Randomization

1 Motivation

1.1 Randomized dimensionality reduction

As discussed in the notes on the SVD, dimensionality reduction is an important tool in data analysis. Dimensionality reduction via PCA is optimal in the sense that it preserves as much variance as possible in the data. However, it is computationally expensive. It also requires having all the data available beforehand, which is impractical for real-time applications. For such cases we need a non-adaptive alternative to PCA that chooses the projection before seeing the data. A possibility is to project the data onto a small number of random directions, by taking inner products with unit-norm vectors with random orientations. Note that this is not a projection strictly speaking because the directions are not orthogonal. Perhaps surprisingly, this is often quite effective, as illustrated in Figure 1.[1] The randomized maps do not preserve the variance as much as PCA, but they do a pretty good job. In these notes we will provide a mathematical analysis of this phenomenon.

1.2 Compressed sensing

As mentioned in previous notes, magnetic resonance imaging (MRI) is a popular medical-imaging technique that measures the response of the atomic nuclei of body tissues to high-frequency radio waves when placed in a strong magnetic field. MRI measurements can be modeled as samples from the 2D or 3D Fourier series of the object that is being imaged, for example a slice of a human leg. An important challenge in MRI is to reduce measurement time: long acquisition times are expensive and bothersome for the patients, especially if they are children or seriously ill.

Gathering less measurements, or equivalently undersampling the Fourier coefficients of the image of interest, results in shorter data-acquisition times, but poses the challenge of recovering the image from undersampled data. As we saw in the notes on the frequency domain, regular undersampling results in severe aliasing in the image domain. As shown in Figure 2, this is a result of adding a circular shift of the image with itself. The aliasing is extremely difficult to remove, because most the features from the shifted image are very similar to those of the unshifted image. Interestingly, random undersampling results in noise-like aliasing, as illustrated in the figure. This kind of aliasing is very different to the image features, and is therefore easier to remove. The problem of estimating signals from underdetermined random measurements is known as compressed sensing.

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

Figure 1: Dimensionality reduction of a data set of seeds with seven features (area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove) onto two dimensions. The seeds belong to three different varieties of wheat: Kama, Rosa and Canadian. The images compare dimensionality reduction via PCA with a randomized approach, where we project onto two random directions. Each color represents a variety of wheat.

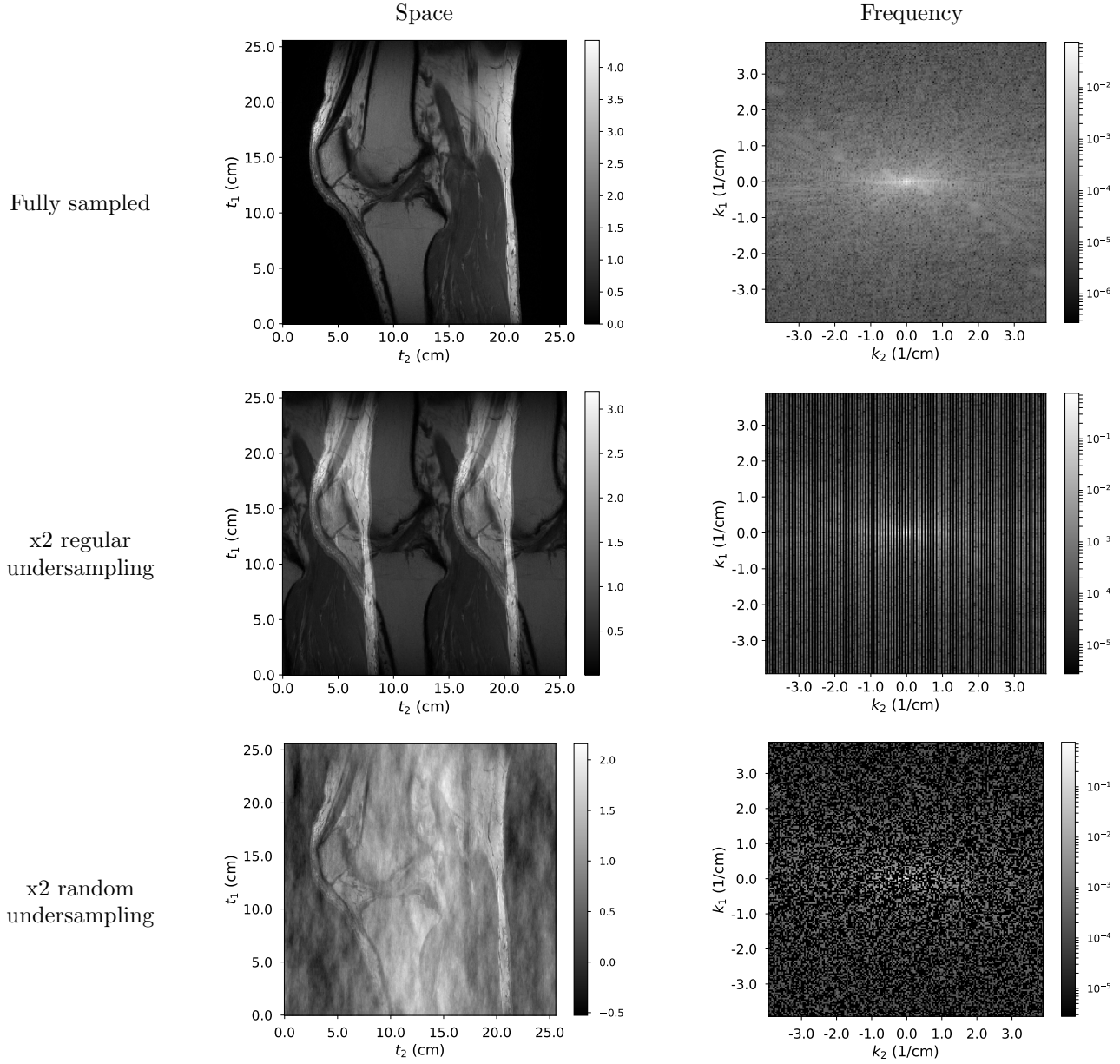

Figure 2: Sagittal section of a human knee measured by magnetic-resonance imaging. The images on the left show the image recovered from the Fourier-domain sampling patterns shown on the right. Regular (center) and random (bottom) undersampling patterns produce very different aliasing artifacts.

2 Gaussian random variables

Our analysis of randomization techniques relies heavily on Gaussian random variables and vectors. In this section we describe some basic properties. We represent random quantities with bold font.

2.1 The Gaussian distribution

The Gaussian or normal random variable is arguably the most popular random variable in statistical modeling and signal processing. The reason is that sums of independent random variables often converge to Gaussian distributions, a phenomenon characterized by the central limit theorem. As a result any quantity that results from the additive combination of several unrelated factors will tend to have a Gaussian distribution. For example, in signal processing and engineering, noise is often modeled as Gaussian. Figure 3 shows the pdfs of Gaussian random variables with different means and variances. When a Gaussian has mean zero and unit variance, we call it a standard Gaussian.

Definition 2.1 (Gaussian)**.** *The pdf of a Gaussian or normal random variable with mean μ and standard deviation σ is given by*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

An important property of Gaussian random variables is that scaling and shifting Gaussians preserves their distribution.

Lemma 2.2. *If $\mathbf{x}$ is a Gaussian random variable with mean μ and standard deviation σ, then for any $a, b \in \mathbb{R}$*

$$\mathbf{y} := a\mathbf{x} + b \tag{2}$$

is a Gaussian random variable with mean $a\mu + b$ and standard deviation $|a|\,\sigma$.

Proof. We assume $a > 0$ (the argument for $a < 0$ is very similar), to obtain

$$\begin{aligned}
F_{\mathbf{y}}(y) &= \mathrm{P}(\mathbf{y} \leq y) && (3)\\
&= \mathrm{P}(a\mathbf{x} + b \leq y) && (4)\\
&= \mathrm{P}\left(\mathbf{x} \leq \frac{y-b}{a}\right) && (5)\\
&= \int_{-\infty}^{\frac{y-b}{a}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x && (6)\\
&= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(w-a\mu-b)^2}{2a^2\sigma^2}}\,\mathrm{d}w \qquad \text{by the change of variables } w = ax + b. && (7)
\end{aligned}$$

Differentiating with respect to y yields

$$f_{\mathbf{y}}(y) = \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(w-a\mu-b)^2}{2a^2\sigma^2}} \tag{8}$$

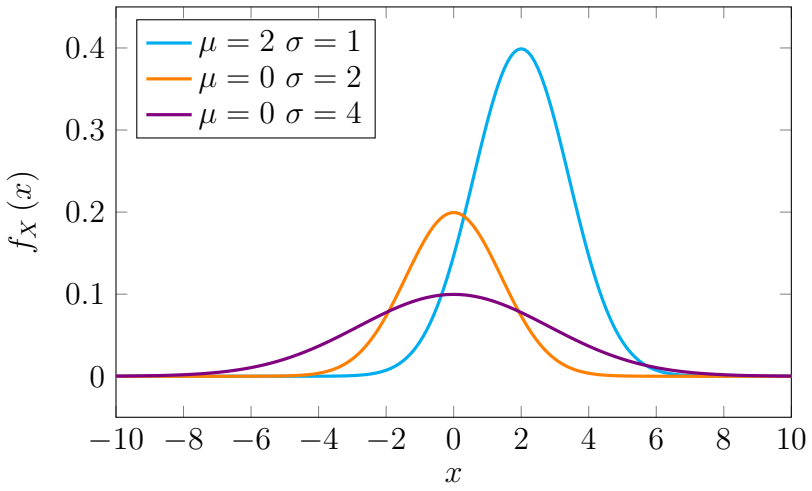

Figure 3: Gaussian random variable with different means and standard deviations.

so $\mathbf{y}$ is indeed a standard Gaussian random variable with mean $a\mu + b$ and standard deviation $|a|\,\sigma$. □

2.2 The multidimensional Gaussian distribution

Gaussian random vectors are a multidimensional generalization of Gaussian random variables. They are parametrized by a vector and a matrix that correspond to their mean and covariance matrix.

Definition 2.3 (Gaussian random vector)**.** *A Gaussian random vector $\vec{\mathbf{x}}$ of dimension d is a random vector with joint pdf*

$$f_{\vec{\mathbf{x}}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})\right), \tag{9}$$

where $|\Sigma|$ denotes the determinant of Σ. The mean vector $\vec{\mu} \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, which is symmetric and positive definite, parametrize the distribution.

When the covariance matrix of a Gaussian vector is diagonal, its components are all independent.

Lemma 2.4 (Uncorrelation implies mutual independence for Gaussian random variables)**.** *If all the components of a Gaussian random vector $\vec{\mathbf{x}}$ are uncorrelated, then they are also mutually independent.*

Proof. If all the components are uncorrelated then the covariance matrix is diagonal

$$\Sigma_{\vec{\mathbf{x}}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}, \tag{10}$$

where σ_i is the standard deviation of the ith component. Now, the inverse of this diagonal matrix is just

$$\Sigma_{\vec{\mathbf{x}}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_d^2} \end{bmatrix}, \tag{11}$$

and its determinant is $|\Sigma| = \prod_{i=1}^{d} \sigma_i^2$ so that

$$f_{\vec{\mathbf{x}}}\left(\vec{x}\right) = \frac{1}{\sqrt{\left(2\pi\right)^d |\Sigma|}} \exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \tag{12}$$

$$= \prod_{i=1}^{d} \frac{1}{\sqrt{(2\pi)}\sigma_i} \exp\left(-\frac{\left(\vec{x}_i-\mu_i\right)^2}{2\sigma_i^2}\right) \tag{13}$$

$$= \prod_{i=1}^{d} f_{\vec{\mathbf{x}}_i}\left(\vec{x}_i\right). \tag{14}$$

Since the joint pdf factors into the product of the marginals, the components are all mutually independent. □

When the covariance matrix of a Gaussian vector is the identity and its mean is zero, then its entries are independent identically-distributed (iid) standard Gaussians with mean zero and unit variance. We refer to such vectors as standard Gaussian vectors.

A fundamental property of Gaussian random vectors is that performing linear transformations on them always yields vectors with joint distributions that are also Gaussian. This is a multidimensional generalization of Lemma 2.2. We omit the proof, which is similar to that of Lemma 2.2.

Theorem 2.5 (Linear transformations of Gaussian random vectors are Gaussian)**.** *Let $\vec{\mathbf{x}}$ be a Gaussian random vector of dimension d with mean $\vec{\mu}_{\vec{\mathbf{x}}}$ and covariance matrix $\Sigma_{\vec{\mathbf{x}}}$. For any matrix $A \in \mathbb{R}^{m\times d}$ and $\vec{b} \in \mathbb{R}^m$, $\vec{\mathbf{y}} = A\vec{\mathbf{x}} + \vec{b}$ is a Gaussian random vector with mean $\vec{\mu}_{\vec{\mathbf{x}}} := A\vec{\mu}_{\vec{\mathbf{x}}} + \vec{b}$ and covariance matrix $\Sigma_{\vec{\mathbf{y}}} := A\Sigma_{\vec{\mathbf{x}}}A^T$, as long as $\Sigma_{\vec{\mathbf{y}}}$ is full rank.*

We have already this result in our derivation of the Wiener filter, where we assume that the DFT coefficients of iid Gaussian noise are also iid Gaussian. Indeed, by Theorem 2.5 the DFT of an iid Gaussian vector $\vec{\mathbf{z}}$ with zero mean and variance σ^2 is a Gaussian vector with covariance matrix $F_{[N]}\sigma^2 I F_{[N]}^* = N\sigma^2 I$ and mean zero. Similarly, any signal representation obtained using

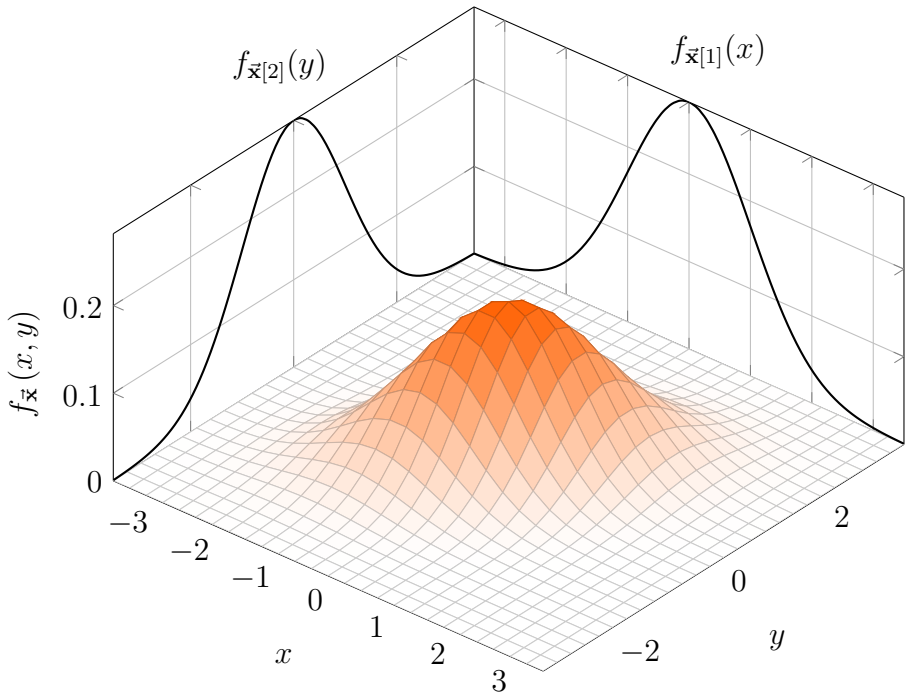

Figure 4: Joint pdf of a two-dimensional Gaussian vector $\vec{\mathbf{x}}$ and marginal pdfs of its two entries.

an orthogonal matrix maps iid Gaussian noise to iid Gaussian noise. We leveraged this fact when designing thresholding-based denoising techniques.

Theorem 2.5 also implies that subvectors of Gaussian vectors are also Gaussian. Figure 4 show the joint pdf of a two-dimensional Gaussian vector together with the marginal pdfs of its entries.

Another consequence of Theorem 2.5 is that standard Gaussian vectors are *isotropic*. The vectors do not favor any direction in their ambient space. No matter how you rotate them, their distribution stays the same. More precisely, for any orthogonal matrix U, if $\vec{\mathbf{x}}$ is a standard Gaussian vector, then by Theorem 2.5 $U\vec{\mathbf{x}}$ has the same distribution, since its mean equals $U\vec{0} = \vec{0}$ and its covariance matrix equals $UIU^T = UU^T = I$. Note that this is a stronger statement than saying that its variance is the same in every direction, which is true for any vector with uncorrelated entries.

2.3 Gaussian random vectors in high dimensions

In the previous section we establish that the direction of standard Gaussian vectors is isotropic. We now consider their magnitude. As we can see in Figure 4, in low dimensions the joint pdf of Gaussian vectors is mostly concentrated around the origin. This is *not* the case as the dimension of the ambient space grows. On the contrary, the norm of the random vector concentrates rapidly around the square root of its dimension, as we observe in the numerical experiment described in Figure 5. The squared ℓ_2-norm of a standard d-dimensional Gaussian vector $\vec{x}$ is the sum of the squares of d independent standard Gaussian random variables. This random quantity is known as a χ^2 random variable.

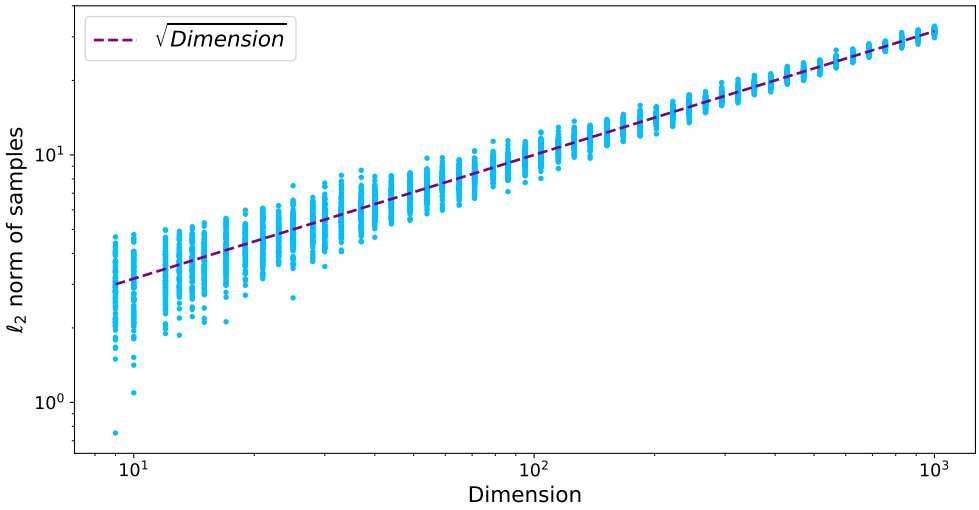

Figure 5: ℓ_2 norms of 100 independent samples from standard Gaussian random vectors in different dimensions. The norms of the samples concentrate around the square root of the dimension.

Definition 2.6 (χ^2 random variable)**.** *A χ^2 (chi squared) random variable with d degrees of freedom is defined as*

$$\mathbf{y} := \sum_{i=1}^{d} \mathbf{x}_i^2 \tag{15}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are standard Gaussians.

Figure 6 shows the pdf of χ^2 random variables for different values of d. As suggested by the numerical experiments, the densities concentrate around $\sqrt{d}$ as d increases.

The mean of the squared ℓ_2-norm of a standard Gaussian vector of dimension d indeed equals $\sqrt{d}$, since

$$\mathrm{E}\left(||\vec{\mathbf{x}}||_2^2\right) = \mathrm{E}\left(\sum_{i=1}^{d} \vec{\mathbf{x}}[i]^2\right) \tag{16}$$

$$= \sum_{i=1}^{d} \mathrm{E}\left(\vec{\mathbf{x}}[i]^2\right) \tag{17}$$

$$= d. \tag{18}$$

The standard deviation determines how much the squared ℓ_2-norm deviates from this value. The following lemma shows that it equals $\sqrt{2d}$.

Lemma 2.7 (Variance of the squared ℓ_2 norm of a standard Gaussian vector)**.** *Let $\vec{\mathbf{x}}$ be a standard Gaussian random vector of dimension d. The variance of $||\vec{\mathbf{x}}||_2^2$ is $2d$.*

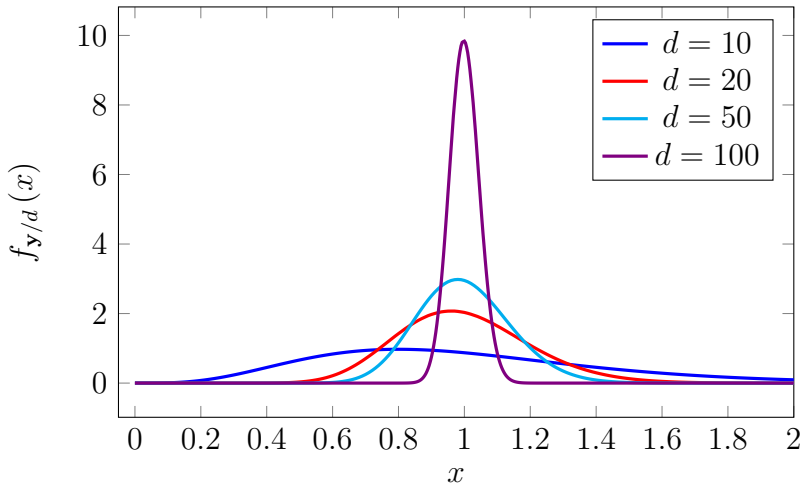

Figure 6: Pdfs of $\mathbf{y}/d$ for different values of d, where $\mathbf{y}$ is a χ^2 random variable with d degrees of freedom.

Proof. Recall that $\mathrm{Var}\left(||\vec{\mathbf{x}}||_2^2\right) = \mathrm{E}\left(\left(||\vec{\mathbf{x}}||_2^2\right)^2\right) - \mathrm{E}\left(||\vec{\mathbf{x}}||_2^2\right)^2$. The result follows from

$$\begin{aligned}
\mathrm{E}\left(\left(||\vec{\mathbf{x}}||_2^2\right)^2\right) &= \mathrm{E}\left(\left(\sum_{i=1}^{d} \vec{\mathbf{x}}[i]^2\right)^2\right) && (19)\\
&= \mathrm{E}\left(\sum_{i=1}^{d}\sum_{j=1}^{d} \vec{\mathbf{x}}[i]^2\vec{\mathbf{x}}[j]^2\right) && (20)\\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{E}\left(\vec{\mathbf{x}}[i]^2\vec{\mathbf{x}}[j]^2\right) && (21)\\
&= \sum_{i=1}^{d} \mathrm{E}\left(\vec{\mathbf{x}}[i]^4\right) + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \mathrm{E}\left(\vec{\mathbf{x}}[i]^2\right)\mathrm{E}\left(\vec{\mathbf{x}}[j]^2\right) && (22)\\
&= 3d + d(d-1) \quad \text{since the 4th moment of a standard Gaussian equals 3} && (23)\\
&= d(d+2). && (24)
\end{aligned}$$

□

The result implies that as d grows, the relative deviation of the squared norm of a standard Gaussian vector from its mean decreases proportionally to $\sqrt{2/d}$. Geometrically, the probability density concentrates close to the surface of a sphere with radius $\sqrt{d}$. For our analysis of randomized techniques in the subsequent section, we need to characterize this phenomenon rigorously, through probabilistic non-asymptotic bounds on the deviation from the mean. The following theorem shows how to obtain such a bound by applying Markov's inequality.

Theorem 2.8 (Chebyshev tail bound for the ℓ_2 norm of a standard Gaussian vector)**.** *Let $\vec{\mathbf{x}}$ be a standard Gaussian random vector of dimension d. For any $\epsilon > 0$ we have*

$$P\left(d\left(1-\epsilon\right) < ||\vec{\mathbf{x}}||_2^2 < d\left(1+\epsilon\right)\right) \geq 1 - \frac{2}{d\epsilon^2}. \tag{25}$$

Proof. The bound is a consequence of Markov's inequality, which quantifies the intuitive idea that if a random variable is nonnegative and small then the probability that it takes large values must be small.

Theorem 2.9 (Markov's inequality, proof in Section 5.1)**.** *Let* $\mathbf{x}$ *be a nonnegative random variable. For any positive constant* $a > 0$*,*

$$\mathrm{P}\left(\mathbf{x} \geq a\right) \leq \frac{\mathrm{E}\left(\mathbf{x}\right)}{a}. \tag{26}$$

Let $\mathbf{y} := ||\vec{\mathbf{x}}||_2^2$,

$$\begin{aligned}
\mathrm{P}\left(|\mathbf{y} - d| \geq d\epsilon\right) &= \mathrm{P}\left(\left(\mathbf{y} - \mathrm{E}\left(\mathbf{y}\right)\right)^2 \geq d^2\epsilon^2\right) && (27)\\
&\leq \frac{\mathrm{E}\left(\left(\mathbf{y} - \mathrm{E}\left(\mathbf{y}\right)\right)^2\right)}{d^2\epsilon^2} \qquad \text{by Markov's inequality} && (28)\\
&= \frac{\mathrm{Var}\left(\mathbf{y}\right)}{d^2\epsilon^2} && (29)\\
&= \frac{2}{d\epsilon^2} \qquad \text{by Lemma 2.7.} && (30)
\end{aligned}$$

When Markov's inequality is applied to bound the deviation from the mean like this, it is usually called Chebyshev's inequality. □

The bound in Theorem 2.8 only relies on the variance. As a result, it is quite weak. The probability of the square deviating by ϵd is inverse proportional to the product $d\epsilon^2$. Unfortunately, this is not sufficient for our purposes. The following theorem provides a better bound by implicitly exploiting the fact that the higher moments of a standard Gaussian are well behaved.

Theorem 2.10 (Chernoff tail bound for the ℓ_2 norm of a standard Gaussian vector)**.** *Let* $\vec{\mathbf{x}}$ *be a standard Gaussian random vector of dimension* d*. For any* $\epsilon \in (0, 1)$ *we have*

$$P\left(d\left(1 - \epsilon\right) < ||\vec{\mathbf{x}}||_2^2 < d\left(1 + \epsilon\right)\right) \geq 1 - 2\exp\left(-\frac{d\epsilon^2}{8}\right). \tag{31}$$

Proof. Let $\mathbf{y} := ||\vec{\mathbf{x}}||_2^2$. The result is implied by

$$P\left(\mathbf{y} > d\left(1 + \epsilon\right)\right) \leq \exp\left(-\frac{d\epsilon^2}{8}\right), \tag{32}$$

$$P\left(\mathbf{y} < d\left(1 - \epsilon\right)\right) \leq \exp\left(-\frac{d\epsilon^2}{8}\right). \tag{33}$$

We present the proof of (32). The proof of (33) is essentially the same and is presented in

Section 5.3. Let $t > 0$ be an arbitrary positive number, and note that

$$
\begin{aligned}
P\left(\mathbf{y} > a\right) &= P\left(\exp\left(t\mathbf{y}\right) > \exp\left(at\right)\right) && (34)\\
&\leq \exp\left(-at\right) \mathrm{E}\left(\exp\left(t\mathbf{y}\right)\right) \qquad \text{by Markov's inequality} && (35)\\
&\leq \exp\left(-at\right) \mathrm{E}\left(\exp\left(\sum_{i=1}^{d} t\mathbf{x_i}^2\right)\right) && (36)\\
&\leq \exp\left(-at\right) \prod_{i=1}^{d} \mathrm{E}\left(\exp\left(t\mathbf{x_i}^2\right)\right) \quad \text{by independence of } \mathbf{x_1}, \ldots, \mathbf{x_d} && (37)\\
&= \frac{\exp\left(-at\right)}{\left(1-2t\right)^{\frac{d}{2}}}, && (38)
\end{aligned}
$$

where the last step is a consequence of the following lemma.

Lemma 2.11 (Proof in Section 5.2)**.** *If* $\mathbf{x}$ *is a standard Gaussian and* $t < 1/2$*,*

$$
\mathrm{E}\left(\exp\left(t\mathbf{x}^2\right)\right) = \frac{1}{\sqrt{1-2t}}. \tag{39}
$$

Note that the lemma implies a bound on the higher-order moments of a standard Gaussian $\mathbf{x}$, since

$$
\begin{aligned}
\mathrm{E}\left(\exp\left(t\mathbf{x}^2\right)\right) &= \mathrm{E}\left(\sum_{i=0}^{\infty} \frac{\left(t\mathbf{x}^2\right)^i}{i!}\right) && (40)\\
&= \sum_{i=0}^{\infty} \frac{\mathrm{E}\left(t^i\left(\mathbf{x}^{2i}\right)\right)}{i!}. && (41)
\end{aligned}
$$

Bounds that exploit the behavior of higher-order moments to control tail probabilities through the expectation of an exponential are often called Chernoff bounds.

We set $a := d\left(1+\epsilon\right)$ and

$$
t := \frac{1}{2} - \frac{1}{2\left(1+\epsilon\right)}, \tag{42}
$$

by minimizing over $t \in (0, 1/2)$ in (38). This gives

$$
\begin{aligned}
P\left(\mathbf{y} > d\left(1+\epsilon\right)\right) &\leq \left(1+\epsilon\right)^{\frac{d}{2}} \exp\left(-\frac{d\epsilon}{2}\right) && (43)\\
&= \exp\left(-\frac{d}{2}\left(\epsilon - \log\left(1+\epsilon\right)\right)\right) && (44)\\
&\leq \exp\left(-\frac{d\epsilon^2}{8}\right), && (45)
\end{aligned}
$$

where the last step follows from the fact that 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 $g\left(0\right) = 0$). □

The dimension must be quite high for these bounds to be meaningful: at least larger than $1/\epsilon^2$.

2.4 Projection onto a fixed subspace

Characterizing the projection of a standard Gaussian on a fixed subspace is an important component of our analysis of randomized dimensionality reduction. In the lecture notes on SVD, it already proved useful in analyzing the training error incurred by linear regression. We have established that the probability density of standard Gaussian errors is isotropic and has variance d. Intuitively, if we project the density onto a subspace of dimension k, we would expect to capture a fraction of the variance equal to k/d, so the projection should have variance equal to k. In this section we establish that this is indeed the case.

Given our definition of Gaussian vector, the projection of a d-dimensional Gaussian vector $\vec{\mathbf{x}}$ onto a subspace of dimension $k < d$ is not a Gaussian vector. Let U be a matrix with columns containing an orthonormal basis of a subspace. We have $\mathcal{P}_{\mathcal{S}}\left(\vec{\mathbf{x}}\right) = UU^T\vec{\mathbf{x}}$. The covariance matrix of the projection equals

$$\Sigma_{\mathcal{P}_{\mathcal{S}}\left(\vec{\mathbf{x}}\right)} = UU^T\Sigma_{\vec{\mathbf{x}}}UU^T \tag{46}$$

$$= UU^T, \tag{47}$$

which is not full rank (its rank equals k). However, the coefficients expressing the projection in terms of the basis vectors of the subspace $U^T\vec{\mathbf{x}}$, are Gaussian. Their covariance equals

$$\Sigma_{U^T\vec{\mathbf{x}}} = U^T\Sigma_{\vec{\mathbf{x}}}U \tag{48}$$

$$= I, \tag{49}$$

so the coefficient vector is a standard Gaussian of dimension k. Since

$$||\mathcal{P}_{\mathcal{S}}\left(\vec{\mathbf{x}}\right)||_2^2 = (UU^T\vec{\mathbf{x}})^T UU^T\vec{\mathbf{x}} \tag{50}$$

$$= \left|\left|U^T\vec{\mathbf{x}}\right|\right|_2^2, \tag{51}$$

applying Theorem 2.10 we confirm that the ℓ_2 norm of the projection concentrates around $\sqrt{k}$.

Corollary 2.12. *Let $\mathcal{S}$ be a k-dimensional subspace of $\mathbb{R}^d$ and $\vec{\mathbf{x}}$ a d-dimensional standard Gaussian vector. For any $\epsilon \in (0, 1)$*

$$\sqrt{k\left(1-\epsilon\right)} \leq ||\mathcal{P}_{\mathcal{S}}\,\vec{\mathbf{x}}||_2 \leq \sqrt{k\left(1+\epsilon\right)} \tag{52}$$

with probability at least $1 - 2\exp\left(-k\epsilon^2/8\right)$.

3 Randomized dimensionality reduction

In this section, we analyze the use of randomized linear maps to achieve dimensionality reduction. The randomized linear map consists of multiplication with a random matrix $\mathbf{A}$ of dimensions $d \times k$, where d is the ambient dimension and k is the reduced dimension. We build the matrix by sampling each entry independently from a standard Gaussian distribution.

Dimensionality-reduction techniques are useful if they preserve the information that we are interested in. In many cases, we would like to conserve the distances between different data points.

This allows us to apply algorithms such as nearest neighbors in the lower-dimensional space. Our goal in this section is to study how likely it is for a $k \times d$ Gaussian random matrix to preserves the distance between a set of points in $\mathbb{R}^d$.

We begin by showing that applying a Gaussian random matrix to a deterministic vector yields a Gaussian random vector.

Lemma 3.1. *Let* $\mathbf{A}$ *be an* $k \times d$ *matrix with iid standard Gaussian entries. If* $\vec{v} \in \mathbb{R}^d$ *is a deterministic vector with unit* ℓ_2 *norm, then* $\mathbf{A}\vec{v}$ *is a* k*-dimensional iid Gaussian vector.*

Proof. By Theorem 2.5, $(\mathbf{A}\vec{v})[i]$, $1 \leq i \leq k$ is Gaussian, since it is the inner product between $\vec{v}$ and the ith row $\mathbf{A}_{i,:}$ (interpreted as a vector in $\mathbb{R}^d$), which is a standard Gaussian vector. The mean of the entry is zero because the mean of $\mathbf{A}_{i,:}$ is zero and the variance equals

$$\begin{aligned}
\text{Var}\left(\mathbf{A}_{i,:}^T \vec{v}\right) &= \vec{v}^T \Sigma_{\mathbf{A}_{i,:}} \vec{v} && (53)\\
&= \vec{v}^T I \vec{v} && (54)\\
&= ||\vec{v}||_2^2 && (55)\\
&= 1, && (56)
\end{aligned}$$

so the entries of $\mathbf{A}\vec{v}$ are all standard Gaussians. They are also independent because each is just a function of a specific row, and all the rows in the matrix are mutually independent. □

A crucial question is to what extent applying the random map affects the norm of the deterministic vector. The following lemma shows that the norm of the deterministic vector is well preserved with high probability if we scale the random matrix by $1/\sqrt{k}$.

Lemma 3.2. *Let* $\mathbf{A}$ *be a* $k \times d$ *matrix with iid standard Gaussian entries. For any* $\vec{v} \in \mathbb{R}^d$ *with unit norm and any* $\epsilon \in (0, 1)$

$$\sqrt{1-\epsilon} \leq \left|\left| \frac{1}{\sqrt{k}} \mathbf{A}\vec{v} \right|\right|_2 \leq \sqrt{1+\epsilon} \tag{57}$$

with probability at least $1 - 2\exp\left(-k\epsilon^2/8\right)$.

Proof. The result follows from Theorem 2.10 and Lemma 3.1. □

The result immediately implies that the random map approximately preserves the distance between two fixed points. If the difference between the vectors equals $\vec{y}$. By the lemma– setting $\vec{v} := \vec{y}/\left|\left|y\right|\right|_2$– the distance between the mapped points satisfies

$$\sqrt{1-\epsilon}\left|\left|y\right|\right|_2 \leq \left|\left| \frac{1}{\sqrt{k}} \mathbf{A}y \right|\right|_2 \leq \sqrt{1+\epsilon}\left|\left|y\right|\right|_2 \tag{58}$$

with high probability, as long as k is sufficiently large. The bounds do not immediately apply to a set of points, as opposed to just two, but we can extend them by leveraging the union bound. This yields the Johnson-Lindenstrauss lemma, which provides a lower bound on the probability of preserving the distances that scales as the inverse of the number of points. As a result, a map achieving small distortion can be found in logarithmic time. The result is striking because the lower bound on k does not depend on the ambient dimension d, and its dependence on the number of points in the data set is only logarithmic. The proof is based on [3].

Lemma 3.3 (Johnson-Lindenstrauss lemma)**.** *Let* $\mathbf{A}$ *be a* $k \times d$ *matrix with iid standard Gaussian entries. Let* $\vec{x}_1, \ldots, \vec{x}_p \in \mathbb{R}^d$ *be any fixed set of* p *deterministic vectors. For any pair* $\vec{x}_i, \vec{x}_j$ *and any* $\epsilon \in (0,1)$

$$(1-\epsilon)\left|\left|\vec{x}_i - \vec{x}_j\right|\right|_2^2 \leq \left|\left|\frac{1}{\sqrt{k}}\mathbf{A}\vec{x}_i - \frac{1}{\sqrt{k}}\mathbf{A}\vec{x}_j\right|\right|_2^2 \leq (1+\epsilon)\left|\left|\vec{x}_i - \vec{x}_j\right|\right|_2^2, \tag{59}$$

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

$$k \geq \frac{16\log(p)}{\epsilon^2}. \tag{60}$$

Proof. To prove the result we control the action of the matrix on the normalized difference of the vectors

$$\vec{v}_{ij} := \frac{\vec{x}_i - \vec{x}_j}{\left|\left|\vec{x}_i - \vec{x}_j\right|\right|_2}, \tag{61}$$

which has unit ℓ_2-norm unless $\vec{x}_i = \vec{x}_j$ (in which case the norm of the difference is preserved exactly). We denote the event that the norm of the action of $\mathbf{A}$ on $\vec{v}_{ij}$ concentrates around k by

$$\mathcal{E}_{ij} = \left\{k(1-\epsilon) < \left|\left|\mathbf{A}\vec{v}_{ij}\right|\right|_2^2 < k(1+\epsilon)\right\} \quad 1 \leq i < p,\ i < j \leq p.$$

Lemma 3.2 implies that each of the $\mathcal{E}_{ij}$ hold with high probability as long as condition (60) holds

$$\mathrm{P}\left(\mathcal{E}_{ij}^c\right) \leq \frac{2}{p^2}. \tag{62}$$

However, this is not enough. Our event of interest is the *intersection* of all the $\mathcal{E}_{ij}$. Unfortunately, the events are dependent (since the vectors are hit by the same matrix), so we cannot just multiply their individual probabilities. Instead, we apply the union bound to control the complement of the intersection.

Theorem 3.4 (Union bound, proof in Section 5.4)**.** *Let* $S_1, S_2, \ldots, S_n$ *be a collection of events in a probability space. Then*

$$\mathrm{P}\left(\cup_i S_i\right) \leq \sum_{i=1}^{n} \mathrm{P}\left(S_i\right). \tag{63}$$

The number of events in the intersection is $\binom{p}{2} = p(p-1)/2$, because that is the number of different pairs of vectors in the set $\{\vec{x}_1, \ldots, \vec{x}_p\}$. The union bound yields

$$\begin{aligned}
\mathrm{P}\left(\bigcap_{i,j}\mathcal{E}_{ij}\right) &= 1 - \mathrm{P}\left(\bigcup_{i,j}\mathcal{E}_{ij}^c\right) && (64)\\
&\geq 1 - \sum_{i,j}\mathrm{P}\left(\mathcal{E}_{ij}^c\right) && (65)\\
&\geq 1 - \frac{p(p-1)}{2}\frac{2}{p^2} && (66)\\
&\geq \frac{1}{p}. && (67)
\end{aligned}$$

Note that the application of the union bound is the reason we need an exponential bound in Theorem 2.10. □

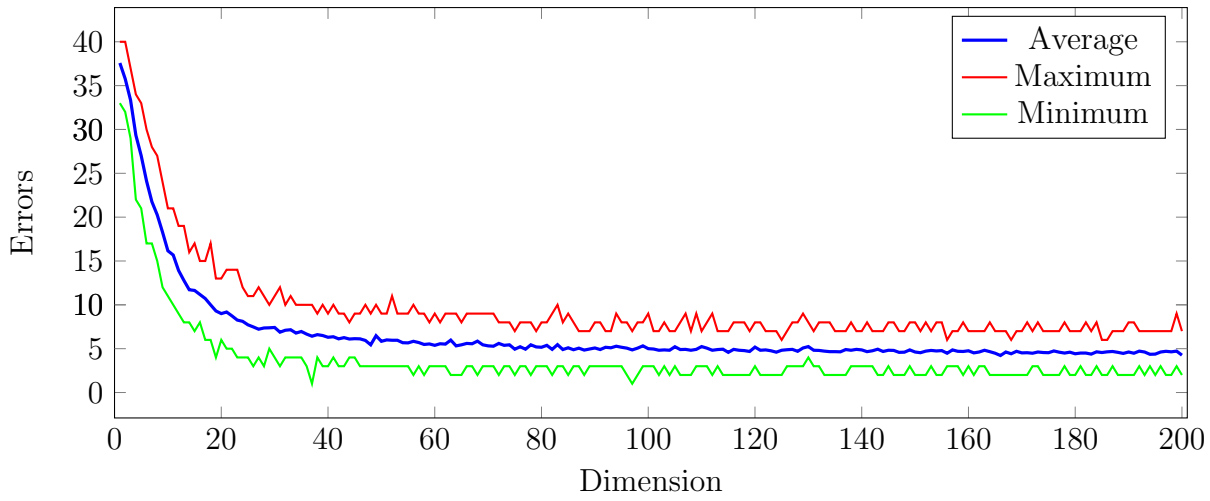

Figure 7: Average, maximum and minimum number of errors (over 50 tries) for nearest-neighbor classification after a randomized dimensionality reduction for different dimensions.

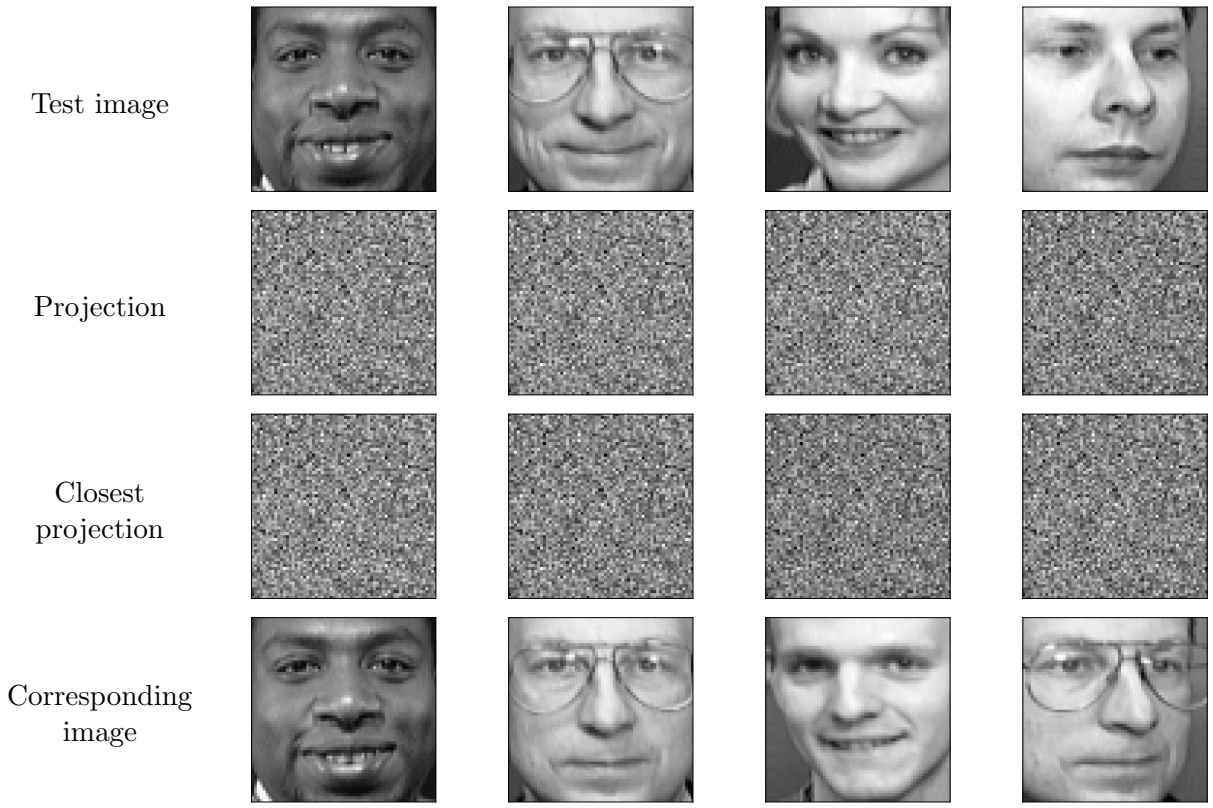

Figure 8: Results of nearest-neighbor classification combined with randomized dimensionality reduction of dimension 50 for four of the people in Example 3.5. The assignments of the first two examples are correct, but the other two are wrong.

Example 3.5 (Nearest neighbors after random projection)**.** The nearest-neighbors algorithm performs classification based on distances between feature vectors. If the training set contains n examples, the method requires computing n distances in an d-dimensional space (where d is the number of features) to classify each new example. The computational cost is $\mathcal{O}(nd)$, so if we need to classify p points the total cost is $\mathcal{O}(ndp)$. If we perform a random projection of each of the points onto a lower-dimensional space k before classifying them, then the computational cost is:

- kdn operations to project the training data using a $k \times d$ iid standard Gaussian matrix.
- kdp operations to project each point in the test set using the same matrix.
- knp to perform nearest-neighbor classification in the lower-dimensional space.

The overall cost is $\mathcal{O}(kp \max\{d, n\})$, which is a significant reduction from $\mathcal{O}(ndp)$. It is also more efficient than the PCA-based approach described in the lecture notes on the SVD, which includes an additional $\mathcal{O}(dn \min\{d, n\})$ step to compute the SVD.

Figure 7 shows the accuracy of the algorithm on the same data as Example 4.3 in Lecture Notes 1. At dimension $k = 50$ we achieve a similar average precision as in the ambient dimension (5 errors out of 40 test images compared to 4 out of 40). Figure 8 shows some examples of the projected data represented in the original d-dimensional space along with their nearest neighbors in the k-dimensional space. △

4 Compressed sensing

The goal of compressed sensing is to recover signals from a small number of linear measurements. If the signals are modeled as vectors of dimension d, it is of course impossible to recover an arbitrary signal from less than d measurements. However, signals are often highly structured. For example, images are approximately sparse when represented using wavelets, as described in the lecture notes on signal representations. If we are interested in a class of signals that can be represented with only $s < d$ parameters, it seems plausible that recovery could be possible from less than d measurements. In order to perform a mathematical analysis of compressed sensing, we consider the problem of estimating sparse vectors from underdetermined linear vectors. This is a highly simplified scenario, but it provides valuable insights about compressed sensing in general.

4.1 The restricted-isometry property

Compressed sensing is impossible to achieve using regularly undersampled frequency measurements. The reason is *aliasing*, as we discussed in the lecture notes on the frequency domain. The bottom two rows Figure 9 illustrate this with a simple example where two different sparse vectors produce exactly the same measurements. The figure also shows that this does not occur for randomized measurements. In fact, the random data seem to preserve the structure of the sparse vector quite well, which is consistent with the experiment in Figure 2. A map that preserves the norm of every vector in a given set is called a *restricted isometry*, because it is almost an isometry

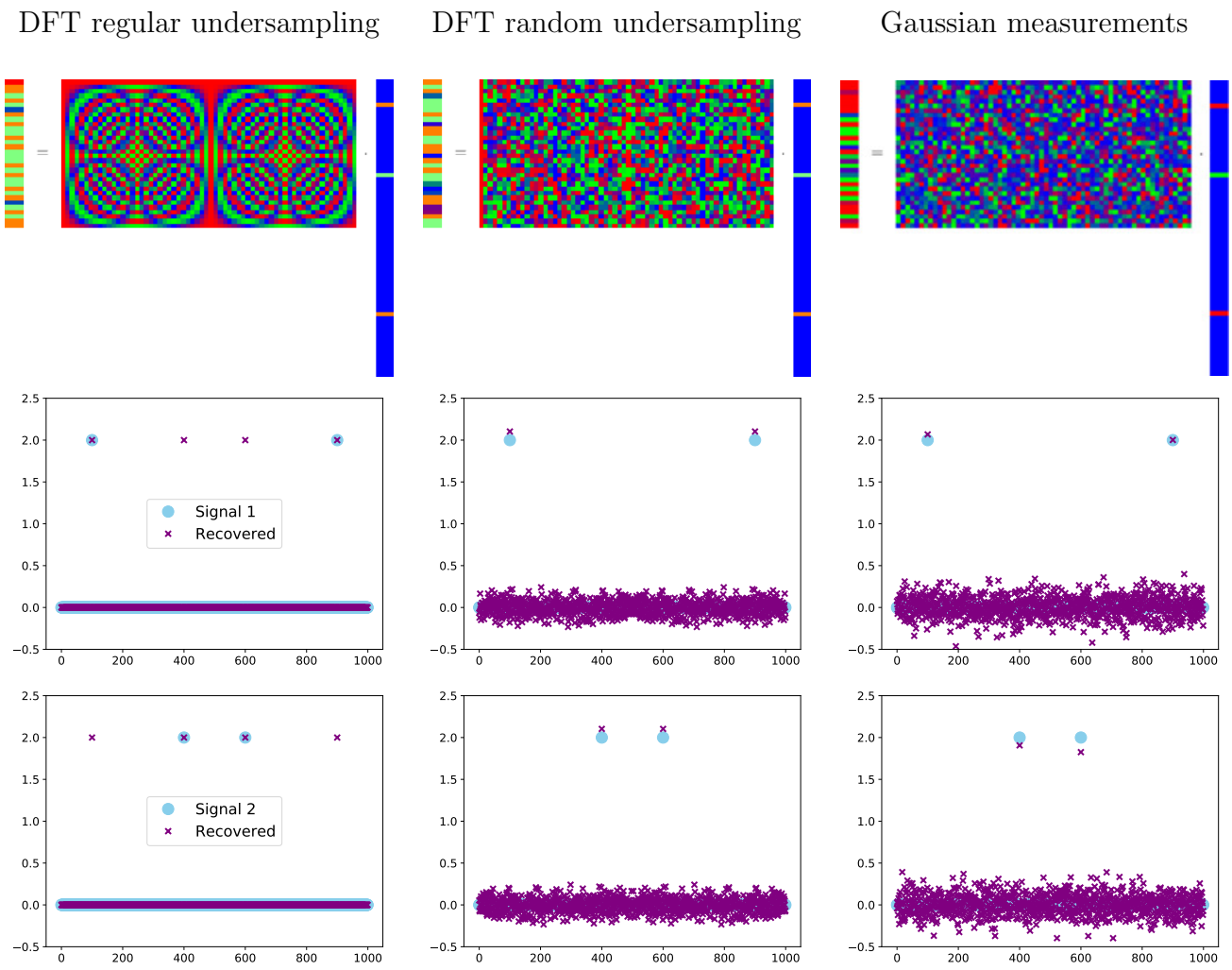

Figure 9: The top row illustrates different linear operators that can be used to obtain compressed measurements of a sparse vector: a regularly-undersampled DFT matrix, a randomly-undersampled DFT matrix and a Gaussian matrix (we only show the real-part of the DFT submatrices). The second and third rows show the result of computing $A^*A\vec{x}$ for two different 2-sparse signals, where A is the linear measurement operator. Due to aliasing, the regularly-undersampled Fourier measurements are the same for the two sparse signals.

when restricted to act upon the elements of the set. Here we fix the set to be vectors with a fixed sparsity level.

Definition 4.1 (Restricted-isometry property)**.** *An $m \times d$ matrix A, where $m < d$, satisfies the restricted-isometry property with constant ϵ if for any s-sparse d-dimensional vector $\vec{x}$*

$$(1-\epsilon)\,||\vec{x}||_2 \leq ||A\vec{x}||_2 \leq (1+\epsilon)\,||\vec{x}||_2\,. \tag{68}$$

If a matrix A satisfies the restricted-isometry property (RIP) for a sparsity level $2s$ then for any pair of vectors $\vec{x}_1$ and $\vec{x}_2$ with sparsity level s, the distance between their corresponding measurements $\vec{y}_1$ and $\vec{y}_2$ is lower bounded by the difference between the two vectors

$$\begin{aligned} ||\vec{y}_2 - \vec{y}_1||_2 &= ||A\,(\vec{x}_1 - \vec{x}_2)||_2 &(69)\\ &\geq (1-\epsilon)\,||\vec{x}_2 - \vec{x}_1||_2\,. &(70)\end{aligned}$$

In particular, this means that the problem of recovering sparse signals from such measurements is well posed in the following sense: if the data are generated from an s-sparse signal, there cannot be another s-sparse signal that produces similar data. Unfortunately, verifying that a matrix satisfies the restricted-isometry property is not computationally tractable (essentially, one has to check all possible sparse submatrices). However, we can show that the RIP holds with high probability for random matrices. The following theorem establishes this for Gaussian iid matrices. The proof for random Fourier measurements is more complicated [2, 4].

Theorem 4.2 (Restricted-isometry property for Gaussian matrices)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times d}$ be a random matrix with iid standard Gaussian entries. $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies the restricted-isometry property for a constant ϵ with probability $1 - \frac{C_2}{d}$ as long as the number of measurements*

$$m \geq \frac{C_1 s}{\epsilon^2} \log\left(\frac{d}{s}\right) \tag{71}$$

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

If we ignore logarithmic factors, the theorem establishes that for the RIP to hold, we need a number of random measurements that is proportional to the sparsity, and not to the ambient dimension!

4.2 Proof of Theorem 4.2

If we fix the nonzero entries of the sparse signal, then the RIP reduces to a statement about the smallest and largest singular values of a fixed submatrix of the Gaussian measurement matrix.

Lemma 4.3. *Let $T \subset \{1, \ldots, d\}$ be a set of s indices. Any matrix $A \in \mathbb{R}^{m\times d}$, $m < d$, satisfies*

$$\sigma_s(A_T) \leq ||A\vec{x}||_2 \leq \sigma_1(A_T) \tag{72}$$

for all vectors $\vec{x} \in \mathbb{R}^d$ with support restricted to T. A_T is the $m \times s$ submatrix of A that contains the columns indexed by T; $\sigma_1(A_T)$ and $\sigma_s(A_T)$ are its largest and smallest singular value respectively. In addition, there exist vectors $\vec{v}_1$ and $\vec{v}_s$ such that the upper and lower bound respectively are tight.

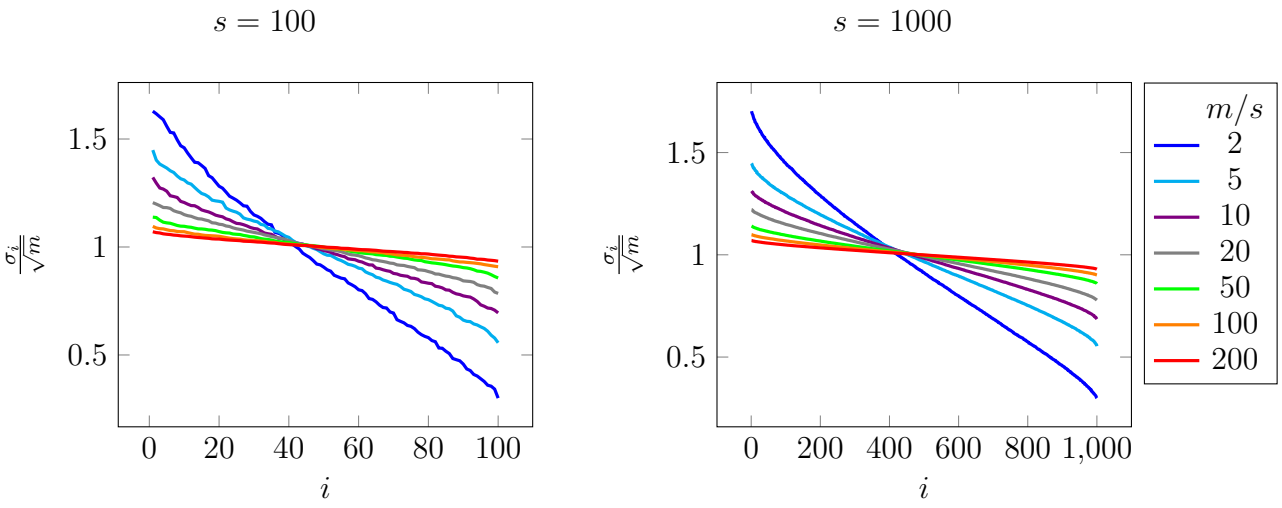

Figure 10: Singular values of $m \times s$ matrices with iid standard Gaussian entries for different values of m and s.

Proof. For any vector $\vec{x} \in \mathbb{R}^d$ with support restricted to T, $A\vec{x} = A_T \vec{x}_T$, where $\vec{x}_T \in \mathbb{R}^s$ is the subvector of $\vec{x}$ that contains its nonzero entries. The result follows immediately from Theorem 2.4 in the notes on the SVD. The vectors that make the bounds tight are the first and last right singular vectors of A_T. □

Our strategy will be to control the singular values of a fixed submatrix, and then apply the union bound to extend the bounds to all possible submatrices.

We need to analyze the singular values of tall matrices of dimension $m \times s$ with iid standard Gaussian entries. Numerically, we observe that if we keep s fixed and increase m, all s singular values converge to $\sqrt{m}$, as shown in Figure 10. This implies that the matrix is approximately equal to $\sqrt{m}\, UV^T$ for two orthogonal matrices U and V. It is therefore an approximately orthogonal matrix if we scale it by $1/\sqrt{m}$. Geometrically, if we generate a fixed number of standard Gaussian vectors at increasing ambient dimensions, the vectors will tend to be almost orthogonal with high probability as the dimension grows.

The following result establishes a non-asymptotic bound on the singular values using a covering number argument from [1] that can be applied to other distributions and situations. See also [5] for some excellent notes on high-dimensional probability techniques in this spirit. We defer the proof to Section 4.3.

Theorem 4.4 (Singular values of a Gaussian matrix)**.** *Let* $\mathbf{M}$ *be a* $m \times s$ *matrix with iid standard Gaussian entries such that* $m > s$*. For any fixed* $\epsilon > 0$*, the singular values of* $\mathbf{M}$ *satisfy*

$$\sqrt{m}\,(1-\epsilon) \leq \boldsymbol{\sigma_s} \leq \boldsymbol{\sigma_1} \leq \sqrt{m}\,(1+\epsilon) \tag{73}$$

with probability at least $1 - 2\left(\frac{12}{\epsilon}\right)^s \exp\left(-\frac{m\epsilon^2}{32}\right)$.

Setting $\mathbf{M} := \mathbf{A}_T$, where $\mathbf{A}_T$ is a fixed $m \times s$ submatrix of $\mathbf{A}$, the result implies that the smallest and largest singular values $\boldsymbol{\sigma}_s$ and $\boldsymbol{\sigma}_m$ of $\mathbf{A}_T$ satisfy

$$\sqrt{m}\,(1-\epsilon) \leq \boldsymbol{\sigma_s} \leq \boldsymbol{\sigma_1} \leq \sqrt{m}\,(1+\epsilon)\,. \tag{74}$$

As a result, for any vector $\vec{x}$ with support T

$$\sqrt{1-\epsilon}\,||\vec{x}||_2 \leq \frac{1}{\sqrt{m}}\,||\mathbf{A}\vec{x}||_2 \leq \sqrt{1+\epsilon}\,||\vec{x}||_2\,. \tag{75}$$

This is not enough for our purposes, we need this to hold for *all* supports of size s, i.e. on all possible combinations of s columns selected from the d columns in $\mathbf{A}$. A simple bound on the binomial coefficient yields the following bound on the number of different supports of size s

$$\binom{d}{s} \leq \left(\frac{ed}{s}\right)^s. \tag{76}$$

By the union bound (Theorem 3.4), we consequently have that the bounds (75) hold for any s-sparse vector with probability at least

$$\begin{aligned} 1 - 2\left(\frac{ed}{s}\right)^s \left(\frac{12}{\epsilon}\right)^s \exp\left(-\frac{m\epsilon^2}{32}\right) &= 1 - \exp\left(\log 2 + s + s\log\left(\frac{d}{s}\right) + s\log\left(\frac{12}{\epsilon}\right) - \frac{m\epsilon^2}{2}\right) \\ &\leq 1 - \frac{C_2}{d} \end{aligned} \tag{77}$$

for some constant C_2 as long as m satisfies (71).

4.3 Proof of Theorem 4.4

To establish the bounds on the singular values, we need to show that for any vector with unit ℓ_2 norm

$$\sqrt{m}\,(1-\epsilon) < ||\mathbf{M}\vec{v}||_2 < \sqrt{m}\,(1+\epsilon)\,. \tag{78}$$

This is reminiscent of the proof of the Johnson-Lindenstrauss lemma. Can we prove this for a fixed vector and use the union bound to extend the result to all unit-norm vectors? Unfortunately the answer is no. The set of all unit-ℓ_2-norm vectors in $\mathbb{R}^s$, which is usually referred to as the s-dimensional sphere $\mathcal{S}^{s-1}$, has *infinite* cardinality, so the union bound cannot help us. Instead, we apply a more sophisticated strategy:

- First, we show that the bounds hold for a finite subset of $\mathcal{S}^{s-1}$, called an ϵ-net, which *covers* the sphere, in the sense that all the points in $\mathcal{S}^{s-1}$ are close to at least one of the elements of the set.
- Second, we show that the bounds can be extended to any point that is close enough to one of the points of the ϵ-net, which completes the proof.

Bounds on the ϵ-net

We begin by defining ϵ-nets. Figure 11 shows an ϵ-net for the two-dimensional sphere $\mathcal{S}^1$.

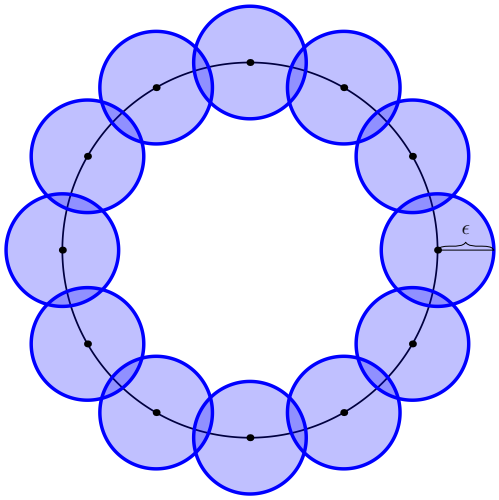

Figure 11: ϵ-net for the two-dimensional sphere $\mathcal{S}^1$, which is just a circle.

Definition 4.5 (ϵ-net)**.** *An ϵ-net of a set $\mathcal{X} \subseteq \mathbb{R}^s$ is a subset $\mathcal{N}_\epsilon \subseteq \mathcal{X}$ such that for every vector $\vec{x} \in \mathcal{X}$ there exists $\vec{y} \in \mathcal{N}_\epsilon$ for which*

$$||\vec{x} - \vec{y}||_2 \leq \epsilon. \tag{79}$$

The smallest possible number of points in the ϵ-net of a set is called its covering number.

Definition 4.6 (Covering number)**.** *The covering number $\mathcal{N}(\mathcal{X}, \epsilon)$ of a set $\mathcal{X}$ at scale ϵ is the minimal cardinality of an ϵ-net of $\mathcal{X}$, or equivalently the minimal number of balls of radius ϵ with centers in $\mathcal{X}$ required to cover $\mathcal{X}$.*

The following theorem, proved in Section 5.5, provides a bound for the covering number of the k-dimensional sphere $\mathcal{S}^{s-1}$.

Theorem 4.7 (Covering number of a sphere)**.** *The covering number of the s-dimensional sphere $\mathcal{S}^{s-1}$ at scale ϵ satisfies*

$$\mathcal{N}\left(\mathcal{S}^{s-1}, \epsilon\right) \leq \left(\frac{2+\epsilon}{\epsilon}\right)^s \leq \left(\frac{3}{\epsilon}\right)^s. \tag{80}$$

Let $\epsilon_1 := \epsilon/4$ and $\epsilon_2 := \epsilon/2$. Consider an ϵ_1-net $\mathcal{N}_{\epsilon_1}$ of $\mathcal{S}^{s-1}$. We define the event

$$\mathcal{E}_{\vec{v},\epsilon_2} := \left\{m\left(1-\epsilon_2\right)||\vec{v}||_2^2 \leq ||\mathbf{M}\vec{v}||_2^2 \leq m\left(1+\epsilon_2\right)||\vec{v}||_2^2\right\}. \tag{81}$$

By Lemma 3.2 for any fixed $\vec{v} \in \mathbb{R}^s$ $\mathrm{P}\left(\mathcal{E}_{\vec{v},\epsilon_2}^c\right) \leq 2\exp\left(-m\epsilon^2/32\right)$, so by the union bound

$$\begin{aligned}
\mathrm{P}\left(\cup_{\vec{v}\in\mathcal{N}_{\epsilon_1}} \mathcal{E}_{\vec{v},\epsilon_2}^c\right) &\leq \sum_{\vec{v}\in\mathcal{N}_{\epsilon_1}} \mathrm{P}\left(\mathcal{E}_{\vec{v},\epsilon_2}^c\right) && (82)\\
&\leq |\mathcal{N}_{\epsilon_1}|\,\mathrm{P}\left(\mathcal{E}_{\vec{v},\epsilon_2}^c\right) && (83)\\
&\leq 2\left(\frac{12}{\epsilon}\right)^s \exp\left(-\frac{m\epsilon^2}{32}\right). && (84)
\end{aligned}$$

Extension of the bounds to the rest of the sphere

Now, to finish the proof we need to show that if $\cup_{\vec{v}\in\mathcal{N}_{\epsilon_1}}\mathcal{E}^c_{\vec{v},\epsilon_2}$ holds then the bound holds for every element in $\mathcal{S}^{s-1}$. Our main tools are the triangle inequality and Theorem 2.4 in the notes on the SVD, which states that

$$\sigma_1 = \max_{||\vec{y}||_2=1} ||\mathbf{M}\vec{y}||_2 \,, \tag{85}$$

$$\sigma_s = \min_{||\vec{y}||_2=1} ||\mathbf{M}\vec{y}||_2 \,. \tag{86}$$

For any arbitrary vector $\vec{x} \in \mathcal{S}^{s-1}$ on the sphere there exists a vector in the $\epsilon/4$-covering set $\vec{v} \in \mathcal{N}(\mathcal{X}, \epsilon_1)$ such that $||\vec{x} - \vec{v}||_2 \leq \epsilon/4$. By the triangle inequality this implies

$$||\mathbf{M}\vec{x}||_2 \leq ||\mathbf{M}\vec{v}||_2 + ||\mathbf{M}(\vec{x} - \vec{v})||_2 \tag{87}$$

$$\leq \sqrt{m}\left(1 + \frac{\epsilon}{2}\right) + ||\mathbf{M}(\vec{x} - \vec{v})||_2 \qquad \text{assuming } \cup_{\vec{v}\in\mathcal{N}_{\epsilon_1}}\mathcal{E}^c_{\vec{v},\epsilon_2} \text{ holds} \tag{88}$$

$$\leq \sqrt{m}\left(1 + \frac{\epsilon}{2}\right) + \boldsymbol{\sigma_1}||\vec{x} - \vec{v}||_2 \qquad \text{by (85)} \tag{89}$$

$$\leq \sqrt{m}\left(1 + \frac{\epsilon}{2}\right) + \frac{\boldsymbol{\sigma_1}\epsilon}{4}. \tag{90}$$

By (85) $\boldsymbol{\sigma_1}$ is the smallest upper bound on $||\mathbf{M}\vec{x}||_2$ for all $\vec{x}$ on the sphere, so the bound in equation (90) cannot be smaller, i.e.

$$\boldsymbol{\sigma_1} \leq \sqrt{m}\left(1 + \frac{\epsilon}{2}\right) + \frac{\boldsymbol{\sigma_1}\epsilon}{4}, \tag{91}$$

which implies

$$\boldsymbol{\sigma_1} \leq \sqrt{m}\left(\frac{1 + \epsilon/2}{1 - \epsilon/4}\right) \tag{92}$$

$$= \sqrt{m}\left(1 + \epsilon - \frac{\epsilon(1-\epsilon)}{4-\epsilon}\right) \tag{93}$$

$$\leq \sqrt{m}(1 + \epsilon)\,. \tag{94}$$

The lower bound on $\boldsymbol{\sigma_s}$ follows from a similar argument combined with (94). By the triangle inequality

$$||\mathbf{M}\vec{x}||_2 \geq ||\mathbf{M}\vec{v}||_2 - ||\mathbf{M}(\vec{x} - \vec{v})||_2 \tag{95}$$

$$\geq \sqrt{m}\left(1 - \frac{\epsilon}{2}\right) - ||A(\vec{x} - \vec{v})||_2 \qquad \text{assuming } \cup_{\vec{v}\in\mathcal{N}_{\epsilon_1}}\mathcal{E}^c_{\vec{v},\epsilon_2} \text{ holds} \tag{96}$$

$$\geq \sqrt{m}\left(1 - \frac{\epsilon}{2}\right) - \boldsymbol{\sigma_1}||\vec{x} - \vec{v}||_2 \qquad \text{by (85)} \tag{97}$$

$$\geq \sqrt{m}\left(1 - \frac{\epsilon}{2}\right) - \frac{\epsilon}{4}\sqrt{m}(1 + \epsilon) \qquad \text{by (94)} \tag{98}$$

$$= \sqrt{m}(1 - \epsilon)\,. \tag{99}$$

By (86) $\boldsymbol{\sigma_s}$ is the largest lower bound on $||\mathbf{M}\vec{x}||_2$ for all $\vec{x}$ on the sphere, so $\boldsymbol{\sigma_s} \geq \sqrt{m}(1 - \epsilon)$ as long as $\cup_{\vec{v}\in\mathcal{N}_{\epsilon_1}}\mathcal{E}^c_{\vec{v},\epsilon_2}$ holds.

5 Proofs

5.1 Proof of Theorem 2.9

Consider the indicator variable $1_{\mathbf{x} \geq a}$. We have

$$\mathbf{x} - a\, 1_{\mathbf{x} \geq a} \geq 0. \tag{100}$$

In particular its expectation is nonnegative (as it is the sum or integral of a nonnegative quantity over the positive real line). By linearity of expectation and the fact that $1_{\mathbf{x} \geq a}$ is a Bernoulli random variable with expectation $\mathrm{P}(\mathbf{x} \geq a)$ we have

$$\mathrm{E}(\mathbf{x}) \geq a\, \mathrm{E}(1_{\mathbf{x} \geq a}) = a\, \mathrm{P}(\mathbf{x} \geq a). \tag{101}$$

5.2 Proof of Lemma 2.11

$$\mathrm{E}\left(\exp\left(t\mathbf{x}^2\right)\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) \exp\left(tu^2\right) \mathrm{d}u \tag{102}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(1-2t)\, u^2}{2}\right) \mathrm{d}u \qquad \text{finite for } 1 - 2t > 0 \tag{103}$$

$$= \frac{1}{\sqrt{2\pi\,(1-2t)}} \int_{-\infty}^{\infty} \exp\left(-\frac{v^2}{2}\right) \mathrm{d}v \qquad \text{change of variables } v = \sqrt{1-2t}u$$

$$= \frac{1}{\sqrt{1-2t}}. \tag{104}$$

5.3 Proof of (33)

A very similar argument to the one that yields (37) gives

$$P(\mathbf{y} < a') = P\left(\exp\left(-t'\mathbf{y}\right) > \exp\left(-a't'\right)\right) \tag{105}$$

$$\leq \exp\left(a't'\right) \prod_{i=1}^{d} \mathrm{E}\left(\exp\left(-t'\mathbf{x_i}^2\right)\right). \tag{106}$$

Setting $t' = t$ in (39), we have

$$\mathrm{E}\left(\exp\left(-t'\mathbf{x}^2\right)\right) = \frac{1}{\sqrt{1+2t'}}. \tag{107}$$

This implies

$$P(\mathbf{y} < a') \leq \frac{\exp\left(a't'\right)}{(1+2t')^{\frac{d}{2}}}. \tag{108}$$

Setting

$$t' := -\frac{1}{2} + \frac{1}{2(1-\epsilon)}, \tag{109}$$

$$a' := d(1-\epsilon) \tag{110}$$

we have

$$P(\mathbf{y} < d(1-\epsilon)) \leq (1-\epsilon)^{\frac{d}{2}} \exp\left(\frac{d\epsilon}{2}\right) \tag{111}$$

$$= \exp\left(-\frac{d}{2}(-\epsilon - \log(1-\epsilon))\right). \tag{112}$$

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

$$P(\mathbf{y} < d(1-\epsilon)) \leq \exp\left(-\frac{d\epsilon^2}{2}\right) \tag{113}$$

$$\leq \exp\left(-\frac{d\epsilon^2}{8}\right). \tag{114}$$

5.4 Proof of Theorem 3.4

Let us define the sets:

$$\tilde{S}_i = S_i \cap \cap_{j=1}^{i-1} S_j^c. \tag{115}$$

It is straightforward to show by induction that $\cup_{j=1}^{n} S_j = \cup_{j=1}^{n} \tilde{S}_j$ for any n, so $\cup_i S_i = \cup_i \tilde{S}_i$. The sets $\tilde{S}_1$, $\tilde{S}_2$, ... are disjoint by construction, so

$$\mathrm{P}(\cup_i S_i) = \mathrm{P}\left(\cup_i \tilde{S}_i\right) = \sum_i \mathrm{P}\left(\tilde{S}_i\right) \tag{116}$$

$$\leq \sum_i \mathrm{P}(S_i) \quad \text{because } \tilde{S}_i \subseteq S_i. \tag{117}$$

5.5 Proof of Theorem 4.7

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

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

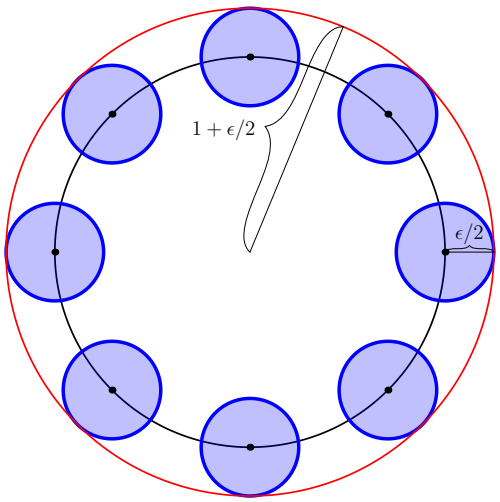

Figure 12: Sketch of the proof of Theorem 4.7 in two dimensions. $\mathcal{B}^s_{1+\epsilon/2}\left(\vec{0}\right)$ is the big red circle. The smaller shaded circles correspond to $\mathcal{B}^s_{\epsilon/2}\left(\vec{x}\right)$ for each $\vec{x}$ in the ϵ-net.

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 ϵ apart and they are all inside the ball of radius $1+\epsilon/2$ centered at the origin $\vec{0}$ because $\mathcal{N}_\epsilon \subseteq \mathcal{S}^{s-1}$. This means that

$$\text{Vol}\left(\mathcal{B}^k_{1+\epsilon/2}\left(\vec{0}\right)\right) \geq \text{Vol}\left(\cup_{\vec{x}\in\mathcal{N}_\epsilon}\mathcal{B}^k_{\epsilon/2}\left(\vec{x}\right)\right) \tag{118}$$

$$= |\mathcal{N}_\epsilon|\,\text{Vol}\left(\mathcal{B}^k_{\epsilon/2}\left(\vec{0}\right)\right) \tag{119}$$

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

$$\text{Vol}\left(\mathcal{B}^k_r\left(\vec{0}\right)\right) = r^k\,\text{Vol}\left(\mathcal{B}^k_1\left(\vec{0}\right)\right), \tag{120}$$

so (118) implies

$$(1+\epsilon/2)^k \geq |\mathcal{N}_\epsilon|\,(\epsilon/2)^k. \tag{121}$$

References

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

[2] 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.

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

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

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