = 50$, $\sigma_n^2 = 0.1$ and $M = 400$. m_{cv} varies from 10 to 80 with step 10 and we have $m = M - m_{\text{cv}}$. For OMP-CV, m measurements are used for reconstruction, while m_{cv} measurements are used for CV. For OMP-residual, m measurements are used for reconstruction and the termination is based on residual with the accurate noise level given. In addition, the recovery performance, where all M measurements are used for reconstruction using OMP-residual, is given for comparison. We average 1000 repetitions for experiments of each parameter setting.

The experiment result, plotted in Figure 6, shows the best performance of OMP-CV lies in the region where m_{cv} is neither too small nor too large. OMP-CV outperforms OMP-residual except when m_{cv} is very small, indicating that CV-based termination is better than residual-based termination, even if the latter uses an exact noise level. Additionally, note that with the same number of measurements at hand (M measurements for both OMP-CV and OMP-residual), OMP-CV can achieve recovery performance similar to OMP-residual with parameters appropriately set, even though prior knowledge is required for OMP-residual. In this sense, OMP-CV outperforms OMP-residual.

5.3.3 The Performance of OMP-CV under Different Noise Level

This experiment investigates the recovery performance of OMP-CV under different noise level. In addition to the reference performance of OMP-residual, the performance of OMP with no prior knowledge, referred to as *OMP-without-prior-knowledge*, is given as baseline. For OMP-without-prior-knowledge, the stopping criteria [17] is to terminate the iteration either when iteration times exceeds M or when

$$\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_2 < 10^{-5}\|\mathbf{y}\|_2. \tag{35}$$

The number of total measurements M is fixed to be 400. For OMP-CV, $m = 352$ measurements are used for reconstruction while $m_{\text{cv}} = 48$ measurements for cross validation. For OMP-residual and OMP-without-prior-knowledge, all M measurements are used for reconstruction. The noise level σ_n^2 varies from 0.02 to 0.2 with step 0.02. Other parameters are $N = 1000$, and $k = 50$. The experiment result is shown in Figure 7.

From Figure 7 we can see that OMP-CV and OMP-residual have similar performance under different noise level even when the prior knowledge is available for OMP-residual. We also observe that the performance of OMP largely deteriorates when information of noise

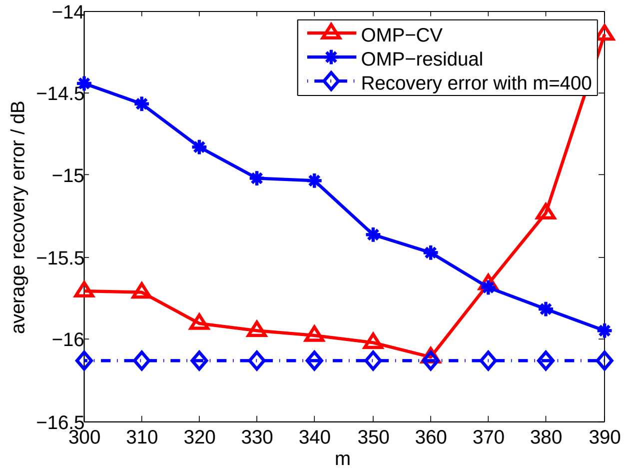


Figure 6: The trade Off between m and m_{cv}. Parameters are fixed as $N = 1000$, $k = 50$, $\sigma_n^2 = 0.1$, and $M = 400$. m_{cv} varies from 10 to 80 with step 10 while $m = M - m_{\mathrm{cv}}$. OMP-CV outperforms OMP except when m_{cv} is too small. In addition, using same number of measurements, OMP-CV has recovery performance similar to OMP-residual (recovery error with $m = 400$) with parameters appropriately set even though the prior knowledge is required for OMP-residual.

level is not available (OMP-without-prior-knowledge). In this sense, OMP-CV outperforms OMP-residual.

6 Conclusion

This paper presents a theoretical study of CV in compressive sensing, providing analysis of general CV problems as well as analysis of the OMP-CV algorithm. As a highly practical algorithm, OMP-CV could reconstruct the signal without prior knowledge such as sparsity or noise level; its performance is supported in this paper both theoretically and empirically. Additionally, our results on general CV problems could also be used to understand other CS-based reconstruction algorithms.

CV sacrifices a small amount of measurements to estimate the reconstruction error. In a nutshell, this technique makes it possible for greedy algorithms to reconstruct the signal without prior knowledge like the sparsity or noise level. In future work, we would like to extend our analysis on CV by studying the use of CV in other greedy sparse recovery algorithms.

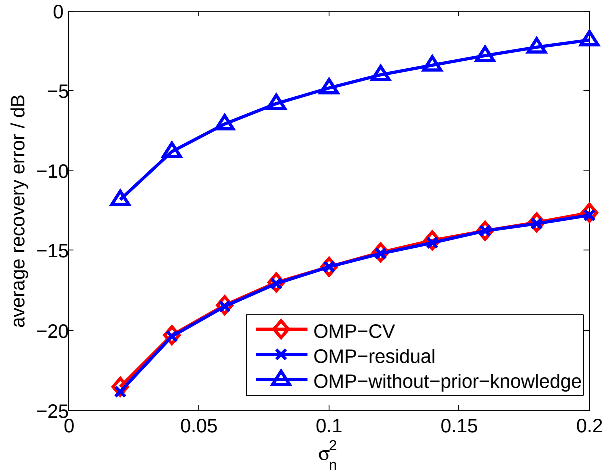


Figure 7: The performance of OMP-CV, OMP-residual, and OMP-without-prior-knowledge under different noise level. Parameters are $N = 1000$, $k = 50$, $M = 500$, and $m_{\text{cv}} = 48$. σ_n^2 varies from 0.02 to 0.2 with step 0.02. OMP-CV and OMP-residual have similar performance under different noise level even when the prior knowledge is available for OMP-residual. In addition, the performance of OMP-without-prior-knowledge is very bad where no prior knowledge is given to decide the stopping criteria of OMP.

7 Acknowledgement

The auther Jinye Zhang would like to thank to Wensi You for her encouragement, without which this work could hardly be completed. He would also like to thank Xinyue Shen for her helpful suggestions, which made the research process much smoother.

Appendix

A proof of Lemma 4

PROOF It follows from Lemma 1 that,

$$\|\mathbf{A}_S^\dagger \mathbf{A}_T \mathbf{x}_T\|_2 = \|(\mathbf{A}_S' \mathbf{A}_S)^{-1} \mathbf{A}_S' \mathbf{A}_T \mathbf{x}_T\|_2 \leq \frac{1}{1-\delta_{|S|}} \|\mathbf{A}_S' \mathbf{A}_T \mathbf{x}_T\|_2. \tag{36}$$

Furthermore with Lemma 3 one has,

$$\|\mathbf{A}_S' \mathbf{A}_T \mathbf{x}_T\|_2 \leq \delta_{|S|+|T|} \|\mathbf{x}_T\|_2. \tag{37}$$

Take inequality (37) into inequality (36) to complete the proof. ■

B proof of Lemma 6

Proof Rewrite the CV residual as,

$$
\begin{aligned}
\epsilon_{\rm cv} &= \|\mathbf{y}_{\rm cv} - \mathbf{A}_{\rm cv}\hat{\mathbf{x}}\|_2^2 \\
&= \|\mathbf{A}_{\rm cv}\mathbf{x} + \mathbf{n}_{\rm cv} - \mathbf{A}_{\rm cv}\hat{\mathbf{x}}\|_2^2 \\
&= \|\mathbf{A}_{\rm cv}(\mathbf{x} - \hat{\mathbf{x}}) + \mathbf{n}_{\rm cv}\|_2^2 \\
&= \sum_{i=1}^{m_{\rm cv}} \left(\sum_{j=1}^{N} A_{cv_{ij}} \Delta x_j + n_{cv_i} \right)^2 \\
&= \sum_{i=1}^{m_{\rm cv}} \epsilon_{cv_i}^2,
\end{aligned} \tag{38}
$$

where Δx_i is the i-th element of $\Delta\mathbf{x} = \mathbf{x} - \hat{\mathbf{x}}$ and $\epsilon_{cv_i} = \sum_{j=1}^{N} A_{cv_{ij}} \Delta x_j + n_{cv_i}$.

Calculate the mean and variance of ϵ_{cv_i} as,

$$
\mathrm{E}(\epsilon_{\rm cv_i}) = \mathrm{E}\left(\mathrm{n_{cv_i}} + \sum_{\rm j=1}^{\rm N} \mathrm{A_{cv_{ij}}} \Delta \mathrm{x_j} \right) = \sum_{\rm j=1}^{\rm N} \Delta \mathrm{x_j} \mathrm{E}(\mathrm{A_{cv_{ij}}}) + \mathrm{E}(\mathrm{n_{cv_i}}) = 0, \tag{39}
$$

and

$$
\begin{aligned}
\mathrm{Var}(\epsilon_{\rm cv_i}) &= \mathrm{E}(\epsilon_{\rm cv_i}^2) = \mathrm{E}\left[(\mathrm{n_{cv_i}} + \sum_{\rm j=1}^{\rm N} \mathrm{A_{cv_{ij}}} \Delta \mathrm{x_j})^2 \right] \\
&= \sum_{j=1}^{N} \Delta x_j^2 \mathrm{E}(\mathrm{A_{cv_{ij}}^2}) + \mathrm{E}(\mathrm{n_{cv_i}^2}) = \frac{1}{\rm m} \left(\sum_{\rm j=1}^{\rm N} \Delta \mathrm{x_j^2} + \sigma_{\rm n}^2 \right) \\
&= \frac{1}{m}(\varepsilon_{\rm x} + \sigma_{\rm n}^2).
\end{aligned} \tag{40}
$$

Being linear combination of Gaussian variables, ϵ_{cv_i} is Gaussian distributed, i.e. $\epsilon_{cv_i} \sim \mathcal{N}(0, \frac{1}{m}(\varepsilon_{\rm x} + \sigma_{\rm n}^2))$. Furthermore, $\epsilon_{cv_i}^2$ obeys χ_1^2 distribution with mean $\frac{1}{m}(\varepsilon_{\rm x} + \sigma_{\rm n}^2)$ and variance $\frac{2}{m^2}(\varepsilon_{\rm x} + \sigma_{\rm n}^2)^2$. With the assumption that $m_{\rm cv}$ is large, the Central Limit Theorem gives that

$$
\epsilon_{\rm cv} = \sum_{i=1}^{m_{\rm cv}} \epsilon_{cv_i}^2 \sim \mathcal{N}\left(\frac{m_{\rm cv}}{m}(\varepsilon_{\rm x} + \sigma_{\rm n}^2), \frac{2m_{\rm cv}}{m^2}(\varepsilon_{\rm x} + \sigma_{\rm n}^2)^2 \right), \tag{41}
$$

which completes the proof. ■

C proof of Lemma 7

Proof Write $\Delta\epsilon_{\text{cv}}$ as,

$$
\begin{aligned}
&\Delta\epsilon_{\text{cv}} \\
=&\epsilon^p_{\text{cv}} - \epsilon^q_{\text{cv}} \\
=&\sum_{i=1}^{m_{\text{cv}}} \left[\left(\sum_{j=1}^{N} A_{cv_{ij}} \Delta x^p_j + n_{cv_i} \right)^2 - \left(\sum_{j=1}^{N} A_{cv_{ij}} \Delta x^q_j + n_{cv_i} \right)^2 \right] \\
=&\sum_{i=1}^{m_{\text{cv}}} \left[\left(\sum_{j=1}^{N} A_{cv_{ij}} (\Delta x^p_j + \Delta x^q_j) + 2n_{cv_i} \right) \left(\sum_{j=1}^{N} A_{cv_{ij}} (\Delta x^p_j - \Delta x^q_j) \right) \right] \\
=&\sum_{i=1}^{m_{\text{cv}}} \left[\left(\sum_{j=1}^{N} A_{cv_{ij}} (\Delta x^p_j + \Delta x^q_j) \right) \left(\sum_{k=1}^{N} A_{cv_{ik}} (\Delta x^p_k - \Delta x^q_k) \right) + 2n_{cv_i} \left(\sum_{k=1}^{N} A_{cv_{ik}} (\Delta x^p_k - \Delta x^q_k) \right) \right] \\
=&\sum_{i=1}^{m_{\text{cv}}} R_i,
\end{aligned}
\tag{42}
$$

where

$$
R_i = \left[\left(\sum_{j=1}^{N} A_{cv_{ij}} (\Delta x^p_j + \Delta x^q_j) \right) \left(\sum_{k=1}^{N} A_{cv_{ik}} (\Delta x^p_k - \Delta x^q_k) \right) + 2n_{cv_i} \left(\sum_{k=1}^{N} A_{cv_{ik}} (\Delta x^p_k - \Delta x^q_k) \right) \right].
$$

The mean and variance of R_i are calculated as follows. R_i is the linear combination of random variables of three kinds: $A^2_{cv_{ij}}$, $A_{cv_{ij}} A_{cv_{ik}}$ and $n_{cv_i} A_{cv_{ij}}$, whose means and variances are,

$$
\mathrm{E}(\mathrm{A}^2_{\mathrm{cv_{ij}}}) = \frac{1}{\mathrm{m}} \quad \mathrm{Var}(\mathrm{A}^2_{\mathrm{cv_{ij}}}) = \frac{2}{\mathrm{m}^2}, \tag{43}
$$

$$
\mathrm{E}(\mathrm{A_{cv_{ij}}} \mathrm{A_{cv_{ik}}}) = 0 \quad \mathrm{Var}(\mathrm{A_{cv_{ij}}} \mathrm{A_{cv_{ik}}}) = \frac{1}{\mathrm{m}^2}, \tag{44}
$$

$$
\mathrm{E}(\mathrm{n_{cv_i}} \mathrm{A_{cv_{ij}}}) = 0 \quad \mathrm{Var}(\mathrm{n_{cv_i}} \mathrm{A_{cv_{ij}}}) = \frac{\sigma^2_{\mathrm{n}}}{\mathrm{m}^2}. \tag{45}
$$

Also, any two of these three types of random variable are mutually independent, i.e.

$$
\mathrm{Cov}(\mathrm{A}^2_{\mathrm{cv_{ij}}}, \mathrm{A_{cv_{ij}}} \mathrm{A_{cv_{ik}}}) = 0, \tag{46}
$$

$$
\mathrm{Cov}(\mathrm{A_{cv_{ij}}} \mathrm{A_{cv_{ik}}}, \mathrm{n_{cv_i}} \mathrm{A_{cv_{ij}}}) = 0, \tag{47}
$$

$$
\mathrm{Cov}(\mathrm{A}^2_{\mathrm{cv_{ij}}}, \mathrm{n_{cv_i}} \mathrm{A_{cv_{ij}}}) = 0, \tag{48}
$$

where $\mathrm{Cov}(\cdot)$ denotes the covariance function.

Therefore, the mean of R_i is

$$\begin{aligned}\mathrm{E(R_i)} &= \frac{1}{m}\sum_{j=1}^{N}(\Delta x_j^p + \Delta x_j^q)(\Delta x_j^p - \Delta x_j^q) \\ &= \frac{1}{m}\sum_{j=1}^{N}\left[(\Delta x_j^p)^2 - (\Delta x_j^q)^2\right] \\ &= \frac{1}{m}(\varepsilon_{\mathrm{x}}^p - \varepsilon_{\mathrm{x}}^q).\end{aligned} \tag{49}$$

As for variance,

$$\begin{aligned}&\mathrm{Var}(R_i) \\ =&\mathrm{Var}\left(\sum_{j=1}^{N} A_{cv_{ij}}(\Delta x_j^p + \Delta x_j^q)\sum_{k=1}^{N} A_{cv_{ik}}(\Delta x_k^p - \Delta x_k^q)\right) + 2n_{cv_i}\sum_{k=1}^{N} A_{cv_{ik}}(\Delta x_k^p - \Delta x_k^q) \\ =&\mathrm{Var}\left(\sum_{j=1}^{N} A_{cv_{ij}}(\Delta x_j^p + \Delta x_j^q)\sum_{k=1}^{N} A_{cv_{ik}}(\Delta x_k^p - \Delta x_k^q)\right) + 4\mathrm{Var}\left(n_{cv_i}\sum_{k=1}^{N} A_{cv_{ik}}(\Delta x_k^p - \Delta x_k^q)\right).\end{aligned} \tag{50}$$

Both terms of (50) are calculated respectively as,

$$\begin{aligned}\text{first term} &= \mathrm{Var}\left(\sum_{j=1}^{N} A_{cv_{ij}}(\Delta x_j^p + \Delta x_j^q)\sum_{k=1}^{N} A_{cv_{ik}}(\Delta x_k^p - \Delta x_k^q)\right) \\ &= \frac{2}{m^2}\sum_{j=1}^{N}(\Delta x_j^p + \Delta x_j^q)^2(\Delta x_j^p - \Delta x_j^q)^2 \\ &+ \frac{1}{m^2}\sum_{j=1}^{N}\sum_{k=j+1}^{N}\Big((\Delta x_j^p + \Delta x_j^q)(\Delta x_k^p - \Delta x_k^q) + (\Delta x_k^p + \Delta x_k^q)(\Delta x_j^p - \Delta x_j^q)\Big)^2 \\ &= \frac{2}{m^2}\sum_{j=1}^{N}\Big((\Delta x_j^p)^2 - (\Delta x_j^q)^2\Big)^2 + \frac{4}{m^2}\sum_{j=1}^{N}\sum_{k=j+1}^{N}\Big(\Delta x_j^p\Delta x_k^p - \Delta x_j^q\Delta x_k^q\Big)^2 \\ &= \frac{2}{m^2}\left[\left(\sum_{j=1}^{N}(\Delta x_j^p)^2\right)^2 + \left(\sum_{j=1}^{N}(\Delta x_j^q)^2\right)^2 - 2\left(\sum_{j=1}^{N}\Delta x_j^p\Delta x_j^q\right)^2\right],\end{aligned} \tag{51}$$

and

$$\text{second term} = 4\mathrm{Var}\left(n_{cv_i}\sum_{k=1}^{N} A_{cv_{ik}}(\Delta x_k^p - \Delta x_k^q)\right) = \frac{4\sigma_{\mathrm{n}}^2}{m^2}\sum_{j=1}^{N}(\Delta x_j^p - \Delta x_j^q)^2. \tag{52}$$

Hence we have

$$
\begin{aligned}
&\mathrm{Var}(R_i)\\
=&\frac{2}{m^2}\left[\left(\sum_{j=1}^{N}(\Delta x_j^p)^2\right)^2+\left(\sum_{j=1}^{N}(\Delta x_j^q)^2\right)^2-2\left(\sum_{j=1}^{N}\Delta x_j^p\Delta x_j^q\right)^2\right]+\frac{4}{m^2}\sigma_{\mathrm{n}}^2\sum_{j=1}^{N}(\Delta x_j^p-\Delta x_j^q)^2\\
=&\frac{2}{m^2}\left[(\|\Delta\mathbf{x}^p\|_2^2)^2+(\|\Delta\mathbf{x}^q\|_2^2)^2-2\langle\Delta\mathbf{x}^p,\Delta\mathbf{x}^q\rangle^2+2\sigma_{\mathrm{n}}^2\|\Delta\mathbf{x}^p-\Delta\mathbf{x}^q\|_2^2\right].
\end{aligned}
\tag{53}
$$

Furthermore, it follows from the Central Limit Theorem that,

$$
\Delta\epsilon_{\mathrm{cv}}=\sum_{i=1}^{m_{\mathrm{cv}}}R_i\sim\mathcal{N}(\mu,\sigma^2), \tag{54}
$$

where

$$
\mu=\frac{m_{\mathrm{cv}}}{m}(\varepsilon_{\mathrm{x}}^p-\varepsilon_{\mathrm{x}}^q)
$$

and

$$
\sigma^2=\frac{2m_{\mathrm{cv}}}{m^2}\{(\varepsilon_{\mathrm{x}}^p)^2+(\varepsilon_{\mathrm{x}}^q)^2-2\langle\Delta\mathbf{x}^p,\Delta\mathbf{x}^q\rangle^2+2\sigma_{\mathrm{n}}^2\|\Delta\mathbf{x}^p-\Delta\mathbf{x}^q\|_2^2\}.
$$

Recall that $\Delta\mathbf{x}_{\mathrm{g}}^p=[\Delta\mathbf{x}^{p\prime},\sigma_{\mathrm{n}}]'$, $\Delta\mathbf{x}_{\mathrm{g}}^q=[\Delta\mathbf{x}^{q\prime},\sigma_{\mathrm{n}}]'$, $\varepsilon_{\mathrm{g}}^p=\|\Delta\mathbf{x}_{\mathrm{g}}^p\|_2^2$, $\varepsilon_{\mathrm{g}}^q=\|\Delta\mathbf{x}_{\mathrm{g}}^q\|_2^2$, and $\rho_{\mathrm{g}}=\frac{\langle\Delta\mathbf{x}_{\mathrm{g}}^p,\Delta\mathbf{x}_{\mathrm{g}}^q\rangle}{\|\Delta\mathbf{x}_{\mathrm{g}}^p\|_2\|\Delta\mathbf{x}_{\mathrm{g}}^q\|_2}$. The probability distribution of ϵ_{cv} can be rewritten as

$$
\Delta\epsilon_{\mathrm{cv}}^p\sim\mathcal{N}(\mu,\sigma^2), \tag{55}
$$

■

where

$$
\mu=\frac{m_{\mathrm{cv}}}{m}(\varepsilon_{\mathrm{g}}^p-\varepsilon_{\mathrm{g}}^q)
$$

and

$$
\sigma^2=\frac{2m_{\mathrm{cv}}}{m^2}[(\varepsilon_{\mathrm{g}}^p)^2+(\varepsilon_{\mathrm{g}}^q)^2-2\rho_{\mathrm{g}}^2\varepsilon_{\mathrm{g}}^p\varepsilon_{\mathrm{g}}^q].
$$

D proof of Theorem 4

Proof The proof of Theorem 4 generally consists of three steps. Firstly, the connection between recovered signals in two iterations of OMP-CV is studied so that our general CV techniques could be properly utilized. Second, the knowledge of connection between recovered signals is combined with general CV techniques to calculate the CV comparison success probability. Finally, analyze such probability in both situations of Theorem 4 to complete the proof.

D.1 Step1: Connection between Two Recovered Signals

Consider recovered signals generated in the p-th and q-th iteration, $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^q$. Assume next that $p < q$. The recovered signals can be written as,

$$\hat{\mathbf{x}}^p_{T^p} = \mathbf{A}^\dagger_{T^p}\mathbf{y}, \tag{56}$$

$$\hat{\mathbf{x}}^q_{T^q} = \mathbf{A}^\dagger_{T^q}\mathbf{y}. \tag{57}$$

The connection between $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^q$ can be described as:

Lemma 8 *Let $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^q$ be recovered signals of the p-th and q-th iteration in OMP-CV and $p < q$. $\hat{\mathbf{x}}^q_{T^q}$ can be written as,*

$$\hat{\mathbf{x}}^q_{T^q} = \mathbf{A}^\dagger_{T^q}\mathbf{y} = \begin{bmatrix} \mathbf{A}^\dagger_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n} - \mathbf{A}_{T^{q-p}}\boldsymbol{\delta}_{T^{q-p}}) \\ \boldsymbol{\delta}_{T^{q-p}} \end{bmatrix} + \mathbf{x}_{T^q}, \tag{58}$$

where $\boldsymbol{\delta}_{T^{q-p}} = (\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}})^{-1}\mathbf{A}'_{T^{q-p}}P_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n})$ and T^{q-p} denotes the set of indices $T^q \backslash T^p$. [2]

Proof The proof of Lemma 8 is in Appendix E. ■

Furthermore, we can calculate $\Delta\mathbf{x}^p$ and $\Delta\mathbf{x}^q$:

$$\Delta\mathbf{x}^p = \mathbf{x}_T - \hat{\mathbf{x}}^p_T = \begin{bmatrix} -\mathbf{A}^\dagger_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n}) \\ \mathbf{x}_{T^{q-p}} \\ \mathbf{x}_{(T^q)^c} \end{bmatrix}, \tag{59}$$

$$\Delta\mathbf{x}^q = \mathbf{x}_T - \hat{\mathbf{x}}^q_T = \begin{bmatrix} -\mathbf{A}^\dagger_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n} - \mathbf{A}_{T^{q-p}}\boldsymbol{\delta}_{T^{q-p}}) \\ -\boldsymbol{\delta}_{T^{q-p}} \\ \mathbf{x}_{(T^q)^c} \end{bmatrix}, \tag{60}$$

where $(T^q)^c$ denotes the index set $T \backslash T^q$. Then, $\boldsymbol{\delta}_{T^{q-p}}$ can be bounded as:

Lemma 9

$$\|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 \le \eta(\|\mathbf{x}_{(T^q)^c}\|_2^2 + \sigma_{\mathrm{n}}^2) \tag{61}$$

where η is related to RIPs of the sensing matrix $\mathbf{A}$. With the hypothesis $\delta_d \le 0.1$ we have $\eta \le 0.0127$.

Proof The proof of Lemma 9 can be found in Appendix F. ■

[2] For simplicity, the element order of $\hat{\mathbf{x}}^q_{T^q}$ is reordered. Similar operations are conducted to the analysis below.

D.2 Step2: Combining with General CV Techniques

As stated in Theorem 2,

$$\frac{1}{\lambda^2} = \frac{2}{m_{\mathrm{cv}}}\left[1 + 2(1-\rho_{\mathrm{g}}^2)\frac{\varepsilon_{\mathrm{g}}^q\varepsilon_{\mathrm{g}}^p}{(\varepsilon_{\mathrm{g}}^q - \varepsilon_{\mathrm{g}}^p)^2}\right]. \tag{62}$$

We next calculate

$$(1-\rho_{\mathrm{g}}^2)\frac{\varepsilon_{\mathrm{g}}^q\varepsilon_{\mathrm{g}}^p}{(\varepsilon_{\mathrm{g}}^q - \varepsilon_{\mathrm{g}}^p)^2} = \frac{\varepsilon_{\mathrm{g}}^q\varepsilon_{\mathrm{g}}^p - \langle \Delta\mathbf{x}_{\mathrm{g}}^q, \Delta\mathbf{x}_{\mathrm{g}}^p\rangle^2}{(\varepsilon_{\mathrm{g}}^q - \varepsilon_{\mathrm{g}}^p)^2}. \tag{63}$$

It follows from Lemma 8 that

$$\varepsilon_{\mathrm{g}}^p = \sigma_{\mathrm{n}}^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2 + \|\mathbf{x}_{T^{q-p}}\|_2^2 + \|\mathbf{A}_{T^p}^{\dagger}(\mathbf{A}_{(T^p)^c}\mathbf{x}_{(T^p)^c} + \mathbf{n})\|_2^2, \tag{64}$$

$$\varepsilon_{\mathrm{g}}^q = \sigma_{\mathrm{n}}^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2 + \|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{A}_{T^p}^{\dagger}[\mathbf{A}_{(T^p)^c}\mathbf{x}_{(T^p)^c} + \mathbf{n} - \mathbf{A}_{T^{q-p}}(\mathbf{x}_{T^{q-p}} + \boldsymbol{\delta}_{T^{q-p}})]\|_2^2. \tag{65}$$

Make the following notation,

Definition 6

$$\mathbf{a} \triangleq \mathbf{A}_{T^p}^{\dagger}(\mathbf{A}_{(T^p)^c}\mathbf{x}_{(T^p)^c} + \mathbf{n}), \tag{66}$$

$$\mathbf{b} \triangleq \mathbf{A}_{T^p}^{\dagger}[\mathbf{A}_{(T^p)^c}\mathbf{x}_{(T^p)^c} + \mathbf{n} - \mathbf{A}_{T^{q-p}}(\mathbf{x}_{T^{q-p}} + \boldsymbol{\delta}_{T^{q-p}})], \tag{67}$$

$$\mathbf{c} \triangleq \mathbf{A}_{T^p}^{\dagger}(\mathbf{A}_{T^{q-p}}(\mathbf{x}_{T^{q-p}} + \boldsymbol{\delta}_{T^{q-p}})). \tag{68}$$

where $\mathbf{a} = \mathbf{b} + \mathbf{c}$. With Lemma 4, one has,

$$\|\mathbf{a}\|_2^2 \le (\frac{\delta_{k+1}}{1-\delta_p})^2(\|\mathbf{x}_{(T^p)^c}\|_2^2 + \sigma_{\mathrm{n}}^2), \tag{69}$$

$$\|\mathbf{b}\|_2^2 \le (\frac{\delta_{k+1}}{1-\delta_p})^2(\|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2 + \sigma_{\mathrm{n}}^2), \tag{70}$$

$$\|\mathbf{c}\|_2^2 \le (\frac{\delta_q}{1-\delta_p})^2\|\mathbf{x}_{T^{q-p}} + \boldsymbol{\delta}_{T^{q-p}}\|_2^2. \tag{71}$$

Then, $\varepsilon_{\mathrm{g}}^p$ and $\varepsilon_{\mathrm{g}}^q$ can be rewritten as

$$\varepsilon_{\mathrm{g}}^p = \sigma_{\mathrm{n}}^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2 + \|\mathbf{x}_{T^{q-p}}\|_2^2 + \|\mathbf{a}\|_2^2, \tag{72}$$

$$\varepsilon_{\mathrm{g}}^q = \sigma_{\mathrm{n}}^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2 + \|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{b}\|_2^2. \tag{73}$$

Furthermore,

$$\begin{aligned}
\varepsilon_{\mathrm{g}}^p\varepsilon_{\mathrm{g}}^q - \langle \Delta\mathbf{x}_{\mathrm{g}}^p, \Delta\mathbf{x}_{\mathrm{g}}^q\rangle^2 =& (\|\mathbf{x}_{T^{q-p}}\|_2^2 + \|\mathbf{a}\|_2^2)(\|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{b}\|_2^2) - (\boldsymbol{\delta}_{T^{q-p}} \cdot \mathbf{x}_{T^{q-p}} - \mathbf{a}\cdot\mathbf{b})_2^2 \\
&+ (\|\mathbf{x}_{T^{q-p}} + \boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{a}-\mathbf{b}\|_2^2)(\sigma_{\mathrm{n}}^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2) \\
&\le (\|\mathbf{x}_{T^{q-p}}\|_2^2 + \|\mathbf{a}\|_2^2)(\|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{b}\|_2^2) \\
&+ (\|\mathbf{x}_{T^{q-p}} + \boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{a}-\mathbf{b}\|_2^2)(\sigma_{\mathrm{n}}^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2) \\
&\le \left[\|\mathbf{x}_{T^{q-p}}\|_2^2 + (\frac{\delta_{k+1}}{1-\delta_p})^2(\|\mathbf{x}_{(T^p)^c}\|_2^2 + \sigma_{\mathrm{n}}^2)\right] \\
&\left[\|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 + (\frac{\delta_{k+1}}{1-\delta_p})^2(\|\boldsymbol{\delta}_{T^{q-p}}\|_2^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2 + \sigma_{\mathrm{n}}^2)\right] \\
&+ (1 + (\frac{\delta_q}{1-\delta_p})^2)\|\mathbf{x}_{T^{q-p}} + \boldsymbol{\delta}_{T^{q-p}}\|_2^2(\sigma_{\mathrm{n}}^2 + \|\mathbf{x}_{(T^q)^c}\|_2^2).
\end{aligned} \tag{74}$$

Meanwhile,

$$\varepsilon_{\mathrm{g}}^{p}-\varepsilon_{\mathrm{g}}^{q}=\|\mathbf{x}_{T^{q-p}}\|_2^2-\|\boldsymbol{\delta}_{T^{q-p}}\|_2^2+\|\mathbf{a}\|_2^2-\|\mathbf{b}\|_2^2. \tag{75}$$

D.3 Step3: Analyses of Both Situations in Theorem 4

So far, we have a general knowledge of how λ could be represented by $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^q$. Next, we analyze both situations of Theorem 4 respectively. Let $\hat{\mathbf{x}}^o$ be the oracle output.

Situation 1: $T\subset T^o$ and $T\backslash T^p\neq\emptyset$

In this situation, $p<o$ and $\|\mathbf{x}_{(T^o)^c}\|_2^2=0$. Consider recovered signals $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^o$. On one hand, write q as o in Equation (74) we have,

$$\begin{aligned}\varepsilon_{\mathrm{g}}^{p}\varepsilon_{\mathrm{g}}^{o}-\langle\Delta\mathbf{x}_{\mathrm{g}}^{p},\Delta\mathbf{x}_{\mathrm{g}}^{o}\rangle^2\leq&\left(1+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\right)^2\|\mathbf{x}_{T^{o-p}}\|_2^2\|\boldsymbol{\delta}_{T^{o-p}}\|_2^2\\&+\left(1+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\right)\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\sigma_{\mathrm{n}}^2\left(\|\mathbf{x}_{T^{o-p}}\|_2^2+\|\boldsymbol{\delta}_{T^{o-p}}\|_2^2\right)\\&+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^4\sigma_{\mathrm{n}}^4+\left(1+\left(\frac{\delta_o}{1-\delta_p}\right)^2\right)\|\mathbf{x}_{T^{o-p}}+\boldsymbol{\delta}_{T^{o-p}}\|_2^2\sigma_{\mathrm{n}}^2.\end{aligned} \tag{76}$$

Then it follows Lemma 9 that

$$\begin{aligned}\varepsilon_{\mathrm{g}}^{p}\varepsilon_{\mathrm{g}}^{o}-\langle\Delta\mathbf{x}_{\mathrm{g}}^{p},\Delta\mathbf{x}_{\mathrm{g}}^{o}\rangle^2\leq&\left(1+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\right)^2\eta\|\mathbf{x}_{T^{o-p}}\|_2^2\sigma_{\mathrm{n}}^2\\&+\left(1+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\right)\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\sigma_{\mathrm{n}}^2\left(\|\mathbf{x}_{T^{o-p}}\|_2^2+\eta\sigma_{\mathrm{n}}^2\right)\\&+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^4\sigma_{\mathrm{n}}^4+\left(1+\left(\frac{\delta_o}{1-\delta_p}\right)^2\right)\left(\|\mathbf{x}_{T^{o-p}}\|_2^2+\eta\sigma_{\mathrm{n}}^2\right)\sigma_{\mathrm{n}}^2.\end{aligned} \tag{77}$$

Recall that,

$$\alpha^p=\frac{\|\mathbf{x}_{T\backslash T^p}\|_2}{\sigma_{\mathrm{n}}}=\frac{\|\mathbf{x}_{T^{o-p}}\|_2}{\sigma_{\mathrm{n}}}, \tag{78}$$

then we have,

$$\varepsilon_{\mathrm{g}}^{p}\varepsilon_{\mathrm{g}}^{o}-\langle\Delta\mathbf{x}_{\mathrm{g}}^{p},\Delta\mathbf{x}_{\mathrm{g}}^{o}\rangle^2\leq\frac{\beta_1}{2}(\alpha^p)^2\sigma_{\mathrm{n}}^4+\frac{\beta_2}{2}\sigma_{\mathrm{n}}^4, \tag{79}$$

where

$$\begin{aligned}\beta_1&=2\left[\left(1+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\right)^2\eta+\left(1+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\right)\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2+\left(1+\left(\frac{\delta_o}{1-\delta_p}\right)^2\right)\right]\\\beta_2&=2\left[\left(1+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\right)\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2\eta+\left(\frac{\delta_{k+1}}{1-\delta_p}\right)^4+\left(1+\left(\frac{\delta_o}{1-\delta_p}\right)^2\right)\eta\right].\end{aligned}$$

On the other hand, write q as o in Equation (75) we have,

$$\begin{aligned}\varepsilon_{\rm g}^p - \varepsilon_{\rm g}^o &= \|\mathbf{x}_{T^{o-p}}\|_2^2 - \|\boldsymbol{\delta}_{T^{o-p}}\|_2^2 + \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2 \\ &= (\|\mathbf{x}_{T^{o-p}}\|_2^2 + \|\mathbf{A}_{T^p}^\dagger \mathbf{A}_{T^{o-p}}\mathbf{x}_{T^{o-p}}\|_2^2) - (\|\boldsymbol{\delta}_{T^{o-p}}\|_2^2 + \|\mathbf{A}_{T^p}^\dagger \mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}\|_2^2) \\ &\quad + 2\langle \mathbf{A}_{T^p}^\dagger \mathbf{A}_{T^{o-p}}(\mathbf{x}_{T^{o-p}} + \boldsymbol{\delta}_{T^{o-p}}), \mathbf{A}_{T^p}^\dagger \mathbf{n}\rangle \\ &\geq \|\mathbf{x}_{T^{o-p}}\|_2^2 - (1 + (\frac{\delta_o}{1-\delta_p})^2)\|\boldsymbol{\delta}_{T^{o-p}}\|_2^2 - 2\frac{\delta_o \delta_{p+1}}{(1-\delta_p)^2}\|\mathbf{x}_{T^{o-p}} + \boldsymbol{\delta}_{T^{o-p}}\|_2 \sigma_{\rm n} \\ &\geq (\alpha^p)^2\sigma_{\rm n}^2 - (1 + (\frac{\delta_o}{1-\delta_p})^2)\eta\sigma_{\rm n}^2 - 2\frac{\delta_o\delta_{p+1}}{(1-\delta_p)^2}((\alpha^p)^2\sigma_{\rm n} + \sqrt{\eta}\sigma_{\rm n})\sigma_{\rm n} \\ &\geq \left[(\alpha^p)^2 - \beta_3\alpha^p - \beta_4\right]\sigma_{\rm n}^2,\end{aligned} \tag{80}$$

where $\beta_3 = 2\frac{\delta_o\delta_{p+1}}{(1-\delta_p)^2}$ and $\beta_4 = (1 + (\frac{\delta_o}{1-\delta_p})^2)\eta + 2\frac{\delta_o\delta_{p+1}}{(1-\delta_p)^2}\sqrt{\eta}$. Notice that $\varepsilon_{\rm g}^p - \varepsilon_{\rm g}^o > 0$, then we have,

$$\varepsilon_{\rm g}^p - \varepsilon_{\rm g}^o \geq \max(\left[(\alpha^p)^2 - \beta_3\alpha^p - \beta_4\right]\sigma_{\rm n}^2, 0). \tag{81}$$

Put Equation (79) and (81) together we finally reach

$$\frac{\varepsilon_{\rm g}^p\varepsilon_{\rm g}^o - \langle \Delta\mathbf{x}_{\rm g}^p, \Delta\mathbf{x}_{\rm g}^o\rangle^2}{(\varepsilon_{\rm g}^p - \varepsilon_{\rm g}^o)^2} \leq \frac{\frac{\beta_1}{2}(\alpha^p)^2 + \frac{\beta_2}{2}}{\max([(\alpha^p)^2 - \beta_3\alpha^p - \beta_4], 0)^2}. \tag{82}$$

Take above inequality into Equation (62),

$$\frac{1}{\lambda^2} \leq \frac{2}{m_{\rm cv}}\left[1 + \frac{\beta_1(\alpha^p)^2 + \beta_2}{\max([(\alpha^p)^2 - \beta_3\alpha^p - \beta_4], 0)^2}\right]. \tag{83}$$

Furthermore,

$$\begin{aligned}\lambda^2 &\geq \frac{m_{\rm cv}}{2}\frac{\left[(\alpha^p)^2 - \beta_3\alpha^p - \beta_4\right]^2}{\max([(\alpha^p)^2 - \beta_3\alpha^p - \beta_4], 0)^2 + (\beta_1(\alpha^p)^2 + \beta_2)} \\ &= \frac{m_{\rm cv}}{2}\left[1 - \frac{\beta_1(\alpha^p)^2 + \beta_2}{\max([(\alpha^p)^2 - \beta_3\alpha^p - \beta_4], 0)^2 + (\beta_1(\alpha^p)^2 + \beta_2)}\right] \\ &= \frac{m_{\rm cv}}{2}[1 - {\rm g}(\alpha^{\rm p})] \\ &\approx \frac{m_{\rm cv}}{2}\left[1 - \frac{\beta_1}{(\alpha^p)^2 + \beta_1}\right],\end{aligned} \tag{84}$$

where

$${\rm g}(\alpha^{\rm p}) = [\beta_1(\alpha^{\rm p})^2 + \beta_2]/[\beta_1(\alpha^{\rm p})^2 + \beta_2 + \max((\alpha^{\rm p})^2 - \beta_3\alpha^{\rm p} - \beta_4, 0)^2].$$

The last approximation hold because $\beta_1 \gg \beta_2, \beta_3, \beta_4$. According to Theorem 2, $\Phi(\lambda)$ is the probability that $\epsilon_{\rm cv}^o < \epsilon_{\rm cv}^p$. Therefore, one may take square root of both sides of Equation (84)to complete the proof of the first part of Theorem 4.

Situation 2: $T \subset T^o$ and $T \backslash T^p = \emptyset$

In this situation, please notice that $\rho_{\rm g}$ of $\hat{\mathbf{x}}^o$ and $\hat{\mathbf{x}}^p$ is extremely closed to 1. Therefore, we would like to first calculate the lower bound of $\rho_{\rm g}$ and then complete the proof using

Theorem 3. As p may exceed o, we first assume $p < o$ and then analyze the other situation. Recall that

$$\rho_{\rm g} = \frac{\langle \Delta \mathbf{x}_{\rm g}^o, \Delta \mathbf{x}_{\rm g}^p \rangle}{\|\Delta \mathbf{x}_{\rm g}^o\|_2 \|\Delta \mathbf{x}_{\rm g}^p\|_2}. \tag{85}$$

On one hand,

$$\begin{aligned}
\langle \Delta \mathbf{x}_{\rm g}^o, \Delta \mathbf{x}_{\rm g}^p \rangle &= \langle \mathbf{A}_{T^p}^{\dagger}\mathbf{n}, \mathbf{A}_{T^p}^{\dagger}(\mathbf{n} - \mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}})\rangle + \sigma_{\rm n}^2 \\
&= \|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2^2 - \langle \mathbf{A}_{T^p}^{\dagger}\mathbf{n}, \mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}\rangle + \sigma_{\rm n}^2 \\
&\geq \|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2^2 - \|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2 \|\mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}\|_2 + \sigma_{\rm n}^2 \\
&\geq -\|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2 \|\mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}\|_2 + \sigma_{\rm n}^2 \\
&\geq \sigma_{\rm n}^2 (1 - \frac{\delta_{p+1}\delta_{o+1}}{(1-\delta_p)^2}\sqrt{\eta}),
\end{aligned} \tag{86}$$

where the last inequality holds due to Lemma 4. On the other hand,

$$\begin{aligned}
\varepsilon_{\rm g}^p \varepsilon_{\rm g}^o &= (\|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2^2 + \sigma_{\rm n}^2)(\|\mathbf{A}_{T^p}^{\dagger}(\mathbf{n} - \mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}})\|_2^2 + \|\boldsymbol{\delta}_{T^{o-p}}\|_2^2 + \sigma_{\rm n}^2) \\
&\leq \left(\left(\frac{\delta_{p+1}}{1-\delta_p}\right)^2 \sigma_{\rm n}^2 + \sigma_{\rm n}^2\right)\left(\left(\frac{\delta_{o+1}}{1-\delta_p}\right)^2 \sigma_{\rm n}^2 + \eta\sigma_{\rm n}^2 + \sigma_{\rm n}^2\right) \\
&= \sigma_{\rm n}^4 \left(\left(\frac{\delta_{p+1}}{1-\delta_p}\right)^2 + 1\right)\left(\left(\frac{\delta_{o+1}}{1-\delta_p}\right)^2 + \eta + 1\right).
\end{aligned} \tag{87}$$

Then we have

$$\rho_{\rm g} = \frac{\langle \Delta \mathbf{x}_{\rm g}^o, \Delta \mathbf{x}_{\rm g}^p \rangle}{\|\Delta \mathbf{x}_{\rm g}^o\|_2 \|\Delta \mathbf{x}_{\rm g}^p\|_2} \geq \beta_5, \tag{88}$$

where $\beta_5 = \frac{(1-\frac{\delta_{p+1}\delta_{o+1}}{(1-\delta_p)^2}\sqrt{\eta})}{\sqrt{((\frac{\delta_{p+1}}{1-\delta_p})^2+1)((\frac{\delta_{o+1}}{1-\delta_p})^2+\eta+1)}}$.

For the other situation, i.e. $o < p$, just switch the position of o and p to obtain the same lower bound of $\rho_{\rm g}$.

With the lower bound of $\rho_{\rm g}$, it follows from Theorem 3 that the CV comparison success probability of $\hat{\mathbf{x}}^p$ and $\hat{\mathbf{x}}^o$ is equal or higher than $\Phi(\lambda_0)$ if

$$\frac{\varepsilon_{\rm g}^p}{\varepsilon_{\rm g}^o} \geq 2C_0 + 1 + 2\sqrt{C_0^2 + C_0}, \tag{89}$$

where $C_0 = (1-\rho_{\rm g}^2)\frac{\lambda_0^2}{m_{\rm cv}-2\lambda_0^2} \leq (1-\beta_5^2)\frac{\lambda_0^2}{m_{\rm cv}-2\lambda_0^2}$.

Let $C_1 = 2C_0 + 1 + 2\sqrt{C_0^2 + C_0}$. Suppose among all recovered signals that recovered all indices in the support T, there are n signals whose recovery error is larger than $C_1\varepsilon_{\rm g}^o$, i.e. satisfying Inequility (89). Let S denote the set containing iteration indices of these n recovered signals, i.e.,

$$S = \{i \mid \varepsilon_{\rm g}^i \geq C_1 \varepsilon_{\rm g}^o\}. \tag{90}$$

The probability that $\epsilon_{\rm cv}^o$ is larger than any CV residuals of these n recovered signals is smaller than $n[1-\Phi(\lambda_0)]$, i.e.,

$$\mathrm{P}(\exists i \in S \;\; s.t. \;\; \epsilon_{\rm cv}^i \leqslant \epsilon_{\rm cv}^o) < n[1-\Phi(\lambda_0)]. \tag{91}$$

Since this is the probability smaller than which it holds that a recovered signal with recovery error larger $C_1\varepsilon_{\text{g}}^o$ to be the OMP-CV output, i.e.

$$\text{P}(\{\hat{\mathbf{x}}^i \text{ is the OMP-CV output}\} \bigcap \{i \in S\}) < \text{P}(\exists i \in S \;\text{ s.t. }\; \epsilon_{\text{cv}}^i \leqslant \epsilon_{\text{cv}}^o), \tag{92}$$

and since n does not exceed $(d-k)$, it holds with probability larger than $\{1-(d-k)[1-\Phi(\lambda_0)]\}$ that OMP-CV output has recovery error smaller than $C_1\varepsilon_{\text{g}}^o$. This completes the proof of the second part of Theorem 4. ∎

E proof of Lemma 8

Proof Make the notation $\mathbf{\Gamma}_i = (\mathbf{A}'_{T^i}\mathbf{A}_{T^i})^{-1}$, $i = p, q$. $\mathbf{\Gamma}_q$ can be written in $\mathbf{\Gamma}_p$ as,

$$\begin{aligned} \mathbf{\Gamma}_q &= \begin{bmatrix} \mathbf{A}'_{T^p}\mathbf{A}_{T^p} & \mathbf{A}'_{T^p}\mathbf{A}_{T^{q-p}} \\ \mathbf{A}'_{T^{q-p}}\mathbf{A}_{T^p} & \mathbf{A}'_{T^{q-p}}\mathbf{A}_{T^{q-p}} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{\Gamma}_p + \mathbf{\Gamma}_p\boldsymbol{\alpha}_p\boldsymbol{\theta}_p\boldsymbol{\alpha}'_p\mathbf{\Gamma}_p & -\mathbf{\Gamma}_p\boldsymbol{\alpha}_p\boldsymbol{\theta}_p \\ -\boldsymbol{\theta}_p\boldsymbol{\alpha}'_p\mathbf{\Gamma}_p & \boldsymbol{\theta}_p \end{bmatrix}, \end{aligned} \tag{93}$$

where $\boldsymbol{\alpha}_p = \mathbf{A}'_{T^p}\mathbf{A}_{T^{q-p}}$, $\boldsymbol{\theta}_p = (\mathbf{A}'_{T^{q-p}}\mathbf{A}_{T^{q-p}} - \boldsymbol{\alpha}'_p\mathbf{\Gamma}_p\boldsymbol{\alpha}_p)^{-1}$. Next, write $\mathbf{A}^\dagger_{T^q}$ in $\mathbf{A}^\dagger_{T^p}$ as,

$$\begin{aligned} \mathbf{A}^\dagger_{T^q} &= (\mathbf{A}'_{T^q}\mathbf{A}_{T^q})^{-1}\mathbf{A}'_{T^q} = \mathbf{\Gamma}_q\mathbf{A}'_{T^q} \\ &= \begin{bmatrix} \mathbf{\Gamma}_p + \mathbf{\Gamma}_p\boldsymbol{\alpha}_p\boldsymbol{\theta}_p\boldsymbol{\alpha}'_p\mathbf{\Gamma}_p & -\mathbf{\Gamma}_p\boldsymbol{\alpha}_p\boldsymbol{\theta}_p \\ -\boldsymbol{\theta}_p\boldsymbol{\alpha}'_p\mathbf{\Gamma}_p & \boldsymbol{\theta}_p \end{bmatrix} \begin{bmatrix} \mathbf{A}'_{T^p} \\ \mathbf{A}'_{T^{q-p}} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}^\dagger_{T^p} - \mathbf{A}^\dagger_{T^p}\mathbf{A}_{T^{q-p}}\boldsymbol{\theta}_p\mathbf{A}'_{T^{q-p}}P_{T^p} \\ \theta_p\mathbf{A}'_{T^{q-p}}P_{T^p} \end{bmatrix}, \end{aligned} \tag{94}$$

where $P_{T^p} = I - \mathbf{A}_{T^p}\mathbf{A}^\dagger_{T^p}$. In addition,

$$\boldsymbol{\theta}_p = (\mathbf{A}'_{T^{q-p}}\mathbf{A}_{T^{q-p}} - \boldsymbol{\alpha}'_p\mathbf{\Gamma}_p\boldsymbol{\alpha}_p)^{-1} = (\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}})^{-1}. \tag{95}$$

Finally, we reach that,

$$\begin{aligned} \hat{\mathbf{x}}^q_{T^q} &= \mathbf{A}^\dagger_{T^q}\mathbf{y} = \mathbf{A}^\dagger_{T^q}(\mathbf{A}_{T^q}\mathbf{x}_{T^q} + \mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n}) \\ &= \begin{bmatrix} \mathbf{A}^\dagger_{T^p} - \mathbf{A}^\dagger_{T^p}\mathbf{A}_{T^{q-p}}\theta_p\mathbf{A}'_{T^{q-p}}P_{T^p} \\ \boldsymbol{\theta}_p\mathbf{A}'_{T^{q-p}}P_{T^p} \end{bmatrix} (\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n}) + \mathbf{x}_{T^q} \\ &= \begin{bmatrix} \mathbf{A}^\dagger_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n} - \mathbf{A}_{T^{q-p}}\boldsymbol{\delta}_{T^{q-p}}) \\ \boldsymbol{\delta}_{T^{q-p}} \end{bmatrix} + \mathbf{x}_{T^q}, \end{aligned} \tag{96}$$

where $\boldsymbol{\delta}_{T^{q-p}} = (\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}})^{-1}\mathbf{A}'_{T^{q-p}}P_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c} + \mathbf{n})$. ∎

F proof of Lemma 9

Proof Let $\mathbf{u} \in \mathbb{R}^{q-p}$ be an arbitrary vector. With Lemma 1 we have,

$$\sqrt{1-\delta_{q-p}}\|P_{T^p}\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2 \leq \|\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2 \leq \sqrt{1+\delta_{q-p}}\|P_{T^p}\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2. \tag{97}$$

Furthermore with Lemma 5 we have,

$$\sqrt{1-\left(\frac{\delta_q}{1-\delta_q}\right)^2}\|\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2 \leq \|P_{T^p}\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2 \leq \|\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2. \tag{98}$$

Again with Lemma 1 again we have,

$$\sqrt{1-\delta_{q-p}}\|\mathbf{u}\|_2 \leq \|\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2 \leq \sqrt{1+\delta_{q-p}}\|\mathbf{u}\|_2. \tag{99}$$

Combining Inequality (97), Inequality(98), and Inequality(99) as,

$$(1-\delta_{q-p})^2\left(1-\left(\frac{\delta_q}{1-\delta_q}\right)^2\right)\|\mathbf{u}\|_2^2 \leq \|\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}}\mathbf{u}\|_2 \leq (1+\delta_{q-p})^2\|\mathbf{u}\|_2^2. \tag{100}$$

The above inequality shows that the singular values of matrix $\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}}$ lie between $(1-\delta_{q-p})\sqrt{(1-(\frac{\delta_q}{1-\delta_q})^2)}$ and $(1+\delta_{q-p})$. Hence,

$$\frac{1}{(1+\delta_{q-p})^2}\|\mathbf{u}\|_2^2 \leq \|(\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}})^{-1}\mathbf{u}\|_2^2 \leq \frac{1}{(1-\delta_{q-p})^2(1-(\frac{\delta_q}{1-\delta_q})^2)}\|\mathbf{u}\|_2^2. \tag{101}$$

Finally,

$$\begin{aligned}
&\|\delta_{T^{q-p}}\|_2^2\\
=&\|(\mathbf{A}'_{T^{q-p}}P_{T^p}\mathbf{A}_{T^{q-p}})^{-1}\mathbf{A}'_{T^{q-p}}P_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c}+\mathbf{n})\|_2^2\\
\leq&\frac{1}{(1-\delta_{q-p})^2(1-(\frac{\delta_q}{1-\delta_q})^2)}\|\mathbf{A}'_{T^{q-p}}P_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c}+\mathbf{n})\|_2^2\\
=&\frac{1}{(1-\delta_{q-p})^2(1-(\frac{\delta_q}{1-\delta_q})^2)}\|\mathbf{A}'_{T^{q-p}}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c}+\mathbf{n})-\mathbf{A}'_{T^{q-p}}\mathbf{A}_{T^p}\mathbf{A}^\dagger_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c}+\mathbf{n})\|_2^2\\
\leq&\frac{1}{(1-\delta_{q-p})^2(1-(\frac{\delta_q}{1-\delta_q})^2)}\left(\|\mathbf{A}'_{T^{q-p}}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c}+\mathbf{n})\|_2^2+\|\mathbf{A}'_{T^{q-p}}\mathbf{A}_{T^p}\mathbf{A}^\dagger_{T^p}(\mathbf{A}_{(T^q)^c}\mathbf{x}_{(T^q)^c}+\mathbf{n})\|_2^2\right)\\
\leq&\frac{1}{(1-\delta_{q-p})^2(1-(\frac{\delta_q}{1-\delta_q})^2)}\left(\delta^2_{|(T^q)^c|+q-p+1}+\left(\frac{\delta_q\delta_{|(T^q)^c|+p+1}}{1-\delta_p}\right)^2\right)(\|\mathbf{x}_{(T^q)^c}\|_2^2+\sigma_{\mathrm{n}}^2)\\
=&\eta(\|\mathbf{x}_{(T^q)^c}\|_2^2+\sigma_{\mathrm{n}}^2),
\end{aligned} \tag{102}$$

where

$$\eta = \frac{1}{(1-\delta_{q-p})^2\left(1-\left(\frac{\delta_q}{1-\delta_q}\right)^2\right)}\left(\delta^2_{|(T^q)^c|+q-p+1}+\left(\frac{\delta_q\delta_{|(T^q)^c|+p+1}}{1-\delta_p}\right)^2\right).$$

If, e.g., $\delta_d \leq 0.1$, one has,

$$\eta \leq 0.0127, \tag{103}$$

which completes the proof. ■

References

[1] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candès, "Compressive sampling," in *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pp. 1433–1452, 2006.

[3] R. Venkataramani and Y. Bresler, "Sub-nyquist sampling of multiband signals: perfect reconstruction and bounds on aliasing error," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 3, pp. 1633–1636, IEEE, 1998.

[4] E. J. Candes and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.

[5] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.

[6] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of computational mathematics*, vol. 9, no. 3, pp. 317–334, 2009.

[7] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 58, no. 2, pp. 1094–1121, 2012.

[8] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2230–2249, 2009.

[9] D. Needell and J. A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[10] P. Boufounos, M. F. Duarte, and R. G. Baraniuk, "Sparse signal reconstruction from noisy compressive measurements using cross validation," in *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*, pp. 299–303, IEEE, 2007.

[11] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach.* Prentice/Hall International Englewood Cliffs, NJ, 1982.

[12] S. Geisser, *Predictive interference: an introduction*, vol. 55. CRC Press, 1993.

[13] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, pp. 1137–1145, 1995.

[14] R. Ward, "Compressed sensing with cross validation," *Information Theory, IEEE Transactions on*, vol. 55, no. 12, pp. 5773–5782, 2009.

[15] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5406–5425, 2006.

[16] M. A. Davenport, P. T. Boufounos, and R. G. Baraniuk, "Compressive domain interference cancellation," tech. rep., DTIC Document, 2009.

[17] D. Donoho, V. Stodden, and Y. Tsaig, "Sparselab. 2007."