On Low Rank Matrix Approximations with Applications to Synthesis Problem in Compressed Sensing

Anatoli Juditsky[*] Fatma Kılınç Karzan[†] Arkadi Nemirovski[‡]

May 23, 2011

Abstract

We consider the *synthesis problem* of Compressed Sensing – given s and an $M \times n$ matrix A, extract from A an $m \times n$ submatrix A_m, with m as small as possible, which is *s-good*, that is, every signal x with at most s nonzero entries can be recovered from observation $A_m x$ by ℓ_1 minimization: $x = \mathrm{argmin}_u\{\|u\|_1 : A_m u = A_m x\}$. We show that under reasonable assumptions the synthesis problem can be reformulated as the problem of *entry-wise* approximation, within a given accuracy, of $n \times n$ matrix $W = Y^T A$, with $Y \in \mathbb{R}^{M \times n}$ given, by a matrix of the form $Y_m^T A_m$, with A_m comprised of m rows of A. We propose randomized algorithms for efficiently solving the latter problem with accuracy guaranties $\mathbf{E}\{\|W - Y_m^T A_m\|_\infty\} \leq O(1)L(Y,A)\sqrt{\frac{\ln(n)}{m}}$. Here $L(Y,A)$ is an easy-to-specify quantity which in good cases is a moderate absolute constant (e.g., $L(A,A) = 1$ when A is the Hadamard matrix, and similarly for the matrix of Fourier transform on any finite Abelian group). We also supply derandomized versions of the approximation algorithms which do not require random sampling of matrices and attain the same accuracy bounds. We further demonstrate that in terms of approximation accuracy our algorithms are optimal up to logarithmic in n factors. Finally, we provide preliminary numerical results on the performance of our algorithms for the synthesis problem.

1 Introduction

Let $A \in \mathbb{R}^{m \times n}$ be a matrix with $m < n$. Compressed Sensing focuses on recovery of a *sparse* signal $x \in \mathbb{R}^n$ from its noisy observations

$$y = Ax + e,$$

[*] LJK, Université J. Fourier, B.P. 53, 38041 Grenoble Cedex 9, France, `Anatoli.Juditsky@imag.fr`

[†] Georgia Institute of Technology, Atlanta, Georgia 30332, USA, `fkilinc@isye.gatech.edu`

[‡] Georgia Institute of Technology, Atlanta, Georgia 30332, USA, `nemirovs@isye.gatech.edu`. Research of the second and the third authors was supported by the Office of Naval Research grant # N000140811104 and the NSF grant DMS-0914785.

where e is an observation noise such that $\|e\| \leq \delta$ for certain known norm on $\mathbb{R}^m$ and some given δ. The standard recovering routine is

$$\widehat{x} \in \operatorname*{Argmin}_{w}\{\|w\|_1 : \|Aw - y\| \leq \delta.\}.$$

We call the matrix A *s-good* if whenever the true signal x is s-sparse (i.e., has at most s nonzero entries) and there is no observation errors ($\delta = 0$), x is the unique optimal solution to the optimization program $\min\{\|w\|_1 : Aw = Ax\}$.

To the best of our knowledge, nearly the strongest verifiable sufficient condition for A to be s-good is as follows (cf [7]):

$$\text{There exists } Y \in \mathbb{R}^{m\times n} \text{ such that } \|I_n - Y^T A\|_\infty < \frac{1}{2s} \tag{1}$$

(here and in what follows $\|X\|_\infty = \max\limits_{i,j} |X_{ij}|$, X_{ij} being the elements of X).[1]

In this paper we consider the *synthesis problem* of Compressed Sensing as follows:

> *Given s and an $M \times n$ matrix A, extract from it an $m \times n$ submatrix A_m, certified to be s-good, with m as small as possible.*

One can think, e.g., of a spatial or planar n-point grid $\mathcal{E}$ of possible locations of signal sources and an M-element grid $\mathcal{S}$ of possible locations of sensors. A sensor in a given location measures a known, depending on the location, linear form of the signals emitted at the nodes of $\mathcal{E}$, and the goal is to place a given number $m \ll M$ of sensors at the nodes of $\mathcal{S}$ in order to be able to recover the location of sources via the ℓ_1-minimization, conditioned that there are s sources at most. Since the property of s-goodness is difficult to verify, we will look for a submatrix of the original matrix A for which the s-goodness can be certified by the sufficient condition (1). Suppose that along with A we know an $M \times n$ matrix Y_M which certifies that the "level of goodness" of A is at least s, that is, we have

$$\|I_n - Y_M^T A\|_\infty \leq \mu < \frac{1}{2s}. \tag{2}$$

Then we can approach the synthesis problem as follows:

> Given $M \times n$ matrices Y_M and A and a tolerance $\epsilon > 0$, we want to extract from A m rows (the smaller is m, the better) to get an $m \times n$ matrix A_m which, along with properly chosen $Y_m \in \mathbb{R}^{m\times n}$, satisfies the relation $\|Y_M^T A - Y_m^T A_m\|_\infty \leq \epsilon$.

Choosing $\epsilon < \frac{1}{2s} - \mu$ and invoking (2), we ensure that the output A_m of the above procedure is s-good. This simple observation motivates our interest to the problem of approximating a given matrix by a matrix of specified (low rank) in the *uniform norm.*

[1] We address the reader to [7] for details concerning the derivation, the link to the necessary and sufficient condition of s-goodness and its comparison to traditional non-verifiable sufficient conditions for s-goodness based on Restricted Isometry or Restricted Eigenvalue Property and a verifiable sufficient condition based on mutual incoherence.

Note that in the existing literature on low rank approximation of matrices the emphasis is on efficient construction when the approximation error is measured in the Frobenius norm (for the Frobenius norm $\|A\|_F = \left(\sum_{i,j} A_{ij}^2\right)^{1/2}$). Though the Singular Value Decomposition (SVD) gives the best rank k approximation in terms of all the norms that are invariant under rotation (e.g., the Frobenius norm and the spectral norm), its computational cost may be prohibitive for applications involving large matrices. Recently, the properties of fast low rank approximations in the Frobenius norm based on the randomized sampling of rows (or columns) of the matrix (see, e.g., [4, 5]) or random sampling of a few individual entries (see [1] and references therein) has been studied extensively. Another randomized fast approximation based on the preprocessing by the Fast Fourier Transform or Fast Hadamard Transform has been studied in [8]. Yet we do not know explicit bounds available from the previous literature which concern numerically efficient low rank approximations in the uniform norm.

The only known to us result on low rank approximation of matrices in uniform norm is the one in [9]; it states then if $W = Y^T A \in \mathbb{R}^{m\times n}$ and the rows in Y, A are of Euclidean length at most D, then, for every k, W admits a k-rank approximation $W_k = Y_k^T A_k$, where Y_k and A_k are $k \times m$ and $k \times n$ matrices with rows which are linear combinations of those in Y, A, respectively, such that $\|W - W_k\|_\infty \leq O(1)D^2\sqrt{\ln(mn)/k}$. This result does not help in the synthesis problem, where we want the rows of A_k to be just rows of A, and not linear combinations of these rows.

The main result of this paper is as follows. Let $W = Y^T A$, where Y and A are known $M \times n$ matrices. We consider the approximation $W_k = Y_k^T A_k$ of W such that the matrices Y_k and A_k of dimension $m_k \times n$, $m_k \leq k \leq M$, are composed of multiples of the rows of the matrices Y and A respectively[2]. We show that a fast (essentially, of numerical complexity $O(kMn^2)$) approximation W_k can be constructed which satisfies

$$\|W - W_k\|_\infty = O(1)L(Y,A)\sqrt{\frac{\ln(n)}{k}},$$

where $L(Y,A) = \sum_i \|y_i\|_\infty \|a_i\|_\infty$ and y_i^T, a_i^T denote the i-th rows of Y and A respectively. Note that for moderate values of $L(Y,A) = O(1)$ and $k < n/2$ this approximation is "quasi-optimal", as we know (cf., e.g. [7, Proposition 4.2]) that (for certain matrices W) the accuracy of such an approximation cannot be better than $O(k^{-1/2})$. Moreover, in Section 2.4, we show that when W is an $n \times n$ identity matrix, as in the case of Compressed Sensing Synthesis problem, the above bound is unimprovable up to a logarithmic factor. See also Section 2.3 for a discussion of how large $L(Y,A)$ can be in the case of A being a Hadamard matrix. We propose two types of construction of fast approximations: we consider the randomized construction, for which the accuracy bounds above hold in expectation (or with significant probability). We also supply "derandomized" versions of the approximation algorithms which do not require random sampling of matrices and attain the same accuracy bounds as the randomized method.

[2] Allowing rows of A_k to be *multiples* of rows of A in our context is the same as to require the rows of A_k to be among the rows of A – the corresponding factors can be moved from rows of A_k to those of Y_k.

2 Low rank approximation in Compressed Sensing

In this section we suppose to be given s and an $M \times n$ matrix A and our objective is to extract from A a submatrix A_k which is composed of, at most, k rows of A, with as small k as possible, which is s-good. We assume that A admits a "goodness certificate" Y. Namely, we are given an $M \times n$ matrix Y such that

$$\mu := \|I_n - Y^T A\|_\infty < \frac{1}{2s}, \tag{3}$$

and we are looking for A_k and the corresponding Y_k such that $\|I_n - Y_k^T A_k\| < \frac{1}{2s}$.

2.1 Random sampling algorithm

The starting point of our developments is the following simple

Lemma 2.1 *Let for $\beta > 0$, let*

$$V_\beta(z) = \beta \ln \left(\sum_{i=1}^d \cosh \left(\frac{z_i}{\beta} \right) \right) - \beta \ln d : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}_+. \tag{4}$$

Then

(i) we have $\|z\|_\infty - \beta \ln(2d) \le V_\beta(z) \le \|z\|_\infty$;

(ii) if $\beta_1 \le \beta_2$ then $V_{\beta_1}(z) \ge V_{\beta_2}(z)$;

(iii) function V_β is convex and continuously differentiable on $\mathbb{R}^d$. Further, its gradient V'_β is Lipschitz-continuous with the constant β^{-1}:

$$\|V'_\beta(z_1) - V'_\beta(z_2)\|_1 \le \beta^{-1} \|z_1 - z_2\|_\infty, \tag{5}$$

and $\|V'_\beta(z)\|_1 \le 1$ for all $z \in \mathbb{R}^d$.

For proof, see Appendix A.

Lemma 2.1 has the following immediate consequence:

Proposition 2.1 *Let $\beta \ge \beta' > 0$ (non-random) and let ξ_1,...,ξ_k be random vectors in $\mathbb{R}^d$ such that $\mathbf{E}\{\xi_1\} = 0$ and $\mathbf{E}\{\xi_i | \xi_1, ..., \xi_{i-1}\} = 0$ a.s. for all $i \in \{2, \dots, k\}$, and $\mathbf{E}\{\|\xi_i\|_\infty^2\} \le \sigma_i^2 < \infty$ for all $i \in \{1, \dots, k\}$, and let $S_k = \sum_{i=1}^k \xi_i$ and $S_0 = 0$. Then for $k \ge 1$*

$$\mathbf{E}\{V_\beta(S_k)\} \le \mathbf{E}\{V_{\beta'}(S_{k-1})\} + \frac{\sigma_k^2}{2\beta}. \tag{6}$$

As a result,

$$\mathbf{E}\left\{\|S_k\|_\infty\right\} \le \sqrt{2\ln(2d) \sum_{i=1}^k \sigma_i^2}. \tag{7}$$

Proof. Let $\beta \geq \beta'$. By applying items (ii) and (iii) of the lemma for $k \geq 1$ we get:

$$\begin{array}{l} V_\beta(S_k) \leq V_\beta(S_{k-1}) + \langle V'_\beta(S_{k-1}),\, \xi_k\rangle + \frac{1}{2\beta}\|\xi_k\|^2_\infty \\ \leq V_{\beta'}(S_{k-1}) + \langle V'_\beta(S_{k-1}),\, \xi_k\rangle + \frac{1}{2\beta}\|\xi_k\|^2_\infty \end{array}$$

When taking the expectation (first conditional to $\xi_1, ..., \xi_{k-1}$), due to $\mathbf{E}\{\xi_k|\xi_1, ..., \xi_{k-1}\} = 0$ a.s. for $k \geq 2$ and then using $\mathbf{E}\{\langle V'_\beta(S_0), \xi_1\rangle\} = 0$ (due to $\mathbf{E}\{\xi_1\} = 0$), we obtain for $k \geq 1$

$$\mathbf{E}\{V_\beta(S_k)\} \leq \mathbf{E}\{V_{\beta'}(S_{k-1})\} + \frac{\mathbf{E}\{\|\xi_k\|^2_\infty\}}{2\beta} \leq \mathbf{E}\{V_{\beta'}(S_{k-1})\} + \frac{\sigma_k^2}{2\beta},$$

which is (6). Now let us set $\beta' = \beta = \sqrt{\frac{\sum_{i=1}^k \sigma_i^2}{2\ln(2d)}}$. Since $V_\beta(0) = 0$ we conclude that

$$\mathbf{E}\{V_\beta(S_k)\} \leq \sum_{i=1}^k \frac{\sigma_i^2}{2\beta}.$$

On the other hand, by item (i) of Lemma 2.1,

$$\mathbf{E}\{\|S_k\|_\infty\} \leq \beta\ln(2d) + \mathbf{E}\{V_\beta(S_k)\} \leq \beta\ln(2d) + \sum_{i=1}^k \frac{\sigma_i^2}{2\beta} \leq \sqrt{2\ln(2d)\sum_{i=1}^k \sigma_i^2}$$

proving (7). □

The random sampling algorithm. Denoting y_i^T and a_i^T, $i = 1, ..., M$, i-th rows of Y and A, respectively, let us set

$$\theta_i = \|y_i\|_\infty\, \|a_i\|_\infty, \quad L = \sum_i \theta_i, \quad \pi_i = \frac{\theta_i}{L}, \quad z_i = \frac{L}{\theta_i} y_i, \tag{8}$$

and let $W = Y^T A$. Observe that

$$\begin{array}{rcl} W & = & \sum_{i=1}^M \pi_i \left(z_i a_i^T\right), \\ \|z_i a_i^T\|_\infty & = & L, \;\; 1 \leq i \leq M, \\ \sum_{i=1}^M \pi_i & = & 1, \;\; \pi_i \geq 0, \;\; 1 \leq i \leq M. \end{array} \tag{9}$$

Now let Ξ be random rank 1 matrix taking values $z_i a_i^T$ with probabilities π_i, and let $\Xi_1, \Xi_2, ...$ be a sample of independent realizations of Ξ. Consider the random matrix

$$W_k = \frac{1}{k}\sum_{\ell=1}^k \Xi_\ell.$$

Then W_k is, by construction, of the form $Y_k^T A_k$, where A_k is a random $m_k \times n$ submatrix of A with $m_k \leq k$.

As an immediate consequence of Proposition 2.1 we obtain the following statement:

Proposition 2.2 *One has*

$$\mathbf{E}\left\{\|W_k - W\|_\infty\right\} \leq 2Lk^{-1/2}\sqrt{2\ln(2n^2)}. \tag{10}$$

In particular, the probability of the event

$$\mathcal{E} = \{\Xi_1, ..., \Xi_k : \|W_k - W\|_\infty \leq 4Lk^{-1/2}\sqrt{2\ln(2n^2)}\}$$

is $\geq 1/2$, and whenever this event takes place, we have in our disposal a matrix Y_k and a $m_k \times n$ submatrix A_k of A with $m_k \leq k$ such that

$$\|I_n - Y_k^T A_k\|_\infty \leq \|I_n - W\|_\infty + \|W_k - W\|_\infty \leq \mu_k := \mu + 4Lk^{-1/2}\sqrt{2\ln(2n^2)}. \tag{11}$$

Proof. By (9) we have $\|z_i a_i^T\|_\infty = L$ for all i, and besides this, treating $\imath$ as random index distributed in $\{1, ..., M\}$ according to probability distribution $\pi = \{\pi_i\}_{i=1}^M$, we have $\mathbf{E}\{z_\imath a_\imath^T\} = W$. It follows that $\|\Xi_\ell - W\|_\infty \leq 2L$ and $\mathbf{E}\{\Xi_\ell - W\} = 0$. If we denote $S_i = \sum_{\ell=1}^i (\Xi_\ell - W)$, when applying Proposition 2.1 we obtain

$$\mathbf{E}\{\|S_k\|_\infty\} \leq 2L\sqrt{2k\ln(2n^2)},$$

and we arrive at (10). □

Discussion. Proposition 2.2 suggests a certain approach to the synthesis problem. Indeed, according to this Proposition, picking at random k rows $a_{i_\ell}^T$, where $i_1, ..., i_k$ are sampled independently from the distribution π, we get with probability at least 1/2 a random $m_k \times n$ matrix A_k, $m_k \leq k$, which is provably s-good with $s = O(1)(L\sqrt{\ln(n)/k} + \mu)^{-1}$. When $L = O(1)$, this is nearly as good as it could be, since the sufficient condition for s-goodness stated in (1) can justify s-goodness of an $m \times n$ sensing matrix with $n > O(1)m$ only when $s \leq O(1)\sqrt{m}$, see [7, Proposition 4.2].

2.2 Derandomization

Looking at the proof of Proposition 2.1, we see that the construction of A_k and Y_k can be derandomized. Indeed, (6) implies that

> *Whenever $S \in \mathbb{R}^{n\times n}$ and $\beta \geq \beta'$ there exists i such that*
>
> $$V_\beta(S + (z_i a_i^T - W)) \leq V_{\beta'}(S) + \frac{2L^2}{\beta}.$$
>
> *Specifically, the above bound is satisfied for every i such that*
>
> $$\langle V'_\beta(S), z_i a_i^T - W\rangle \leq 0,$$
>
> and because $\pi_i \geq 0\ \forall i$ and $\sum_i \pi_i(z_i a_i^T - W) = 0$, the latter inequality is certainly satisfied for some i.

Now assume that given a sequence $\beta_1 \leq \beta_2 \leq ...$ of positive reals, we build a sequence of matrices S_i according to the following rules:

1. $S_0 = 0$;
2. $S_{k+1} = S_k + (v_k a_{\ell_k}^T - W)$ with $\ell_k \in \{1, ..., M\}$ and $v_k \in \mathbb{R}^n$ such that

$$V_{\beta_{k+1}}(S_{k+1}) \leq V_{\beta_k}(S_k) + \delta_{k+1}, \;\; \delta_{k+1} \leq \frac{2L^2}{\beta_{k+1}}, \tag{12}$$

   where by definition $V_{\beta_0}(0) = 0$.

Then for every $k \geq 1$ the matrix $U_k = k^{-1} S_k$ is of the form $Y_k^T A_k - W$, where A_k is a $m_k \times n$ submatrix of A with $m_k \leq k$, and

$$\|S_k\|_\infty \leq \beta_k \ln(2n^2) + \sum_{\ell=1}^{k} \delta_\ell,$$

whence

$$\|Y_k^T A_k - I_n\|_\infty \leq \mu + k^{-1} \left(\beta_k \ln(2n^2) + \sum_{\ell=1}^{k} \delta_\ell \right).$$

In particular, for the choice $\beta_\ell = 2L\sqrt{\frac{\ell}{\ln(2n^2)}}$, $\ell = 1, 2, ...$, we obtain[3]

$$\|Y_k^T A_k - I_n\|_\infty \leq \mu + 4L\sqrt{\frac{\ln(2n^2)}{k}}$$

One can consider at least the following three (numerically efficient) policies for choosing v_k and ℓ_k satisfying (12); we order them according to their computational complexity.

A. Given S_k, we test one by one the options $\ell_k = i$, $v_k = z_i$, $i = 1, ..., M$, until an option satisfying (12) is met (or test all the n options and choose the one which results in the smallest $V_{\beta_{k+1}}(S_{k+1})$). Note that accomplishing a step of this scheme requires $O(Mn^2)$ elementary operations.

A′. In this version of A, we test the options $\ell_k = i$, $v_k = z_i$ when picking i at random, as independent realizations of the random variable $\imath$ taking values $1, ..., M$ with probabilities π_i, until an option with $\langle V'_{\beta_{k+1}}(S_k), z_i a_i^T - W \rangle \leq 0$ is met. Since $\mathbf{E}\left\{ \langle V'_{\beta_{k+1}}(S_k), z_i a_i^T - W \rangle \right\} \leq 0$, we may hope that this procedure will take essentially less steps than the ordered scan through the entire range $1, ..., M$ of values of i.

[3] for a given k, setting $\beta_\ell = L\sqrt{\frac{2k}{\ln(2n^2)}}$, $1 \leq \ell \leq k$, the right hand side in the bound can be reduced to $\mu + 2L\sqrt{\frac{2\ln(2n^2)}{k}}$.

B. Given S_k we solve M one-dimensional convex optimization problems

$$t_i^* \in \operatorname*{Argmin}_{t\in\mathbb{R}_+} V_{\beta_{k+1}}(S_k + tz_ia_i^T - W),\ 1 \le i \le M, \tag{13}$$

then select the one, let its index be i_*, with the smallest value of $V_{\beta_{k+1}}(S_k + t_i^* z_i a_i^T - W)$, and put $v_k = t_{i_*}^* z_{i_*}$, $\ell_k = i_*$.

If the bisection algorithm is used to find t_i^*, solving the problem (13) for one i to the relative accuracy ϵ requires $O(n^2 \ln(1/\epsilon))$ elementary operations. The total numerical complexity of the step of the method is $O(Mn^2 \ln(1/\epsilon))$.

C. Given S_k, we solve M convex optimization problems

$$u_i^* \in \operatorname*{Argmin}_{u\in\mathbb{R}^n} V_{\beta_{k+1}}(S_k + ua_i^T - W),\ \ 1 \le i \le M, \tag{14}$$

then select the one, let its index be i_*, with the smallest value of $V_{\beta_{k+1}}(S_k + u_i^* a_i^T - W)$, and set $v_k = u_i^*$, $\ell_k = i_*$.

Note that due to the structure of V_β to solve (14) it suffices to find a solution to the system

$$\begin{array}{l} \sum_{\ell=1}^n \gamma_\ell \sinh(\alpha_{j\ell} + \gamma_\ell u_j) = 0, \\ \alpha_{j\ell} = \frac{[S_k]_{j\ell} - [W]_{j\ell}}{\beta_k},\ \ \gamma_\ell = \frac{[A]_{\ell i}}{\beta_k},\ \ 1 \le j, \ell \le n. \end{array} \tag{15}$$

Since the equations of the system (15) are independent, one can use bisection to find the component u_j of the solution.[4] Finding a solution of relative accuracy ϵ to each equation then requires $O(n \ln(1/\epsilon))$ arithmetical operations, and the total complexity of solving (14) becomes $O(Mn^2 \ln(1/\epsilon))$.

Selecting Y and W. Note that the numerical schemes of this section should be initialized with matrices Y and $W = Y^TA$. We can do as follows:

1. We start with solving the problem

$$Y \in \operatorname*{Argmin}_{Z=[z_1^T;\ldots;z_M^T]\in\mathbb{R}^{M\times n}} \left\{ \sum_{i=1}^M \|z_i\|_\infty \|a_i^T\|_\infty :\ \|I_n - Z^TA\|_\infty \le \mu \right\},$$

   where μ is a certain fraction of $\frac{1}{2s}$. Assuming the problem is feasible for the chosen μ, we get in this way the "initial point" – the matrix $W = Y^TA$.

2. Then we apply the outlined procedure to find A_k and Y_k. At each step ℓ of this procedure, we get certain $m_\ell \times n$ submatrix A_ℓ of A and a matrix Y_ℓ. When $\|I_n - Y_\ell^T A_\ell\|_\infty$ becomes less than $\frac{1}{2s}$ we terminate. Alternatively, we can solve at each step ℓ an auxiliary problem $\min\limits_{U\in\mathbb{R}^{m_\ell\times n}} \|I_n - U^TA_\ell\|_\infty$ and terminate when the optimal value in this problem becomes less than $\frac{1}{2s}$.

[4] Note that due to the convexity of the left-hand side of the equation in (15), even faster algorithm of Newton family can be used.

2.3 Numerical illustration

Here we report on preliminary numerical experiments with the synthesis problem as posed in the introduction. In our experiment, A is square, specifically, this is the Hadamard matrix H_{11} of order 2048.

> Recall that the Hadamard matrix H_ν, $\nu = 0, 1, \ldots$ is a square matrix of order 2^ν given by the recurrence
>
> $$H_0 = 1,\ H_{s+1} = \left[\begin{array}{cc} H_s & H_s \\ H_s & -H_s \end{array}\right],$$
>
> whence H_ν is a symmetric matrix with entries ±1 and $H_\nu^T H_\nu = 2^\nu I_{2^\nu}$.

The goal of the experiment was to extract from $A = H_{11}$ an $m \times 2048$ submatrix A_m which satisfies the relation (cf. (1))

$$\mathrm{Opt}(A_m) := \min_{Y_m \in \mathbb{R}^{m\times n}} \|I_n - Y_m^T A_m\|_\infty < \frac{1}{2s},\ n = 2048 \tag{16}$$

with $s = 10$; under this requirement, we would like to have m as small as possible. In Compressed Sensing terms, we are trying to solve the synthesis problem with $A = H_{11}$; in low rank approximation terms, we want to approximate I_{2048} in the uniform norm within accuracy < 0.05 by a rank m matrix of the form $Y_m^T A_m$, with the rows of A_m extracted from H_{11}. The advantages of the Hadamard matrix in our context is twofold:

1. The error bound (10) is proportional to the quantity L defined in (8). By the origin of this quantity, we clearly have $\|Y^T A\|_\infty = \|\sum_{i=1}^M y_i a_i^T\|_\infty \leq L$, whence $L \geq 1 - \mu > 1 - \frac{1}{2s} \geq 1/2$ by (3). On the other hand, with $A = H_\nu$ being an Hadamard matrix, setting $Y = 2^{-\nu} H_\nu$, so that $Y^T A = I_{2^\nu}$, we ensure the validity of (3) with $\mu = 0$ and get $L = 1$, that is, μ is as small as it could be, and L is nearly as small as it could be.

2. Whenever A_m is a submatrix of H_ν, the optimization problem in the left hand side of (16) is easy to solve.

Item 2 deserves an explanation. Clearly, the optimization program in (16) reduces to the series of $n = 2048$ LP programs

$$\mathrm{Opt}_i(A_m) = \min_{y \in \mathbb{R}^m} \|e_i - A_m^T y\|_\infty,\ 1 \leq i \leq n, \tag{17}$$

where e_i is the standard basic orth in $\mathbb{R}^n$, and $\mathrm{Opt}(A_m) = \max_i \mathrm{Opt}_i(A_m)$. The point is (for justification, see Appendix B) that *when A_m is an $m \times n$ submatrix of the $n \times n$ Hadamard matrix, $\mathrm{Opt}_i(A_m)$ is independent of i,* so that checking the inequality in (16) requires solving a *single* LP program with m variables rather than solving n LO programs of the same size.

The experiment was organized as follows. As it was already mentioned, we used $\nu = 11$ (that is, $n = 2048$) and $s = 10$ (that is, the desired uniform norm of approximating I_{2048} by $Y_m^T A_m$ was 0.05). We compared two approximation policies:

- "Blind" approximation – we choose a random permutation $\sigma(\cdot)$ of the indices $1, ..., 2048$ and look at the submatrices A^k, $k = 1, 2, ...$ obtained by extracting from H_{11} rows with indices $\sigma(1), \sigma(2), ..., \sigma(k)$ until a submatrix satisfying (16) is met. This is a refinement of the Random sampling algorithm as applied to $A = H_{11}$ and $Y = 2^{-11}A$, which results in $W = I_{2048}$. The refinement is that instead of looking for approximation of $W = I_{2048}$ of the form $\frac{1}{k}\sum_{\ell=1}^{k} z_{i_\ell} a_{i_\ell}^T$, where $i_1, i_2, ...$ are independent realizations of random variable $\imath$ taking values $1, ..., \mu$ with equal probabilities (as prescribed by (8) in the case of $A = H_\nu$), we look for the best approximation of the form $Y_k^T A^k$, where A^k is the submatrix of A with the row indices $\sigma(1), ..., \sigma(k)$.

- "Active" approximation, which is obtained from algorithm $\mathbf{A}'$ by the same refinement as in the previous item.

In our experiments, we ran every policy 6 times. The results were as follows:

"Blind" policy $\mathcal{B}$: the rank of 0.05-approximation of $W = I_{2048}$ varied from 662 to 680.

"Active" policy $\mathcal{A}$: the rank of 0.05-approximation of W varied from 617 to 630.

Note that in both algorithms the resulting matrix A_m is built "row by row", and the certified levels of goodness of the intermediate matrices $A^1, A^2, ...$ are computed. In the below table we indicate, for the most successful (resulting in the smallest m) of the 6 runs of each algorithm, the smallest values of k for which A^k was certified to be s-good, $s = 1, 2, ..., 10$:

s	1	2	3	4	5	6	7	8	9	10
$\mathcal{B}$	15	58	121	197	279	343	427	512	584	662
$\mathcal{A}$	12	47	104	172	246	323	399	469	547	617

Finally, we remark that with A being the Hadamard matrix H_ν, the "no refinement" versions of our policies would terminate according to the criterion $\|I_n - \frac{1}{k}A_k^T A_k\|_\infty < \frac{1}{2s}$, which, on a closest inspection, is nothing but a slightly spoiled version of the goodness test based on mutual incoherence [3][5]. In the experiments we are reporting, this criterion is essentially weaker that the one based on (16): for the best, over the 6 runs of the algorithms $\mathcal{A}$ and $\mathcal{B}$, 10-good submatrices A_m of H_{11} we got the test based on mutual incoherence certifies the levels of goodness as low as 5 (in the case of $\mathcal{B}$) and 7 (in the case of $\mathcal{A}$).

2.4 Lower bound

We have seen that if $Y^T A = W \in \mathbb{R}^{m\times n}$, then the $\|\cdot\|_\infty$-error of the best in this norm approximation of W by a matrix of rank k by selecting rows from Y and A is *at most* $O(1)L(Y, A)\sqrt{\frac{\ln(n)}{k}}$. We intend to demonstrate that in general this bound is unimprovable,

[5] The mutual incoherence test is as follows: given a $k \times n$ matrix $B = [b_1, ..., b_n]$ with nonzero columns, we compute the quantity $\mu(B) = \max\limits_{i\neq j} |b_i^T b_j|/b_i^T b_i$ and claim that B is s-good for all s such that $s < \frac{1+\mu(B)}{2\mu(B)}$. With the Hadamard A, the "no refinement" criterion for our scheme is nothing but $s < \frac{1}{2\mu(A^k)}$.

up to a logarithmic in m and n factor even when we are allowed to use any rank k matrix in the approximation. Specifically, the following result holds:

Proposition 2.3 *When $n \geq 2k$, the $\|\cdot\|_\infty$ error of any approximation of the unit matrix I_n by a matrix of rank k is at least*

$$\frac{1}{2\sqrt{k}}. \tag{18}$$

Proof [cf. [7, Proposition 4.2]] Let $\alpha(n,k)$ be the minimal $\|\cdot\|_\infty$ error of approximation of I_n by a matrix of rank $\leq k$; this function clearly is nondecreasing in n. Let ν be an integer such that $k < \nu \leq n$, and W be an $\nu \times \nu$ matrix of rank $\leq k$ such that $\|I_\nu - W\|_\infty = \alpha := \alpha(\nu,k)$. By variational characterization of singular values, at least $\nu - k$ singular values of $I_\nu - W$ are ≥ 1, whence $\mathrm{Tr}([I_\nu - W][I_\nu - W]^T) \geq \nu - k$. On the other hand, $\|I_\nu - W\|_\infty \leq \alpha$, whence $\mathrm{Tr}([I_\nu - W][I_\nu - W]^T) \leq \nu^2\alpha^2$. We conclude that $\alpha^2 \geq \frac{\nu-k}{\nu^2}$ for all ν with $k < \nu \leq n$, whence $\alpha^2 \geq \frac{1}{4k}$ when $n \geq 2k$. □

References

[1] D. Achlioptas, F. McSherry, Fast computation of low rank matrix approximations *Journal of the ACM*, **54**, 1-19 (2007).

[2] R.I. Arriaga, S. Vempala, An algorithmic theory of learning: Robust concepts and random projection. *Machine learning*, **63**, 161-182 (2006).

[3] Donoho, D., Elad, M., Temlyakov V.N. Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE Trans. Inf. Theory*, **52**, 6-18 (2006).

[4] P. Drineas, R. Kannan, and M.W. Mahoney, Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, **36**, 158-183, (2006).

[5] A. Frieze, R. Kannan, and S. Vempala, Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM*, **51**, 1025-1041 (2004).

[6] G.J.O. Jameson, *Summing and nuclear norms in Banach space theory*, Cambridge University Press, Cambridge, New York (1987).

[7] A. Juditsky, A. Nemirovski, On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. – to appear in *Mathematical Programming Series B*, Special Issue on Machine Learning.
E-print: http://www.optimization-online.org/DB_HTML/2008/09/2087.html

[8] N.H. Nguyen, T.T. Do, and T.D. Tran, A fast and efficient algorithm for low-rank approximation of a matrix. *STOC '09: Proceedings of the* 41^{st} *annual ACM symposium on Theory of computing*, 215-224 (2009).

[9] N. Srebro, A. Shraibman, Rank, trace-norm and max-norm. In P. Auer, R. Meir, editors, *Proc. Annual Conf. Computational Learning Theory*, *Lecture Notes in Artificial Intelligence*, **3559**, 545-560. Springer-Verlag,(2005).

A Proof of Lemma 2.1

Properties (i) and (ii) are immediate consequences of the definition of V_β given in (4). Observe that V_β is convex and continuously differentiable with

$$\left|\frac{d}{dt}\Big|_{t=0} V_\beta(x+th)\right| = \left|\frac{\sum_{i=1}^d \sinh(x_i/\beta)h_i}{\sum_{i=1}^d \cosh(x_i/\beta)}\right| \leq \|h\|_\infty \,\forall h,$$

whence $\|V_\beta'(x)\|_1 \leq 1$ for $x \in \mathbb{R}^d$. Verification of (5) takes one line: V_β is twice continuously differentiable with

$$\frac{d^2}{dt^2}\Big|_{t=0} V_\beta(x+th) = \beta^{-1}\frac{\sum_{i=1}^d \cosh(x_i/\beta)h_i^2}{\sum_{i=1}^d \cosh(x_i/\beta)} - \beta^{-1}\frac{\left(\sum_{i=1}^d \sinh(x_i/\beta)h_i\right)^2}{\left(\sum_{i=1}^d \cosh(x_i/\beta)\right)^2} \leq \beta^{-1}\|h\|_\infty^2.$$

□

B Problems (17) in the case of Hadamard matrix A

We claim that if A_m is an $m \times 2^\nu$ submatrix of the Hadamard matrix H_ν of order $n = 2^\nu$, then the optimal values in all problems (17) are equal to each other. The explanation is as follows. Let G be a finite abelian group of cardinality n. Recall that a *character* of G is a complex-valued function $\xi(g)$ such that $\xi(0) = 1$ and $\xi(g+h) = \xi(g)\xi(h)$ for all $g, h \in G$; from this definition it immediately follows that $|\xi(g)| \equiv 1$. The characters of a finite abelian group G form abelian group G_*, the multiplication being the pointwise multiplication of functions, and this group is isomorphic to G. The Fourier Transform matrix associated with G is the $n \times n$ matrix with rows indexed by $\xi \in G_*$, columns indexed by $g \in G$ and entries $\xi(g)$. For example, the usual DFT matrix of order n corresponds to the cyclic group $G = \mathbb{Z}_n := \mathbb{Z}/n\mathbb{Z}$, while the Hadamard matrix H_ν is nothing but the Fourier Transform matrix associated with $G = [\mathbb{Z}_2]^\nu$ (in this case, all characters take values ±1). For $g \in G$ let $e_g(h)$ stand for the function on G which is equal to 1 at $h = g$ and is equal to 0 at $h \neq g$. Given an m-element subset Q of G_*, consider the submatrix $A = [\xi(g)]_{\substack{\xi\in Q\\ g\in G}}$ of the Fourier Transform matrix, along with n optimization problems

$$\min_{y\in\mathbb{C}^m} \|\Re[e_g - A^T y]\|_\infty = \min_{y_\xi\in\mathbb{C}} \max_{h\in G} |\Re[e_g(h) - \sum_{\xi\in Q} y_\xi \xi(h)]| \qquad (P_g)$$

These problems clearly have equal optimal values, due to

$$\begin{array}{l}\max\limits_{h\in G} |\Re[e_g(h) - \sum_{\xi\in Q} y_\xi \xi(h)]| = \max\limits_{h\in G} |\Re[e_0(h-g) - \sum_{\xi\in Q}[y_\xi \xi(g)]\xi(h-g)]| \\ = \max\limits_{f=h-g\in G} |\Re[e_0(f) - \sum_{\xi\in Q}[y_\xi\xi(g)]\xi(f)|.\end{array}$$

As applied to $G = \mathbb{Z}_2^\nu$, this observation implies that all quantities given by (17) are the same.