Foundations of Data Analysis (MA4800)

Massimo Fornasier

Fakultät für Mathematik
Technische Universität München
massimo.fornasier@ma.tum.de
http://www-m15.ma.tum.de/

Slides of Lecture
June 13 2017
Preliminaries on Linear Algebra (LA)

We give for granted familiarity with the basics of LA taught in standard courses, in particular,

▶ vector spaces, spans and linear combinations, linear bases, linear maps and matrices, eigenvalues, complex numbers, scalar products, theory of symmetric matrices.

For more details we refer to the lecture notes (in German) G. Kemper, *Lineare Algebra fuer Informatik*, TUM, 2017.

of the course taught for Computer Scientists at TUM, or any other standard international text of LA.
Matrix Notations

The index sets $I, J, L, \ldots$ are assumed to be finite.

As soon as the complex conjugate values appear\(^{1}\), the scalar field is restricted to $K \in \{\mathbb{R}, \mathbb{C}\}$.

The set $K^{I \times J}$ is the vector space of matrices $A \in K^{I \times J}$, whose entries are denoted by $A_{ij}$, for $i \in I$ and $j \in J$. Vice versa, numbers $a_{ij} \in K$ may be used to define $A := (a_{ij})_{i \in I, j \in J} \in K^{I \times J}$.

If $A \in K^{I \times J}$, the transposed matrix $A^T \in K^{J \times I}$, and $A_{ji}^T := A_{ij}$. A matrix $A \in K^{I \times I}$ is symmetric if $A^T = A$. The Hermitian transposed matrix $A^H \in K^{J \times I}$ coincides with $\overline{A^T}$. If $K = \mathbb{R}$ then clearly $A^H = A^T$. A Hermitian matrix satisfies $A^H = A$.

Often, we will need to consider the rows $A^{(i)}$ and the columns $A^{(j)}$ of a matrix, defined respectively as vectors $A^{(i)} = (a_{ij})_{j \in J}$ and $A^{(j)} = (a_{ij})_{i \in I} = (A^T)^{(j)}$.

\(^{1}\)In the case $K = \mathbb{R}$, then $\alpha = \overline{\alpha}$, for all $\alpha \in K$. 
Matrix Notations

We assume that the standard matrix-vector and matrix-matrix multiplications are done in the usual way: \((Ax)_i = \sum_{j \in J} a_{ij}x_j\) or \((AB)_{i\ell} = \sum_{j \in J} A_{ij}B_{j\ell}\) as usual, for \(x \in \mathbb{K}^J\), \(A \in \mathbb{K}^{I \times J}\) and \(B \in \mathbb{K}^{J \times L}\).

The Kronecker symbol is defined by

\[
\delta_{ij} = \begin{cases} 
1, & \text{if } i = j \in I, \\
0, & \text{otherwise.}
\end{cases}
\]

The unit vector \(e^{(i)} \in \mathbb{K}^I\) is defined by

\[
e^{(i)} = (\delta_{ij})_{j \in I}
\]

The symbol \(I = (\delta_{ij})_{j \in I, j \in I}\) is used for the identity matrix. Since matrices and index sets do not appear in the same place, the simultaneous use of the symbol \(I\) do not create confusion (example: \(I \in \mathbb{K}^{I \times I}\)).
Matrix Notations

The range of matrix $A \in \mathbb{K}^{I \times J}$ is

$$\text{range}(A) = \{Ax : x \in \mathbb{K}^J\} = \text{span}\{A(i), i \in I\}.$$  

Hence, the range of a matrix is a vector space spanned by its columns.

The Euclidean scalar product in $\mathbb{K}^I$ is given by

$$\langle x, y \rangle = y^H x = \sum_{i \in I} x_i \overline{y}_i,$$  \hspace{1cm} (1)

where for $\mathbb{K} = \mathbb{R}$ the conjugate sign can be ignored.

It often useful and very important to notice that the matrix-vector $Ax$ and matrix-matrix $AB$ multiplications can also be expressed in terms of scalar products of the rows of $A$ with $x$ and of the rows of $A$ with the columns of $B$, i.e., according to our terminology of

$$(Ax)_i = \langle A^{(i)}, \overline{x} \rangle, \quad (AB)_{i \ell} = \langle A^{(i)}, \overline{B^{(\ell)}} \rangle.$$ \hspace{1cm} (2)
Two vectors $x, y \in \mathbb{K}^I$ are orthogonal (and we may write $x \perp y$) if $\langle x, y \rangle = 0$.

We often consider sets which are mutually orthogonal. In this case for $X, Y \subset \mathbb{K}^I$ we say that $X$ is orthogonal to $Y$ and we write $X \perp Y$ if $\langle x, y \rangle = 0$, for all $x \in X$ and $y \in Y$.

When $X$ and $Y$ are linear subspaces their orthogonality can be simply checked by showing that they possess bases which are mutually orthogonal. This is a very important principle we will use often.

A family of vectors $X = \{x_\nu\}_{\nu \in F} \subset \mathbb{K}^I$ is orthogonal if the vectors $x_\nu$ are pairwise orthogonal, i.e., $\langle x_\nu, x_{\nu'} \rangle = 0$ for $\nu \neq \nu'$. The family is additionally called orthonormal if $\langle x_\nu, x_\nu \rangle = 1$ for all $\nu \in F$. 
A matrix $A \in \mathbb{K}^{I \times J}$ is called orthogonal, if the columns of $A$ are orthonormal, equivalently if

$$A^H A = I \in \mathbb{K}^{J \times J}.$$ 

An orthogonal square matrix $A \in \mathbb{K}^{I \times I}$ is called unitary. Differently from just orthogonal matrices, unitary matrices satisfy

$$A^H A = AA^H = I,$$

i.e., $A^H = A^{-1}$ is the inverse matrix of $A$.

Assume that the index sets satisfy either $I \subset J$ or $J \subset I$. Then a (rectangular) matrix $A \in \mathbb{K}^{I \times J}$ is diagonal if $A_{ij} = 0$ for all $i \neq j$. 

Matrix Rank

Proposition

Let \( A \in \mathbb{K}^{I \times J} \). The following statements are equivalent and may be all used as a definition of matrix rank \( r = \text{rank}(A) \).

- (a) \( r = \dim \text{range}(A) \);
- (b) \( r = \dim \text{range}(A^H) \);
- (c) \( r \) is the maximal number of linearly independent rows of \( A \);
- (d) \( r \) is the maximal number of linearly independent columns of \( A \);
- (e) \( r \) is minimal with the property

\[
A = \sum_{i=1}^{r} a_i b_i^H, \text{ where } a_i \in \mathbb{K}^I, b \in \mathbb{K}^J;
\]

- (f) \( r \) is maximal with the property that there exists an invertible \( r \times r \) submatrix of \( A \);
- (g) \( r \) is the number of positive singular values (soon!).
The rank is bounded by the maximal rank, which, for matrices, is always given by $r_{\text{max}} = \min\{\#I, \#J\}$, and this bound is attained by full-rank matrices.

As linear independence may depend on the field $\mathbb{K}$ one may question whether the rank of a real-valued matrix depends on considering it in $\mathbb{R}$ or in $\mathbb{C}$. For matrices it turns out that it does not matter: the rank is the same.

We will often work with matrices of bounded rank $r \leq k$. We denote accordingly with $\mathcal{R}_k = \{A \in \mathbb{K}^{I \times J} : \text{range}(A) \leq k\}$ such a set. Notice that this set is not a vector space (exercise!).
In the following we consider an abstract vector space $V$ over the field $K$. As a typical example we may keep in mind the Euclidean plane $V = \mathbb{R}^2$ endowed with the classical Euclidean norm.

We recall below the axioms of an abstract norm on $V$: a norm is a map $\| \cdot \| : V \to [0, \infty)$ with the following properties

- $\|v\| = 0$ if and only if $v = 0$;
- $\|\lambda v\| = |\lambda|\|v\|$ for all $v \in V$ and $\lambda \in K$;
- $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$ (triangle inequality).

A norm is always continuous as a consequence of the (inverse) triangle inequality:

$$\|\|v\| - \|w\|\| \leq \|v - w\|, \text{ for all } v, w \in V.$$
Scalar products and (pre)-Hilbert spaces

A normed vector space \((V, \| \cdot \|)\) is a pre-Hilbert space if its norm is defined by
\[
\|v\| = \sqrt{\langle v, v \rangle}, \quad v \in V,
\]
where \(\langle \cdot, \cdot \rangle : V \times V \to \mathbb{K}\) is a scalar product on \(V\), i.e., it fulfills the properties
\begin{itemize}
  \item \(\langle v, v \rangle > 0\) for \(v \neq 0\);
  \item \(\langle v, w \rangle = \overline{\langle w, v \rangle}\), for \(v, w \in V\);
  \item \(\langle u + \lambda v, w \rangle = \langle u, w \rangle + \lambda \langle v, w \rangle\) for \(u, v, w \in V\) and \(\lambda \in \mathbb{K}\);
  \item \(\langle w, u + \lambda v \rangle = \langle w, u \rangle + \overline{\lambda} \langle w, v \rangle\) for \(u, v, w \in V\) and \(\lambda \in \mathbb{K}\).
\end{itemize}

The triangle inequality for the norm follows from the Schwarz inequality
\[
|\langle v, w \rangle| \leq \|v\| \|w\|, \quad v, w \in V.
\]
We describe the pre-Hilbert space also by the pair \((V, \langle \cdot, \cdot \rangle)\). A typical example of scalar product over \(K^I\) is the one we introduced in (1), which generates the Euclidean norm on \(K^I\):

\[
\|v\|_2 = \sqrt{\sum_{i \in I} |v_i|^2}, \quad v \in K^I.
\]

As before one can define orthogonality between vectors, orthogonal, and orthonormal sets of vectors.
A Hilbert space is a pre-Hilbert space \((V, \langle \cdot, \cdot \rangle)\), whose topology induced by the associated norm \(\| \cdot \|\) is complete, i.e., any Cauchy sequence in such vector spaces is convergent.

It is important to stress that every finite dimensional pre-Hilbert space \((V, \langle \cdot, \cdot \rangle)\) is actually complete, hence always a Hilbert space.

Thus, those who are not familiar with topological notions (e.g., completeness, Cauchy sequences, etc.), one should just reason in what follows according to their Euclidean geometric intuition.
Recall now that a set $C$ in a vector space is convex if, for all $v, w \in C$ and all $t \in [0, 1]$, $tv + (1 - tw) \in C$.

Given a closed convex set $C$ in a Hilbert space $(V, \langle \cdot, \cdot \rangle)$ one defines the projection of any vector $v$ on $C$ as

$$P_C(v) = \arg \min_{w \in C} \|v - w\|.$$

This definition is well-posed, as the projection is actually unique, and an equivalent defition is given by fulfilling the following inequality

$$\Re \langle z - P_C(v), v - P_C(v) \rangle \leq 0,$$

for all $z \in C$. This is left as an exercise.
Projections onto subspaces: Pythagoras-Fourier Theorem

Of extreme importance for us are the orthogonal projections onto subspaces.

In case $C = W \subset V$ is actually a closed linear subspace of $V$, then the projection onto $W$ can be readily computed as soon as one disposes of an orthonormal basis for $W$. Let $\{w_\nu\}_{\nu \in F}$ be a (countable) orthonormal basis for $W$ then

$$P_W(v) = \sum_{\nu \in F} \langle v, w_\nu \rangle w_\nu, \text{ for all } v \in V. \quad (3)$$

Moreover, it holds the Pythagoras-Fourier Theorem:

$$\|P_W(v)\|^2 = \sum_{\nu \in F} |\langle v, w_\nu \rangle|^2, \text{ for all } v \in V.$$
We will use very much this characterization of the orthogonal projections and the Pythagoras-Fourier Theorem, especially for the case where \( W = V \).

In this case, obviously, \( P_V = I \) and, we have the orthonormal expansion of any vector \( v \in V \),

\[
v = \sum_{\nu \in F} \langle v, w_{\nu} \rangle w_{\nu},
\]

and the norm equivalence

\[
\| v \|^2 = \sum_{\nu \in F} |\langle v, w_{\nu} \rangle|^2.
\]
The effort of defining abstract scalar products and norms allows us now to introduce several norms for matrices.

First of all, we need to introduce the concept of trace, which is the map $\text{tr} : K^{I \times I} \to K$ defined by

$$\text{tr}(A) = \sum_{i \in I} A_{ii},$$

i.e., it is the sum of the diagonal elements of the matrix $A$. 
Trace of a matrix: properties

The trace enjoyes several properties, which we collect in the following:

Proposition (Properties of the trace)

(a) $\text{tr}(AB) = \text{tr}(BA)$, for any $A \in \mathbb{K}^{I \times J}$ and $B \in \mathbb{K}^{J \times I}$;

(b) $\text{tr}(ABC) = \text{tr}(BCA)$, for any $A$, $B$, $C$ matrices of compatible size and indexes; this property is called the circularity property of the trace;

(c) as a consequence of the previous property we obtain the invariance of the trace under unitary transformations, i.e., $\text{tr}(A) = \text{tr}(UAU^H)$ for $A \in \mathbb{K}^{I \times I}$ and any unitary matrix $U \in \mathbb{K}^{I \times I}$;

(d) $\text{tr}(A) = \sum_{i \in I} \lambda_i$, where $\{\lambda_i : i \in I\}$ is the set of eigenvalues of $A$. 
Matrix norms

The Frobenius norm of a matrix is essentially the Euclidean norm computed over the entries of the matrix (considered as a vector):

$$\|A\|_F := \sqrt{\sum_{i \in I, j \in J} |A_{ij}|^2}, \quad A \in \mathbb{K}^{I \times J}.$$  

It is also known as Schur norm or Hilbert-Schmidt norm. This norm is generated by the scalar product (that's why we made the effort of introducing abstract scalar products!)

$$\langle A, B \rangle_F := \sum_{i \in I} \sum_{j \in J} A_{ij} \overline{B_{ij}} = \text{tr}(AB^H) = \text{tr}(B^HA). \quad (4)$$

In particular,

$$\|A\|_F^2 = \text{tr}(AA^H) = \text{tr}(A^HA) \quad (5)$$

holds.
Matrix norms

Let $\| \cdot \|_X$ and $\| \cdot \|_Y$ be vector norms on the vector spaces $X = \mathbb{K}^I$ and $Y = \mathbb{K}^J$, respectively. Then the associated matrix norm is

$$\|A\| := \|A\|_{X \rightarrow Y} := \sup_{z \neq 0} \frac{\|Az\|_Y}{\|z\|_X}, \quad A \in \mathbb{K}^{I \times J}.$$  

If both $\| \cdot \|_X$ and $\| \cdot \|_Y$ coincides with the Euclidean norms $\| \cdot \|_2$ on $X = \mathbb{K}^I$ and $Y = \mathbb{K}^J$, respectively, then the associated matrix norm is called the spectral norm and it is denoted with $\|A\|_\infty$.

As the spectral norm has a central importance in many applications, it is often denoted just by $\|A\|$.
Unitary invariance and submultiplicativity

Both the matrix norms we introduced so far are invariant with respect to unitary transformations, i.e., $\|A\|_F = \|UAV^H\|_F$ and $\|A\|_\infty = \|UAV^H\|_\infty$ for unitary matrices $U, V$.

Moreover, both are submultiplicative, i.e.,

$\|AB\|_F \leq \|A\|_\infty \|B\|_F \leq \|A\|_F \|B\|_F$ and $\|AB\| \leq \|A\| \|B\|$, for matrices $A, B$ of compatible sizes.
We introduce and analyze the singular value decomposition (SVD) of a matrix $A$, which is the factorization of $A$ into the product of three matrices $A = UΣV^H$, where $U$, $V$ are orthogonal matrices of compatible size and the matrix $Σ$ is diagonal with positive real entries.

All these terms, orthogonal, diagonal matrix, have been introduced in the previous lectures.
To gain insight into the SVD, treat the $n = \#I$ rows of a matrix $A \in \mathbb{K}^{I \times J}$ as points in a $d$-dimensional space, where $d = \#J$, and consider the problem of finding the best $k$-dimensional subspace with respect to the set of points.

Here best means minimize the sum of the squares of the perpendicular distances of the points to the subspace. An orthonormal basis for this subspace is built as fundamental directions with maximal variance of the group of high dimensional points, and are called principal components.
Geometrical derivation: principal component analysis
We begin with a special case of the problem where the subspace is 1-dimensional, a line through the origin.

We will see later that the best-fitting $k$-dimensional subspace can be found by $k$ applications of the best fitting line algorithm.

Finding the best fitting line through the origin with respect to a set of points $\{x_i := A(i) \in \mathbb{K}^2 : i \in I\}$ in the Euclidean plane $\mathbb{K}^2$ means minimizing the sum of the squared distances of the points to the line. Here distance is measured perpendicular to the line. The problem is called the best least squares fit.
Pythagoras Theorem, scalar products, and orthogonal projections

In the best least squares fit, one is minimizing the distance to a subspace. Now, consider projecting orthogonally a point $x_i$ onto a line through the origin. Then, by Pythagoras theorem

$$x_{i1}^2 + x_{i2}^2 + \cdots + x_{id}^2 = \text{(length of projection)}^2 + \text{(distance of pt. to line)}^2.$$
Pythagoras Theorem, scalar products, and orthogonal projections

In particular, from the formula above, one has

\[(\text{distance of pt to line})^2 = x_{i1}^2 + x_{i2}^2 + \cdots + x_{id}^2 - (\text{length of projection})^2.\]

To minimize the sum of the squares of the distances to the line, one could minimize \(\sum_i (x_{i1}^2 + x_{i2}^2 + \cdots + x_{id}^2)\) minus the sum of the squares of the lengths of the projections of the points to the line.

However, the first term of this difference is a constant (independent of the line), so minimizing the sum of the squares of the distances is equivalent to maximizing the sum of the squares of the lengths of the projections onto the line.
Pythagoras Theorem, scalar products, and orthogonal projections
For best-fit subspaces, we could maximize the sum of the squared lengths of the projections onto the subspace instead of minimizing the sum of squared distances to the subspace.

Consider the rows of $A$ as $n = \#I$ rows of a matrix $A \in \mathbb{K}^{I \times J}$ as points in a $d$-dimensional space, where $d = \#J$. Consider the best-fit line through the origin.

Let $v$ be a unit vector along this line. The length of the projection of $A^{(i)}$, the $i^{th}$ row of $A$, onto $v$ is, according to our definition of scalar product, $|\langle A^{(i)}, v \rangle|$.

From this, we see that the sum of length squared of the projections is

$$\|Av\|_2^2 = \sum_{i \in I} |\langle A^{(i)}, v \rangle|^2. \quad (6)$$
Best-fit and first singular direction

The best-fit line is the one maximizing $\|A v\|_2^2$ and hence minimizing the sum of the squared distances of the points to the line.

With this in mind, define the first singular vector, $v_1$ of $A$, which is a column vector, as the best-fit line through the origin for the $n$ points in $d$-dimensional space that are the rows of $A$.

Thus

$$v_1 = \arg \max_{\|v\|_2 = 1} \|A v\|_2.$$ 

The value $\sigma_1(A) = \|A v_1\|_2$ is called the first singular value of $A$. Note that $\sigma_1(A)^2$ is the sum of the squares of the projections of the points to the line determined by $v_1$. 
This is the first time we use an optimization criterion to approach a data interpretation problem and to create a connection between linear algebra and optimization.

As we will discuss along this course, optimization plays indeed a huge role in data analysis and we will have to discuss it in more detail later on.
A greedy approach to best $k$-dimensional fit

The greedy approach to find the best-fit 2-dimensional subspace for a matrix $A$, takes $v_1$ as the first basis vector for the 2-dimensional subspace and finds the best 2-dimensional subspace containing $v_1$.

The fact that we are using the sum of squared distances will again help, thanks to Pythagoras Theorem: For every 2-dimensional subspace containing $v_1$, the sum of squared lengths of the projections onto the subspace equals the sum of squared projections onto $v_1$ plus the sum of squared projections along a vector perpendicular to $v_1$ in the subspace.

Thus, instead of looking for the best 2-dimensional subspace containing $v_1$, look for a unit vector, call it $v_2$, perpendicular to $v_1$ that maximizes $\|Av\|_2$ among all such unit vectors.

Using the same greedy strategy to find the best three and higher dimensional subspaces, defines $v_3, v_4, \ldots$ in a similar manner.
Formal definitions

More formally, the second singular vector, $v_2$ is defined by the best-fit line perpendicular to $v_1$:

$$v_2 = \arg \max_{\|v\|_2 = 1, \langle v_1, v \rangle = 0} \|Av\|_2.$$ 

The value $\sigma_2(A) = \|Av_2\|_2$ is called the second singular value of $A$.

The $k^{th}$ singular vector $v_k$ is defined similarly by

$$v_k = \arg \max_{\|v\|_2 = 1, \langle v_1, v \rangle = 0, \ldots, \langle v_{k-1}, v \rangle = 0} \|Av\|_2. \quad (7)$$

and so on.

The process stops when we have found $v_1, \ldots, v_r$ as singular vectors and

$$0 = \arg \max_{\|v\|_2 = 1, \langle v_1, v \rangle = 0, \ldots, \langle v_r, v \rangle = 0} \|Av\|_2.$$
Optimality of the greedy algorithm

If instead of finding $v_1$ that maximized $\|Av\|_2$ and then the best-fit 2-dimensional subspace containing $v_1$, we had found the best-fit 2-dimensional subspace, we might have done better.

Surprisingly enough, this is actually not the case. We now give a simple proof that the greedy algorithm indeed finds the best subspaces of every dimension.

Proposition

Let $A \in \mathbb{K}^{I \times J}$ and $v_1, v_2, \ldots, v_r$ be the singular vectors defined above. For $1 \leq k \leq r$, let $V_k$ be the subspace spanned by $v_1, v_2, \ldots, v_k$. Then for each $k$, $V_k$ is the best-fit $k$-dimensional subspace for $A$. 
We conclude this part with an important property left as an exercise: show that necessarily

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_r(A),$$

i.e., the singular values come always with a natural nonincreasing order.
Note that the vector $Av_i$ is really a list of lengths (with signs) of the projections of the rows of $A$ onto $v_i$. Think of $\sigma_i(A) = \|Av_i\|_2$ as the component of the matrix $A$ along $v_i$.

For this interpretation to make sense, it should be true that adding up the squares of the components of $A$ along each of the $v_i$ gives the square of the whole content of the matrix $A$.

This is indeed the case and is the matrix analogy of decomposing a vector into its components along orthogonal directions.
On matrix norms again

Consider one row, say $A^{(i)}$ of $A$. Since $v_1, v_2, \ldots, v_r$ span the space of all rows of $A$ (exercise!), $\langle A^{(i)}, v \rangle = 0$ for all $v$ orthogonal to $v_1, v_2, \ldots, v_r$.

Thus, for each row $A^{(i)}$, by Pythagoras Theorem,

$$\|A^{(i)}\|_2^2 = \sum_{k=1}^{r} |\langle A^{(i)}, v_k \rangle|^2.$$

Summing over all rows $i \in I$ and recalling (6), we obtain

$$\sum_{i \in I} \|A^{(i)}\|_2^2 = \sum_{k=1}^{r} \sum_{i \in I} |\langle A^{(i)}, v_k \rangle|^2 = \sum_{k=1}^{r} \|Av_k\|_2^2 = \sum_{k=1}^{r} \sigma_k(A)^2.$$

But

$$\sum_{i \in I} \|A^{(i)}\|_2^2 = \sum_{i \in I} \sum_{j \in J} |A_{ij}|^2,$$

is the Frobenius norm of $A$, so that we obtained

$$\|A\|_F = \sqrt{\sum_{k=1}^{r} \sigma_k(A)^2}. \quad (9)$$
On matrix norms again

We might also observe that

\[ \sigma_1(A) = \max_{\|v\|_2=1} \|Av\|_2 = \arg \max_{v \neq 0} \frac{\|Av\|_2}{\|v\|_2} = \|A\|, \quad (10) \]

i.e., the first singular vector corresponds to the spectral norm of \( A \). This allows us to define other and more general norms, called the Schatten-\( p \)-norms for matrices defined for \( 1 \leq p < \infty \) by

\[ \|A\|_p = \left( \sum_{k=1}^{r} \sigma_k(A)^p \right)^{1/p}, \]

and for \( p = \infty \)

\[ \|A\|_\infty = \max_{k=1,\ldots,r} \sigma_k(A). \]

Notice that these definitions gives precisely \( \|A\|_F = \|A\|_2 \) and \( \|A\| = \|A\|_\infty \) as particular Schatten-norms. Of particular relevance in certain applications related to recommender systems is the so-called nuclear norm \( \|A\|_* = \|A\|_1 \) corresponding to the Schatten-1-norm.
We just mentioned above, and left it as an exercise, that the orthogonal basis constituted by the vectors $v_1, v_2, \ldots, v_r$ spans the space of all rows of $A$.

This means that the number $r$ (of them) is actually the rank of the matrix! And this a proof of (g) in the Proposition above characterizing the rank!
The vectors $Av_1, Av_2, \ldots, Av_r$ form yet another fundamental set of vectors associated with $A$: We normalize them to length one by setting

$$u_i = \frac{1}{\sigma_i(A)} Av_i$$

The vectors $u_1, u_2, \ldots, u_r$ are called the left singular vectors of $A$. The $v_i$ are called the right singular vectors.

The SVD theorem (soon!) will fully explain the reason for these terms.
Orthonormality of the left singular vectors

Clearly, the right singular vectors are orthogonal by definition.

We now show that the left singular vectors are also orthogonal.

Theorem
Let $A \in \mathbb{K}^{I \times J}$ be a rank $r$ matrix. The left singular vectors of $A$, $u_1, u_2, \ldots, u_r$ are orthogonal.
SVD in its full glory

Theorem

Let $A \in \mathbb{K}^{I \times J}$ with right singular vectors $v_1, v_2, \ldots, v_r$, left singular vectors $u_1, u_2, \ldots, u_r$, and corresponding singular values $\sigma_1, \ldots, \sigma_r$. Then

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^H$$
Best rank-$k$ approximation

Given the singular value decomposition $A = \sum_{k=1}^{r} \sigma_k u_k v_k^H$ of a matrix $A \in \mathbb{K}^{I \times J}$, we define the $k$-rank truncation by

$$A_k = \sum_{\ell=1}^{k} \sigma_\ell u_\ell v_\ell^H.$$  

It should be clear that $A_k$ is a $k$-rank matrix.

**Lemma**

The rows of $A_k$ are the orthogonal projections of the rows of $A$ onto the subspace $V_k$ spanned by the first $k$ singular vectors of $A$. 
Best rank-$k$ Frobenius approximation

**Theorem**

For any matrix $B$ of rank at most $k$

$$\|A - A_k\|_F \leq \|A - B\|_F$$
Best rank-$k$ spectral approximation

**Lemma**
\[ \|A - A_k\|^2 = \sigma_{k+1}^2. \]

**Theorem**
*For any matrix $B$ of rank at most $k$*

\[ \|A - A_k\| \leq \|A - B\|. \]
Nonconvexity and curse of dimensionality

We introduced the SVD by means of a greedy algorithm, which at each step is supposed to solve a nonconvex optimization problem of the type

\[ v_k = \arg \max_{\|v\|_2=1, \langle v_1, v \rangle = \ldots = \langle v_{k-1}, v \rangle = 0} \|Av\|_2. \] (11)

Well, the issue with this is that it is neither a convex optimization problem nor, in principle, stated on a low dimensional space. In fact our goal is often to deal with “big data” (i.e., many data points, each of very high dimension), which eventually implies large dimensionalities of the matrix \( A \).

How can we then approach the computation of the SVD without incurring in the so called “curse of dimensionality”, which is an emphatic way of describing the computational infeasibility of a certain numerical task?
The Power Method

It is easiest to describe first in the case when $A$ is real-valued, square, and symmetric and has the same right and left singular vectors, namely, $A = \sum_{k=1}^{r} \sigma_k v_k v_k^T$.

In this case we have

$$A^2 = \left( \sum_{k=1}^{r} \sigma_k v_k v_k^T \right) \left( \sum_{\ell=1}^{r} \sigma_{\ell} v_{\ell} v_{\ell}^T \right) = \sum_{k,\ell=1}^{r} \sigma_{\ell} \sigma_k v_k v_k^T v_{\ell} v_{\ell}^T = \sum_{k=1}^{r} \sigma_k^2 v_k v_k^T.$$

Similarly, if we take the $m^{th}$ power of $A$, again all the cross terms are zero and we will get

$$A^m = \sum_{k=1}^{r} \sigma_k^m v_k v_k^T.$$
We had the spectral norm of $A$ and a spectral gap

If we had $\sigma_1 \gg \sigma_2$, we would have

$$\lim_{m \to \infty} \frac{A^m}{\sigma_1^m} = v_1 v_1^T.$$

While it is not so simple to compute $\sigma_1$ (why?!), which corresponds to the spectral norm of $A$, one can easily compute the Frobenius norm $\|A^m\|_F$, so that we can consider $\frac{A^m}{\|A^m\|_F}$ which again converges for $m \to \infty$ to $v_1 v_1^T$ from which $v_1$ may be computed (exercise!).
What if $A$ is not squared?

But then $B = AA^T$ is squared. If again, the SVD of $A$ is $A = \sum_{k=1}^{r} \sigma_k u_k v_k^T$ then

$$B = \sum_{k=1}^{r} \sigma_k^2 u_k u_k^T.$$  

This is the spectral decomposition of $B$. Using the same kind of calculation as above, one obtains

$$B^m = \sum_{k=1}^{r} \sigma_k^{2m} u_k u_k^T.$$  

As $m$ increases, for $k > 1$ the ratio $\sigma_k^{2m}/\sigma_1^{2m}$ goes to zero and $B^m$ gets approximately equal to

$$\sigma_1^{2m} u_1 u_1^T,$$

provided again that $\sigma_1 \gg \sigma_2$. This suggests how to compute both $\sigma_1$ and $u_1$, simply by powering $B$.  

Two relevant issues

- If we do not have a spectral gap $\sigma_1 \gg \sigma_2$ we get into troubles.
- Computing $B^m$ costs $m$ matrix-matrix multiplication, when done in the schoolbook way it costs $O(md^3)$ operations or $O(m \log d)$ when done by successive squaring.
Use a **random** vector application instead!

Instead, we compute

\[ B^m x, \]

where \( x \) is a random unit length vector.

The idea is that the component of \( x \) in the direction of \( u_1 \), i.e.,

\[ x_1 := \langle x, u_1 \rangle \]

is actually bounded away from zero with high probability and would get multiplied by \( \sigma_1^2 \), while the other components of \( x \) along other \( u_i \)'s would be multiplied by \( \sigma_k^2 \ll \sigma_1^2 \) only.

Each increase in \( m \) requires multiplying \( B \) by the vector \( B^{m-1} x \), which we can further break up into

\[ B^m x = A (A^T (B^{m-1} x)). \]

This requires two matrix vector products, involving the matrices \( A^T \) and \( A \). Since \( B^m x \approx \sigma_1^{2k} u_1 (u_1^T x) \) is a scalar multiple of \( u_1 \), \( u_1 \) can be recovered from \( B^m x \) by normalization.
Concentration of measure phenomenon

The concentration of measure phenomenon informally speaking is describing the fact that - in high-dimension - certain events happen with high-probability. In particular, we have

Lemma
Let $x \in \mathbb{R}^d$ be a unit d-dimensional vector of components $x = (x_1, \ldots, x_d)$ with respect to the canonical basis and picked at random from the set $\{x : \|x\|_2 \leq 1\}$. The probability that $|x_1| \geq \alpha > 0$ is at least $1 - 2\alpha \sqrt{d - 1}$. 
Extensions

Remark
Notice that in the previous result essentially shows also that, independently of the dimension $d$, the $x_1 = \langle x, u_1 \rangle$ component of a random unit vector $x$ with respect to any orthonormal basis $\{u_1, \ldots, u_d\}^2$ is bounded away from zero with overwhelming probability.

Remark
In view of the isometrical mapping $(a, b) \to a + ib$ from $\mathbb{R}^2$ to $\mathbb{C}$ the previous result extends to random unit vectors in $\mathbb{C}^d$ simply by modifying the statement as follows: The probability that, for a randomly chosen unit vector $z \in \mathbb{C}^d$ $|z_1| \geq \alpha > 0$ holds is at least $1 - 2\alpha \sqrt{2d - 1}$.

\[\text{Since the sphere is rotation invariant, the arguments above apply to any orthonormal basis by rotating it to coincide with the canonical basis!}\]
Randomized Power Method

By injecting randomization in the process, we are able to use the blessing of the dimensionality (concentration of measure) against the curse of dimensionality.

**Theorem**

Let $A \in \mathbb{K}^{I \times J}$ and $x \in \mathbb{K}^I$ be a random unit length vector. Let $V$ be the space spanned by the left singular vectors of $A$ corresponding to singular values greater than $(1 - \varepsilon)\sigma_1$. Let $m$ be $\Omega\left(\frac{\ln(d/\varepsilon)}{\varepsilon}\right)$. Let $w^*$ be unit vector after $m$ iterations of the power method, namely,

$$w^* = \frac{(AA^H)^m x}{\| (AA^H)^m x \|_2}.$$

The probability that $w^*$ has a component of at least $O\left(\frac{\varepsilon}{\alpha d}\right)$ orthogonal to $V$ is at most $1 - 2\alpha \sqrt{2d - 1}$.

To make the previous statement concrete, let's take $\alpha = \frac{1}{10/2d}$. Then, with probability at least $9/10$ (i.e., very close to 1), after $m$ iteration of order $\Omega\left(\frac{\ln(d/\varepsilon)}{\varepsilon}\right)$ the vector $w^*$ has a component of at least $O(\varepsilon)$ orthogonal to $V$. 
Applications of the SVD: Principal Component Analysis

The traditional use of SVD is in Principal Component Analysis (PCA). PCA is illustrated by an example - customer-product data where there are $n$ customers buying $d$ products.
Let matrix $A$ with elements $A_{ij}$ represent the probability of customer $i$ purchasing product $j$ (or the amount or utility of product $j$ to customer $i$).

One hypothesizes that there are really only $k$ underlying basic factors like age, income, family size, etc. that determine a customer’s purchase behavior.

An individual customer’s behavior is determined by some weighted combination of these underlying factors.
Principal Component Analysis

That is, a \( i \) customer’s purchase behavior can be characterized by a \( k \)-dimensional vector \((u_{\ell,i})_{\ell=1,...,k}\), where \( k \) is much smaller than \( n \) and \( d \).

The components \( u_{\ell,i} \), for \( \ell = 1, \ldots, k \) of the vector are weights for each of the basic factors.

Associated with each basic factor \( \ell \) is a vector of probabilities \( v_{\ell} \), each component of which is the probability of purchasing a given product by someone whose behavior depends only on that factor.

More abstractly, \( A = UV^H \) is an \( n \times d \) matrix that can be expressed as the product of two matrices \( U \) and \( V \) where \( U \) is an \( n \times k \) matrix expressing the factor weights for each customer and \( V \) is a \( k \times d \) matrix expressing the purchase probabilities of products that correspond to that factor:

\[
(A^{(i)})^H = \sum_{\ell=1}^{k} u_{\ell,i} v_{\ell}^H.
\]
One twist is that $A$ may not be exactly equal to $UV^T$, but close to it since there may be noise or random perturbations.

Taking the best rank-$k$ approximation $A_k$ from SVD (as described above) gives us such a $U, V$. In this traditional setting, one assumed that $A$ was available fully and we wished to find $U, V$ to identify the basic factors or in some applications to denoise $A$ (if we think of $A - UV^T$ as noise).
Low rankness and recommender systems

Now imagine that $n$ and $d$ are very large, on the order of thousands or even millions, there is probably little one could do to estimate or even store $A$.

In this setting, we may assume that we are given just given a few elements of $A$ and wish to estimate $A$. If $A$ was an arbitrary matrix of size $n \times d$, this would require $\Omega(nd)$ pieces of information and cannot be done with a few entries.

But again hypothesize that $A$ was a small rank matrix with added noise. If now we also assume that the given entries are randomly drawn according to some known distribution, then there is a possibility that SVD can be used to estimate the whole of $A$.

This area is called collaborative filtering or low-rank matrix completion, and one of its uses is to target an ad to a customer based on one or two purchases.
SVD as a compression algorithm

Suppose $A$ is the pixel intensity matrix of a large image. The entry $A_{ij}$ gives the intensity of the $ij^{th}$ pixel. If $A$ is $n \times n$, the transmission of $A$ requires transmitting $O(n^2)$ real numbers. Instead, one could send $A_k$, that is, the top $k$ singular values $\sigma_1, \sigma_2, \ldots, \sigma_k$ along with the left and right singular vectors $u_1, u_2, \ldots, u_k$, and $v_1, v_2, \ldots, v_k$.

This would require sending $O(kn)$ real numbers instead of $O(n^2)$ real numbers.

If $k$ is much smaller than $n$, this results in savings. For many images, a $k$ much smaller than $n$ can be used to reconstruct the image provided that a very low resolution version of the image is sufficient.

Thus, one could use SVD as a compression method.
SVD as a compression algorithm
Sparsifying bases

It turns out that in a more sophisticated approach, for certain classes of pictures one could use a fixed sparsifying basis so that the top (say) hundred singular vectors are sufficient to represent any picture approximately.

This means that the space spanned by the top hundred sparsifying basis vectors is not too different from the space spanned by the top two hundred singular vectors of a given matrix in that class.

Compressing these matrices by this sparsifying basis can save substantially since the sparsifying basis is transmitted only once and a matrix is transmitted by sending the top several hundred singular values for the sparsifying basis.

We will discuss later on the use of sparsifying bases.
Consider representing a document by a vector each component of which corresponds to the number of occurrences of a particular word in the document.

The English language has on the order of 25,000 words. Thus, such a document is represented by a 25,000-dimensional vector. The representation of a document is called the word vector mode.

A collection of $n$ documents may be represented by a collection of 25,000-dimensional vectors, one vector per document. The vectors may be arranged as rows $n \times 25,000$ matrix.
Singular Vectors and ranking documents

An important task for a document collection is to rank the documents.

A naive method would be to rank in order of the total length of the document (which is the sum of the components of its term-document vector). Clearly, this is not a good measure in many cases.

This naive method attaches equal weight to each term and essentially takes the projection of each term-document vector in the direction of the vector whose components have just value 1.

Is there a better weighting of terms, i.e., a better projection direction which would measure the intrinsic relevance of the document to the collection?
A good candidate is the best-fit direction for the collection of term-document vectors, namely the top (left) singular vector of the term-document matrix.

An intuitive reason for this is that this direction has the maximum sum of squared projections of the collection and so can be thought of as a synthetic term-document vector best representing the document collection.
Ranking in order of the projection of each document’s term vector along the best-fit direction has a nice interpretation in terms of the power method. For this, we consider a different example - that of web with hypertext links.

The World Wide Web can be represented by a directed graph whose nodes correspond to web pages and directed edges to hypertext links between pages.

Some web pages, called authorities, are the most prominent sources for information on a given topic. Other pages called hubs, are ones that identify the authorities on a topic.

Authority pages are pointed to by many hub pages and hub pages point to many authorities.
WWW ranking

One would like to assign hub weights and authority weights to each node of the web.

If there are $n$ nodes, the hub weights form a $n$-dimensional vector $u$ and the authority weights form a $n$-dimensional vector $v$.

Suppose $A$ is the adjacency matrix representing the directed graph: $A_{ij}$ is 1 if there is a hypertext link from page $i$ to page $j$ and 0 otherwise. Given hub vector $u$, the authority vector $v$ could be computed by the formula

$$v_j = \sum_{i=1}^{n} u_i a_{ij},$$

since the right hand side is the sum of the hub weights of all the nodes that point to node $j$. In matrix terms,

$$v = A^T u.$$ 

Similarly, given an authority vector $v$, the hub vector could be computed by $u = Av$. 
WWW ranking

Of course, at the start, we have neither vector.

But the above suggests a power-like iteration: Start with any $v$. Set $u = Av$; then set $v = A^T u$, and repeat the process.

We know from the power method that this converges to the left and right singular vectors. So, after sufficiently many iterations, we may use the left vector $u$ as hub weights vector and project each column of $A$ onto this direction and rank columns (authorities) in order of this projection.

But the projections just form the vector $A^T u$ which equals $v$. So we can just rank by order of $v_j$.

This is the basis of an algorithm called the HITS algorithm which was one of the early proposals for ranking web pages. A different ranking called page rank is widely used. It is based on a random walk on the graph described above, but we do not explore it here.
Given $A \in \mathbb{K}^{I \times J}$ of rank $r$ and its singular value decomposition

$$A = U \Sigma V^H = \sum_{k=1}^{r} \sigma_k u_k v_k^H,$$

we define the Moore-Penrose pseudo-inverse $A^\dagger \in \mathbb{K}^{J \times I}$ as

$$A^\dagger = \sum_{k=1}^{r} \sigma_k^{-1} v_k u_k^H.$$

In matrix form it is

$$A^\dagger = V \Sigma^{-1} U^H,$$

where the inverse $\cdot^{-1}$ is applied only to the nonzero singular values of $A$. 

Pseudo-inverse matrices
Pseudo-inverse matrices

If $A^H A \in \mathbb{K}^{J \times J}$ is invertible (implying $\# I \geq \# J$) and this is true if $r = \# J$ and the columns are linearly independent, then

$$A^\dagger = (A^H A)^{-1} A^H.$$

Indeed

$$(A^H A)^{-1} A^H = (V \Sigma^2 V^H)^{-1} V \Sigma U^H = (V \Sigma^{-2} V^H) V \Sigma U^H = V \Sigma^{-1} U^H = A^\dagger.$$ 

In this case, it should be also noted two fundamental properties of $A^\dagger$. First of all, it is a left inverse of $A$, i.e.,

$$I = (A^H A)^{-1} A^H A = A^\dagger A,$$

where $I \in \mathbb{K}^{r \times r}$. On the other side the matrix

$$P_{\text{range}(A)} = AA^\dagger,$$

corresponds to the orthogonal projection of $\mathbb{K}^I$ onto the range of the matrix $A$, i.e., the span on the columns of $A$. 
Similarly, if $AA^H \in \mathbb{K}^{I \times I}$ is invertible (implying $\# I \leq \# J$) and this is true if $r = \# I$ and the rows are linearly independent, then

$$A^\dagger = A^H (AA^H)^{-1}.$$ 

In this case, $A^\dagger$ is a right inverse of $A$, i.e.,

$$I = AA^H (AA^H)^{-1} = AA^\dagger.$$
Least square problems

- The system $Ax = y$ is overdetermined, i.e., there are more equations than unknowns. In this case the problem might even not be solvable and approximated solutions are needed, i.e., those that minimize the mismatch $\|Ax - y\|_2$;

- The system $Ax = y$ is underdetermined, i.e., there are less equations than unknowns. In this case the problem has an infinite number of solutions and one would wish to determine one of them with special properties, by means of a proper selection criterion. A criterion of selection of proper solutions of the systems $Ax = y$ is often to pick the solution with minimal Euclidean norm $\|x\|_2$. 
Pseudo inverse matrices and Least square problems

Let us first connect least squares problems with the Moore-Penrose pseudo-inverse introduced above.

Proposition

Let $A \in \mathbb{K}^{I \times J}$ and $y \in \mathbb{K}^I$. Define $\mathcal{M} \subset \mathbb{K}^J$ to be the set of minimizers of the map $x \rightarrow \|Ax - y\|_2$. The convex optimization problem

$$\arg \min_{x \in \mathcal{M}} \|x\|_2$$

(12)

has the unique solution $x^* = A^\dagger y$.

Proof at the blackboard ...
Overdetermined systems

Corollary

Let $A \in K^{I \times J}$ with $n = \#I \geq \#J = d$ be of full rank $r = d$, and let $y \in K^I$. The least square problem

$$\arg \min_{x \in K^J} \|Ax - y\|_2,$$

has the unique solution $x = A^\dagger y$.

This result follows from the Proposition above because the minimizer $x^*$ of $\|Ax - y\|_2$ is in this case unique.
Corollary

Let $A \in \mathbb{K}^{I \times J}$ with $n = \#I \leq \#J = d$ be of full rank $r = n$, and let $y \in \mathbb{K}^I$. The least square problem

$$\arg \min_{x \in \mathbb{K}^J} \|x\|_2 \text{ subject to } Ax = y,$$

has the unique solution $x = A^\dagger y$. (14)
Stability of the Singular Value Decomposition

In many situations, there is a data matrix of interest $A$, but one has only a perturbed version of it $\tilde{A} = A + E$ where $E$ is going to be a relatively small quantifiable perturbation. To which extent can we rely on the singular value decomposition of $\tilde{A}$ to estimate the one of $A$?

In many real-life cases, certain sets of data tend to behave as random vectors generated in the vicinity of a lower dimensional linear subspace.

The more data we get the more they take a specific enveloping shape around that subspace. Hence, we could consider many matrices built out of sets of such randomly generated points. How do the SVD’s of such matrices look like?

What happens if we acquire more and more data and the matrices become larger and larger?

Also to be able to respond in a quantitative way to such questions we need to develop a theory of stability of singular value decompositions under small perturbations.
Spectral theorem

$A \in \mathbb{K}^{I \times I}$ is actually a Hermitian matrix, i.e. $A = A^H$. We recall that a nonzero vector $v \in \mathbb{K}^I$ is an eigenvector of $A$ if $Av = \lambda v$ for some scalar $\lambda \in \mathbb{K}$ called the corresponding eigenvalue.

**Theorem (Spectral theorem for Hermitian matrices)**

If $A \in \mathbb{K}^{I \times I}$ and $A = A^H$, then there exists an orthonormal basis $\{v_1, \ldots, v_n\}$ consisting of eigenvectors of $A$ with real corresponding eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ such that

$$A = \sum_{k=1}^{n} \lambda_k v_k v_k^H.$$

The previous representation is called the spectral decomposition of $A$.

In fact, by denoting $u_k = \text{sign} \lambda_k$ and $\sigma_k = |\lambda_k|$, then

$$A = \sum_{k=1}^{n} \lambda_k v_k v_k^H = \sum_{k=1}^{n} \text{sign} \lambda_k |\lambda_k| v_k v_k^H = \sum_{k=1}^{n} \sigma_k u_k v_k^H,$$

which is actually the SVD of $A$. 
Weyl’s bounds

Assume now \( E \) is Hermitian so that \( \tilde{A} = \tilde{A}^H \) holds, as well. As the eigenvalues of such matrices are always real, we can order in a nonincreasing order so that \( \lambda_1 \) is the largest (non necessarily positive!). Then

\[
\lambda_1(\tilde{A}) = \max_{\|v\|_2=1} v^H(A + E)v \\
\leq \max_{\|v\|_2=1} v^HAv + \max_{\|v\|_2=1} v^HEv \\
= \lambda_1(A) + \lambda_1(E)
\]

By using the spectral decomposition of \( A + E \) do prove as an exercise the first equality above.
Weyl’s bounds

Let $v_1$ be the eigenvector of $A$ associated to the top eigenvalue of $A$ and

$$
\lambda_1(\tilde{A}) = \max_{\|v\|_2=1} v^H (A + E) v
$$

$$
\geq v_1^H (A + E) v_1
$$

$$
= \lambda_1(A) + v_1^H E v_1
$$

$$
\geq \lambda_1(A) + \lambda_n(E).
$$

From these estimates we obtain the bounds

$$
\lambda_1(A) + \lambda_n(E) \leq \lambda_1(A + E) \leq \lambda_1(A) + \lambda_1(E).
$$
Weyl’s bounds

This chain of inequalities can be properly extended to the 2nd, 3rd, etc. eigenvalues (we do not perform this extension), leading to the following important result.

Theorem (Weyl)

If $A, E \in \mathbb{K}^{I \times I}$ are two Hermitian matrices, then for all $k = 1, \ldots, n$

$$\lambda_k(A) + \lambda_n(E) \leq \lambda_k(A + E) \leq \lambda_k(A) + \lambda_1(E).$$

This is a very strong result because it does not depend at all from the magnitude of the perturbation matrix $E$. As a corollary

Corollary

If $A, E \in \mathbb{K}^{I \times I}$ are two arbitrary matrices, then for all $k = 1, \ldots, n$

$$|\sigma_k(A + E) - \sigma_k(A)| \leq \|E\|,$$

where $\sigma_k$’s are the corresponding singular vectors and $\|E\|$ is the spectral norm of $E$. 
Mirsky’s bounds

Actually more can be said:

**Theorem (Mirsky)**

*If $A, E \in K^{I \times I}$ are two arbitrary matrices, then*

$$\sqrt{\sum_{k=1}^{n} |\sigma_k(A + E) - \sigma_k(A)|^2} \leq \|E\|_F,$$

*where $\|E\|_F$ is the Frobenius norm of $E$.*

In its original form, Mirsky’s theorem holds for an arbitrary unitarily invariant norm (hence for the spectral norm) and includes Weyl’s theorem as a special case.
To quantify the stability properties of the singular value decomposition under perturbations, we need to introduce the notion of principal angles between subspaces. To introduce the concept we recall that in a Euclidean space $\mathbb{R}^d$ the angle $\theta(v, w)$ of two vectors $v, w \in \mathbb{R}^d$ is defined by

$$
\cos \theta(v, w) = \frac{w^T v}{\|v\|_2 \|w\|_2}.
$$

In particular, for unitary vectors $v, w$ one has $\cos \theta(v, w) = w^T v$. 
Stability of singular spaces: principal angles

Definition
Let \( V, W \in \mathbb{K}^{J \times I} \) be orthogonal matrices spanning \( V \) and \( W \), \( n \)-dimensional subspaces of \( \mathbb{K}^J \) (here we may assume that \( n = \#I \leq \#J = d \))\(^3\). We define the principal angles between \( V \) and \( W \) by
\[
\cos \Theta(V, W) = \Sigma(W^H V),
\]
where \( \Sigma(W^H V) \) is the list of the singular values of the matrix \( W^H V \).

\(^3\)Here we committed an abuse of notation as we denoted with \( V, W \) the matrices of the orthonormal bases of two subspaces \( V, W \). However, this identification here is legitimated as all we need to know about the subspaces are some orthonormal bases of them.
Stability of singular spaces: principal angles

This definition sounds a bit abstract and it is perhaps useful to characterize the principal angles in terms of a geometric description. Recalling the explicit characterization of the orthogonal projection onto a subspace given in (3), we observe that

\[ P_V = VV^H, \quad P_W = WW^H. \]

The following result essentially says that the principal angles quantify how different are the orthogonal projections onto \( V \) and \( W \).

**Proposition**

Let \( V, W \in \mathbb{K}^{J \times I} \) be orthogonal matrices spanning \( V \) and \( W \), \( n \)-dimensional subspaces of \( \mathbb{K}^J \) (here we may assume that \( n = \#I \leq \#J = d \)). Then

\[ \| P_V - P_W \|_F = \| VV^H - WW^H \|_F = \sqrt{2} \| \sin \Theta(V, W) \|_2. \]

Proof at the blackboard ...
Stability of singular spaces: principal angles

The orthogonal space $W_\perp$ to any subspace $W$ of $\mathbb{K}^J$ is defined by $W_\perp = \{ v \in \mathbb{K}^J : \langle v, w \rangle = 0, \quad \forall w \in W \}$. We show yet another description of the singular angles. The projections onto $W$ and $W_\perp$ are related by

$$P_W = I - P_{W_\perp} \text{ or } WW^H = I - W_\perp W_\perp^H. \quad (15)$$

Proposition

Let $V, W \in \mathbb{K}^{J \times I}$ be orthogonal matrices spanning $V$ and $W$, $n$-dimensional subspaces of $\mathbb{K}^J$ (here we may assume that $n = \#I \leq \#J = d$). We denote $W_\perp$ an orthogonal matrix spanning $W_\perp$. Then

$$\| W_\perp^H V \|_F = \| \sin \Theta(V, W) \|_2.$$

Proof at the blackboard ...
Stability of singular spaces: principal angles

We leave as an additional exercise to show that, as a corollary of the previous result, one has also

$$\| P_{W^\perp} P_V \|_F = \| \sin \Theta(V, W) \|_2. \quad (16)$$

Again, one has to use the properties of the trace operator!
Stability of singular spaces: Wedin’s bound

\( A, \tilde{A} \in \mathbb{K}^{I \times J} \) and define \( E = \tilde{A} - A \). We fix
\[ 1 \leq k \leq r_{\text{max}} = \min\{\#I, \#J\} \]
and we consider the SVD

\[
A = \begin{pmatrix} U_1 & U_0 \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_0 \end{pmatrix} \begin{pmatrix} V_1^H \\ V_0^H \end{pmatrix} = U_1 \Sigma_1 V_1^H + U_0 \Sigma_0 V_0^H,
\]

where

\[
V_1 = [v_1, \ldots, v_k], \quad V_0 = [v_{k+1}, \ldots, v_{r_{\text{max}}}], \quad U_1 = [u_1, \ldots, u_k], \quad U_0 = [u_{k+1}, \ldots, u_{r_{\text{max}}}],
\]

and

\[
\Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_k), \quad \Sigma_0 = \text{diag}(\sigma_{k+1}, \ldots, \sigma_{r_{\text{max}}}).
\]

For the same \( k \) we use corresponding notations for an analogous decomposition for \( \tilde{A} \) except that all the symbols have \( \tilde{\cdot} \) on the top.

It is obvious that

\[
A + E = A_1 + A_0 + E = \tilde{A}_1 + \tilde{A}_0 = \tilde{A}.
\]
Let us notice that the range($A_1$) is indeed spanned by the orthonormal system $U_1$, while range($A_1^H$) is spanned by $V_1$. Similar statements hold for $\tilde{A}_1$. We define now two “residual” matrices

$$R_{11} = A\tilde{V}_1 - \tilde{U}_1\tilde{\Sigma} \quad \text{and} \quad R_{21} = A^H\tilde{U}_1 - \tilde{V}_1\tilde{\Sigma}^H.$$
Stability of singular spaces: Wedin’s bound

A connection between $R_{11}$ and $E$ is seen through the rewriting

$$R_{11} = A\tilde{V}_1 - \tilde{U}_1\tilde{\Sigma} = (\tilde{A} - E)\tilde{V}_1 - \tilde{U}_1(\tilde{U}_1^H\tilde{A}\tilde{V}_1) = -E\tilde{V}_1.$$ 

We used that $\tilde{\Sigma} = \tilde{U}_1^H\tilde{A}\tilde{V}_1 = \tilde{U}_1^H\tilde{A}_1\tilde{V}_1$ and that $\tilde{U}_1\tilde{U}_1^H$ is the orthogonal projection onto the range of range($\tilde{A}_1$), hence $\tilde{U}_1\tilde{U}_1^H\tilde{A}_1 = \tilde{A}_1$, and $\tilde{A}_1\tilde{V}_1 = \tilde{A}\tilde{V}_1$. Analogously, it can be shown that

$$R_{21} = -E^H\tilde{U}_1.$$
Theorem (Wedin)

Under the notations so far used in this section, assume that there exist an $\alpha \geq 0$ and $\delta > 0$ such that

$$\sigma_k(\tilde{A}) \geq \alpha + \delta \text{ and } \sigma_{k+1}(A) \leq \alpha.$$  \hspace{1cm} (18)

Then for every unitarily invariant norm $\| \cdot \|_*$ (in particular for the spectral and Frobenius norm)

$$\max\{\| \sin \Theta(\tilde{V}_1, V_1)\|_*, \| \sin \Theta(\tilde{U}_1, U_1)\|_*\} \leq \max\{\| R_{11} \|_*, \| R_{21} \|_*\} \leq \frac{\|E\|_*}{\delta}.$$
Introduction to probability theory

In the probabilistic proof of the convergence of the power method for computing singular vectors of a matrix $A$, we used the following random process

to define a random point on the sphere: we considered an imaginary player $X$ of darts, with no experience, throwing darts - actually at random - at the target, the two dimensional disk $\Omega = B_r$, the disk of radius $r$ and center 0. At every point $\omega \in \Omega$ within the target hit by a dart we assign a point on the boundary of the target (the one dimensional sphere $S^1$) $X(\omega) = \frac{\omega}{\|\omega\|_2} \in S^1$. 
Introduction to probability theory

It is also true that the one dimensional sphere $S^1$ is parametrized by the angle $\theta \in [0, 2\pi)$, hence we can also consider the player $X$ as a map from $\omega \in \Omega$ to $\theta = \arg \frac{\omega}{||\omega||_2} \in [0, 2\pi) \subset \mathbb{R}$

$$X : \Omega \rightarrow \mathbb{R}.$$ 

We defined the probability that the very inexperienced player $X$ hits a point $\omega \in B \subset \Omega$ simply as the ratio of the area of $B$ relative to the entire area of the target $\Omega$:

$$\mathbb{P}(\omega \in B) = \frac{\text{Vol}(B)}{\text{Vol}(\Omega)}.$$ 

What is then the probability of the event

$$B = \{\omega \in \Omega : \theta_1 \leq X(\omega) \leq \theta_2\}?$$
What is then the probability of the event

\[ B = \{ \omega \in \Omega : \theta_1 \leq X(\omega) \leq \theta_2 \} ? \]

The area of \( B \) is the area of the disk sector of vectors with angles between \([\theta_1, \theta_2]\), which is computed by using the polar coordinates:

\[
\text{Vol}(B) = \int_{\theta_1}^{\theta_2} \int_{0}^{r} sdsd\theta = \frac{r^2(\theta_2 - \theta_1)}{2}.
\]

Hence, recalling that \( \text{Vol}(\Omega) = r^2\pi \), we have

\[
\mathbb{P}(\{ \omega \in \Omega : \theta_1 \leq X(\omega) \leq \theta_2 \}) = \frac{\text{Vol}(B)}{\text{Vol}(\Omega)} = \frac{r^2(\theta_2 - \theta_1)}{2\pi r^2} = \frac{\theta_2 - \theta_1}{2\pi}.
\]
What is the average outcome of the (random) player? Well, if we consider $N$ attempts of the player, the empirical average result would be

$$\frac{1}{N} \sum_{i=1}^{N} X(\omega_i),$$

which, for $N \to \infty$ converges to

$$\mathbb{E}X := \int_{\Omega} X(\omega) d\mathbb{P}(\omega) = \frac{1}{2\pi} \int_{0}^{2\pi} \theta d\theta = \frac{(2\pi)^2}{4\pi} = \pi.$$

The function $\phi(\theta) = \frac{1}{2\pi} I_{[0,2\pi]}(\theta)$ is called the probability density of the random player $X$. In general one computes the quantities

$$\mathbb{E}g(X) := \int_{\Omega} g(X(\omega)) d\mathbb{P}(\omega) = \int_{\mathbb{R}} g(\theta) \phi(\theta) d\theta.$$
Introduction to probability theory

If we define $\Sigma$ the set of all possible events (identified with measurable subsets of $\Omega$) the triple $(\Omega, \Sigma, \mathbb{P})$ is called a probability space.

The random player, which we described above with the function $X : \Omega \rightarrow \mathbb{R}$, is an example of random variable.

The average outcome of its playing $E X$ is the expected value of the random variable.

As already mentioned above, this particular random variable has a density function $\phi(\theta)$ which allows to compute interesting quantities (such as the probabilities of certain events) by integrals on the real line.

This relatively intuitive construction is generalized to abstract probability spaces as follows.
Abstract probability theory

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, where $\Sigma$ denotes a $\sigma$-algebra (of the events) on the sample space and $\mathbb{P}$ a probability measure on $(\Omega, \Sigma)$.

The probability of an event $B \in \Sigma$ is denoted by

$$
\mathbb{P}(B) = \int_B d\mathbb{P}(\omega) = \int_{\Omega} I_B(\omega) \, d\mathbb{P}(\omega)
$$

where the characteristic function $I_B(\omega)$ takes the value 1 if $\omega \in B$ and 0 otherwise.

The union bound (or Bonferroni’s inequality, or Boole’s inequality) states that for a collection of events $B_\ell \in \Sigma$, $\ell = 1, \ldots, n$, we have

$$
\mathbb{P}\left( \bigcup_{\ell=1}^{n} B_\ell \right) \leq \sum_{\ell=1}^{n} \mathbb{P}(B_\ell) .
$$

(19)
A random variable $X$ is a real-valued measurable function on $(\Omega, \Sigma)$. Recall that $X$ is called measurable if the preimage $X^{-1}(A) = \{\omega \in \Omega : X(\Omega) \in A\}$ is contained in $\Sigma$ for all Borel measurable subsets $A \subset \mathbb{R}$.

Usually, every reasonable function $X$ will be measurable; in particular, all functions appearing in this lectures.

In what follows we will usually not mention the underlying probability space $(\Omega, \Sigma, \mathbb{P})$ when speaking about random variables. The distribution function $F = F_X$ of $X$ is defined as

$$F(t) = \mathbb{P}(X \leq t), \quad t \in \mathbb{R}. \quad (20)$$
Densities

A random variable \( X \) possesses a *probability density function* \( \phi: \mathbb{R} \to \mathbb{R}_+ \) if

\[
P(a < X \leq b) = \int_a^b \phi(t) \, dt \quad \text{for all } a < b \in \mathbb{R} \quad (21)
\]

Then \( \phi = \frac{d}{dt} F(t) \). The *expectation* or mean of a random variable will be denoted by

\[
\mathbb{E}X = \int_{\Omega} X(\omega) \, d\mathbb{P}(\omega). \quad (22)
\]

If \( X \) has probability density function \( \phi \) then for a function \( g: \mathbb{R} \to \mathbb{R} \),

\[
\mathbb{E}g(X) = \int_{-\infty}^{+\infty} g(t)\phi(t) \, dt \quad (23)
\]

whenever the integral exists.
Moments

The quantities $E X^p, p > 0$ are called moments of $X$, while $E |X|^p$ are called absolute moments. (Sometimes we may omit “absolute”.)

The quantity $E(X - EX)^2 = EX^2 - (EX)^2$ is called variance.

For $1 \leq p < \infty$, $(E|X|^p)^{\frac{1}{p}}$ defines a norm on the $L^p(\Omega, \mathbb{P})$-space of all $p$-integrable random variables, in particular, the triangle inequality

$$
(\mathbb{E}|X + Y|^p)^{\frac{1}{p}} \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} + (\mathbb{E}|Y|^p)^{\frac{1}{p}}
$$

holds for all $p$-integrable random variables $X, Y$ on $(\Omega, \Sigma, \mathbb{P})$.

Hoelder’s inequality states that, for random variables $X, Y$ on a common probability space and $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, we have

$$
|\mathbb{E}XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} (\mathbb{E}|Y|^q)^{\frac{1}{q}}.
$$
Sequences of random variables and convergence

Let $X_n$, $n \in \mathbb{N}$, be a sequence of random variables such that $X_n$ converges to $X$ as $n \to \infty$ in the sense that $\lim_{n \to \infty} X_n(\omega) = X(\omega)$ for all $\omega$.

*Lebesgue’s dominated convergence theorem* states that if there exists a random variable $Y$ with $\mathbb{E}|Y| < \infty$ such that $|X_n| \leq |Y|$ a.s. then $\lim_{n \to \infty} \mathbb{E}X_n = \mathbb{E}X$.

Lebesgue’s dominated convergence theorem has as well an obvious formulation for integrals of sequences of functions.
Fubini’s theorem on the integration of functions of two variables can be formulated as follows.

Let \( f : A \times B \to C \) be measurable, where \((A, \nu)\) and \((B, \mu)\) are measurable spaces.

If \( \int_{A \times B} |f(x, y)| \, d(\nu \otimes \mu)(x, y) < \infty \) then

\[
\int_A \left( \int_B f(x, y) \, d\mu(y) \right) \, d\nu(x) = \int_B \left( \int_A f(x, y) \, d\nu(x) \right) \, d\mu(y).
\]
Moment computations and Cavalieri’s formula

Absolute moments can be computed by means of the following formula.

**Proposition**

*The absolute moments of a random variable $X$ can be expressed as*

$$E|X|^p = p \int_0^\infty \mathbb{P}(|X| \geq t)t^{p-1} \, dt, \quad p > 0.$$ 

**Corollary**

*For a random variable $X$ the expectation satisfies*

$$E X = \int_0^\infty \mathbb{P}(X \geq t) \, dt - \int_0^\infty \mathbb{P}(X \leq -t) \, dt.$$
Tail bounds and Markov inequality

The function $t \to \mathbb{P}(|X| \geq t)$ is called the tail of $X$. A simple but often effective tool to estimate the tail by expectations and moments is the Markov inequality.

**Theorem**

*Let $X$ be a random variable. Then*

$$
\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t}
$$

*for all $t > 0$.*
Remark
As an important consequence we note that for $p > 0$

$$
P(|X| \geq t) = P(|X|^p \geq t^p) \leq t^{-p} \mathbb{E}|X|^p, \quad \text{for all } t > 0.\)

The special case $p = 2$ is referred to as the Chebyshev inequality. Similarly, for $\theta > 0$ we obtain

$$
P(X \geq t) = P(\exp(\theta X) \geq \exp(\theta t)) \leq \exp(-\theta t) \mathbb{E}\exp(\theta X) \quad \text{for all } t \in \mathbb{R}.
$$

The function $\theta \mapsto \mathbb{E}\exp(\theta X)$ is usually called the Laplace transform or the moment generating function of $X$. 
Normal distributions

A normal distributed random variable or Gaussian random variable $X$ has probability density function

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(t - \mu)^2}{2\sigma^2} \right).$$ \hspace{1cm} (25)

Its distribution is often denoted by $N(\mu, \sigma)$.

It has mean $\mathbb{E}X = \mu$ and variance $\mathbb{E}(X - \mu)^2 = \sigma^2$.

A standard Gaussian random variable (or standard normal or simply standard Gaussian), usually denoted $g$, is a Gaussian random variable with $\mathbb{E}g = 0$ and $\mathbb{E}g^2 = 1$. 
Random vectors

A random vector \( X = [X_1, \ldots, X_n]^\top \in \mathbb{R}^n \) is a collection of \( n \) random variables on a common probability space \((\Omega, \Sigma, \mathbb{P})\).

Its expectation is the vector \( \mathbb{E}X = [\mathbb{E}X_1, \ldots, \mathbb{E}X_n]^\top \in \mathbb{R}^n \), while its joint distribution function is defined as

\[
F(t_1, \ldots, t_n) = \mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_n), \quad t_1, \ldots, t_n \in \mathbb{R}.
\]

Similarly to the univariate case, the random vector \( X \) has a joint probability density if there exists a function \( \phi: \mathbb{R}^n \to [0, 1] \) such that for any measurable domain \( D \subset \mathbb{R}^n \)

\[
\mathbb{P}(X \in D) = \int_D \phi(t_1, \ldots, t_n) \, dt_1 \cdots dt_n.
\]

A complex random vector \( Z = X + iY \in \mathbb{C}^n \) is a special case of a \( 2n \)-dimensional real random vector \((X, Y) \in \mathbb{R}^{2n}\).
Independence

A collection of random variables $X_1, \ldots, X_n$ is (stochastically) independent if, for all $t_1, \ldots, t_n \in \mathbb{R}$,

$$
\mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_n) = \prod_{\ell=1}^{n} \mathbb{P}(X_\ell \leq t_\ell).
$$

For independent random variables, we have

$$
\mathbb{E}\left[\prod_{\ell=1}^{n} X_\ell\right] = \prod_{\ell=1}^{n} \mathbb{E}[X_\ell].
$$

(26)

If they have a joint probability density function $\phi$ then the latter factorizes as

$$
\phi(t_1, \ldots, t_n) = \phi_1(t_1) \times \cdots \times \phi_n(t_n)
$$

where the $\phi_1, \ldots, \phi_n$ are the probability density functions of $X_1, \ldots, X_n$. 
Independence

In general, a collection $X_1 \in \mathbb{R}^{n_1}, \ldots, X_m \in \mathbb{R}^{n_m}$ of random vectors are independent if for any collection of measurable sets $A_\ell \subset \mathbb{R}^{n_\ell}$, $\ell \in [m],$

$$P(X_1 \in A_1, \ldots, X_m \in A_m) = \prod_{\ell=1}^{m} P(X_\ell \in A_\ell).$$

If furthermore $f_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{N_\ell}$, $\ell = 1, \ldots, m$, are measurable functions then also the random vectors $f_1(X_1), \ldots, f_m(X_m)$ are independent.

A collection $X_1, \ldots, X_m \in \mathbb{R}^{n}$ of independent random vectors that all have the same distribution is called independent identically distributed (i.i.d.).

A random vector $X'$ will be called an independent copy of $X$ if $X$ and $X'$ are independent and have the same distribution.
The sum $X + Y$ of two independent random variables $X, Y$ having probability density functions $\phi_X, \phi_Y$, has probability density function $\phi_{X+Y}$ given by the convolution

$$\phi_{X+Y}(t) = (\phi_X * \phi_Y)(t) = \int_{-\infty}^{+\infty} \phi_X(u)\phi_Y(t-u) \, du \quad (27)$$
Fubini’s theorem for expectations

The theorem takes the following form. Let \( X, Y \in \mathbb{R}^n \) be two independent random vectors (or simply random variables) and \( f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) be a measurable function such that \( \mathbb{E}|f(X, Y)| < \infty \). Then the functions

\[
  f_1 : \mathbb{R}^n \to \mathbb{R}, \quad f_1(x) = \mathbb{E}f(x, Y), \\
  f_2 : \mathbb{R}^n \to \mathbb{R}, \quad f_2(y) = \mathbb{E}f(X, y)
\]

are measurable, \( \mathbb{E}|f_1(X)| < \infty \) and \( \mathbb{E}|f_2(Y)| < \infty \) and

\[
  \mathbb{E}f_1(X) = \mathbb{E}f_2(Y) = \mathbb{E}f(X, Y). \tag{28}
\]

The random variable \( f_1(X) \) is also called conditional expectation or expectation conditional on \( X \) and will sometimes be denoted by \( \mathbb{E}_Y f(X, Y) \).
Gaussian vectors

A random vector \( \mathbf{g} \in \mathbb{R}^n \) is called a standard Gaussian vector if its components are independent standard normal distributed random variables. More generally, a random vector \( \mathbf{X} \in \mathbb{R}^n \) is said to be a Gaussian vector or multivariate normal distributed if there exists a matrix \( \mathbf{A} \in \mathbb{R}^{n \times k} \) such that \( \mathbf{X} = \mathbf{A} \mathbf{g} + \mu \), where \( \mathbf{g} \in \mathbb{R}^k \) is a standard Gaussian vector and \( \mu \in \mathbb{R}^n \) is the mean of \( \mathbf{X} \).

The matrix \( \Sigma = \mathbf{A} \mathbf{A}^T \) is then the covariance matrix of \( \mathbf{X} \), i.e., \( \Sigma = \mathbb{E}(\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top \). If \( \Sigma \) is non-degenerate, i.e., \( \Sigma \) is positive definite, then \( \mathbf{X} \) has a joint probability density function of the form

\[
\psi(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} \langle \mathbf{x} - \mu, \Sigma^{-1}(\mathbf{x} - \mu) \rangle \right)
\]

In the degenerate case when \( \Sigma \) is not invertible \( \mathbf{X} \) does not have a density.

It is easily deduced from the density that a rotated standard Gaussian \( \mathbf{U} \mathbf{g} \), where \( \mathbf{U} \) is an orthogonal matrix, has the same distribution as \( \mathbf{g} \) itself.
Sum of gaussian variables

If $X_1, \ldots, X_n$ are independent and normal distributed random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma^2_1, \ldots, \sigma^2_n$ then $X = [X_1, \ldots, X_n]^\top$ has a multivariate normal distribution and its sum $Z = \sum_{\ell=1}^n X_\ell$ has the univariate normal distribution with mean $\mu = \sum_{\ell=1}^n \mu_\ell$ and variance $\sigma^2 = \sum_{\ell=1}^n \sigma^2_\ell$, as can be calculated from (27).
Jensen’s inequality

Jensen’s inequality reads as follows.

**Theorem**

*Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a convex function, and let* $\mathbf{X} \in \mathbb{R}^n$ *be a random vector. Then*

$$f(\mathbb{E}\mathbf{X}) \leq \mathbb{E}f(\mathbf{X}).$$

*Note that* $-f$ *is convex if* $f$ *is concave, so that for concave functions* $f$, *Jensen’s inequality reads*

$$\mathbb{E}f(\mathbf{X}) \leq f(\mathbb{E}\mathbf{X}).$$  \hspace{1cm} (29)
Cramér’s Theorem

We often encounter sums of independent mean zero random variables. Deviation inequalities bound the tail of such sums.

We recall that the *moment generating function* of a (real-valued) random variable \( X \) is defined by

\[
\theta \mapsto \mathbb{E} \exp(\theta X),
\]

for all \( \theta \in \mathbb{R} \) whenever the expectation on the right hand side is well-defined.

Its logarithm is the *cumulant generating function*

\[
C_X(\theta) = \ln \mathbb{E} \exp(\theta X).
\]

With the help of these definitions we can formulate Cramér’s theorem.
Cramér’s Theorem

Theorem
Let $X_1, \ldots, X_M$ be a sequence of independent (real-valued) random variables, with cumulant generating functions $C_{X_\ell}$, $\ell \in [M] := \{1, \ldots, M\}$. Then, for $t > 0$,

$$
P \left( \sum_{\ell=1}^{M} X_\ell \geq t \right) \leq \exp \left( \inf_{\theta > 0} \left\{ -\theta t + \sum_{\ell=1}^{M} C_{X_\ell}(\theta) \right\} \right).
$$

Remark
The function

$$
t \mapsto \inf_{\theta > 0} \left\{ -\theta t + \sum_{\ell=1}^{M} X_{X_\ell}(\theta) \right\}
$$

appearing in the exponential is closely connected to a convex conjugate function appearing in convex analysis.
**Hoeffding’s inequality**

**Theorem**

Let $X_1, \ldots, X_M$ be a sequence of independent random variables such that $\mathbb{E}X_\ell = 0$ and $|X_\ell| \leq B_\ell$ almost surely, $\ell \in [M]$. Then

$$
\Pr \left( \sum_{\ell=1}^{M} X_\ell \geq t \right) \leq \exp \left( - \frac{t^2}{2 \sum_{\ell=1}^{M} B_\ell^2} \right),
$$

and consequently,

$$
\Pr \left( \left| \sum_{\ell=1}^{M} X_\ell \right| \geq t \right) \leq 2 \exp \left( - \frac{t^2}{2 \sum_{\ell=1}^{M} B_\ell^2} \right). \quad (30)
$$
Bernstein inequalities

Theorem

Let $X_1, \ldots, X_M$ be independent mean zero random variables such that, for all integers $n \geq 2$,

$$\mathbb{E}|X_\ell|^n \leq \frac{n!}{2} R^{n-2} \sigma_\ell^2 \quad \text{for all } \ell \in [M]$$

(31)

for some constants $R > 0$ and $\sigma_\ell > 0$, $\ell \in [M]$. Then, for all $t > 0$,

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2(\sigma^2 + Rt)}\right),$$

(32)

where $\sigma^2 := \sum_{\ell=1}^{M} \sigma_\ell^2$. 
Bernstein inequalities

Before providing the proof we give two consequences. The first is the Bernstein inequality for bounded random variables.

Corollary

Let $X_1, \ldots, X_M$ be independent random variables with zero mean such that $|X_\ell| < K$ almost surely, for $\ell \in [M]$ and some constant $K > 0$. Further assume $\mathbb{E}|X_\ell|^2 \leq \sigma_\ell^2$ for constants $\sigma_\ell > 0$, $\ell \in [M]$. Then, for all $t > 0$,

$$
\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2(\sigma^2 + \frac{1}{3}Kt)}\right),
$$

(33)

where $\sigma^2 := \sum_{\ell=1}^{M} \sigma_\ell^2$. 
Bernstein inequalities

As a second consequence, we present the Bernstein inequality for subexponential random variables.

**Corollary**

Let $X_1, \ldots, X_M$ be independent mean zero subexponential random variables, that is, $\mathbb{P}(|X_\ell| \geq t) \leq \beta e^{-\kappa t}$ for some constants $\beta, \kappa > 0$ for all $t > 0$, $\ell \in [M]$. Then

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2 \exp\left(-\frac{(\kappa t)^2}{2(2\beta M + \kappa t)}\right), \quad (34)$$