THE ROBUST THEORY OF ADAPTIVE COMPRESSIVE SAMPLING (ACS)

CHRISTOPHER J. HILLAR AND DARREN RHEA

1. Introduction

We need to generalize Theorem 1 below to noise case (deterministic error).

2. The ACS Reconstruction Theorem

We shall assume throughout that patterns $\mathbf{x} \in \mathbb{R}^n$ in the sender region have *k-sparse* (simply called *sparse* if k is understood) causes $\mathbf{a}$ in a fixed *dictionary* $\Psi \in \mathbb{R}^{n \times p}$; that is, each pattern $\mathbf{x} \in \mathbb{R}^n$ can be expressed as $\mathbf{x} = \Psi\mathbf{a}$ for a column vector $\mathbf{a} \in \mathbb{R}^p$ with only $k \ll n \leq p$ nonzero entries. The *sampling matrix* $\Phi \in \mathbb{R}^{m \times n}$ is also assumed to satisfy a *compressive sampling* condition with respect to dictionary Ψ. Specifically, we assume that the matrix $A = \Phi\Psi$ is injective on the set of sparse vectors:

$$A\mathbf{a}_1 = A\mathbf{a}_2 \text{ for } k\text{-sparse } \mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^p \implies \mathbf{a}_1 = \mathbf{a}_2. \tag{1}$$

Clearly, this is a *necessary* condition for recovering sparse causes of sampled patterns. It turns out that condition (1) is also *sufficient*, and this is the main content of the ACS Theorem (Theorem 2).

We next describe very general conditions under which (1) holds for the matrix $A = \Phi\Psi$. As is well-known in the compressive sampling community, assumption (1) is fulfilled with high probability for suitably random[1] Φ and *any* fixed orthogonal (square) matrix Ψ as long as

$$m \geq Ck \log n. \tag{2}$$

Here, C is an absolute constant that does not depend on n or k. In other words, condition (1) holds with high probability when the sampling size m is at least nearly linear in the complexity k of the signals. The logarithmic term in (2) is a mild but necessary penalty for the ambient dimensionality n of the signals. For a more detailed discussion of these facts (including proofs) and their relationship to approximation theory and concentration of measure phenomenon, we refer the reader to [1] and the references therein. When Ψ is not orthogonal (and possibly nonsquare), assumption (1) is still fulfilled with very high probability for random Φ as long as (2) holds and Ψ has certain incoherence properties [2, 5].

[1] Somewhat surprisingly, there is no known deterministic construction of such a Φ even though "most" matrices will work.

Briefly in words, the *adaptive compressive sampling* (ACS) scheme is a (unsupervised) dictionary learning [4] of linearly sampled signals. Again in words, when we say that ACS has *converged* on a sparsity inducing dictionary Θ, we mean that compressed, sparsely encoded signals can be represented as a sparse linear combination of columns of the dictionary Θ, the representation being inferred by a sparse recovery procedure f. The precise mathematical definition is as follows.

Definition 1. *We say that ACS learning has converged on a sparsity-inducing dictionary $\Theta \in \mathbb{R}^{m\times p}$ if the sampling $\mathbf{y} = \Phi\mathbf{x}$ by the compression matrix Φ of the sender region's signal $\mathbf{x} = \Psi\mathbf{a}$ (with $\mathbf{a}$ k-sparse) always satisfies*

$$\mathbf{y} = \Theta\mathbf{b}$$

for a k-sparse vector $\mathbf{b} = \widehat{\mathbf{b}}(\mathbf{y})$ inferred from the convex optimization:[2]

$$\widehat{\mathbf{b}}(\mathbf{y}) = \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ ||\mathbf{y} - \Theta\mathbf{b}||_2^2 + \lambda|\mathbf{b}|_1 \right\}. \tag{3}$$

In other words, ACS converges when the learning has succeeded to make Θ a sparse dictionary of the compressed data $\mathbf{y}$.

As we explain in this section, once ACS has converged, the sparse causes $\mathbf{a}$ of an uncompressed signal $\mathbf{x} = \Psi\mathbf{a}$ are faithfully recovered by the output $\widehat{\mathbf{b}}$ of ACS (see Figure **??** and Theorem 2 below). More precisely, we prove that any procedure which takes input $\mathbf{y} = \Phi\Psi\mathbf{a}$ (with $\mathbf{a}$ sparse) and produces sparse vectors $f(\mathbf{a})$ satisfying

$$\mathbf{y} = \Theta f(\mathbf{a})$$

for a matrix Θ must necessarily recover the sparse causes $\mathbf{a}$ up to a fixed diagonal scaling D and permutation (or relabeling) P; that is, $f(\mathbf{a}) = PD\mathbf{a}$.

This result, the content of Theorem 1 below, is surprising and remarkably general; it says that any dictionary learning scheme producing sparse reconstructions in a compressed space automatically gives faithful transmission of sparse signals – regardless of the original dictionary Ψ or sampling matrix Φ.[3] The recovery method (3) for our choice of procedure is natural in this context because of its efficient use in compressive sampling [3] to recover sparse vectors.

Although self-contained and (mathematically) elementary, our proof of Theorem 1 relies on abstract ideas from Ramsey theory (see Theorem **??** below). Recall that a *permutation matrix* is a $\{0, 1\}$-matrix P with exactly one 1 in each column and exactly one 1 in each row (thus, $P\mathbf{v}$ for a vector $\mathbf{v}$ just permutes its entries).[4]

---

[2]For some fixed $\lambda > 0$. Here, we recall that for a column vector $\mathbf{b} = (b_1, \ldots, b_p)^\top \in \mathbb{R}^p$, the ℓ_1 norm of $\mathbf{b}$ is $|\mathbf{b}|_1 = |b_1| + \cdots + |b_p|$. Also, for a vector $\mathbf{z} = (z_1, \ldots, z_m)^\top \in \mathbb{R}^m$, the ℓ_2 norm of $\mathbf{z}$ is $||\mathbf{z}||_2 = (z_1^2 + \cdots + z_m^2)^{1/2}$.

[3]Of course, as long as $A = \Phi\Psi$ satisfies the necessary condition for recovery (1).

[4]Also, $PP^\top = P^\top P = I$, where I denotes the $p \times p$ identity matrix, and $M^\top$ for a matrix M is its transpose.

Theorem 1. *Suppose that $f : \mathbb{R}^p \to \mathbb{R}^p$ is a map sending k-sparse vectors to k-sparse vectors. Also, suppose that $A \in \mathbb{R}^{m\times p}$ satisfies compressive sampling condition (1) and that $B \in \mathbb{R}^{m\times p}$ is a matrix such that:*

$$A\mathbf{a} = Bf(\mathbf{a}), \quad \text{for all } k\text{-sparse } \mathbf{a} \in \mathbb{R}^p. \tag{4}$$

Then, there exists an invertible diagonal matrix $D \in \mathbb{R}^{p\times p}$ and a permutation matrix $P \in \mathbb{R}^{p\times p}$ such that

$$A = BPD \tag{5}$$

and for all k-sparse $\mathbf{a}$,

$$f(\mathbf{a}) = PD\mathbf{a}. \tag{6}$$

Remark 1. *For those readers with some abstract algebra experience, we remark that our proof of Theorem 1 generalizes easily to the case when $\mathbb{R}$ is replaced by any field such as the rational numbers $\mathbb{Q}$.*

Before proving this (seemingly technical) theorem above, we explain how our main application, the ACS Theorem, follows directly from it.

Theorem 2 (The ACS Theorem)**.** *Suppose that ACS converges on a sparsity-inducing dictionary $\Theta \in \mathbb{R}^{m\times p}$. If $A = \Phi\Psi$ satisfies the compressive sampling condition (1), then there is a fixed invertible diagonal matrix $D \in \mathbb{R}^{p\times p}$ and a fixed permutation matrix $P \in \mathbb{R}^{p\times p}$ such that:*

$$\Phi\Psi = \Theta PD \tag{7}$$

and for all signals $\mathbf{x} = \Psi\mathbf{a}$ (with $\mathbf{a}$ k-sparse) which sample to $\mathbf{y} = \Phi\mathbf{x}$, the ACS output $\mathbf{b} = \widehat{\mathbf{b}}(\mathbf{y})$ from (3) satisfies

$$\mathbf{b} = PD\mathbf{a}. \tag{8}$$

Proof. Let $\mathbf{x} = \Psi\mathbf{a}$ for a sparse vector $\mathbf{a}$, and set $\mathbf{y} = \Phi\mathbf{x}$. By assumption, the sparse output $\mathbf{b} = \widehat{\mathbf{b}}$ of ACS satisfies $\mathbf{y} = \Theta\mathbf{b}$. It follows that

$$\Phi\Psi \cdot \mathbf{a} = \Theta \cdot \widehat{\mathbf{b}}(\mathbf{y}), \quad \text{for all } k\text{-sparse } \mathbf{a}. \tag{9}$$

Set $A = \Phi\Psi$, $B = \Theta$, and $f(\mathbf{a}) = \widehat{\mathbf{b}}(\Phi\Psi\mathbf{a})$. We now use statement (9) and Theorem 1 to conclude directly that (7) and (8) hold. □

Corollary 1. *Suppose $\Phi = I$ so that ACS reduces to sparse coding* [4]*. If ACS converges on a sparsity-inducing dictionary Θ, then it must be a scaled permutation of the original dictionary Ψ.*

Proof. Since Φ is the identity matrix, it automatically satisfies (1). Theorem 2 then says that $\Theta = \Psi D^{-1} P^\top$ for an invertible diagonal matrix D and permutation P. □

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[3] E.J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452. Citeseer, 2006.

[4] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[5] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. Technical Report arXiv:0910.2042, UC Berkeley, Department of Statistics, 2009.

Mathematical Sciences Research Institute, 17 Gauss Way, Berkeley, CA 94720

Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA 94720

E-mail address: chillar@msri.org

Department of Mathematics, University of California, Berkeley, CA 94720