Lecture Notes 3: Randomness

1 Gaussian random variables

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 (see Theorem 1.3 below). 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 1 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 1.1 (Gaussian)**.** *The pdf of a Gaussian or normal random variable with mean μ and standard deviation σ is given by*

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

A Gaussian distribution with mean μ and standard deviation σ is usually denoted by $\mathcal{N}(\mu, \sigma^2)$.

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

Lemma 1.2. *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

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

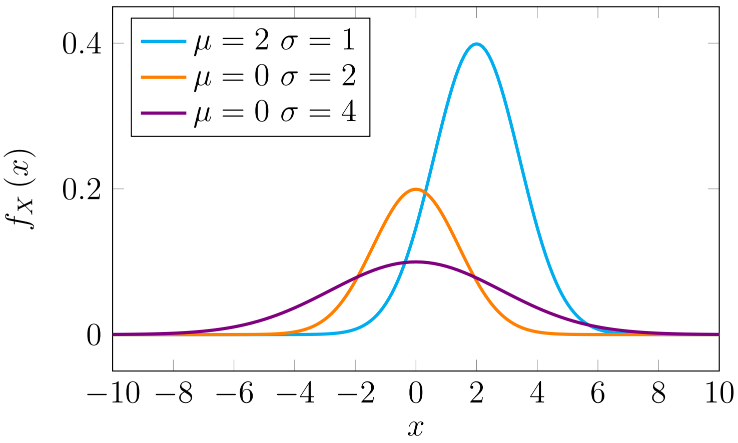

Figure 1: Gaussian random variable with different means and standard deviations.

Differentiating with respect to y yields

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

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

The distribution of the average of a large number of random variables with bounded variances converges to a Gaussian distribution.

Theorem 1.3 (Central limit theorem)**.** *Let $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, ... be a sequence of iid random variables with mean μ and bounded variance σ^2. We define the sequence of averages $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, ..., as*

$$\mathbf{a}_i := \frac{1}{i}\sum_{j=1}^{i} \mathbf{x}_j. \tag{9}$$

The sequence $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, ...

$$\mathbf{b}_i := \sqrt{i}(\mathbf{a}_i - \mu) \tag{10}$$

converges in distribution to a Gaussian random variable with mean 0 and variance σ^2, meaning that for any $x \in \mathbb{R}$

$$\lim_{i\to\infty} f_{\mathbf{b}_i}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}. \tag{11}$$

For large i the theorem suggests that the average $\mathbf{a}_i$ is approximately Gaussian with mean μ and variance $\sigma/\sqrt{n}$. This is verified numerically in Figure 2. Figure 3 shows the histogram of the heights in a population of 25,000 people and how it is very well approximated by a Gaussian random variable[1], suggesting that a person's height may result from a combination of independent factors (genes, nutrition, etc.).

[1] The data is available here.

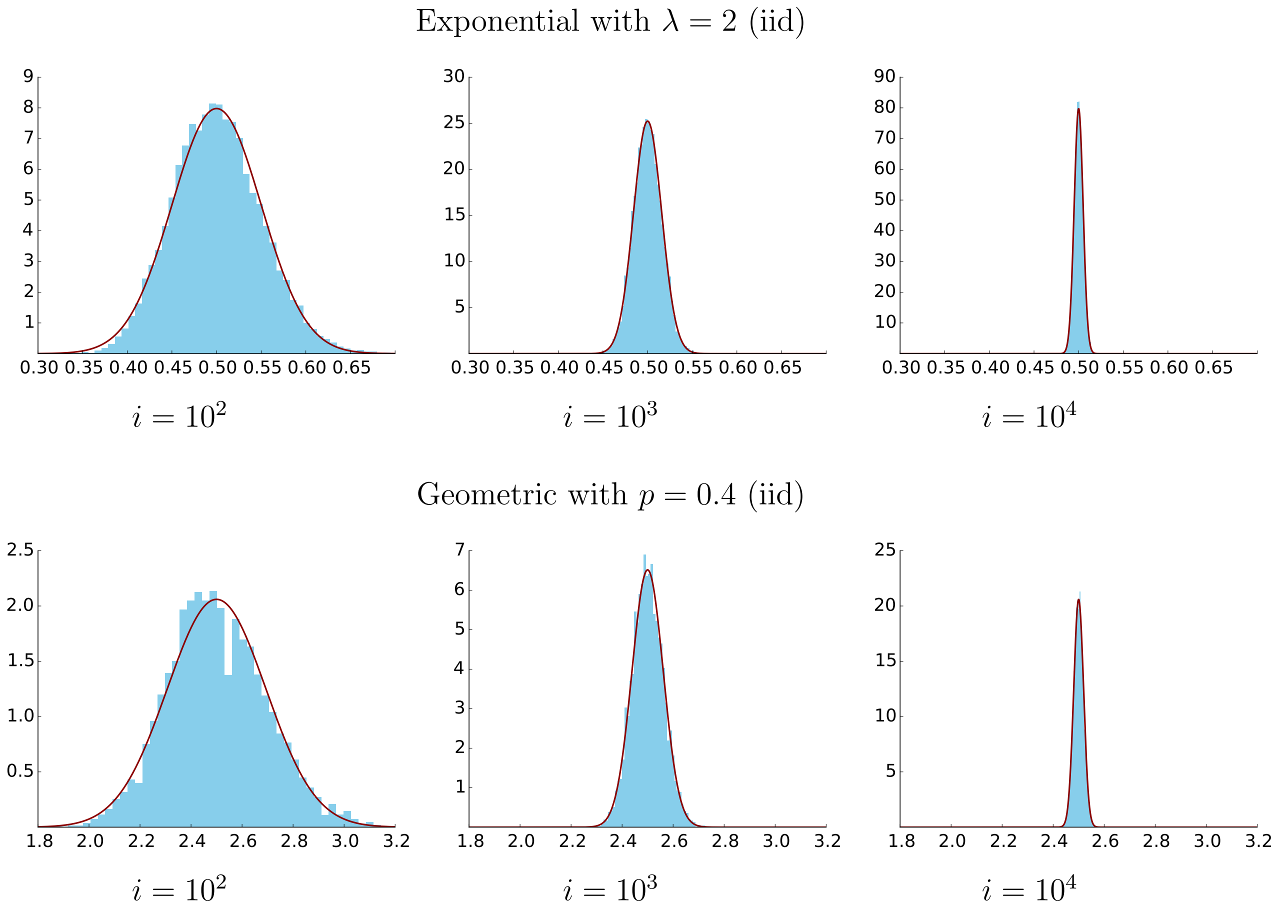

Figure 2: Empirical distribution of the average, defined as in equation (9), of an iid exponential sequence with parameter $\lambda = 2$ (top) and an iid geometric sequence with parameter $p = 0.4$ (bottom). The empirical distribution is computed from 10^4 samples in all cases. The estimate provided by the central limit theorem is plotted in red.

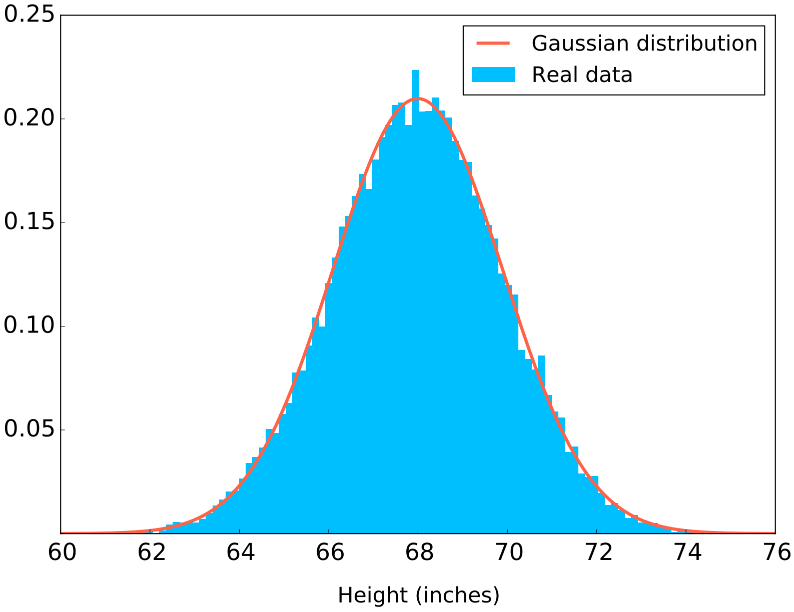

Figure 3: Histogram of heights in a population of 25,000 people (blue) and its approximation using a Gaussian distribution (orange).

2 Gaussian random vectors

2.1 Definition and basic properties

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 2.1 (Gaussian random vector)**.** *A Gaussian random vector $\vec{\mathbf{x}}$ is a random vector with joint pdf ($|\Sigma|$ denotes the determinant of Σ)*

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

where the mean vector $\vec{\mu} \in \mathbb{R}^n$ and the covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, which is symmetric and positive definite, parametrize the distribution. A Gaussian distribution with mean $\vec{\mu}$ and covariance matrix Σ is usually denoted by $\mathcal{N}\left(\vec{\mu}, \Sigma\right)$.

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

Lemma 2.2 (Uncorrelation implies mutual independence for Gaussian random vectors)**.** *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

$$\Sigma_{\vec{\mathbf{x}}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}, \tag{13}$$

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

$$\Sigma_{\vec{\mathbf{x}}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}, \tag{14}$$

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

$$f_{\vec{\mathbf{x}}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right) \tag{15}$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)}\sigma_i} \exp\left(-\frac{(\vec{x}_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{16}$$

$$= \prod_{i=1}^n f_{\vec{\mathbf{x}}_i}(\vec{x}_i). \tag{17}$$

Since the joint pdf factors into a 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 iid standard Gaussians with mean zero and unit variance. We refer to such vectors as iid 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 1.2. We omit the proof, which is similar to that of Lemma 1.2.

Theorem 2.3 (Linear transformations of Gaussian random vectors are Gaussian)**.** *Let $\vec{\mathbf{x}}$ be a Gaussian random vector of dimension n with mean $\vec{\mu}$ and covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$, $\vec{Y} = A\vec{\mathbf{x}} + \vec{b}$ is a Gaussian random vector with mean $A\vec{\mu} + \vec{b}$ and covariance matrix $A\Sigma A^T$.*

An immediate consequence of Theorem 2.3 is 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.

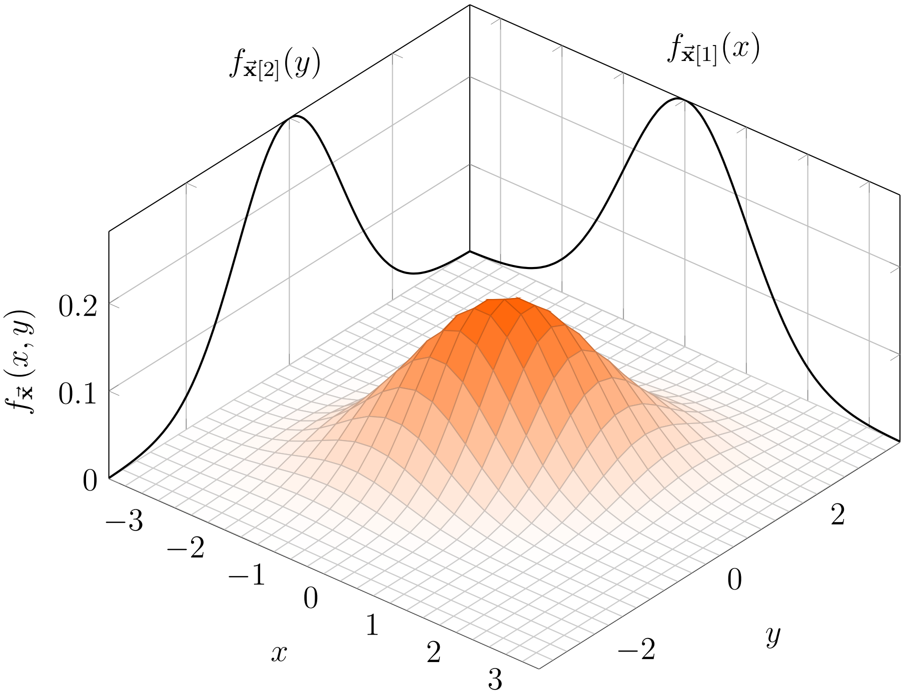

Figure 4: Joint pdf of a two-dimensional Gaussian vector $\vec{\mathbf{x}}$ and marginal pdfs of its two entries.

Another consequence of Theorem 2.3 is that an iid standard Gaussian vector is *isotropic*. This means that the vector does not favor any direction in its ambient space. More formally, no matter how you rotate it, its distribution is the same. More precisely, for any orthogonal matrix U, if $\vec{\mathbf{x}}$ is an iid standard Gaussian vector, then by Theorem 2.3 $U\vec{\mathbf{x}}$ 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.

2.2 Concentration in high dimensions

In the previous section we established that the direction of iid 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. Interestingly, this is not the case as the dimension of the ambient space grows. The squared ℓ_2-norm of an iid standard k-dimensional Gaussian vector $\vec{x}$ is the sum of k independent standard Gaussian random variables, which is known as a χ^2 (chi squared) random variable with k degrees of freedom. As shown in Figure 5, as k grows the pdf of this random variable concentrates around k, which is the mean of the squared ℓ_2-norm:

$$
\begin{aligned}
\mathrm{E}\left(||\vec{\mathbf{x}}||_2^2\right) &= \mathrm{E}\left(\sum_{i=1}^{k} \vec{\mathbf{x}}[i]^2\right) && (18)\\
&= \sum_{i=1}^{k} \mathrm{E}\left(\vec{\mathbf{x}}[i]^2\right) && (19)\\
&= k. && (20)
\end{aligned}
$$

The following lemma shows that the standard deviation of $||\vec{\mathbf{x}}||_2^2$ is $\sqrt{2k}$.

Lemma 2.4 (Variance of the squared ℓ_2 norm of a Gaussian vector). *Let $\vec{\mathbf{x}}$ be an iid Gaussian random vector of dimension k. The variance of $||\vec{\mathbf{x}}||_2^2$ is $2k$.*

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

$$
\begin{aligned}
\mathrm{E}\left(\left(||\vec{\mathbf{x}}||_2^2\right)^2\right) &= \mathrm{E}\left(\left(\sum_{i=1}^{k} \vec{\mathbf{x}}[i]^2\right)^2\right) && (21)\\
&= \mathrm{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} \vec{\mathbf{x}}[i]^2 \vec{\mathbf{x}}[j]^2\right) && (22)\\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{E}\left(\vec{\mathbf{x}}[i]^2 \vec{\mathbf{x}}[j]^2\right) && (23)\\
&= \sum_{i=1}^{k} \mathrm{E}\left(\vec{\mathbf{x}}[i]^4\right) + 2\sum_{i=1}^{k-1}\sum_{j=i}^{k} \mathrm{E}\left(\vec{\mathbf{x}}[i]^2\right)\mathrm{E}\left(\vec{\mathbf{x}}[j]^2\right) && (24)\\
&= 3k + k(k-1) \quad \text{since the 4th moment of a standard Gaussian equals 3} && (25)\\
&= k(k+2). && (26)
\end{aligned}
$$

□

The result implies that as k grows the relative deviation of the norm from its mean decreases proportionally to $1/\sqrt{k}$. Consequently, the squared norm is close to k with increasing probability. This is made precise in the following theorem, which yields a concrete non-asymptotic bound on the probability that it deviates by more than a small constant.

Theorem 2.5 (Chebyshev tail bound for the ℓ_2 norm of an iid standard Gaussian vector). *Let $\vec{\mathbf{x}}$ be an iid standard Gaussian random vector of dimension k. For any $\epsilon > 0$ we have*

$$
P\left(k(1-\epsilon) < ||\vec{\mathbf{x}}||_2^2 < k(1+\epsilon)\right) \geq 1 - \frac{2}{k\epsilon^2}. \qquad (27)
$$

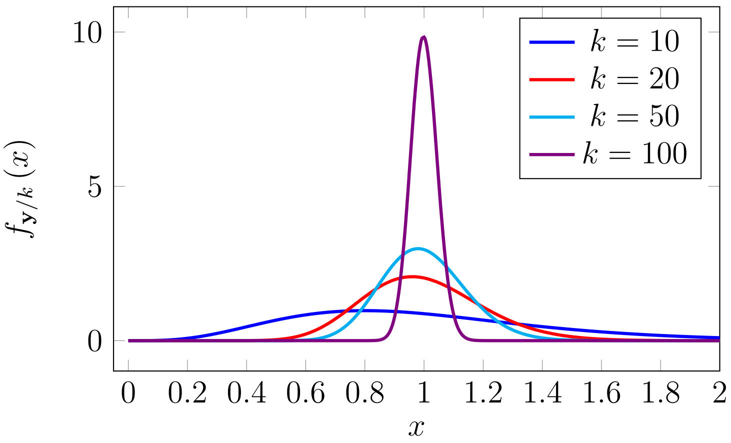

Figure 5: Pdfs of $\mathbf{y}/k$ for different values of k, where $\mathbf{y}$ is a χ^2 random variable with k degrees of freedom.

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

$$\mathrm{P}(\mathbf{x} \geq a) \leq \frac{\mathrm{E}(\mathbf{x})}{a}. \tag{28}$$

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

$$\begin{aligned}
\mathrm{P}(|\mathbf{y} - k| \geq k\epsilon) &= \mathrm{P}\left((\mathbf{y} - \mathrm{E}(\mathbf{y}))^2 \geq k^2\epsilon^2\right) && (29)\\
&\leq \frac{\mathrm{E}\left((\mathbf{y} - \mathrm{E}(\mathbf{y}))^2\right)}{k^2\epsilon^2} \qquad \text{by Markov's inequality} && (30)\\
&= \frac{\mathrm{Var}(\mathbf{y})}{k^2\epsilon^2} && (31)\\
&= \frac{2}{k\epsilon^2} \qquad \text{by Lemma 2.4.} && (32)
\end{aligned}$$

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 2.5 only relies on the variance to bound the probability that the magnitude deviates from its mean. As a result, it is significantly weaker than the following result, which exploits the fact that the higher moments of a standard Gaussian are well behaved.

Theorem 2.7 (Chernoff tail bound for the ℓ_2 norm of an iid standard Gaussian vector)**.** *Let $\vec{\mathbf{x}}$ be an iid standard Gaussian random vector of dimension k. For any $\epsilon \in (0,1)$ we have*

$$P\left(k\left(1-\epsilon\right) < \left|\left|\vec{\mathbf{x}}\right|\right|_2^2 < k\left(1+\epsilon\right)\right) \geq 1 - 2\exp\left(-\frac{k\epsilon^2}{8}\right). \tag{33}$$

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

$$P\left(\mathbf{y} > k\left(1+\epsilon\right)\right) \leq \exp\left(-\frac{k\epsilon^2}{8}\right), \tag{34}$$

$$P\left(\mathbf{y} < k\left(1-\epsilon\right)\right) \leq \exp\left(-\frac{k\epsilon^2}{8}\right). \tag{35}$$

We present the proof of (34). The proof of (35) is essentially the same and is presented in Section 5.3. Let $t > 0$ be an arbitrary positive number, and note that

$$\begin{aligned}
P\left(\mathbf{y} > a\right) &= P\left(\exp\left(t\mathbf{y}\right) > \exp\left(at\right)\right) && (36)\\
&\leq \exp\left(-at\right)\,\mathrm{E}\left(\exp\left(t\mathbf{y}\right)\right) \qquad \text{by Markov's inequality} && (37)\\
&\leq \exp\left(-at\right)\,\mathrm{E}\left(\exp\left(\sum_{i=1}^{k} t\mathbf{x_i}^2\right)\right) && (38)\\
&\leq \exp\left(-at\right)\prod_{i=1}^{k}\mathrm{E}\left(\exp\left(t\mathbf{x_i}^2\right)\right) \qquad \text{by independence of } \mathbf{x_1},\ldots,\mathbf{x_k} && (39)\\
&= \frac{\exp\left(-at\right)}{\left(1-2t\right)^{\frac{k}{2}}}, && (40)
\end{aligned}$$

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

Lemma 2.8 (Proof in Section 5.2)**.** *For $\mathbf{x}$ standard Gaussian and $t < 1/2$,*

$$\mathrm{E}\left(\exp\left(t\mathbf{x}^2\right)\right) = \frac{1}{\sqrt{1-2t}}. \tag{41}$$

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

$$\begin{aligned}
\mathrm{E}\left(\exp\left(t\mathbf{x}^2\right)\right) &= \mathrm{E}\left(\sum_{i=0}^{\infty}\frac{\left(t\mathbf{x}^2\right)^i}{i!}\right) && (42)\\
&= \sum_{i=0}^{\infty}\frac{\mathrm{E}\left(t^i\left(\mathbf{x}^{2i}\right)\right)}{i!}. && (43)
\end{aligned}$$

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 $a := k(1+\epsilon)$ and

$$t := \frac{1}{2} - \frac{1}{2(1+\epsilon)}, \tag{44}$$

by minimizing over $t \in (0, 1/2)$ in (40). This gives

$$\begin{aligned} P(\mathbf{y} > k(1+\epsilon)) &\leq (1+\epsilon)^k \, 2 \exp\left(-\frac{k\epsilon}{2}\right) & (45) \\ &= \exp\left(-\frac{k}{2}(\epsilon - \log(1+\epsilon))\right) & (46) \\ &\leq \exp\left(-\frac{k\epsilon^2}{8}\right), & (47) \end{aligned}$$

where the last step follows from the fact that the function $g(x) := x - \frac{x^2}{4} - \log(1+x)$ is nonnegative between 0 and 1 (the derivative is nonnegative and $g(0) = 0$). □

In the next section we apply this result to characterize the projection of an iid standard Gaussian vector on a subspace.

2.3 Projection onto a fixed subspace

In Example 7.4 of Lecture Notes 1 we observed that the ℓ_2-norm of the projection of iid standard Gaussian noise onto a fixed subspace is proportional to the square root of the dimension of that subspace. In this section we make this precise, using that the fact that the coefficients of the projection in an orthonormal basis of the subspace are themselves iid standard Gaussians.

Lemma 2.9 (Projection of an iid Gaussian vector onto a subspace)**.** *Let $\mathcal{S}$ be a k-dimensional subspace of $\mathbb{R}^n$ and $\vec{\mathbf{z}} \in \mathbb{R}^n$ a vector of iid standard Gaussian noise. $||\mathcal{P}_{\mathcal{S}}\, \vec{\mathbf{z}}||_2^2$ is a χ^2 random variable with k degrees of freedom, i.e. it has the same distribution as the random variable*

$$\mathbf{y} := \sum_{i=1}^{k} \mathbf{x_i}^2 \tag{48}$$

where $\mathbf{x_1}, \ldots, \mathbf{x_k}$ are iid standard Gaussian random variables.

Proof. Let UU^T be a projection matrix for the subspace $\mathcal{S}$, where the columns of $U \in \mathbb{R}^{n \times k}$

are orthonormal. We have

$$
\begin{aligned}
||\mathcal{P}_{\mathcal{S}}\,\vec{\mathbf{z}}||_2^2 &= \left|\left|UU^T\vec{\mathbf{z}}\right|\right|_2^2 && (49)\\
&= \vec{\mathbf{z}}^T UU^T UU^T \vec{\mathbf{z}} && (50)\\
&= \vec{\mathbf{z}}^T UU^T \vec{\mathbf{z}} && (51)\\
&= \vec{\mathbf{w}}^T \vec{\mathbf{w}} && (52)\\
&= \sum_{i=1}^{k} \vec{\mathbf{w}}[i]^2, && (53)
\end{aligned}
$$

where by Theorem 2.3 the random vector $\vec{\mathbf{w}} := U^T\vec{\mathbf{z}}$ is Gaussian with mean zero and covariance matrix

$$
\begin{aligned}
\Sigma_{\vec{\mathbf{w}}} &= U^T \Sigma_{\vec{\mathbf{z}}} U && (54)\\
&= U^T U && (55)\\
&= I, && (56)
\end{aligned}
$$

so the entries are independent standard Gaussians. □

Since the coefficients are standard Gaussians, we can bound the deviation of their norm using Theorem 2.7.

Theorem 2.10. *Let $\mathcal{S}$ be a k-dimensional subspace of $\mathbb{R}^n$ and $\vec{\mathbf{z}} \in \mathbb{R}^n$ a vector of iid Gaussian noise. For any $\epsilon \in (0,1)$*

$$\sqrt{k\,(1-\epsilon)} \leq ||\mathcal{P}_{\mathcal{S}}\,\vec{z}||_2 \leq \sqrt{k\,(1+\epsilon)} \qquad (57)$$

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

Proof. The result follows from Theorem 2.7 and Lemma 2.9. □

3 Gaussian matrices

3.1 Randomized projections

As we discussed in Section 3.3 of Lecture Notes 2, dimensionality reduction via PCA consists of projecting the data on low-dimensional subspaces that are optimal in the sense that they preserve most of the energy. The principal directions are guaranteed to lie in the directions of maximum variation of the data, but finding them requires computing the SVD, which can be computationally expensive 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 seeing the data. A simple method to achieve this is to project the data using a random linear map, represented by a random matrix $\mathbf{A}$

built by sampling each entry independently from a standard Gaussian distribution. The following lemma shows that the distribution of the result of applying such a matrix to a fixed deterministic vector is Gaussian.

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

Proof. By Theorem 2.3, $(\mathbf{A}\vec{v})[i]$, $1 \leq i \leq a$ is Gaussian, since it is the inner product between $\vec{v}$ and the ith row $\mathbf{A}_{i,:}$ (interpreted as a vector in $\mathbb{R}^b$), which is an iid standard Gaussian vector. The mean of the entry is zero because the mean of $\mathbf{A}_{i,:}$ is zero and the variance equals

$$\begin{aligned} \operatorname{Var}\left(\mathbf{A}_{i,:}^T \vec{v}\right) &= \vec{v}^T \Sigma_{\mathbf{A}_{i,:}} \vec{v} && (58) \\ &= \vec{v}^T I \vec{v} && (59) \\ &= \left|\left|\vec{v}\right|\right|_2^2 && (60) \\ &= 1, && (61) \end{aligned}$$

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

A direct consequence of this result is a non-asymptotic bound on the ℓ_2 norm of $\mathbf{A}\vec{v}$ for any fixed deterministic vector $\vec{v}$.

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

$$\sqrt{a(1-\epsilon)} \leq \left|\left|\mathbf{A}\vec{v}\right|\right|_2 \leq \sqrt{a(1+\epsilon)} \qquad (62)$$

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

Proof. The result follows from Theorem 2.7 and Lemma 3.1. □

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 us 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 in a non-asymptotic sense. The result is striking because the lower bound on k – 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 p in the data set is only logarithmic. The proof is based on the arguments in [3].

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

$$(1-\epsilon) \left|\left|\vec{x}_i - \vec{x}_j\right|\right|_2^2 \leq \left|\left|\frac{1}{\sqrt{k}}\mathbf{A}\vec{x}_i - \frac{1}{\sqrt{k}}\mathbf{A}\vec{x}_j\right|\right|_2^2 \leq (1+\epsilon) \left|\left|\vec{x}_i - \vec{x}_j\right|\right|_2^2, \tag{63}$$

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

$$k \geq \frac{16 \log (p)}{\epsilon^2}. \tag{64}$$

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

$$\vec{v}_{ij} := \frac{\vec{x}_i - \vec{x}_j}{\left|\left|\vec{x}_i - \vec{x}_j\right|\right|_2}, \tag{65}$$

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 (64) holds

$$\mathrm{P}\left(\mathcal{E}_{ij}^c\right) \leq \frac{2}{p^2}. \tag{66}$$

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*

$$\mathrm{P}\left(\cup_i S_i\right) \leq \sum_{i=1}^{n} \mathrm{P}\left(S_i\right). \tag{67}$$

The number of events in the intersection is $\binom{p}{2} = p(p-1)/2$, because that is the number

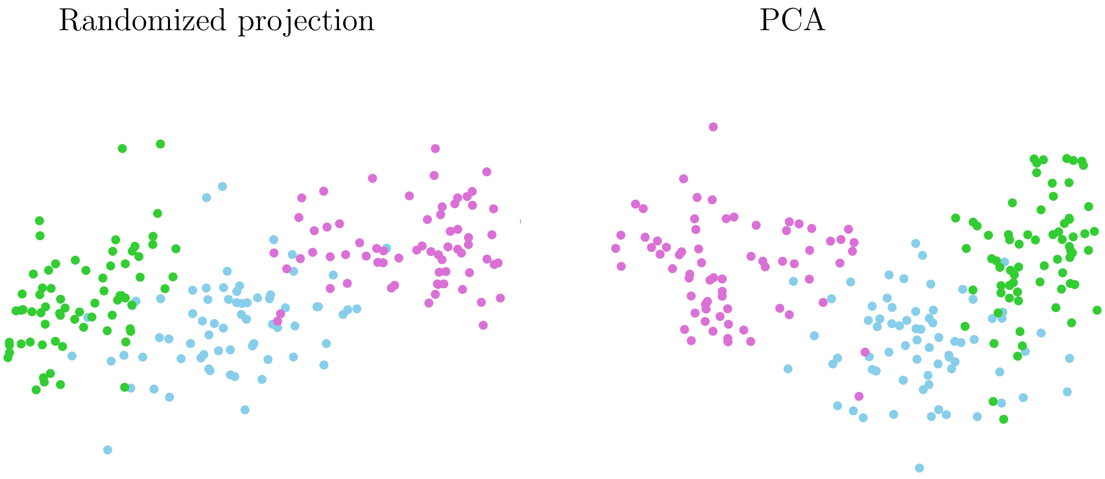

Figure 6: Approximate projection of 7-dimensional vectors describing different wheat seeds onto two random directions. Each color represents a variety of wheat.

of different pairs of vectors in the set $\{\vec{x}_1, \ldots, \vec{x}_p\}$. The union bound yields

$$\begin{aligned}
\mathrm{P}\left(\bigcap_{i,j} \mathcal{E}_{ij}\right) &= 1 - \mathrm{P}\left(\bigcup_{i,j} \mathcal{E}_{ij}^c\right) && (68)\\
&\geq 1 - \sum_{i,j} \mathrm{P}\left(\mathcal{E}_{ij}^c\right) && (69)\\
&\geq 1 - \frac{p\,(p-1)}{2}\,\frac{2}{p^2} && (70)\\
&\geq \frac{1}{p}. && (71)
\end{aligned}$$

□

Example 3.5 (Dimensionality reduction via randomized projections)**.** We consider the same data as in Example 3.6 of Lecture Notes 2. Each data point corresponds to a seed with seven features: area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. The seeds belong to three different varieties of wheat: Kama, Rosa and Canadian.[2] The objective is to project the data onto 2D for visualization. In Figure 6 we compare the result of randomly projecting the data by applying a 2 × 7 iid standard Gaussian matrix, with the result of PCA-based dimensionality reduction. In terms of keeping the different types of wheat separated, the randomized projection preserves the structure in the data as effectively as PCA. △

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

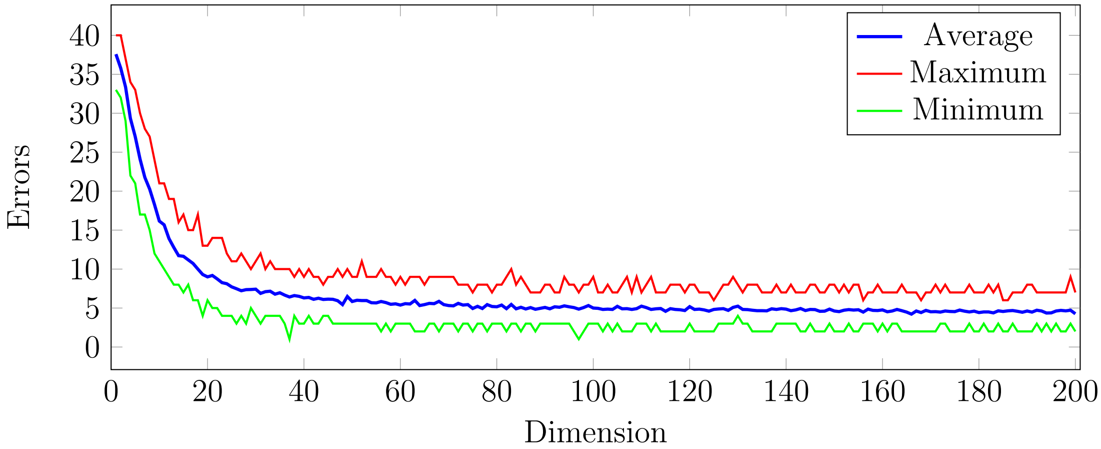

Figure 7: Average, maximum and minimum number of errors (over 50 tries) for nearest-neighbor classification after a randomized dimensionality reduction for different dimensions.

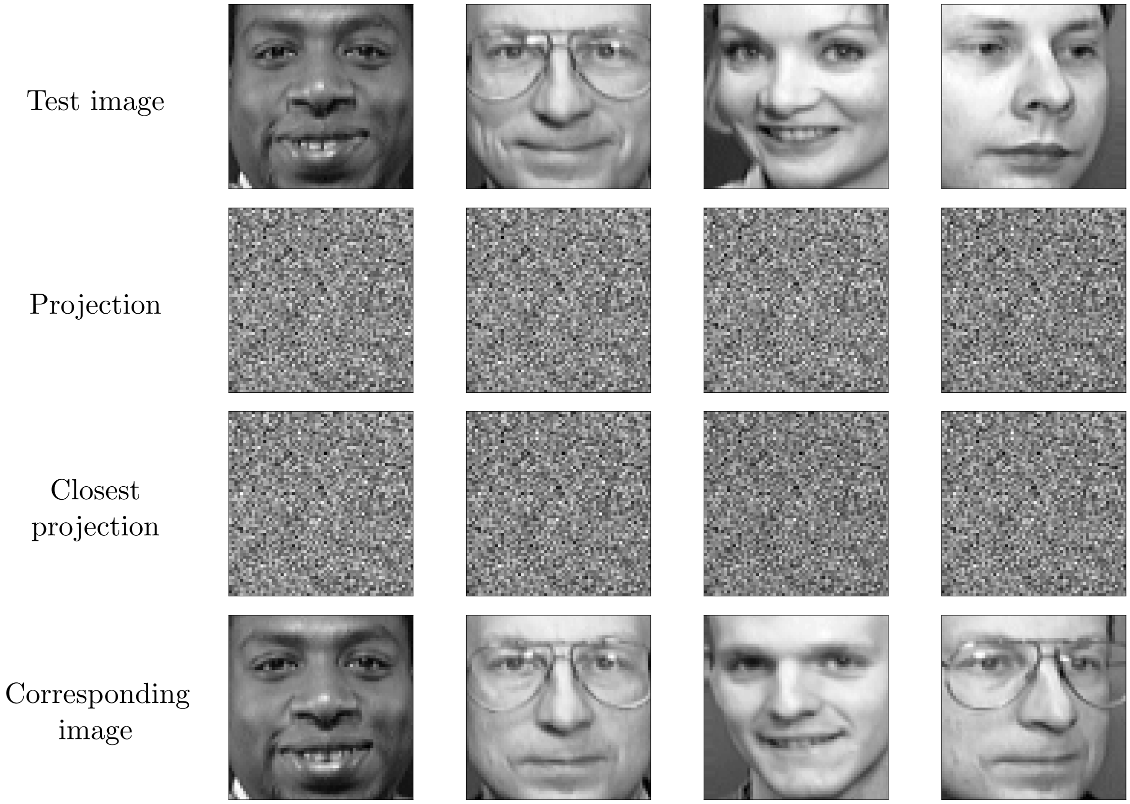

Figure 8: Results of nearest-neighbor classification combined with randomized dimensionality reduction of dimension 50 for four of the people in Example 3.6. The assignments of the first two examples are correct, but the other two are wrong.

Example 3.6 (Nearest neighbors after random projection)**.** The nearest neighbors algorithm for classification (Algorithm 4.2 in Lecture Notes 1) requires computing n distances in an m-dimensional space (where m is the number of features) to classify each new example. The computational cost is $\mathcal{O}(nm)$, so if we need to classify p points the total cost is $\mathcal{O}(nmp)$. 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:

- kmn operations to project the training data using a $k \times m$ iid standard Gaussian matrix.
- kmp 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 $\mathcal{O}(kp \max\{m, n\})$, which is a significant reduction from $\mathcal{O}(nmp)$. It is also more efficient than the PCA-based approach of Example 3.5 in Lecture Notes 2, which includes an additional $\mathcal{O}(mn \min\{m, n\})$ step to compute an SVD.

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

3.2 Singular values

In this section we analyze the singular values of matrices with iid standard Gaussian entries. In particular we consider an $k \times n$ matrix $\mathbf{A}$ where $n > k$. Numerically, we observe that as n grows, all k singular values converge to $\sqrt{n}$, as shown in Figure 9. As a result,

$$\mathbf{A} \approx U\left(\sqrt{n}\, I\right) V^T = \sqrt{n}\, U V^T, \tag{72}$$

i.e. $\mathbf{A}$ is close to an orthogonal matrix. Geometrically, this implies that if we generate a fixed number of iid Gaussian vectors at increasing ambient dimensions, the vectors will tend to be almost orthogonal 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 [6] for some excellent notes on high-dimensional probability techniques in this spirit.

Theorem 3.7 (Singular values of a Gaussian matrix)**.** *Let* $\mathbf{A}$ *be a* $n \times k$ *matrix with iid standard Gaussian entries such that* $n > k$*. For any fixed* $\epsilon > 0$*, the singular values of* $\mathbf{A}$ *satisfy*

$$\sqrt{n}\,(1 - \epsilon) \leq \boldsymbol{\sigma}_{\mathbf{k}} \leq \boldsymbol{\sigma}_{\mathbf{1}} \leq \sqrt{n}\,(1 + \epsilon) \tag{73}$$

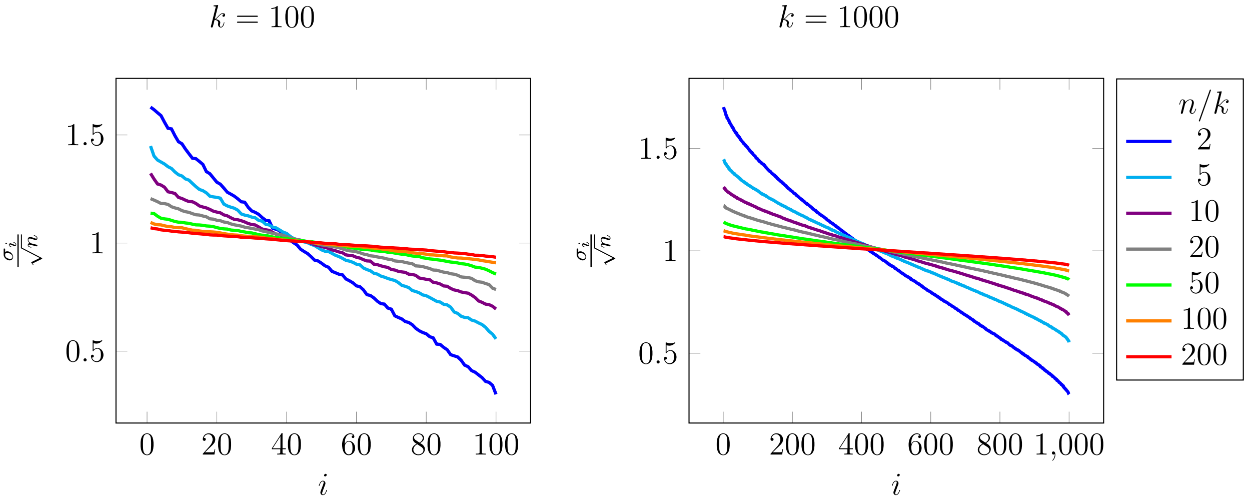

Figure 9: Singular values of $n \times k$ matrices with iid standard Gaussian entries for different values of k and n.

with probability at least $1 - 1/k$ *as long as*

$$n > \frac{64k}{\epsilon^2} \log \frac{12}{\epsilon}. \tag{74}$$

By Theorem 2.7 in Lecture Notes 2, the bounds on the singular values are equivalent to the following bounds

$$\sqrt{n}\,(1-\epsilon) < ||\mathbf{A}\vec{v}||_2 < \sqrt{n}\,(1+\epsilon) \tag{75}$$

where $\vec{v}$ is *any* vector in the k-dimensional sphere $\mathcal{S}^{k-1}$, which contains the unit-ℓ_2-norm vectors in $\mathbb{R}^k$. This set has infinite cardinality, so we cannot apply the union bound to establish the bounds as in the proof of the Johnson-Lindenstrauss lemma. To overcome this obstacle, we consider a set, called an ϵ-net, which covers the sphere in the sense that every other point is not too far from one of its elements. We prove that the bounds hold on the net using the union bound and then establish that as a result they hold for the whole sphere.

Definition 3.8 (ϵ-net)**.** *An ϵ-net of a set $\mathcal{X} \subseteq \mathbb{R}^k$ 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*

$$||\vec{x} - \vec{y}||_2 \leq \epsilon. \tag{76}$$

Figure 10 shows an ϵ-net for the two-dimensional sphere $\mathcal{S}^1$. The smallest possible number of points in the ϵ-net of a set is called its covering number.

Definition 3.9 (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 $\mathcal{X}$.*

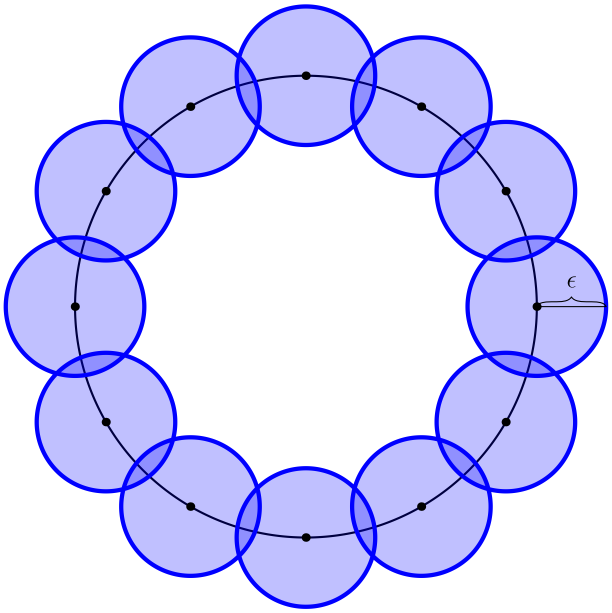

Figure 10: ϵ-net for the two-dimensional sphere $\mathcal{S}^1$, which is just a circle.

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

Theorem 3.10 (Covering number of a sphere)**.** *The covering number of the n-dimensional sphere $\mathcal{S}^{k-1}$ at scale ϵ satisfies*

$$\mathcal{N}\left(\mathcal{S}^{k-1},\epsilon\right) \leq \left(\frac{2+\epsilon}{\epsilon}\right)^k \leq \left(\frac{3}{\epsilon}\right)^k. \tag{77}$$

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

$$\mathcal{E}_{\vec{v},\epsilon_2} := \left\{n\left(1-\epsilon_2\right)||\vec{v}||_2^2 \leq ||\mathbf{A}\vec{v}||_2^2 \leq n\left(1+\epsilon_2\right)||\vec{v}||_2^2\right\}. \tag{78}$$

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

$$\begin{align}
\mathrm{P}\left(\cup_{\vec{v}\in\mathcal{N}_{\epsilon_1}}\mathcal{E}^c_{\vec{v},\epsilon_2}\right) &\leq \sum_{\vec{v}\in\mathcal{N}_{\epsilon_1}}\mathrm{P}\left(\mathcal{E}^c_{\vec{v},\epsilon_2}\right) \tag{79}\\
&\leq \left|\mathcal{N}_{\epsilon_1}\right|\mathrm{P}\left(\mathcal{E}^c_{\vec{v},\epsilon_2}\right) \tag{80}\\
&\leq 2\left(\frac{12}{\epsilon}\right)^k \exp\left(-\frac{n\epsilon^2}{32}\right) \tag{81}\\
&\leq \frac{1}{k} \quad \text{if (74) holds.} \tag{82}
\end{align}$$

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 $\mathcal{S}^{k-1}$, not only for those in the ϵ_1-net. For any arbitrary vector

$\vec{x} \in \mathcal{S}^{k-1}$ 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

$$\begin{aligned} ||\mathbf{A}\vec{x}||_2 &\leq ||\mathbf{A}\vec{v}||_2 + ||\mathbf{A}(\vec{x} - \vec{v})||_2 && (83) \\ &\leq \sqrt{n}\left(1 + \frac{\epsilon}{2}\right) + ||\mathbf{A}(\vec{x} - \vec{v})||_2 \qquad \text{assuming } \cup_{\vec{v} \in \mathcal{N}_{\epsilon_1}} \mathcal{E}^c_{\vec{v},\epsilon_2} \text{ holds} && (84) \\ &\leq \sqrt{n}\left(1 + \frac{\epsilon}{2}\right) + \boldsymbol{\sigma_1} ||\vec{x} - \vec{v}||_2 \qquad \text{by Theorem 2.7 in Lecture Notes 2} && (85) \\ &\leq \sqrt{n}\left(1 + \frac{\epsilon}{2}\right) + \frac{\boldsymbol{\sigma_1}\epsilon}{4}. && (86) \end{aligned}$$

By Theorem 2.7 in Lecture Notes 2 $\boldsymbol{\sigma_1}$ is the smallest upper bound on $||\mathbf{A}\vec{x}||_2$ for all $\vec{x}$ in the sphere, so the bound in equation (86) cannot be smaller:

$$\boldsymbol{\sigma_1} \leq \sqrt{n}\left(1 + \frac{\epsilon}{2}\right) + \frac{\boldsymbol{\sigma_1}\epsilon}{4}, \qquad (87)$$

so that

$$\begin{aligned} \boldsymbol{\sigma_1} &\leq \sqrt{n}\left(\frac{1 + \epsilon/2}{1 - \epsilon/4}\right) && (88) \\ &= \sqrt{n}\left(1 + \epsilon - \frac{\epsilon(1 - \epsilon)}{4 - \epsilon}\right) && (89) \\ &\leq \sqrt{n}(1 + \epsilon). && (90) \end{aligned}$$

The lower bound on $\boldsymbol{\sigma_k}$ follows from a similar argument combined with (90). By the triangle inequality

$$\begin{aligned} ||\mathbf{A}\vec{x}||_2 &\geq ||\mathbf{A}\vec{v}||_2 - ||\mathbf{A}(\vec{x} - \vec{v})||_2 && (91) \\ &\geq \sqrt{n}\left(1 - \frac{\epsilon}{2}\right) - ||A(\vec{x} - \vec{v})||_2 \qquad \text{assuming } \cup_{\vec{v} \in \mathcal{N}_{\epsilon_1}} \mathcal{E}^c_{\vec{v},\epsilon_2} \text{ holds} && (92) \\ &\geq \sqrt{n}\left(1 - \frac{\epsilon}{2}\right) - \boldsymbol{\sigma_1} ||\vec{x} - \vec{v}||_2 \qquad \text{by Theorem 2.7 in Lecture Notes 2} && (93) \\ &\geq \sqrt{n}\left(1 - \frac{\epsilon}{2}\right) - \frac{\epsilon}{4}\sqrt{n}(1 + \epsilon) \qquad \text{by (90)} && (94) \\ &= \sqrt{n}(1 - \epsilon). && (95) \end{aligned}$$

By Theorem 2.7 in Lecture Notes 2 $\boldsymbol{\sigma_k}$ is the largest lower bound on $||\mathbf{A}\vec{x}||_2$ for all $\vec{x}$ on the sphere, so $\boldsymbol{\sigma_k} \geq \sqrt{n}(1 - \epsilon)$ as long as $\cup_{\vec{v} \in \mathcal{N}_{\epsilon_1}} \mathcal{E}^c_{\vec{v},\epsilon_2}$ holds.

4 Randomized singular-value decomposition

4.1 Fast SVD

In this section we describe an algorithm to compute the SVD of a low-rank matrix very efficiently assuming that we have access to an orthonormal basis of its column space.

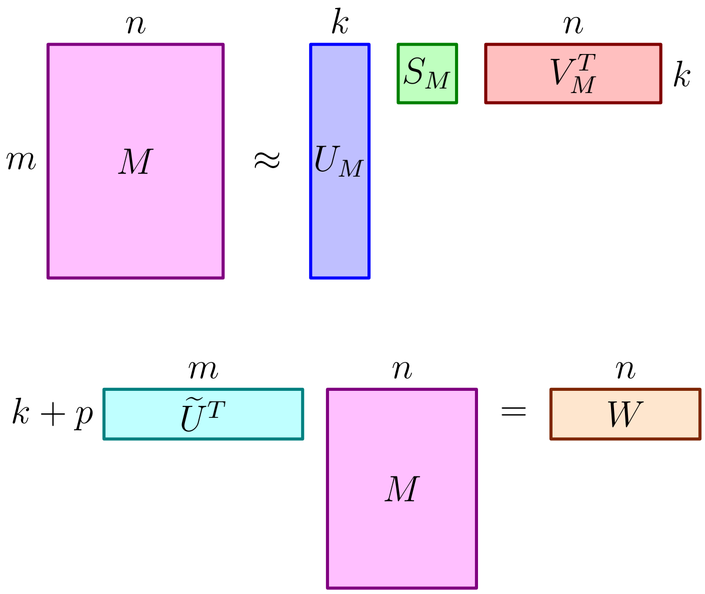

Figure 11: Sketch of the matrices in Algorithm 4.1.

Algorithm 4.1 (Fast SVD)**.** *Given a matrix $M \in \mathbb{R}^{m\times n}$ which is well-approximated as a k-rank matrix:*

1. *Find a matrix $\widetilde{U} \in \mathbb{R}^{m\times(k+p)}$ with $k+p$ orthonormal columns that approximately span the column space of M.*
2. *Compute $W \in \mathbb{R}^{(k+p)\times n}$ defined by $W := \widetilde{U}^T M$.*
3. *Compute the SVD of $W = U_W S_W V_W^T$.*
4. *Output $U := (\widetilde{U} U_W)_{:,1:k}$, $S := (S_W)_{1:k,1:k}$ and $V := (V_W)_{:,1:k}$ as the SVD of M.*

Figure 11 shows the dimensions of the matrices. Computing the SVD of W has complexity $\mathcal{O}(k^2 n)$, so overall the complexity of the algorithm is governed by the second step which has complexity $\mathcal{O}(kmn)$. This is a dramatic reduction from $\mathcal{O}(mn \min\{m,n\})$.

The following lemma establishes that the algorithm works in the idealized situation where the matrix is exactly low rank and we have access to an orthonormal basis of its column space.

Lemma 4.2. *Algorithm 4.1 outputs the SVD of a matrix $M \in \mathbb{R}^{m\times n}$ as long as M is rank k and $\widetilde{U}$ spans its column space.*

Proof. If $\widetilde{U}$ spans the column space of M then

$$M = \widetilde{U}\widetilde{U}^T M \tag{96}$$

$$= \widetilde{U} W \tag{97}$$

$$= \widetilde{U} U_W S_W V_W^T, \tag{98}$$

where $U := \widetilde{U} U_W$ is an $m \times k$ matrix with orthonormal columns since

$$\begin{aligned} U^T U &= U_W^T \widetilde{U}^T \widetilde{U} U_W && (99) \\ &= U_W^T U_W && (100) \\ &= I, && (101) \end{aligned}$$

$S_W \in \mathbb{R}^{k \times k}$ is a diagonal matrix with nonnegative entries and $V_W \in \mathbb{R}^{n \times k}$ has orthonormal columns. We conclude that the output of Algorithm 4.1 is a valid SVD of M. □

The following sections describe two methods for estimating the column space based on randomization (i.e., step 1 of Algorithm 4.1). In practice, most matrices of interest will only be approximately low rank. In that case, the performance of the column-approximation algorithms depends on the gap between the k first singular values and the $k+1$th. To enhance the performance, a popular preprocessing procedure is to use *power iterations* to increase the gap. Instead of M, the idea is to apply Step 1 of Algorithm 4.1 to

$$\widetilde{M} := \left(M M^T\right)^q M, \tag{102}$$

for a small integer q. Expressing M in terms of its SVD, we have

$$\begin{aligned} \widetilde{M} &= \left(U_M S_M^2 U_M^T\right)^q U_M S_M V_M^T && (103) \\ &= U_M S_M^{2q+1} V_M^T. && (104) \end{aligned}$$

$\widetilde{M}$ has the same singular vectors as M but its singular values are raised to the power of $2q+1$, so the gap between small and large singular values is amplified. The idea is very similar to the power method for computing eigenvectors (Algorithm 4.4 in Lecture Notes 2). More details on the power iterations will be given in the next two subsections.

4.2 Randomized column-space approximation

Random projections make it possible to estimate the column space of a low-rank matrix very efficiently. Here we just outline the method and provide some intuition. We refer to [4,5] for a more detailed description and theoretical guarantees.

Algorithm 4.3 (Randomized column-space approximation)**.** *Given a matrix $M \in \mathbb{R}^{m \times n}$ which is well-approximated as a k-rank matrix:*

1. *Create an $n \times (k+p)$ iid standard Gaussian matrix $\mathbf{A}$, where p is a small integer (e.g. 5).*

2. *Compute the $m \times (k+p)$ matrix $\mathbf{B} = M\mathbf{A}$.*

3. *Compute an orthonormal basis for* $\operatorname{col}(\mathbf{B})$ *and output them as a matrix $\widetilde{\mathbf{U}} \in \mathbb{R}^{m \times (k+p)}$.*

Figure 12: Four randomly selected frames from the video in Example 4.7.

Consider the matrix

$$\begin{aligned}\mathbf{B} &= M\mathbf{A} && (105)\\ &= U_M S_M V_M^T \mathbf{A} && (106)\\ &= U_M S_M \mathbf{C}. && (107)\end{aligned}$$

If M is exactly low rank, by Theorem 2.3 $\mathbf{C}$ is a $k \times (k+p)$ iid standard Gaussian matrix since $V_M^T V_M = I$. In that case the column space of $\mathbf{B}$ is the same as that of M because $\mathbf{C}$ is full rank with high probability. When M is only approximately low rank, then $\mathbf{C}$ is a $\min\{m,n\} \times (k+p)$ iid standard Gaussian matrix. Surprisingly, for p equal to a small integer, the product with $\mathbf{C}$ conserves the subspace corresponding to the largest k singular values with high probability, as long as the $k+1$th singular values is sufficiently small. See the seminal paper [4] for more details.

We can improve the accuracy of Algorithm 4.3 by using power iterations as detailed below.

Algorithm 4.4 (Power Iterations for Randomized column-space approximation)**.** *Given a matrix $M \in \mathbb{R}^{m \times n}$ which is well-approximated as a k-rank matrix:*

1. *Create an $n \times (k+p)$ iid standard Gaussian matrix $\mathbf{A}$, where p is a small integer (e.g. 5).*
2. *Compute the $m \times (k+p)$ matrix $\mathbf{B} = M\mathbf{A}$ and let $\widetilde{\mathbf{U}}_0$ denote the matrix formed by orthonormalizing the columns of $\mathbf{B}$.*
3. *For i from 1 to q (inclusive) :*
   (a) *Let $\widetilde{\mathbf{F}}_i$ be formed by orthonormalizing the columns of $M^T \widetilde{\mathbf{U}}_{i-1}$.*
   (b) *Let $\widetilde{\mathbf{U}}_i$ be formed by orthonormalizing the columns of $M\widetilde{\mathbf{F}}_i$.*
4. *Output $\widetilde{\mathbf{U}}_q$.*

Example 4.5 (Randomized SVD of a video)**.** In this example we consider a data set that consists of a video with 160 1080×1920 frames. Four sample frames are show in 12. We interpret each frame as a vector in $\mathbb{R}^{20{,}736{,}000}$. The matrix obtained by stacking these vectors as columns is approximately low rank due to the correlation between the frames,

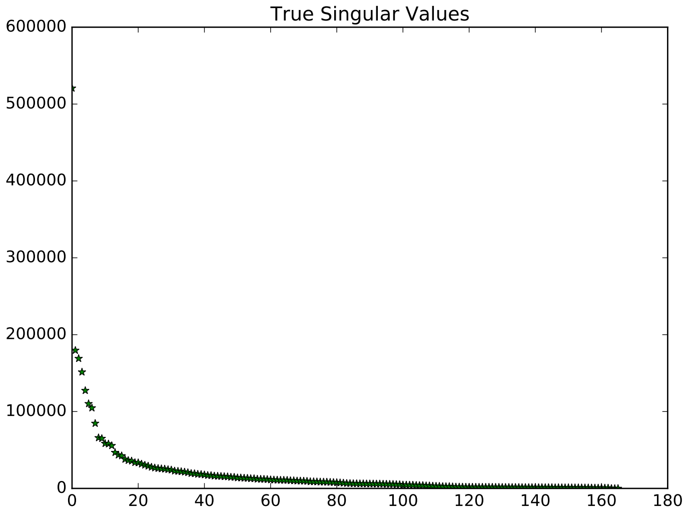

Figure 13: Singular values of the $160 \times 20,736,000$ matrix in Example 4.7.

as can be seen in Figure 13. Computing the SVD of this matrix takes 12 seconds on our machine. Applying Algorithm 4.1 combined with Algorithm 4.4 we obtain a rank-10 approximation in 5.8 seconds. If we consider a larger video with 691 frames, computing the SVD takes 281.1 seconds, whereas the randomized algorithm only takes 10.4 seconds. Figure 14 shows a comparison between the *true* left singular values and the estimate produced by the randomized algorithm, using power iterations with parameter $q = 2$ and setting $p = 7$ in Algorithm 4.4. △

4.3 Random column selection

An alternative procedure for estimating the column space of a low-rank matrix is to randomly select a subset of columns and obtain an orthonormal basis from them.

Algorithm 4.6 (Randomized column-space approximation)**.** *Given a matrix $M \in \mathbb{R}^{m\times n}$ which is well-approximated as a k-rank matrix:*

1. *Select a random subset of column indices $\boldsymbol{\mathcal{I}} := \{\mathbf{i_1}, \mathbf{i_2}, \ldots, \mathbf{i_{k'}}\}$ with $k' \geq k$.*
2. *Compute an orthonormal basis for the columns of the submatrix corresponding to $\boldsymbol{\mathcal{I}}$:*

$$M_{\boldsymbol{\mathcal{I}}} := \begin{bmatrix} M_{:,\mathbf{i_1}} & M_{:,\mathbf{i_2}} & \cdots & M_{:,\mathbf{i_{k'}}} \end{bmatrix} \tag{108}$$

and output them as a matrix $\widetilde{\mathbf{U}} \in \mathbb{R}^{m\times k'}$.

The random submatrix $\mathbf{M}_{\boldsymbol{\mathcal{I}}}$ can be expressed in terms of the left singular vectors and singular values of M, and a submatrix of the right-singular-vector matrix,

$$M_{\boldsymbol{\mathcal{I}}} = U_M S_M \left(V_M\right)_{\boldsymbol{\mathcal{I}}}. \tag{109}$$

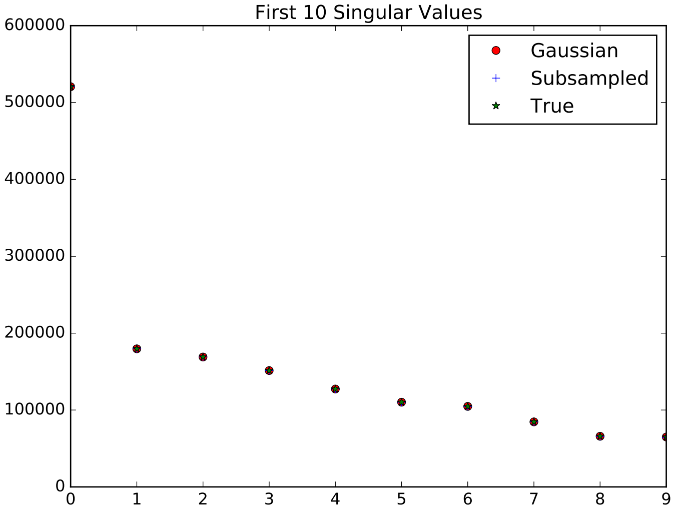

Figure 14: First 10 singular values of the $160 \times 20,736,000$ matrix in Example 4.7 and their estimate using Algorithm 4.1 combined with Algorithm 4.4 and Algorithm 4.6 with power iteration. The estimates are almost perfect.

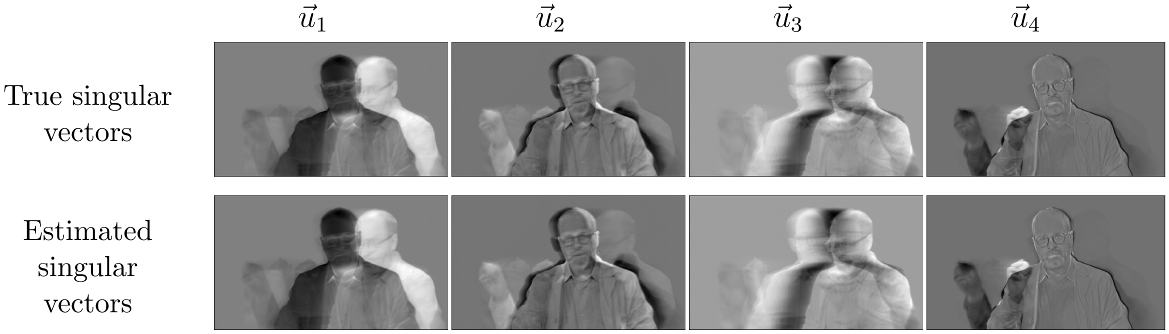

Figure 15: The first 4 left singular vectors of the movie computed using the standard SVD algorithm, and the randomized Algorithm 4.4.

In order for the algorithm to capture the column space of M, we need $(V_M)_{\mathcal{I}}$ to be full rank (and ideally to not have very small singular values). Moreover, if the matrix is only approximately low rank, we need the right singular vectors corresponding to large singular values to be *spread out* so that they are not missed when we subsample. Let us illustrate this with a simple example. Consider the rank-2 matrix

$$M := \begin{bmatrix} -3 & 2 & 2 & 2 \\ 3 & 2 & 2 & 2 \\ -3 & 2 & 2 & 2 \\ 3 & 2 & 2 & 2 \end{bmatrix} \tag{110}$$

and its SVD

$$M = U_M \, S_M V_M^T = \begin{bmatrix} 0.5 & -0.5 \\ 0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 6.9282 & 0 \\ 0 & 6 \end{bmatrix} \begin{bmatrix} 0 & 0.577 & 0.577 & 0.577 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{111}$$

$$\tag{112}$$

Any submatrix of columns that do not include the first one will have a column space that only consists of the first left singular vector. For example if $\mathcal{I} = \{2, 3\}$

$$M_{\mathcal{I}} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} 6.9282 \begin{bmatrix} 0.577 & 0.577 \end{bmatrix}. \tag{113}$$

Column subsampling tends to ignore left singular vectors corresponding to sparse (or approximately sparse) right singular vectors. Depending on the application, this may not be a disadvantage: sparse right singular vectors arise due to columns in M that are almost orthogonal to every other column and can consequently be interpreted as outliers. In contrast, column subsampling preserves the part of the column space corresponding to spread-out right singular vectors with high probability (see [2] for a theoretical characterization of this phenomenon).

To use power iteration with column subsampling, define $\mathbf{B} := \widetilde{\mathbf{U}}$ obtained from 4.6, and apply steps 3,4 from Algorithm 4.4. Although power iteration can be helpful, it will not alleviate the issue with sparse columns described above.

Example 4.7 (Randomized SVD of a video using subset of columns)**.** In this example we consider the same data set as in Example 4.7. We estimate a rank-10 approximation of the $160 \times 20,736,000$ matrix by randomly selecting $k' = 17$ columns as in Algorithm 4.6 and then applying Algorithm 4.1. The running time is 5.2 seconds, even faster than if we use Algorithm 4.1. Figure 14 shows a comparison between the *true* left singular values and the estimate produced by the randomized algorithm, using power iterations with parameter $q = 2$. In Figure 16 we show the top 4 estimated singular vectors compared to the true ones. The approximation is very precise, indicating that the first 10 right singular vectors of the matrix are not sparse or approximately sparse. △

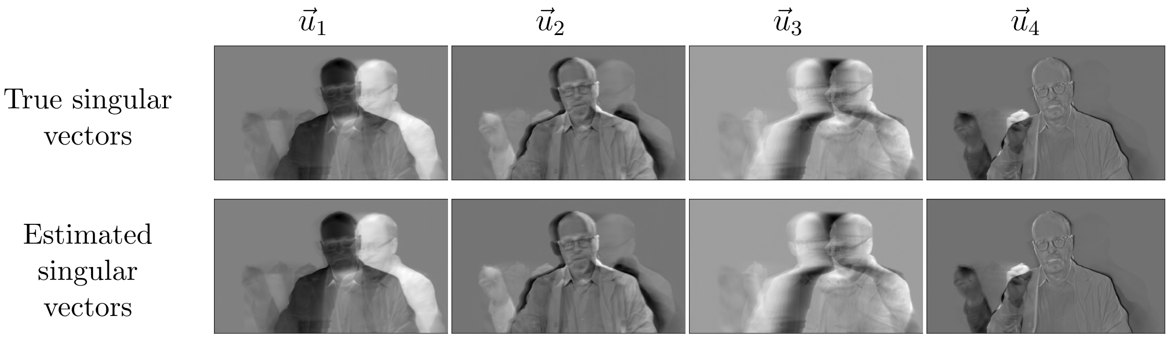

Figure 16: The first 4 left singular vectors of the movie computed using the standard SVD algorithm, and the randomized Algorithm 4.6.

5 Proofs

5.1 Proof of Theorem 2.6

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

$$\mathbf{x} - a\, 1_{\mathbf{x} \geq a} \geq 0. \tag{114}$$

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

$$\mathrm{E}(\mathbf{x}) \geq a\, \mathrm{E}(1_{\mathbf{x} \geq a}) = a\, \mathrm{P}(\mathbf{x} \geq a). \tag{115}$$

5.2 Proof of Lemma 2.8

$$\begin{aligned}
\mathrm{E}\left(\exp\left(t\mathbf{x}^2\right)\right) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) \exp\left(tu^2\right) \mathrm{d}u && (116)\\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(1-2t)\,u^2}{2}\right) \mathrm{d}u \qquad \text{finite for } 1-2t>0 && (117)\\
&= \frac{1}{\sqrt{2\pi\,(1-2t)}} \int_{-\infty}^{\infty} \exp\left(-\frac{v^2}{2}\right) \mathrm{d}v \qquad \text{change of variables } v = (2-t)\,u &&\\
&= \frac{1}{\sqrt{1-2t}}. && (118)
\end{aligned}$$

5.3 Proof of (35)

A very similar argument to the one that yields (39) gives

$$P\left(\mathbf{y} < a'\right) = P\left(\exp\left(-t'\mathbf{y}\right) > \exp\left(-a't'\right)\right) \tag{119}$$

$$\leq \exp\left(a't'\right) \prod_{i=1}^{k} \mathrm{E}\left(\exp\left(-t'\mathbf{x_i}^2\right)\right). \tag{120}$$

Setting $t' = t$ in (41), we have

$$\mathrm{E}\left(\exp\left(-t'\mathbf{x}^2\right)\right) = \frac{1}{\sqrt{1+2t'}}. \tag{121}$$

This implies

$$P\left(\mathbf{y} < a'\right) \leq \frac{\exp\left(a't'\right)}{\left(1+2t'\right)^{\frac{k}{2}}}. \tag{122}$$

Setting

$$t' := -\frac{1}{2} + \frac{1}{2\left(1-\epsilon\right)}, \tag{123}$$

$$a' := k\left(1-\epsilon\right) \tag{124}$$

we have

$$P\left(\mathbf{y} < k\left(1-\epsilon\right)\right) \leq \left(1-\epsilon\right)^{\frac{k}{2}} \exp\left(\frac{k\epsilon}{2}\right) \tag{125}$$

$$= \exp\left(-\frac{k}{2}\left(-\epsilon - \log\left(1-\epsilon\right)\right)\right). \tag{126}$$

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

$$P\left(\mathbf{y} < k\left(1-\epsilon\right)\right) \leq \exp\left(-\frac{k\epsilon^2}{2}\right) \tag{127}$$

$$\leq \exp\left(-\frac{k\epsilon^2}{8}\right). \tag{128}$$

5.4 Proof of Theorem 3.4

Let us define the sets:

$$\tilde{S}_i = S_i \cap \cap_{j=1}^{i-1} S_j^c. \tag{129}$$

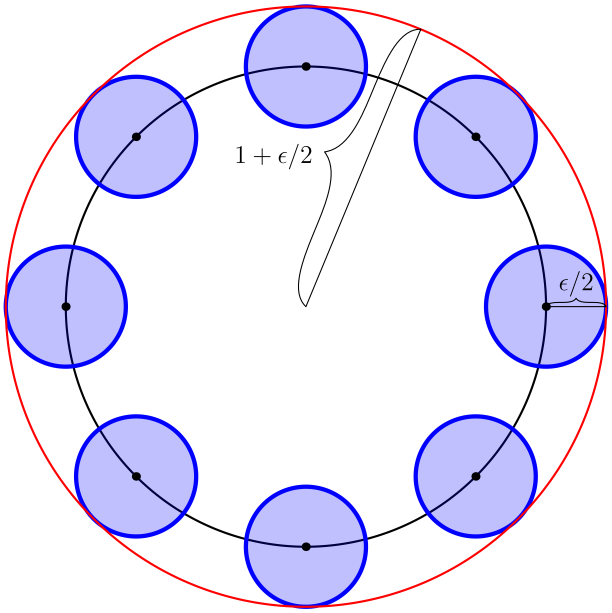

Figure 17: Sketch of the proof of Theorem 3.10 in two dimensions. $\mathcal{B}^k_{1+\epsilon/2}\left(\vec{0}\right)$ is the big red circle. The smaller shaded circles correspond to $\mathcal{B}^k_{\epsilon/2}\left(\vec{x}\right)$ for each $\vec{x}$ in the ϵ-net.

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

$$\mathrm{P}\left(\cup_i S_i\right) = \mathrm{P}\left(\cup_i \tilde{S}_i\right) = \sum_i \mathrm{P}\left(\tilde{S}_i\right) \tag{130}$$

$$\leq \sum_i \mathrm{P}\left(S_i\right) \quad \text{because } \tilde{S}_i \subseteq S_i. \tag{131}$$

5.5 Proof of Theorem 3.10

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

- We initialize $\mathcal{N}_\epsilon$ to the empty set.
- We choose a point $\vec{x} \in \mathcal{S}^{k-1}$ such that $\left|\left|\vec{x}-\vec{y}\right|\right|_2 > \epsilon$ for any $\vec{y} \in \mathcal{N}_\epsilon$. We add $\vec{x}$ to $\mathcal{N}_\epsilon$ until there are no points in $\mathcal{S}^{k-1}$ that are ϵ 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 ϵ apart and they are all inside the ball of radius $1 + \epsilon/2$ centered at the origin $\vec{0}$ because $\mathcal{N}_\epsilon \subseteq \mathcal{S}^{k-1}$. This means that

$$\begin{aligned} \operatorname{Vol}\left(\mathcal{B}^k_{1+\epsilon/2}\left(\vec{0}\right)\right) &\geq \operatorname{Vol}\left(\cup_{\vec{x}\in\mathcal{N}_\epsilon}\mathcal{B}^k_{\epsilon/2}\left(\vec{x}\right)\right) && (132)\\ &= \left|\mathcal{N}_\epsilon\right| \operatorname{Vol}\left(\mathcal{B}^k_{\epsilon/2}\left(\vec{0}\right)\right) && (133)\end{aligned}$$

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

$$\operatorname{Vol}\left(\mathcal{B}^k_r\left(\vec{0}\right)\right) = r^k \operatorname{Vol}\left(\mathcal{B}^k_1\left(\vec{0}\right)\right), \tag{134}$$

so (132) implies

$$\left(1 + \epsilon/2\right)^k \geq \left|\mathcal{N}_\epsilon\right| \left(\epsilon/2\right)^k. \tag{135}$$

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] J. Chiu and L. Demanet. Sublinear randomized algorithms for skeleton decompositions. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1361–1383, 2013.

[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] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[5] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.

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