Lecture notes 5 February 29, 2016

# Random projections

## 1 Introduction

Random projections are a useful tool in the analysis and processing of high-dimensional data. We will analyze two applications that use random projections to compress information embedded in high-dimensional spaces: dimensionality reduction and compressed sensing.

## 2 Dimensionality reduction

The goal of dimensionality-reduction techniques is to project high-dimensional data onto a lower-dimensional space while preserving as much information as possible. In these notes we will denote the dimension of the high-dimensional space by $n$ and the dimension of the low-dimensional subspace by $m$. These methods are a basic tool in data analysis; some applications include visualization (especially if we project onto $\mathbb{R}^2$ or $\mathbb{R}^3$), denoising and decreasing the computational cost of processing the data. Indeed, the complexity of many algorithms depends directly on the ambient dimension and in big-data regimes applying even very simple machine-learning algorithms such as nearest-neighbour classification or least-squares regression may have a huge computational cost.

We will focus on dimensionality reduction via linear projections.

**Definition 2.1** (Linear projection)**.** *The linear projection of $x \in \mathbb{R}^n$ onto a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ is the point of $\mathcal{S}$ that is closest to $x$, i.e. the solution to the optimization problem*

$$\begin{aligned} &\text{minimize} \quad && \|x - u\|_2 && (1)\\ &\text{subject to} \quad && u \in \mathcal{S}. && (2) \end{aligned}$$

The following simple lemma explains how to compute a projection using an orthonormal basis of the subspace that we want to project onto. The proof is in Section A.1 of the appendix.

**Lemma 2.2.** *Let $U$ be a matrix whose columns are an orthonormal basis of a subspace $\mathcal{S} \subseteq \mathbb{R}^n$.*

$$\mathcal{P}_{\mathcal{S}}(x) = UU^T x. \tag{3}$$

Once we fix a projection and a corresponding matrix $U$ the lower-dimensional representation of a vector $x \in \mathbb{R}^n$ is $U^T x \in \mathbb{R}^m$, so the dimensionality is reduced from $n$ to $m$. An interesting problem is how to choose the low-dimensional subspace parametrized by $U$. The following sections describe two popular alternatives: principal component analysis and random projections.

## 2.1 Principal component analysis

Principal component analysis (PCA) is an adaptive dimensionality-reduction technique in which we first determine the directions of maximum variation in a dataset and then project onto them. PCA is based on the singular-value decomposition (SVD). The proof of the following fundamental result can be found in any graduate linear algebra textbook.

**Theorem 2.3.** *Without loss of generality let $m \leq n$. Every rank $r$ real matrix $A \in R^{m \times n}$ has a unique singular-value decomposition of the form (SVD)*

$$\begin{aligned} A &= \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_m \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \cdots \\ v_m^T \end{bmatrix} && (4) \\ &= USV^T, && (5) \end{aligned}$$

*where the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$ are nonnegative real numbers, the matrix $U \in R^{m \times m}$ containing the left singular vectors is orthogonal, and the matrix $V \in R^{m \times n}$ containing the right singular vectors is a submatrix of an orthogonal matrix (i.e. its columns form an orthonormal set).*

**Algorithm 2.4** (Principal component analysis)**.** *Given $k$ data vectors $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k \in \mathbb{R}^n$, we apply the following steps.*

1. *Center the data,*

$$x_i = \tilde{x}_i - \frac{1}{k} \sum_{i=1}^{k} \tilde{x}_i, \qquad 1 \leq i \leq n. \tag{6}$$

2. *Group the centered data in a data matrix $X \in \mathbb{R}^{n \times k}$*

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix}. \tag{7}$$

3. *Compute the SVD of $X$ and extract the left singular vectors corresponding to the $m$ largest singular values. These are the first $m$ principal components.*

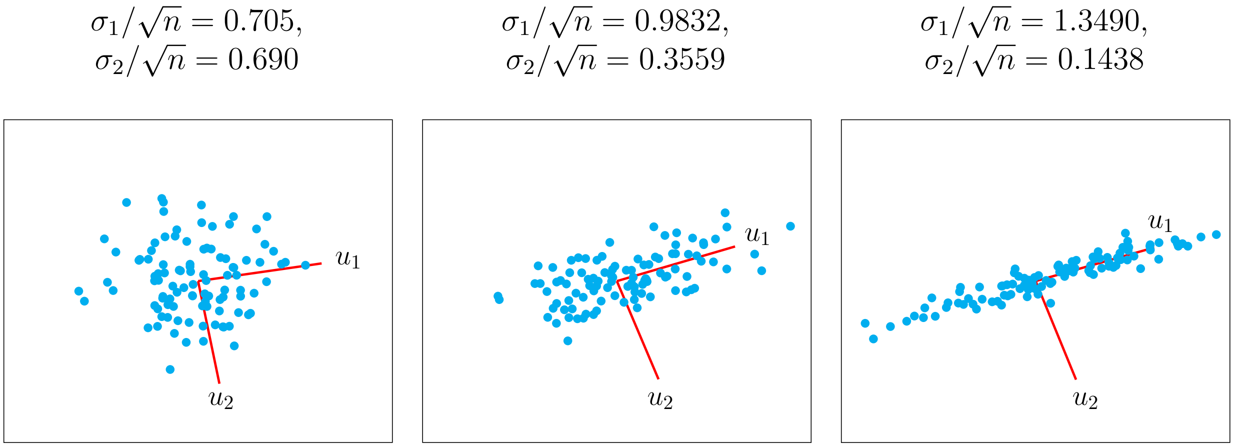


**Figure 1:** PCA of a dataset with $n = 100$ 2D vectors with different configurations. The two first singular values reflect how much energy is preserved by projecting onto the two first principal components.

Once the data are centered, the energy of their projection onto different directions in the ambient space reflects the variation of the dataset along those directions. PCA selects the directions that maximize the $\ell_2$ norm of the projection and are mutually orthogonal. The span of the first $m$ principal components is the subspace the best approximates the data in terms of $\ell_2$-norm error, as established by the following theorem proved in Section A.2 of the appendix.

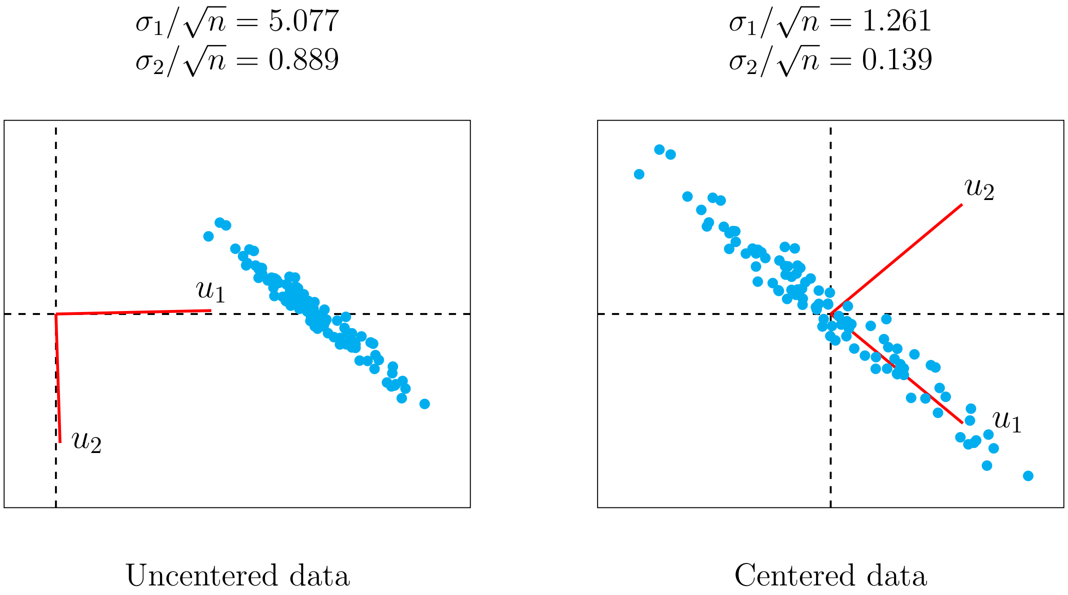
**Theorem 2.5.** *For any matrix $X \in \mathbb{R}^{n\times k}$ with left singular vectors $u_1, u_2, \ldots, u_n$ corresponding to the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$,*

$$\sum_{i=1}^{k} \left|\left|\mathcal{P}_{\operatorname{span}(u_1,u_2,\ldots,u_n)}\, x_i\right|\right|_2^2 \geq \sum_{i=1}^{k} \left|\left|\mathcal{P}_{\mathcal{S}}\, x_i\right|\right|_2^2 , \tag{8}$$

*for any subspace $\mathcal{S}$ of dimension $m \leq \min\{n, k\}$.*

Figure 1 illustrates PCA in 2D. Note how the singular values are proportional to the energy that lies in the direction of the corresponding principal component.

Figure 2 illustrates the importance of centering before applying PCA. Theorem 2.5 still holds if the data are not centered. However, the norm of the projection onto a certain direction no longer reflects the variation of the data. In fact, if the data are concentrated around a point that is far from the origin, the first principal component will tend be aligned in that direction. This makes sense as projecting onto that direction captures more energy. As a result, the principal components do not capture the directions of maximum variation *within* the cloud of data.



$\sigma_1/\sqrt{n} = 5.077$
$\sigma_2/\sqrt{n} = 0.889$

$\sigma_1/\sqrt{n} = 1.261$
$\sigma_2/\sqrt{n} = 0.139$

Uncentered data

Centered data

**Figure 2:** PCA applied to $n = 100$ 2D data points. On the left the data are not centered. As a result the dominant principal component $u_1$ lies in the direction of the mean of the data and PCA does not reflect the actual structure. Once we center, $u_1$ becomes aligned with the direction of maximal variation.

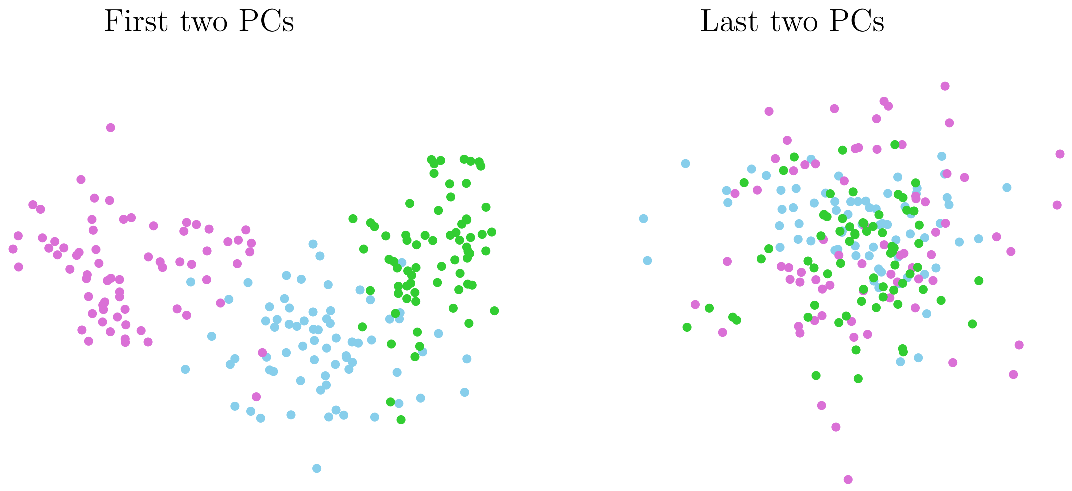
The file *seeds_dataset.txt* contains different geometric attributes (area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove) of seeds belonging to three different varieties of wheat: Kama, Rosa and Canadian[1]. Figure 3 shows the projection of the data onto the first two and the last two principal components. The structure of the data is much better conserved in the first case, which allows to visualize the difference between the three seeds very clearly. Note however that the first principal components only guarantee that the energy in the projection will be preserved, not that the projection will be good for tasks such as classification.

## 2.2 Random projections

To apply PCA we need to process all of the data points beforehand in order to compute the projection. This may be too computationally costly if the dataset is very large or not possible at all if the aim is to project a stream of data in real time. For such cases we need a *non-adaptive* alternative to PCA that chooses the projection before actually seeing the data.

A simple method that tends to work well is to project onto a random subspace. In particular, if the data points are $x_1, x_2, \ldots \in \mathbb{R}^n$, we can obtain a random projection by multiplying the

[1] The data can be found at `https://archive.ics.uci.edu/ml/datasets/seeds`



First two PCs

Last two PCs

**Figure 3:** Projection of 7-dimensional vectors describing different wheat seeds onto the first two (left) and the last two (right) principal components of the dataset. Each color represents a variety of wheat.

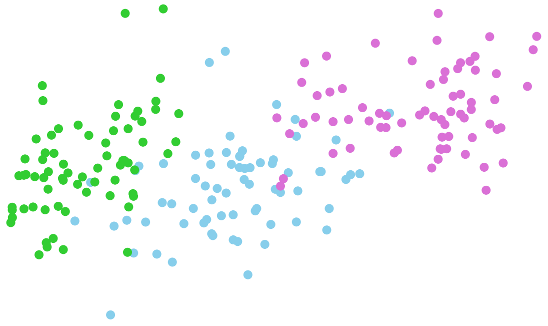


**Figure 4:** Approximate projection of 7-dimensional vectors describing different wheat seeds onto two random directions. Each color represents a variety of wheat.

data with a random matrix $A \in \mathbb{R}^{m \times n}$ to obtain $Ax_1, Ax_2, \ldots \in \mathbb{R}^m$. Strictly speaking, this is a linear projection only if $A$ is a projection matrix with orthonormal rows ($U^T$ in Lemma 2.2). However, in many cases the rows of random matrices are approximately orthogonal and consequently this procedure yields an approximate projection.

Figure 4 shows the result of applying such an approximate projection to the same data used in Figure 3 onto two random directions by multiplying the data with a $2 \times 7$ random Gaussian matrix. The structure of the data seems to be as well conserved by the projection as in the case of PCA.

Dimensionality-reduction techniques are useful if they preserve the information that we are interested in. In many cases, we would like the projection to conserve the distances between the different data points. This allows to apply algorithms such as nearest neighbors in the lower-dimensional space. The following lemma guarantees that random projections do not distort the distances between points with a certain probability. The result is striking because the lower bound on $m$– the dimension of the approximate projection– does not depend on $n$– the ambient dimension of the data– and its dependence on the number of points in the dataset is only logarithmic. Although we prove the result for a matrix with Gaussian entries, the result can be extended to other matrices that can be applied more efficiently [1].

**Lemma 2.6** (Johnson-Lindenstrauss lemma)**.** *Let $\mathcal{S} := \{x_1, \ldots, x_k\}$ in $\mathbb{R}^n$. There exists a random function $f$ such that for any pair of points $x_i, x_j$*

$$(1-\epsilon) \left|\left|x_i - x_j\right|\right|_2^2 \leq \left|\left|f(x_i) - f(x_j)\right|\right|_2^2 \leq (1+\epsilon) \left|\left|x_i - x_j\right|\right|_2^2, \tag{9}$$

*with probability at least $\frac{1}{k}$ as long as*

$$m \geq \frac{8 \log(k)}{\epsilon^2}. \tag{10}$$

*The random function is of the form*

$$f(x) := \frac{1}{\sqrt{m}} A x \tag{11}$$

*and $A$ is a matrix with iid Gaussian entries with zero mean and unit variance.*

*Proof.* To establish the result we use the following proposition, proved in Section A.4 of the appendix, which establishes that the norm of a fixed vector is preserved by the random projection with high probability.

**Proposition 2.7.** *Let $f$ be defined as in Lemma 2.6. For any fixed vector $v \in \mathbb{R}^n$*

$$\mathrm{P}\left((1-\epsilon) \left|\left|v\right|\right|_2^2 \leq \left|\left|f(v)\right|\right|_2^2 \leq (1+\epsilon) \left|\left|v\right|\right|_2^2\right) \geq 1 - 2\exp\left(-\frac{m\epsilon^2}{8}\right). \tag{12}$$

We define the events

$$\mathcal{E}_{ij} = \left\{ (1-\epsilon) \left|\left|x_i - x_j\right|\right|_2^2 \leq \left|\left|f\left(x_i - x_j\right)\right|\right|_2^2 \leq (1+\epsilon) \left|\left|x_i - x_j\right|\right|_2^2 \right\}, \quad i \neq j, 1 \leq i, j \leq k.$$

By Proposition 2.7 applied to $v := x_i - x_j$ and condition (10)

$$\mathrm{P}\left(\mathcal{E}_{ij}^c\right) \leq \frac{2}{k^2}. \tag{13}$$

There are $\binom{k}{2}$ different pairs of points. The union bound yields

$$\begin{aligned} \mathrm{P}\left(\bigcap_{i,j} \mathcal{E}_{ij}\right) &= 1 - \mathrm{P}\left(\bigcup_{i,j} \mathcal{E}_{ij}^c\right) && (14) \\ &\geq 1 - \sum_{i,j} \mathrm{P}\left(\mathcal{E}_{ij}^c\right) && (15) \\ &\geq 1 - \binom{k}{2} \frac{2}{k^2} && (16) \\ &\geq \frac{1}{k} && (17) \end{aligned}$$

where the last inequality follows from condition (10). This completes the proof. □

# 3 Compressed sensing

Compressed sensing allows to recover sparse signals from randomized measurements by minimizing their $\ell_1$ norm. In this section we first illustrate the application of these ideas to MRI and then provide a theoretical analysis.

## 3.1 Compressed sensing in MRI

Magnetic resonance imaging (MRI) is a popular medical imaging technique used in radiology. MRI data can be modeled as samples from the 2D or 3D Fourier transform of the object that is being imaged, for example a slice of a human brain. An estimate of the corresponding image can be obtained by computing the inverse Fourier transform of the data, as shown in Figure 5.

An important challenge in MRI is to reduce measurement time, which can be achieved by undersampling the data. Consider a 1D version of the problem, where the signal is an $n$-dimensional vector $x$ and the DFT matrix is denoted by $F$. We model the undersampled data $y$ as

$$y = F_\Omega \, x, \tag{18}$$

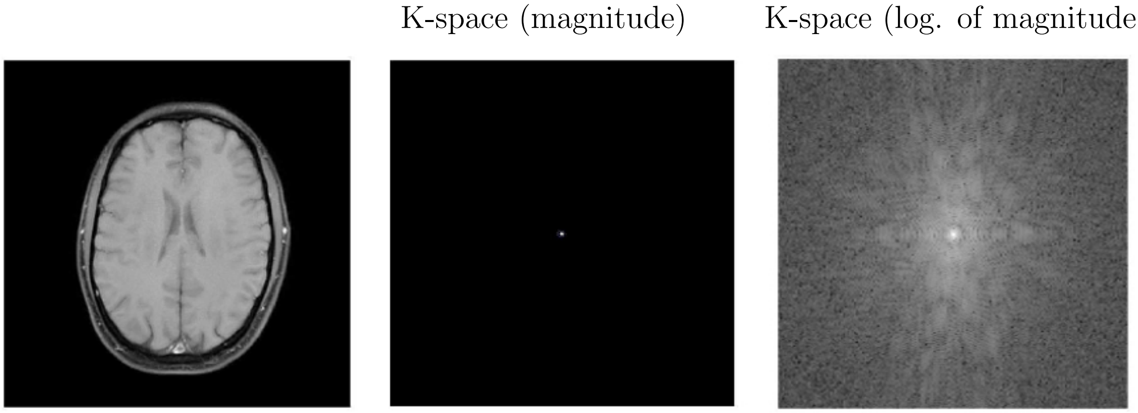


**Figure 5:** Image of a brain obtained by MRI, along with the magnitude of its 2D-Fourier or k-space representation and the logarithm of this magnitude.

where $F_\Omega$ is a submatrix of $F$ obtained by choosing $m$ rows indexed by the set $\Omega$. Since $m < n$ the system is underdetermined and has infinite solutions.

A possible estimate for the image is the solution $x_{\ell_2}$ with minimum $\ell_2$ norm satisfying $y = F_\Omega \, x_{\ell_2}$. By Lemma 3.1 in Lecture Notes 4 this estimate equals

$$x_{\ell_2} := F_\Omega^T \left( F_\Omega F_\Omega^T \right)^{-1} x, \tag{19}$$

where $x$ is the original signal. Equivalently, $x_{\ell_2}$ is the projection of $x$ onto the row space of $F_\Omega$. This projection provides some insight as to the effect of different sampling strategies. Figure 6 shows $x_{\ell_2}$ for two undersampling patterns in 2D: regular undersampling in one direction and random undersampling. The corresponding artifacts are very different. Regular undersampling produces coherent aliasing– the reconstruction is a superposition of shifted copies of the image– whereas random undersampling produces aliasing that essentially looks like noise.

MRI signals are sparse in different transform domains. For example, the brain image in Figure 5 is sparse in the wavelet domain. This is good news if we are trying to recover an image from undersampled data. We cannot estimate more than $m$ parameters from $m$ measurements, so in principle it is hopeless to try to estimate an $n$-dimensional signal from data given by 18. However, if the signal is sparse in some domain and can be parametrized by $s$ coefficients for instance, where $s < m$, then recovery may be possible.

As we discussed in the previous lecture, $\ell_1$-norm minimization is an effective tool for obtaining sparse solutions to underdetermined linear equations. Figure 7 shows the result of applying $\ell_1$-norm minimization to recover an image from the data corresponding to the images shown

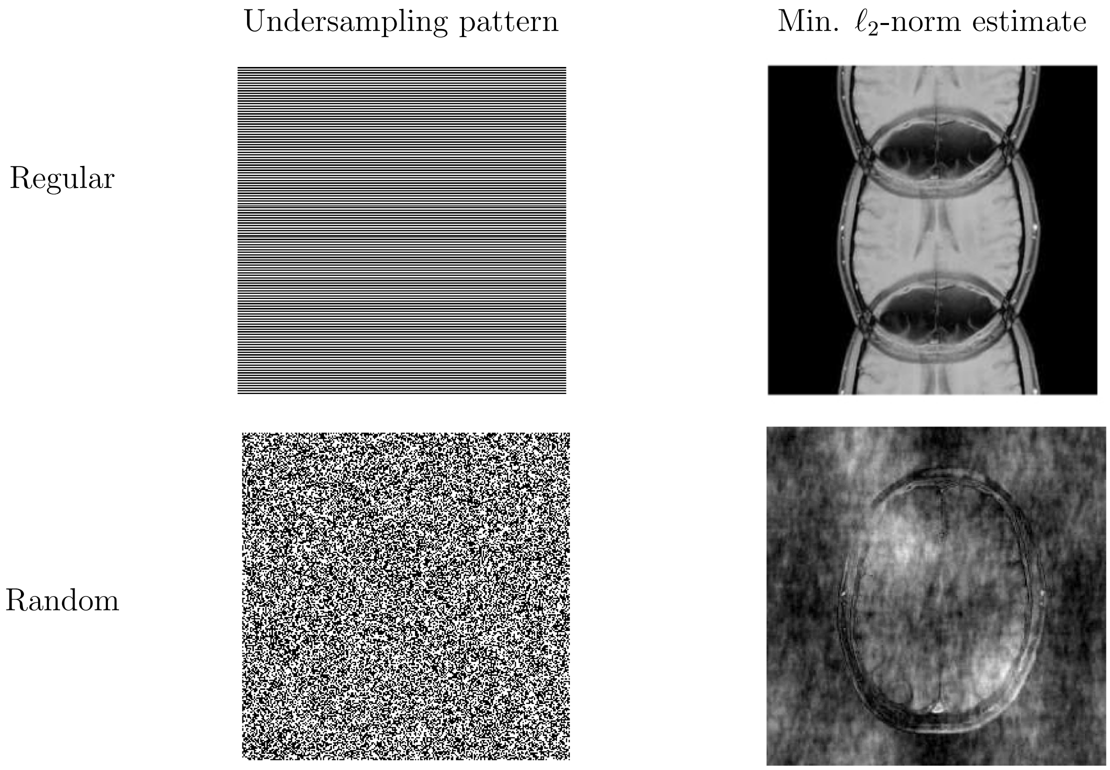


**Figure 6:** Two different sampling strategies in 2D k space: regular undersampling in one direction (top) and random undersampling (bottom). The original data is the same as in Figure 5. On the right we see the corresponding minimum-$\ell_2$-norm estimate for each undersampling pattern.

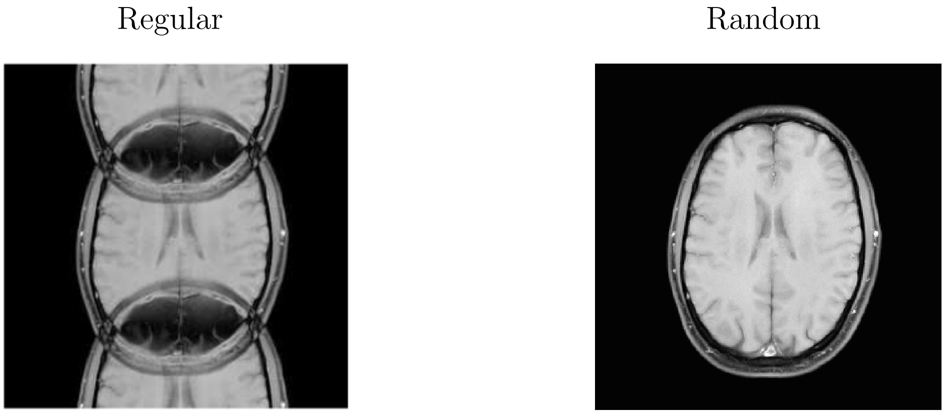


**Figure 7:** Minimum-$\ell_1$-norm for the two undersampling patterns shown in Figure 6.

in Figure 6. Since the image is assumed to be sparse in the wavelet domain, we solve

$$\text{minimize} \quad ||\tilde{c}||_1 \tag{20}$$

$$\text{subject to} \quad y = F_\Omega W \tilde{c} \tag{21}$$

where $W$ is the matrix corresponding to an orthonormal wavelet transform. For regular undersampling, then the estimate is essentially the same as the minimum-$\ell_2$-norm estimate. This is not surprising, since the minimum-$\ell_2$-norm estimate is also sparse in the wavelet domain because it is equal to a superposition of two shifted copies of the image. In contrast, $\ell_1$-norm minimization recovers the original image perfectly when coupled with random projections. Intuitively, $\ell_1$-norm minimization *cleans up* the noisy aliasing caused by random undersampling. In the next section we provide a theoretical characterization of this phenomenon.

## 3.2 Exact recovery

In our theoretical analysis we will focus on one-dimensional sparse signals, but most of the ideas extend to the case where the signal is multidimensional or sparse in a transform domain instead. The measurement model is given by

$$y = Ax \tag{22}$$

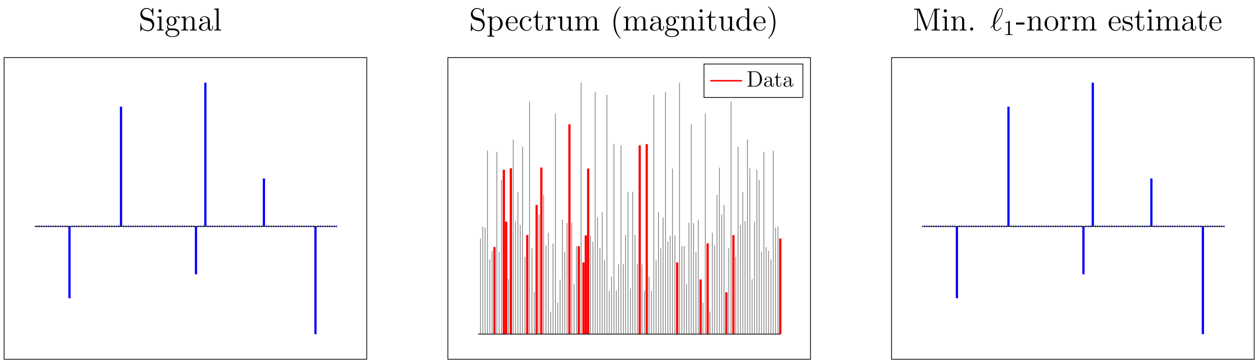
where $x \in \mathbb{R}^n$ is the sparse signal which has $s$ nonzeros, $y \in \mathbb{R}^m$ the data and $A \in \mathbb{R}^{m \times n}$ the random matrix that models the measurement process. In MRI the rows of $A$ are chosen at random from a DFT matrix. Here we will mostly assume that the entries are sampled independently at random from a Gaussian distribution. This model is not of great practical interest because Gaussian measurements do not arise in applications and multiplication with dense Gaussian matrices is computationally expensive. However, the Gaussian assumption makes the proofs significantly simpler and allows to illustrate ideas that are readily applicable to more useful cases such as random Fourier measurements.

Clearly, without any assumptions on $x$ the underdetermined linear system (22) cannot be solved to retrieve $x$ if $m < n$. It is necessary to use the sparsity assumption. If $s < n$ and the sparsity pattern $T$ (the indices of the entries of $x$ that are nonzero) is known, then the problem is actually overdetermined. Indeed, let $x_T$ be the subvector of nonzero entries of $x$ and $A_T \in \mathbb{R}^{m \times s}$ the submatrix of $A$ consisting of the columns indexed by $T$. We have

$$y = Ax = A_T x_T. \tag{23}$$

As long as the columns of $A_T$ are independent and $m \geq s$ we can easily find the original vector by applying any left inverse of $A_T$ to the data. In particular, choosing the pseudoinverse,

$$A^\dagger y = \left(A_T^T A_T\right)^{-1} A_T^T y = x. \tag{24}$$


Signal

Spectrum (magnitude)

Min. $\ell_1$-norm estimate

**Figure 8:** Minimizing the $\ell_1$ norm of the estimate (right) allows to estimate the original signal (left) exactly from a small number of random samples of its spectrum (center).

The bad news is that we don't know $T$! Trying every possible submatrix of $s$ columns is not feasible computationally for even very small values of $s$. Instead, we apply $\ell_1$ norm minimization to favor sparse estimates. We recover $x$ by solving the convex program

$$\begin{aligned} \text{minimize} \quad & \|\tilde{x}\|_1 && (25) \\ \text{subject to} \quad & y = A\tilde{x}. && (26) \end{aligned}$$

Figure 8 shows an example where this procedure allows to reconstruct a sparse signal from random Fourier data. The following theorem establishes that $\ell_1$-norm minimization achieves exact recovery of $x$ with high probability as long as the number of measurements $m$ is proportional to the sparsity of the signal $s$ up to logarithmic factors. The result is remarkable. Random measurements allow to sample the signal at a rate that is essentially independent of the ambient dimension and only depends on how compressible it is.

**Theorem 3.1.** *Assume there exists a signal $x \in \mathbb{R}^n$ with $s$ nonzeros such that*

$$Ax = y \tag{27}$$

*for a random matrix $A \in \mathbb{R}^{m\times n}$ with iid Gaussian entries with zero mean and unit variance. The solution to Problem* (106) *is equal to $x$ with probability at least $1 - \frac{1}{n}$ as long as the number of measurements satisfies*

$$m \geq Cs \log n, \tag{28}$$

*for a fixed numerical constant $C$.*

An important generalization of this result establishes that $A$ can be formed by taking random rows from any unitary matrix $U \in \mathbb{C}^{n\times n}$ (unitary is a fancy word for a matrix with orthonormal columns). In that case, the number of measurements that are necessary for exact recovery also depends on the coherence $\mu$ of the measurements, defined as

$$\mu(U) := \sqrt{n} \max_{1\leq i\leq n, 1\leq j\leq m} |U_i j|\,. \tag{29}$$

Intuitively the coherence measures how localized or *spiky* the rows of $U$ are. If the rows are too localized, they might miss entries in the sparse vector. Recall that we don't know the support $T$ beforehand, otherwise we would definitely use measurements that are localized on $T$! Exact recovery via $\ell_1$-norm minimization is achieved with high probability for this kind of measurements if

$$m \geq C\mu\left(U\right) s \log n. \tag{30}$$

We refer to [6] for a proof of this result (see also [4]). In the case of the DFT $\mu = 1$ since the rows are complex exponentials divided by $\sqrt{n}$.

To prove Theorem 3.1 we use the following result, which establishes that the existence of a certain vector implies exact recovery. We defer the proof to Section A.6 in the appendix.

**Lemma 3.2.** *Let $T$ be the indices of the nonzero entries in $x$. If $A_T$ is full rank and there exists a vector $v \in \mathbb{R}^m$ such that*

$$\left(A^T v\right)_i = \operatorname{sign}\left(x_i\right) \qquad \textit{if } x_i \neq 0 \tag{31}$$

$$\left|\left|\left(A^T v\right)_i\right|\right|_\infty < 1 \qquad \textit{if } x_i = 0 \tag{32}$$

*then $x$ is the unique solution to Problem* (106)*.*

By this lemma, all we need to do to establish exact recovery is show that for any subset $T$ with cardinality $s$ there exists a vector $v$ such that $q := A^T v$ is equal to the sign of $x$ on $T$ and has magnitude strictly bounded by one on $T^c$. $v$ is commonly known as a *dual certificate* in the literature. The reason is that it certifies optimality and is feasible for the dual problem of Problem (106), derived in Section A.7 of the appendix.

**Lemma 3.3** (Dual problem)**.** *The dual problem to Problem* (106) *is*

$$\text{maximize} \quad y^T \tilde{v} \tag{33}$$

$$\text{subject to} \quad \left|\left|A^T \tilde{v}\right|\right|_\infty \leq 1, \tag{34}$$

*where the dual variable $\tilde{v}$ has dimension $n$.*

The dual certificate is feasible for this problem and achieves a cost-function value of $\left|\left|x\right|\right|_1$, since

$$y^T v = x^T A^T v \tag{35}$$

$$= \sum_{i \in T} x_i \operatorname{sign}\left(x_i\right) \tag{36}$$

$$= \left|\left|x\right|\right|_1. \tag{37}$$

By weak duality (Corollary 4.8 in Lecture Notes 2) this implies that $x$ is a solution to the primal problem (note that this does not prove that it is the unique solution but Lemma 3.2 does). In the next section we will construct the dual certificate that establishes Theorem 3.1.

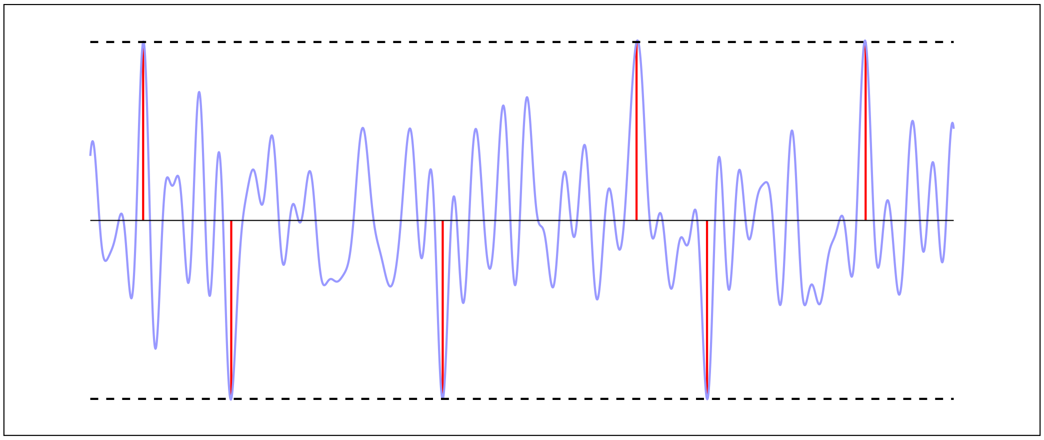


**Figure 9:** Subgradient $q_{\ell_2} := A^T v_{\ell_2}$ (blue) corresponding to the minimum-$\ell_2$-norm dual certificate $v_{\ell_2}$ for a compressed-sensing problem with random Fourier measurements. $q_{\ell_2}$ interpolates the sign of the signal (red) on its support.

## 3.3 Dual certificate

In this section we analyze a dual-certificate candidate and show that it satisfies the conditions in Lemma 3.2 and hence proves Theorem 3.1 with high probability. In particular, we set $v_{\ell_2}$ to be the solution to the following minimization problem

$$\begin{aligned} &\text{minimize} \quad && \left\|\tilde{v}\right\|_2 && (38)\\ &\text{subject to} \quad && A_T^T \tilde{v} = \operatorname{sign}(x_T). && (39)\end{aligned}$$

In words, $v_{\ell_2}$ is the minimum-$\ell_2$-norm vector such that $q_{\ell_2} := A^T v_{\ell_2}$ interpolates the sign of $x$ on $T$. Ideally we would like to minimize the $\ell_\infty$ norm of the subgradient $A^T v$ on $T^c$ instead of its $\ell_2$ norm, but the former problem does not have a closed-form solution. From Lemma 3.1 in Lecture Notes 4 the solution to Problem (38) is

$$v_{\ell_2} = A_T \left(A_T^T A_T\right)^{-1} \operatorname{sign}(x_T) \tag{40}$$

as long as $A_T$ is full rank. Having an explicit expression for the solution allows us to analyze the construction for any arbitrary support and sign pattern. We will do this under the assumption that the entries of the measurement matrix $A$ are Gaussian, but this technique has also been applied to Fourier [7] and other random measurements [4] (see also [6] for a more recent proof technique based on approximate dual certificates that provides better guarantees). Figure 9 shows $q_{\ell_2}$ for the case of random Fourier measurements.

To characterize $v_{\ell_2}$ first we need to establish that $A_T$ is full rank, since otherwise $A_T^T A_T$ would not have an inverse. The following proposition shows that the random matrix preserves the norm of sparse vectors supported on a fixed set $T$ with a certain probability. We defer the proof to Section 3.4.

**Proposition 3.4.** *Fix a set $T \subset \{1, 2, \ldots, n\}$ such that $|T| \leq s$. For any unit-norm vector $x$ with support $T$*

$$1 - \epsilon \leq \frac{1}{\sqrt{m}} ||Ax||_2 \leq 1 + \epsilon \tag{41}$$

*with probability at least*

$$1 - 2 \left( \frac{12}{\epsilon} \right)^s \exp \left( - \frac{m \epsilon^2}{32} \right). \tag{42}$$

Setting $\epsilon = 1/2$ implies that the minimum singular value of $A_T$ is lower bounded,

$$\sigma_{\min} \left( A_T \right) \geq \frac{\sqrt{m}}{2}, \tag{43}$$

with probability at least $1 - \exp \left( - \frac{Cm}{s} \right)$ for a certain constant $C$. This implies $A_T^T A_T$ is invertible and consequently that condition (31) is satisfied,

$$\begin{aligned} \left( q_{\ell_2} \right)_T &= A_T^T A_T \left( A_T^T A_T \right)^{-1} \operatorname{sign} \left( x_T \right) && (44) \\ &= \operatorname{sign} \left( x_T \right). && (45) \end{aligned}$$

All is left is to bound $q_{\ell_2}$ on $T^c$. We define

$$w := A_T \left( A_T^T A_T \right)^{-1} \operatorname{sign} \left( x_T \right). \tag{46}$$

Since the entries of $A$ are all independent, and $w$ only depends on $A^T$, it is independent of any column of $A_{T^c}$. This means that for each $i \in T^c$

$$\begin{aligned} \left( q_{\ell_2} \right)_i &= A_i^T A_T \left( A_T^T A_T \right)^{-1} \operatorname{sign} \left( x_T \right) && (47) \\ &= A_i^T w && (48) \end{aligned}$$

where $A_i$ and $w$ are independent.

By (43), we can bound the norm of $w$

$$||w||_2 \leq \frac{||\operatorname{sign} \left( x_T \right)||_2}{\sigma_{\min} \left( A_T \right)} \leq 2 \sqrt{\frac{s}{m}} \tag{49}$$

with probability $1 - \exp \left( - \frac{Cm}{s} \right)$.

Conditioned on $w$, $A_i^T w$ is Gaussian with mean 0 and variance $||w||_2^2$. We have

$$\begin{aligned} \mathrm{P} \left( \left| A_i^T w \right| \geq 1 | w = w' \right) &\leq \mathrm{P} \left( |u| > \frac{1}{||w'||_2} \right) && (50) \\ &\leq 2 \exp \left( - \frac{1}{2 \, ||w'||_2^2} \right), && (51) \end{aligned}$$

where $u$ has mean 0 and variance 1. We have applied the following deviation bound, proved in Section A.9 of the appendix.

**Lemma 3.5.** *For a Gaussian random variable $u$ with zero mean and unit variance and any $t > 0$*

$$\mathrm{P}\left(|u| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2}\right). \tag{52}$$

Let us define the event

$$\mathcal{E} := \left\{ ||w||_2 \leq 2\sqrt{\frac{s}{m}} \right\}. \tag{53}$$

(51) implies that

$$\mathrm{P}\left(\left|A_i^T w\right| \geq 1 \middle| \mathcal{E}\right) \leq 2\exp\left(-\frac{m}{8s}\right). \tag{54}$$

As a result,

$$\begin{aligned} \mathrm{P}\left(\left|A_i^T w\right| \geq 1\right) &\leq \mathrm{P}\left(\left|A_i^T w\right| \geq 1 \middle| \mathcal{E}\right) + \mathrm{P}\left(\mathcal{E}^c\right) & (55) \\ &\leq \exp\left(-\frac{Cm}{s}\right) + 2\exp\left(-\frac{m}{8s}\right). & (56) \end{aligned}$$

By the union bound,

$$\mathrm{P}\left(\bigcup_{i \in T^c} \left\{\left|A_i^T w\right| \geq 1\right\}\right) \leq n\left(\exp\left(-\frac{Cm}{s}\right) + 2\exp\left(-\frac{m}{8s}\right)\right). \tag{57}$$

We can consequently choose a constant $C'$ so that if the number of measurements satisfies

$$m \geq C' s \log n \tag{58}$$

we have exact recovery with probability $1 - \frac{1}{n}$.

## 3.4 Bounds on singular values of random matrices

In this section we provide the proof of Proposition 3.4, which illustrates a useful method to bound the action of a linear operator on a set of infinite cardinality. Let $\mathcal{X}_T$ be the set of unit-norm vectors $x$ with support $T$. By Proposition 2.7 we have that for any fixed unit-norm vector $v$

$$(1-\epsilon) \leq \frac{1}{m} ||Av||_2^2 \leq (1+\epsilon) \tag{59}$$

with probability $1 - 2\exp\left(-\frac{m\epsilon^2}{8}\right)$. This does not immediately imply the result that we want: the bounds must hold for all vectors in $\mathcal{X}_T$ and we cannot just apply the union bound because the set has infinite cardinality. To overcome this obstacle, we first apply the union bound on a subset of $\mathcal{X}_T$ called an $\epsilon$-net instead.

**Definition 3.6** (Net)**.** *An $\epsilon$-net of a set $\mathcal{X}$ is a subset $\mathcal{N}_\epsilon \subseteq \mathcal{X}$ such that for every point $y \in \mathcal{X}$ there exists $z \in \mathcal{N}_\epsilon$ for which*

$$||y - z||_2 \leq \epsilon. \tag{60}$$

**Definition 3.7** (Covering number)**.** *The covering number $\mathcal{N}(\mathcal{X}, \epsilon)$ of a set $\mathcal{X}$ at scale $\epsilon$ is the minimal cardinality of an $\epsilon$-net of $\mathcal{X}$, or equivalently the minimal number of balls of radius $\epsilon$ with centers in $\mathcal{X}$ required to cover $\mathcal{X}$.*

The following proposition, proved in Section A.8 of the appendix, provides a bound for the covering number of the $s$-dimensional sphere $\mathcal{S}^{s-1}$.

**Proposition 3.8** (Covering number of a sphere)**.** *The covering number of the $s$-dimensional sphere $\mathcal{S}^{s-1}$ at scale $\epsilon$ satisfies*

$$\mathcal{N}\left(\mathcal{S}^{k-1}, \epsilon\right) \leq \left(\frac{2+\epsilon}{\epsilon}\right)^s. \tag{61}$$

It is not difficult to see that every vector $x' \in \mathbb{R}^n$ with support $T$ can be mapped to a vector $x'_T \in \mathbb{R}^s$ by choosing the entries in $T$ and that $||x'||_2 = ||x'_T||_2$. As a result, Proposition 3.8 provides a bound on the covering number of $\mathcal{X}_T$.

**Corollary 3.9.** *If $|T| \leq s$, the covering number of $\mathcal{X}_T$ satisfies*

$$\mathcal{N}(\mathcal{X}_T, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^s. \tag{62}$$

Let $\epsilon_1 := \epsilon/4$ and $\epsilon_2 := \epsilon/2$. Consider an $\epsilon_1$-net $\mathcal{N}_{\epsilon_1}$ of $\mathcal{X}_T$. We define the event

$$\mathcal{E}_{u,\epsilon_2} := \left\{ (1-\epsilon_2)\,||u||_2^2 \leq \left|\left|\frac{1}{k} A u\right|\right|_2^2 \leq (1+\epsilon_2)\,||u||_2^2 \right\}. \tag{63}$$

By Proposition 2.7 and the union bound we have

$$\begin{aligned} \mathrm{P}\left(\bigcup_{u \in \mathcal{N}_{\epsilon_1}} \mathcal{E}^c_{u,\epsilon_2}\right) &\leq \sum_{u \in \mathcal{N}_{\epsilon_1}} \mathrm{P}\left(\mathcal{E}^c_{u,\epsilon_2}\right) && (64) \\ &\leq |\mathcal{N}_{\epsilon_1}|\, \mathrm{P}\left(\mathcal{E}^c_{u,\epsilon_2}\right) && (65) \\ &\leq 2\left(\frac{12}{\epsilon}\right)^s \exp\left(-\frac{m\epsilon^2}{32}\right). && (66) \end{aligned}$$

Now, to finish the proof we need to show that the bound on the elements of the net can be used to bound every other element. Let

$$\frac{\sigma_{\max}(A_T)}{\sqrt{m}} := 1 + \alpha \tag{67}$$

where $\sigma_{\max}(A_T)$ denotes the largest singular value of $A_T$. For all $x' \in \mathcal{X}_T$

$$\frac{1}{\sqrt{m}} \left|\left|Ax'\right|\right|_2 = \frac{1}{\sqrt{m}} \left|\left|A_T x'_T\right|\right|_2 \tag{68}$$

$$\leq 1 + \alpha. \tag{69}$$

For any $x' \in \mathcal{X}_T$, there is a $u \in \mathcal{N}(\mathcal{X}, \epsilon_1)$ such that $\left|\left|x' - u\right|\right|_2 \leq \epsilon/4$. This implies

$$\frac{1}{\sqrt{m}} \left|\left|Ax'\right|\right|_2 \leq \frac{1}{\sqrt{m}} \left|\left|Au\right|\right|_2 + \frac{1}{\sqrt{m}} \left|\left|A(x' - u)\right|\right|_2 \tag{70}$$

$$\leq 1 + \frac{\epsilon}{2} + \frac{(1+\alpha)\,\epsilon}{4}. \tag{71}$$

Since by definition $\sigma_{\max}(A_T)$ is the smallest upper bound for the norm of $\left|\left|A_T x'_T\right|\right|_2 = \left|\left|Ax'\right|\right|_2$ such that $\left|\left|x'_T\right|\right|_2 = 1$ (which holds because $x' \in \mathcal{X}_T$), we have

$$1 + \alpha \leq 1 + \frac{\epsilon}{2} + \frac{(1+\alpha)\,\epsilon}{4}, \tag{72}$$

so that

$$\alpha \leq \frac{3\,\epsilon}{4 - \epsilon} \leq \epsilon. \tag{73}$$

The lower bound on the singular value follows immediately

$$\frac{1}{\sqrt{m}} \left|\left|Ax\right|\right|_2 \geq \frac{1}{\sqrt{m}} \left(\left|\left|Au\right|\right|_2 - \left|\left|A(x - u)\right|\right|_2\right) \tag{74}$$

$$\geq 1 - \frac{\epsilon}{2} - \frac{(1+\alpha)\,\epsilon}{4} \tag{75}$$

$$= 1 - \frac{\epsilon}{2} - \frac{(1+\epsilon)\,\epsilon}{4} \tag{76}$$

$$\geq 1 - \epsilon. \tag{77}$$

## 3.5 Robustness to noise

In the previous sections we have shown that $\ell_1$-norm minimization allows to recover sparse signals exactly with high probability in the absence of noise. In this section we consider recovery from data that is perturbed by noise, as is usually the case in practice. In more detail, given an $s$-sparse signal $x \in \mathbb{R}^n$ we have access to

$$y = Ax + z \tag{78}$$

where $A \in \mathbb{R}^{m \times n}$ is a random matrix and $z \in \mathbb{R}^m$ is an additive noise term. A necessary condition for *any* recovery method to be stable is that $A$ should preserve the energy of the sparse signal; if the signal is in the null space (or almost in the null space) of the measurement operator then we cannot hope to reconstruct it in the presence of noise. A matrix that satisfies the restricted isometry property (RIP) preserves the energy of *any* random vector; it essentially behaves as an isometry when acting upon these class of vectors.

**Definition 3.10** (Restricted isometry property)**.** *If a matrix $M$ satisfies the restricted isometry property with constant $\epsilon_s$ then for any $s$-sparse vector $x$*

$$(1-\epsilon_s)\left|\left|x\right|\right|_2 \leq \left|\left|Mx\right|\right|_2 \leq (1+\epsilon_s)\left|\left|x\right|\right|_2 . \tag{79}$$

The RIP guarantees that the problem of recovering sparse vectors from the corresponding linear measurements is well posed. If $M$ satisfies the RIP for a sparsity level $2s$ then there cannot be two sparse vectors $x_1$ and $x_2$ that are very different and yet produce similar measurements $y_1$ and $y_2$, since

$$\begin{aligned} \left|\left|y_2-y_1\right|\right|_2 &= M\left(x_2-x_1\right) && (80)\\ &\geq (1-\epsilon_{2s})\left|\left|x_2-x_1\right|\right|_2 . && (81)\end{aligned}$$

Gaussian matrices satisfy the restricted isometry property with high probability, as established in the following theorem. Random Fourier measurements also satisfy the RIP [8, 11].

**Theorem 3.11** (Restricted isometry property for Gaussian matrices)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with iid Gaussian entries with zero mean and unit variance. $\frac{1}{\sqrt{m}}A$ satisfies the restricted isometry property with a constant $\epsilon_s$ with probability $1-\frac{C_2}{n}$ as long as the number of measurements*

$$m \geq \frac{C_1 s}{\epsilon_s^2}\log\left(\frac{n}{s}\right) \tag{82}$$

*for two fixed constants $C_1, C_2 > 0$.*

*Proof.* By Proposition 3.4 we have that for a fixed support $T$,

$$(1-\epsilon)\left|\left|x\right|\right|_2 \leq \frac{1}{\sqrt{m}}\left|\left|Ax\right|\right|_2 \leq (1+\epsilon)\left|\left|x\right|\right|_2 \tag{83}$$

for any $x$ with support $T$ with probability at least

$$1-2\left(\frac{12}{\epsilon}\right)^s \exp\left(-\frac{m\epsilon^2}{32}\right). \tag{84}$$

There are

$$\binom{n}{s} \leq \left(\frac{en}{s}\right)^s \tag{85}$$

possible supports so by the union bound the result holds with probability at least

$$1 - 2\left(\frac{en}{s}\right)^s \left(\frac{12}{\epsilon}\right)^s \exp\left(-\frac{m\epsilon^2}{32}\right) = 1 - \exp\left(\log 2 + s + s\log\left(\frac{n}{s}\right) + s\log\left(\frac{12}{\epsilon}\right) - \frac{m\epsilon^2}{2}\right)$$
$$\leq 1 - \frac{C_2}{n} \tag{86}$$

for some constant $C_2$ as long as $m$ satisfies (82). □

The RIP not only establishes that the recovery problem is well posed at least in principle, it also implies that $\ell_1$-norm minimization achieves stable recovery in the presence of noise. Let us assume that we know an upper bound for the $\ell_2$ norm of the noise in model (78). We can then relax the equality constraint in Problem (106) to an inequality constraint that takes into account the noise level.

$$\text{minimize} \quad ||\tilde{x}||_1 \tag{87}$$
$$\text{subject to} \quad ||A\tilde{x} - y||_2 \leq \sigma. \tag{88}$$

If the RIP constant for a matrix is $\epsilon_{2s} < \sqrt{2} - 1$ and $x$ is $s$-sparse the solution $\hat{x}$ to the relaxed problem satisfies

$$||\hat{x} - x||_2 \leq C_0 \, \sigma \tag{89}$$

for a certain constant. In fact, even if the original vector $x$ is not sparse the solution will will satisfy

$$||\hat{x} - x||_2 \leq C_0 \, \epsilon_0 + C_1 \frac{||x - x_s||_1}{\sqrt{s}} \tag{90}$$

where $x_s$ contains the $s$ entries of $x$ with largest magnitude. We refer to [5] for the proof of this result.

# 4 Sampling

In this section we describe an application of compressed sensing to recovering signals that have a sparse spectrum.

## 4.1 Nyquist-Shannon sampling theorem

Consider a bandlimited signal $g \in \mathbb{L}_2([0,1])$, which is bandlimited. Its spectrum is equal to zero beyond a certain cut-off frequency $f$,

$$g(t) := \sum_{k=-f}^{f} c_k \exp(i2\pi kt). \tag{91}$$

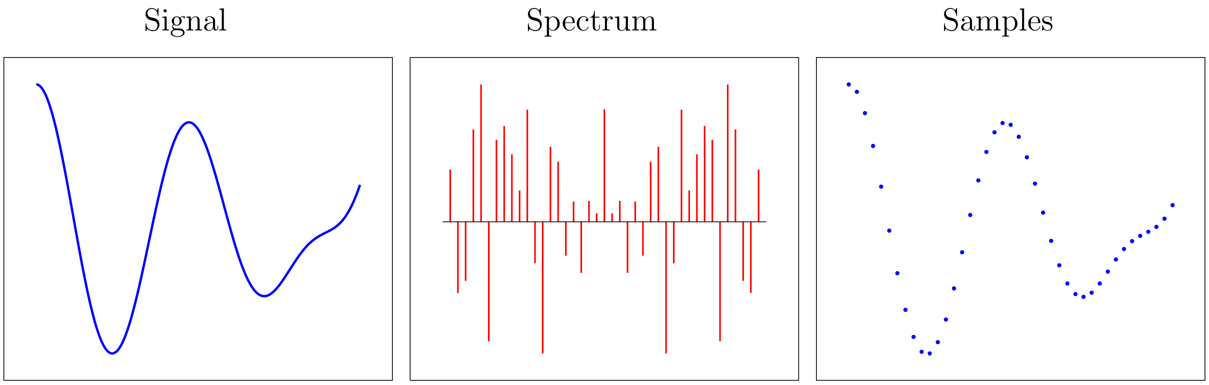


**Figure 10:** Bandlimited signal (left), corresponding spectrum (center) and regular samples (right).

Our aim is to estimate the signal from samples taken at regular intervals $g(0)$, $g\left(\frac{1}{n}\right)$, $g\left(\frac{2}{n}\right)$, …, $g\left(\frac{n-1}{n}\right)$, where the sampling rate is determined by $n$. Such situations arise when we need to store an analog signal such as music or speech digitally. Two important questions are:

1. What sampling rate is necessary to preserve all the information in the signal?
2. How can we reconstruct the original signal from the sample?

Figure 10 shows an example of a bandlimited signal, its spectrum and the corresponding samples.

In order to simplify notation we define

$$a_{-f:f}(t)_k := \begin{bmatrix} \exp(-i2\pi(-f)t) \\ \exp(-i2\pi(-f+1)t) \\ \cdots \\ \exp(-i2\pi(f-1)t) \\ \exp(-i2\pi f t) \end{bmatrix}. \tag{92}$$

We can now easily express $g$ in terms of $a_{-f:f}$ and the vector of Fourier coefficients $c$,

$$g(t) := \sum_{k=-f}^{f} c_k \exp(i2\pi kt) = a_{-f:f}(t)^* \, c. \tag{93}$$

For a fixed sampling rate the samples provide a system of $n$ linear equations

$$F^* c = \begin{bmatrix} a_{-f:f}\left(0\right)^* \\ a_{-f:f}\left(\frac{1}{n}\right)^* \\ a_{-f:f}\left(\frac{2}{n}\right)^* \\ \cdots \\ a_{-f:f}\left(\frac{n-1}{n}\right)^* \end{bmatrix} c = \begin{bmatrix} g\left(0\right) \\ g\left(\frac{1}{n}\right) \\ g\left(\frac{2}{n}\right) \\ \cdots \\ g\left(\frac{n-1}{n}\right) \end{bmatrix}. \tag{94}$$

Since there are $2f+1$ unknowns we need $n \geq 2f+1$ to guarantee that the linear system has a unique solution. Setting $n = 2f+1$, we can check that the vectors

$$\frac{1}{\sqrt{n}} a_{-f:f}\left(0\right), \; \frac{1}{\sqrt{n}} a_{-f:f}\left(\frac{1}{n}\right), \ldots, \frac{1}{\sqrt{n}} a_{-f:f}\left(\frac{n-1}{n}\right) \tag{95}$$

form an orthonormal basis (they actually form a DFT matrix!), so applying the adjoint matrix allows to invert the system

$$c = \frac{1}{n} F F^* c = \frac{1}{n} F \begin{bmatrix} g\left(0\right) \\ g\left(\frac{1}{n}\right) \\ g\left(\frac{2}{n}\right) \\ \cdots \\ g\left(\frac{n-1}{n}\right) \end{bmatrix} = \frac{1}{n} \sum_{j=0}^{n} g\left(\frac{j}{n}\right) a_{-f:f}\left(\frac{j}{n}\right). \tag{96}$$

In order to interpret this linear inversion in traditional signal processing terms, let us define the periodized sinc or Dirichlet kernel

$$D_f\left(t\right) := \frac{1}{n} \sum_{k=-f}^{f} e^{-i2\pi k t} \tag{97}$$

$$= \frac{\sin\left(\pi n t\right)}{n \sin\left(\pi t\right)}. \tag{98}$$

A plot of this function is shown in Figure 11. We can easily express the result of shifting $D$ by $\tau$ in terms of $a_{-f:f}$

$$D_f\left(t-\tau\right) = \frac{1}{n} \sum_{k=-f}^{f} e^{-i2\pi k (t-\tau)} \tag{99}$$

$$= \frac{1}{n} a_{-f:f}\left(t\right)^* a_{-f:f}\left(\tau\right). \tag{100}$$

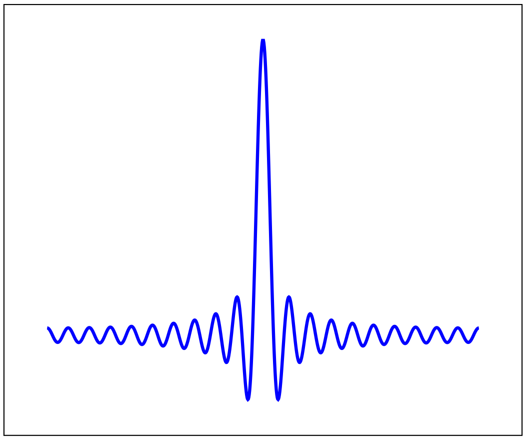


**Figure 11:** Periodized sinc or Dirichlet kernel.

This allows us to establish that obtaining the Fourier coefficients through linear inversion and then using them to reconstruct the signal is equivalent to interpolation with shifted sinc functions, as sketched in Figure 12.

$$
\begin{aligned}
g\left(t\right) &= a_{-f:f}\left(t\right)^{*} c && (101)\\
&= \frac{1}{n}\sum_{j=0}^{n} g\left(\frac{j}{n}\right) a_{-f:f}\left(t\right)^{*} a_{-f:f}\left(\frac{j}{n}\right) \qquad \text{by (96)} && (102)\\
&= \sum_{j=0}^{n} g\left(\frac{j}{n}\right) D_f\left(t-\frac{j}{n}\right). && (103)
\end{aligned}
$$

We conclude that the sampling rate should be twice the cut-off frequency and that reconstruction can be carried out by interpolating the samples with a sinc function. This is known as the Nyquist-Shannon sampling theorem.

## 4.2 Compressive sampling

We now consider the problem of sampling a signal $g \in \mathbb{L}_2\left([0,1]\right)$ with a sparse spectrum

$$g\left(t\right) := \sum_{k\in\mathcal{S}} c_k \exp\left(i2\pi kt\right) \qquad (104)$$

and then recovering it from its samples. The signal consists of $s$ frequency components and has a cut-off frequency of $f$. According to the Shannon-Nyquist sampling theorem, we need $2f+1$ samples in order to preserve the information in the signal and recover it applying

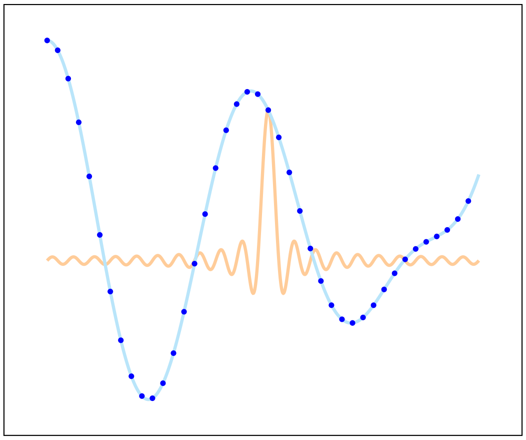


**Figure 12:** Reconstruction via linear inversion is equivalent to interpolation with a sinc function.

sinc interpolation. In contrast, compressed sensing theory implies that recovery via $\ell_1$-norm minimization is possible from $\mathcal{O}\left(s \log\left(2f+1\right)\right)$ randomized samples of the form

$$F_\Omega^* \, c = \begin{bmatrix} \cancel{a_{-f:f}\left(0\right)^*} \\ a_{-f:f}\left(\frac{1}{n}\right)^* \\ \cancel{a_{-f:f}\left(\frac{2}{n}\right)^*} \\ \cdots \\ a_{-f:f}\left(\frac{n-1}{n}\right)^* \end{bmatrix} c = \begin{bmatrix} \cancel{g\left(0\right)} \\ g\left(\frac{1}{n}\right) \\ \cancel{g\left(\frac{2}{n}\right)} \\ \cdots \\ g\left(\frac{n-1}{n}\right) \end{bmatrix} := y_\Omega, \tag{105}$$

where $\Omega$ is a random subset of $\{1, 2, \ldots, 2f+1\}$. In detail, we can estimate the Fourier coefficients $c$ by solving

$$\text{minimize} \quad \left|\left|\tilde{c}\right|\right|_1 \tag{106}$$

$$\text{subject to} \quad y_\Omega = F_\Omega^* \tilde{c} \tag{107}$$

and then reconstruct $g$ using (96). If $s \ll 2f+1$ this procedure allows to reduce the number of measurements significantly with respect to traditional Nyquist sampling. Figure 13 shows an example which compares the two sampling schemes.

# References

The proof of the Johnson-Lindenstrauss lemma is based on [9]. The proofs of exact recovery via $\ell_1$-norm minimization and of the restricted isometry property of Gaussian matrices are

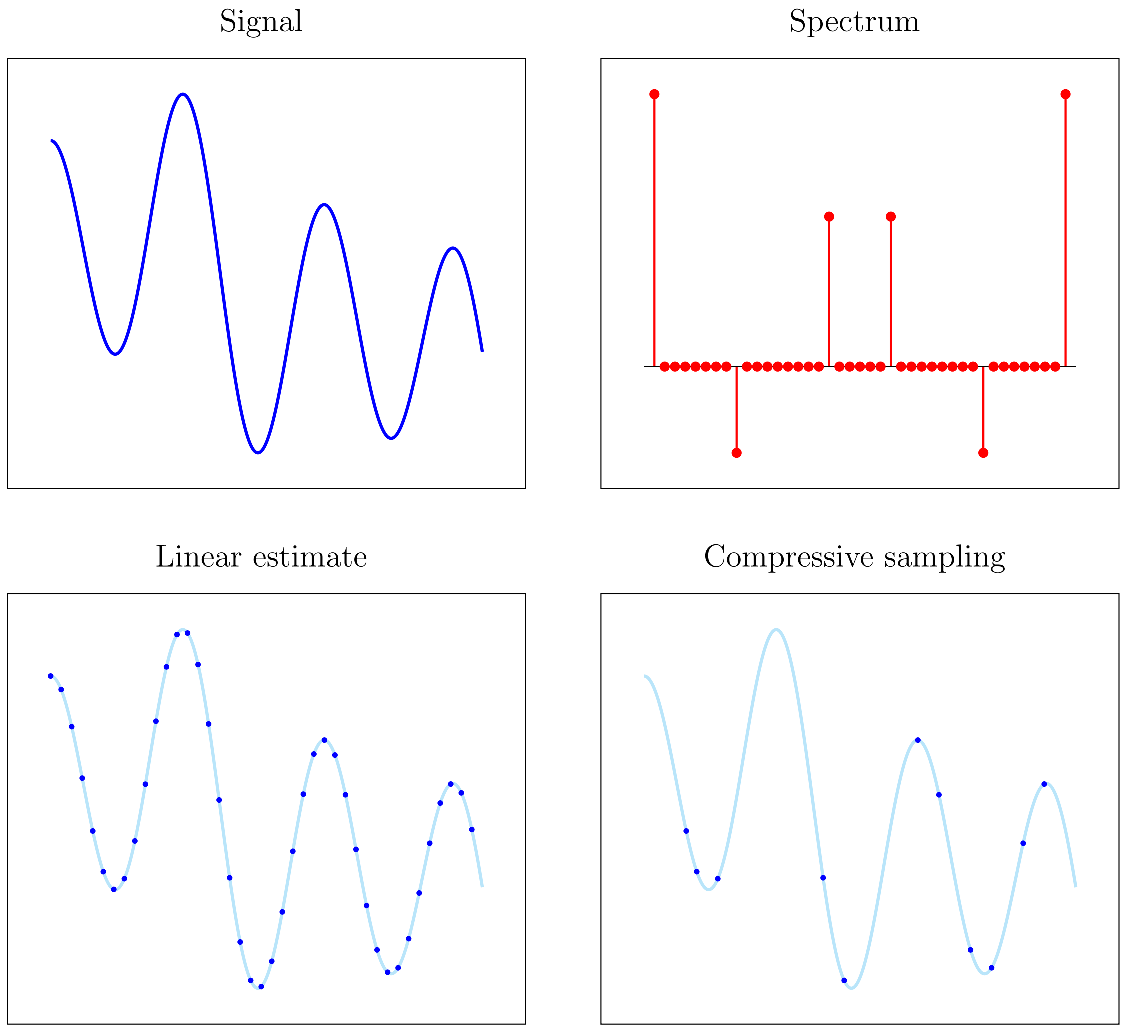


**Figure 13:** Example showing the samples that are necessary to recover a signal (top left) with a sparse spectrum (top right). Compressive sampling (bottom right) requires significantly less samples than traditional Nyquist sampling (bottom left).

based on arguments in [3] and [2]. For further reading on mathematical tools used to analyze compressed sensing we refer to [12] and [10].

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[2] 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.

[3] E. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589, 2013.

[4] E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[5] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[6] E. J. Candès and Y. Plan. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

[7] E. J. Candès, J. K. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[8] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions in Information Theory*, 52:5406–5425, 2006.

[9] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[10] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*, volume 1. 2013.

[11] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

# A Proofs

## A.1 Proof of Lemma 2.2

$x - UU^T x$ is orthogonal to any vector $y := Uc \in \mathcal{S}$ since

$$\begin{aligned} y^T (x - UU^T x) &= c^T U^T x - c^T U^T U U^T x && (108) \\ &= c^T U^T x - c^T U^T x && (109) \\ &= 0. && (110) \end{aligned}$$

By Pythagoras's theorem this implies

$$\begin{aligned} \|x - y\|_2^2 &= \left\| x - UU^T x + UU^T x - y \right\|_2^2 && (111) \\ &= \left\| x - UU^T x \right\|_2^2 + \left\| UU^T x - y \right\|_2^2 && (112) \end{aligned}$$

which is minimized by $y = UU^T x$.

## A.2 Proof of Theorem 2.5

We will need the following lemma, proved in Section A.3.

**Lemma A.1.** *For any matrix $X \in \mathbb{R}^{n \times k}$, where $k > n$, with left singular vectors $u_1, u_2, \ldots, u_n$ corresponding to the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$,*

$$\begin{aligned} \sigma_1 &= \max_{\|u\|_2 = 1} \left\| X^T u \right\|_2, && (113) \\ u_1 &= \arg \max_{\|u\|_2 = 1} \left\| X^T u \right\|_2, && (114) \\ \sigma_m &= \max_{\substack{\|u\|_2 = 1 \\ u \perp u_1, \ldots, u_{m-1}}} \left\| X^T u \right\|_2, \quad 2 \leq m \leq n, && (115) \\ u_m &= \arg \max_{\substack{\|u\|_2 = 1 \\ u \perp u_1, \ldots, u_{m-1}}} \left\| X^T u \right\|_2, \quad 2 \leq m \leq n. && (116) \end{aligned}$$

We will prove the result by induction on $m$. The base case $m = 1$ follows immediately from (114). To complete the proof we need to show that if the result is true for $m - 1 \geq 1$ (the induction hypothesis) then it also holds for $m$.

Let $\mathcal{S}$ be an arbitrary subspace of dimension $m$. We choose an orthonormal basis for the subspace $b_1, b_2, \ldots, b_m$ such that $b_m$ is orthogonal to $u_1, u_2, \ldots, u_{m-1}$. Note that such a vector must exist because otherwise $\mathcal{S}$ cannot have dimension $m$.

By the induction hypothesis,

$$\sum_{i=1}^{m-1} \left|\left|X^T u_i\right|\right|_2^2 = \sum_{i=1}^{k} \left|\left|\mathcal{P}_{\text{span}(u_1,u_2,\ldots,u_{m-1})}\, x_i\right|\right|_2^2 \tag{117}$$

$$\geq \sum_{i=1}^{k} \left|\left|\mathcal{P}_{\text{span}(b_1,b_2,\ldots,b_{m-1})}\, x_i\right|\right|_2^2 \tag{118}$$

$$= \sum_{i=1}^{m-1} \left|\left|X^T b_i\right|\right|_2^2 . \tag{119}$$

By (116)

$$\sum_{i=1}^{n} \left|\left|\mathcal{P}_{\text{span}(u_k)}\, x_i\right|\right|_2^2 = \left|\left|X^T u_k\right|\right|_2^2 \tag{120}$$

$$\geq \left|\left|X^T b_k\right|\right|_2^2 . \tag{121}$$

Combining (119) and (121) we conclude

$$\sum_{i=1}^{k} \left|\left|\mathcal{P}_{\text{span}(u_1,u_2,\ldots,u_m)}\, x_i\right|\right|_2^2 = \sum_{i=1}^{m} \left|\left|X^T u_i\right|\right|_2^2 \tag{122}$$

$$\geq \sum_{i=1}^{m} \left|\left|X^T b_i\right|\right|_2^2 \tag{123}$$

$$\geq \sum_{i=1}^{k} \left|\left|\mathcal{P}_{\mathcal{S}}\, x_i\right|\right|_2^2 . \tag{124}$$

## A.3 Proof of Lemma A.1

The left singular vectors are an orthonormal basis of $\mathbb{R}^n$, so we can represent any unit-norm vector $h_m$ that is orthogonal to $u_m, \ldots, u_{m-1}$ as

$$h_k = \sum_{i=m}^{n} \alpha_i u_i \tag{125}$$

where

$$\left|\left|h_m\right|\right|_2^2 = \sum_{i=m}^{n} \alpha_i^2 = 1. \tag{126}$$

Note that $h_1$ is just an arbitrary unit-norm vector.

Now we have

$$\left|\left|X^T h_m\right|\right|_2^2 = \sum_{i=1}^{n} \sigma_i \left( \sum_{j=m}^{n} \alpha_j u_i^T u_j \right)^2 \tag{127}$$

$$= \sum_{i=m}^{n} \sigma_i \alpha_i^2 \quad \text{because } u_1, \ldots, u_m \text{ is an orthonormal basis} \tag{128}$$

$$\leq \sigma_m \sum_{i=m}^{n} \alpha_i^2 \quad \text{because } \sigma_m \geq \sigma_{m+1} \geq \ldots \geq \sigma_n \tag{129}$$

$$= \sigma_m \quad \text{by (126).} \tag{130}$$

This establishes (113) and (115). To prove (114) and (116) we just need to show that $u_m$ achieves the maximum

$$\left|\left|X^T u_m\right|\right|_2^2 = \sum_{i=1}^{n} \sigma_i \left(u_i^T u_m\right)^2 \tag{131}$$

$$= \sigma_m. \tag{132}$$

## A.4 Proof of Proposition 2.7

The proof relies on the following concentration bounds for $\chi^2$ or chi-square random variables, proved in Section A.5 of the appendix.

**Proposition A.2** (Tail bound for chi-square random variables)**.** *Let $Z$ be a $\chi^2$ or chi-square random variable with $m$ degrees of freedom, i.e.*

$$Z = \sum_{i=1}^{m} X_i^2 \tag{133}$$

*where $X_1, \ldots, X_m$ are independent Gaussian random variables with zero mean and unit variance. For any $\epsilon > 0$ we have*

$$P\left(Z > m\left(1+\epsilon\right)\right) \leq \exp\left(-\frac{m\epsilon^2}{8}\right), \tag{134}$$

$$P\left(Z < m\left(1-\epsilon\right)\right) \leq \exp\left(-\frac{m\epsilon^2}{2}\right). \tag{135}$$

Let $a_i := v^T A_i$ where $A_i$ is the $i$th row of $A$ and hence an $n$-dimensional random vector with zero mean and covariance matrix equal to the identity $I$. $a_i$ is a Gaussian random variable with zero mean and variance $||v||_2^2$. This follows from the following well-known lemma (we omit the proof).

**Lemma A.3** (Linear transformation of a Gaussian vector). *Let $M \in \mathbb{R}m \times n$. If an $n$-dimensional Gaussian vector $v$ has mean $\mu$ and covariance matrix $\Sigma$, $Mv$ is an $m$-dimensional Gaussian vector with mean $M\mu$ and covariance matrix $M\Sigma M^T$.*

As a result, we have that

$$Z := \frac{||Av||_2^2}{||v||_2^2} \tag{136}$$

$$= \sum_{i=1}^{m} \frac{a_i^2}{||v||_2^2} \tag{137}$$

is a chi-square random variable with $m$ degrees of freedom. Let us define the events

$$\mathcal{E}_1 := \{||f(v)||_2 > (1+\epsilon)\, ||v||_2\}, \tag{138}$$

$$\mathcal{E}_2 := \{||f(v)||_2 < (1-\epsilon)\, ||v||_2\}. \tag{139}$$

Inequalities (134) and (135) applied to $Z$ imply

$$P(\mathcal{E}_1) = P\left(Z > m(1+\epsilon)^2\right) \tag{140}$$

$$\leq \exp\left(-\frac{m\epsilon^2}{8}\right), \tag{141}$$

$$P(\mathcal{E}_2) = P\left(Z < m(1-\epsilon)^2\right) \tag{142}$$

$$\leq \exp\left(-\frac{m\epsilon^2}{2}\right). \tag{143}$$

By the union bound

$$P(\mathcal{E}_1 \cup \mathcal{E}_2) \leq P(\mathcal{E}_1) + P(\mathcal{E}_2) \tag{144}$$

$$\leq 2\exp\left(-\frac{m\epsilon^2}{8}\right). \tag{145}$$

## A.5 Proof of Proposition A.2

Proof of (134) The concentration bound is established by applying Markov's inequality after exponentiation, as is commonly done to prove Chernoff bounds.

**Proposition A.4** (Markov's inequality). *Let $X$ be a nonnegative random variable. For any positive constant $a > 0$,*

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}. \tag{146}$$

Markov's inequality is proved in Section A.10 below. Let $t > 0$ be an arbitrary positive number, we have

$$\begin{aligned} P\left(Z > a\right) &= P\left(\exp\left(tZ\right) > \exp\left(at\right)\right) && (147)\\ &\leq \exp\left(-at\right)\,\mathrm{E}\left(\exp\left(tZ\right)\right) \qquad \text{by Markov's inequality} && (148)\\ &\leq \exp\left(-at\right)\,\mathrm{E}\left(\exp\left(\sum_{i=1}^{m} tX_i^2\right)\right) && (149)\\ &\leq \exp\left(-at\right)\prod_{i=1}^{m}\mathrm{E}\left(\exp\left(tX_i^2\right)\right) \qquad \text{by independence of } X_1, \ldots, X_m. && (150) \end{aligned}$$

Some calculus yields

$$\begin{aligned} \mathrm{E}\left(\exp\left(tX^2\right)\right) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{u^2}{2}\right)\exp\left(tu^2\right)\,\mathrm{d}\,u && (151)\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{\left(1-2t\right)u^2}{2}\right)\,\mathrm{d}\,u && (152)\\ &= \frac{1}{\sqrt{2\pi\left(1-2t\right)}}\int_{-\infty}^{\infty}\exp\left(-\frac{v^2}{2}\right)\,\mathrm{d}\,v \qquad \text{change of variables } v = \left(2-t\right)u\\ &= \frac{1}{\sqrt{1-2t}} && (153) \end{aligned}$$

Combining (150) and (153),

$$P\left(Z > a\right) \leq \frac{\exp\left(-at\right)}{\left(1-2t\right)^{\frac{m}{2}}}. \tag{154}$$

Setting

$$\begin{aligned} t &:= \frac{1}{2} - \frac{1}{2\left(1+\epsilon\right)}, && (155)\\ a &:= m\left(1+\epsilon\right) && (156) \end{aligned}$$

we have

$$\begin{aligned} P\left(Z > m\left(1+\epsilon\right)\right) &\leq \left(1+\epsilon\right)^{m} 2\exp\left(-\frac{m\epsilon}{2}\right) && (157)\\ &= \exp\left(-\frac{m}{2}\left(\epsilon - \log\left(1+\epsilon\right)\right)\right). && (158) \end{aligned}$$

The function $g\left(x\right) := x - \frac{x^2}{4} - \log\left(1+x\right)$ is nonnegative between 0 and 1 (the derivative is nonnegative and $g\left(0\right) = 0$). This implies

$$P\left(Z > m\left(1+\epsilon\right)\right) \leq \exp\left(-\frac{m\epsilon^2}{8}\right). \tag{159}$$

Proof of (135)

A very similar argument to the one that yields (150) gives

$$P\left(Z < a'\right) = P\left(\exp\left(-t'Z\right) > \exp\left(-a't'\right)\right) \tag{160}$$

$$\leq \exp\left(a't'\right) \prod_{i=1}^{m} \mathrm{E}\left(\exp\left(-t'X_i^2\right)\right). \tag{161}$$

Setting $t' = t$ in (153), we have

$$\mathrm{E}\left(\exp\left(-t'X^2\right)\right) = \frac{1}{\sqrt{1+2t'}}. \tag{162}$$

This implies

$$P\left(Z < a'\right) \leq \frac{\exp\left(a't'\right)}{\left(1+2t'\right)^{\frac{m}{2}}}. \tag{163}$$

Setting

$$t' := -\frac{1}{2} + \frac{1}{2\left(1-\epsilon\right)}, \tag{164}$$

$$a' := m\left(1-\epsilon\right) \tag{165}$$

we have

$$P\left(Z < m\left(1-\epsilon\right)\right) \leq \left(1-\epsilon\right)^{\frac{m}{2}} \exp\left(\frac{m\epsilon}{2}\right) \tag{166}$$

$$= \exp\left(-\frac{m}{2}\left(-\epsilon - \log\left(1-\epsilon\right)\right)\right). \tag{167}$$

The function $h\left(x\right) := -x - \frac{x^2}{2} - \log\left(1-x\right)$ is nonnegative between 0 and 1 (the derivative is nonnegative and $g\left(0\right) = 0$). We conclude that

$$P\left(Z < m\left(1-\epsilon\right)\right) \leq \exp\left(-\frac{m\epsilon^2}{2}\right). \tag{168}$$

## A.6 Proof of Lemma 3.2

Note that

$$q := A^T v \tag{169}$$

is a subgradient of the $\ell_1$ norm at $x$ by Proposition 3.5 in Lecture Notes 2. For any feasible vector $\tilde{x}$, $h := x - \tilde{x}$ satisfies $Ah = 0$. Since $q$ is a subgradient at $x$ we immediately have

$$\begin{aligned}
||\tilde{x}||_1 &= ||x + h||_1 && (170)\\
&\geq ||x||_1 + q^T h && (171)\\
&= ||x||_1 + v^T A h && (172)\\
&= ||x||_1 . && (173)
\end{aligned}$$

This proves that $x$ is a solution but not that it is the only solution. For this we need to use the strict inequality (32) and the assumption that $A_T$ is full rank. If $A_T$ is full rank then $h_{T^c} \neq 0$ unless $h = 0$ because otherwise $h_T$ would be a nonzero vector in the null space of $A_T$. This together with (32) implies

$$||h_{T^c}||_1 > q^T h_{T^c}. \tag{174}$$

Let $\mathcal{P}_T\left(\cdot\right)$ denote a projection that sets to zero all entries of a vector except the ones indexed by $T$. We have

$$\begin{aligned}
||\tilde{x}||_1 &= ||x + \mathcal{P}_T\left(h\right)||_1 + ||h_{T^c}||_1 && \text{because } x \text{ is supported on } T && (175)\\
&> ||x||_1 + q^T \mathcal{P}_T\left(h\right) + q^T \mathcal{P}_{T^c}\left(h\right) && \text{by (174)} && (176)\\
&= ||x||_1 + q^T h && && (177)\\
&= ||x||_1 . && && (178)
\end{aligned}$$

Since this holds for any arbitrary feasible vector $\tilde{x}$ we conclude that $x$ must be the unique solution.

## A.7 Proof of Lemma 3.3

The Lagrangian is equal to

$$\begin{aligned}
\mathcal{L}\left(\tilde{x}, \tilde{v}\right) &= ||\tilde{x}||_1 + \tilde{v}^T\left(y - A\tilde{x}\right) && (179)\\
&= ||\tilde{x}||_1 - \left(A^T \tilde{v}\right)^T \tilde{x} + y^T \tilde{v}. && (180)
\end{aligned}$$

To obtain the Lagrange dual function we minimize with respect to $\tilde{x}$,

$$\min_{\tilde{x}} \mathcal{L}\left(\tilde{x}, \tilde{v}\right) = \begin{cases} y^T \tilde{v} & \text{if } \left|\left|A^T \tilde{v}\right|\right|_\infty \leq 1, \\ -\infty & \text{otherwise.} \end{cases} \tag{181}$$

The dual problem consists of maximizing this Lagrange dual function.

## A.8 Proof of Proposition 3.8

We construct an $\epsilon$ covering set $\mathcal{N}_\epsilon \subseteq \mathcal{S}^{s-1}$ recursively:

- We initialize $\mathcal{N}_\epsilon$ to the empty set.
- We choose a point $x \in \mathcal{S}^{s-1}$ such that $||x-y||_2 > \epsilon$ for any $y \in \mathcal{C}$. We add $x$ to $\mathcal{N}_\epsilon$ until there are no points in $\mathcal{S}^{s-1}$ that are $\epsilon$ away from any point in $\mathcal{N}_\epsilon$.

This algorithm necessarily ends in a finite number of steps because the $n$-dimensional sphere is compact (otherwise we would have an infinite sequence such that no subsequence converges).

Now, let us consider the balls of radius $\epsilon/2$ centered at each of the points in $\mathcal{N}_\epsilon$. These balls do not intersect since their centers are at least $\epsilon$ apart and they are all inside the ball of radius $1+\epsilon/2$ centered at the origin because $\mathcal{C} \subseteq \mathcal{S}^{s-1}$. This means that

$$\begin{aligned} \operatorname{Vol}\left(\mathcal{B}^s_{1+\epsilon/2}\left(0\right)\right) &\geq \operatorname{Vol}\left(\cup_{x\in\mathcal{N}_\epsilon}\mathcal{B}^s_{\epsilon/2}\left(x\right)\right) && (182)\\ &= \left|\mathcal{N}_\epsilon\right| \operatorname{Vol}\left(\mathcal{B}^s_{\epsilon/2}\left(0\right)\right) && (183) \end{aligned}$$

where $\mathcal{B}^s_r\left(x\right)$ is the ball of radius $r$ centered at $x$. By multivariable calculus

$$\operatorname{Vol}\left(\mathcal{B}^s_r\left(0\right)\right) = r^s \operatorname{Vol}\left(\mathcal{B}^s_1\left(0\right)\right), \tag{184}$$

so (182) implies

$$\left(1+\epsilon/2\right)^s \geq \left|\mathcal{N}_\epsilon\right| \left(\epsilon/2\right)^s. \tag{185}$$

## A.9 Proof of Lemma 3.5

By symmetry of the Gaussian probability density function, we just need to bound the probability that $u > t$. Applying Markov's inequality (see Proposition A.4) we have

$$\begin{aligned} \mathrm{P}\left(u \geq t\right) &= \mathrm{P}\left(\exp\left(ut\right) \geq \exp\left(t^2\right)\right) && (186)\\ &\leq \mathrm{E}\left(\exp\left(ut-t^2\right)\right) && (187)\\ &= \exp\left(-\frac{t^2}{2}\right)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{\left(u-t\right)^2}{2}\right)\mathrm{d}\,u && (188)\\ &= \exp\left(-\frac{t^2}{2}\right). && (189) \end{aligned}$$

## A.10 Proof of Proposition A.4

Consider the indicator variable $1_{X \geq a}$. Clearly the random variable

$$X - a\,1_{X \geq a} \geq 0. \tag{190}$$

So in particular its expectation is non-negative (as it is the sum or integral of a non-negative quantity over the positive real line). By linearity of expectation and the fact that $1_{X \geq a}$ is a Bernoulli random variable with expectation $\mathrm{P}\left(X \geq a\right)$ we have

$$\mathrm{E}\left(X\right) \geq a\,\mathrm{E}\left(1_{X \geq a}\right) = a\,\mathrm{P}\left(X \geq a\right). \tag{191}$$