Basic Mathematics
A Machine Learning Perspective

S.V.N. “Vishy” Vishwanathan
vishy@axiom.anu.edu.au

National ICT of Australia
and
Australian National University

Thanks to Alex Smola for initial version of slides
Overview

- Functional Analysis
- Linear Algebra
- Matrix Theory
- Probability
Metric Spaces

Metric Space:
A pair \((X, d)\), where \(X\) is a set and \(d : X \times X \rightarrow \mathbb{R}^+_0\) is a metric space if \(\forall x, y, z \in X\)

- \(d(x, y) = 0\) iff \(x = y\)
- \(d(x, y) = d(y, x)\) (Symmetry)
- \(d(x, z) \leq d(x, y) + d(y, z)\) (Triangle inequality)

Examples:

Euclidean space
For all \(x, y \in \mathbb{R}^n\) we define \(d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}\)

\(\ell^p\)-space
Space of sequences with \(d(x, y) = (\sum_{i=1}^{\infty} |x_i - y_i|^p)^{\frac{1}{p}}\)

Hilbert space
Space of sequences with \(d(x, y) = \sqrt{\sum_{i=1}^{\infty} (x_i - y_i)^2}\)
Balls, Open and Closed Sets

**Ball:**

Given \( x_0 \in \mathcal{X} \) and \( r > 0 \) we define

\[
B(x_0, r) = \{ x \in \mathcal{X} \mid d(x, x_0) < r \} \quad \text{(Open ball)}
\]

\[
\overline{B}(x_0, r) = \{ x \in \mathcal{X} \mid d(x, x_0) \leq r \} \quad \text{(Closed ball)}
\]

**Open set:**

A subset \( M \) of a metric space \( \mathcal{X} \) is open if it contains an open ball about each of its points.

**Closed set:**

If the complement of \( M \) is open it is called a closed set.

**Examples:**

- The set \((a, b) \subset \mathbb{R}\) is an open set.
- The set \([a, b] \subset \mathbb{R}\) is a closed set.
- The set \((a, b] \subset \mathbb{R}\) is neither open nor closed.
Cauchy Sequences

Convergence:
A sequence $\{x_i\} \in X$ is said to converge if for any $\epsilon$ there exists a $x$ and a $n_0$ such that for all $n \geq n_0$ we have $d(x_n, x) \leq \epsilon$

Cauchy Series:
A sequence $\{x_i\} \in X$ is a Cauchy if for any $\epsilon$ there exists a $n_0$ such that for all $m, n \geq n_0$ we have $d(x_m, x_n) \leq \epsilon$

Completeness:
- A space $X$ is complete if the limits of every Cauchy series are elements of $X$
- We call $\bar{X}$ the completion of $X$, i.e. the union of $X$ and the limits of all Cauchy series in $X$
- The real line $\mathbb{R}$ and complex plane $\mathbb{C}$ are complete
- The set $\mathbb{Q}$ of rationals is not complete!
Vector Spaces

Vector Space:
A set \( X \) such that \( \forall x, y \in X \) and \( \forall \alpha \in \mathbb{R} \) we have

- \( x + y \in X \) (Addition)
- \( \alpha x \in X \) (Multiplication)

Examples:
- Rational numbers \( \mathbb{Q} \) over the rational field
- Real numbers \( \mathbb{R} \)
- Also true for \( \mathbb{R}^n \)

Counterexamples:
- \( f : [0, 1] \rightarrow [0, 1] \) does not form a vector space!
- \( \mathbb{Z} \) is not a vector space over the real field
- The alphabet \( \{a, \ldots, z\} \) is not a vector space! (How do you define + and \( \times \) operators?)
Normed Space:
A pair \((X, \| \cdot \|)\), where \(X\) is a vector space and \(\| \cdot \| : X \to \mathbb{R}_0^+\) is a normed space if \(\forall x, y \in X\) and all \(\alpha \in \mathbb{R}\) it satisfies

- \(\| x \| = 0\) if and only if \(x = 0\)
- \(\| \alpha x \| = |\alpha| \| x \|\) (Scaling)
- \(\| x + y \| \leq \| x \| + \| y \|\) (Triangle inequality)

A norm not satisfying the first condition is called a pseudo norm

Norm and Metric:
A norm induces a metric via \(d(x, y) := \| x - y \|\)

Banach Space:
A complete (in the metric defined by the norm) vector space \(X\) together with a norm \(\| \cdot \|\)
Banach Spaces: Examples

$\ell^m_p$ Spaces:
Take $\mathbb{R}^m$ endowed with the norm $\|x\| := \left(\sum_{i=1}^{m} |x_i|^p\right)^{1/p}$
where $p > 0$

$\ell^p$ Spaces:
- These are subspaces of $\mathbb{R}^N$ with $\|x\| := \left(\sum_{i=1}^{\infty} |x_i|^p\right)^{1/p}$
- The sum might not converge for all series
- For instance $x_i = \frac{1}{i}$ is in $\ell_2$ but not in $\ell_1$

Function Spaces $L_p(\mathcal{X})$:
- For a continuous function $f : \mathcal{X} \rightarrow \mathbb{R}$ define
  $\|f\| := \left(\int_{\mathcal{X}} |f(x)|^p dx\right)^{1/p}$
- Might not be well defined for all functions
- We will see more about $L_2$ functions later in the course
Hilbert Spaces

Inner Product Space:
A pair \((X, \| \cdot \|)\), where \(X\) is a vector space and \(\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{R}_0^+\) is a inner product space if \(\forall x, y, z \in X\) and all \(\alpha \in \mathbb{R}\) it satisfies

- \(\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle\) (Additivity)
- \(\langle \alpha x, y \rangle = \alpha \langle x, y \rangle\) (Linearity)
- \(\langle x, y \rangle = \langle y, x \rangle\) (Symmetry)
- \(\langle x, x \rangle = 0 \iff x = 0\)

Dot Product and Norm:
A dot product induces a norm via \(\| x \| := \sqrt{\langle x, x \rangle}\)

Hilbert Space:
A complete (in the metric induced by the dot product) vector space \(X\), endowed with a dot product \(\langle \cdot, \cdot \rangle\)
Hilbert Spaces: Examples

Euclidean Spaces:
Take $\mathbb{R}^m$ endowed with the dot product $\langle x, y \rangle := \sum_{i=1}^{m} x_i y_i$

$l_2$ Spaces:
- Infinite series of real numbers
- We define a dot product as $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$

Function Spaces $L_2(\mathcal{X})$:
- For continuous functions $f, g : \mathcal{X} \rightarrow \mathbb{C}$ define
  $\langle f, g \rangle := \int_{\mathcal{X}} \overline{f(x)} g(x) dx$
- We take the complex conjugate of $f$ and replace the sum by an integral

Polarization Inequality:
To recover the dot product from the norm compute
$\| x + y \|^2 - \| x \|^2 - \| y \|^2 = 2 \langle x, y \rangle$
Matrices

Matrix:
A real matrix $M \in \mathbb{R}^{m \times n}$ is a linear map from $\mathbb{R}^m$ to $\mathbb{R}^n$.

Symmetry:
- A symmetric matrix $M \in \mathbb{R}^{m \times m}$ satisfies $M_{ij} = M_{ji}$.
- An anti-symmetric matrix satisfies $M_{ij} = -M_{ji}$.

Range and Null Space:
For $M \in \mathbb{R}^{m \times n}$
- Its range is $\{ y \in \mathbb{R}^m \mid y = Mx \text{ for some } x \in \mathbb{R}^n \}$.
- Its null space is $\{ x \in \mathbb{R}^n \mid Mx = 0 \}$.
- We have the relation $n = \dim(\text{null space}) + \dim(\text{range})$. 
**Rank**

**Definition:**
If \( M \in \mathbb{R}^{m \times n} \), \( \text{rank} \ M \) is the largest number of columns of \( M \) that constitute a linearly independent set.

**Characteristics:**
The following are equivalent for a rank \( k \) matrix \( M \)

- Exactly \( k \) rows (columns) of \( M \) are linearly independent
- Dimension of range of \( M \) is is \( k \)
- \( \exists \) a \( k\)-by-\( k \) sub-matrix of \( M \) with non-zero determinant
- All \( (k + 1)\)-by-\( (k + 1) \) sub-matrices have determinant \( 0 \)

**Properties:**
- For \( M \in \mathbb{R}^{m \times n} \), \( \text{rank}(M) \leq \min\{m, n\} \)
- Deleting rows/columns can only decrease the rank
- For \( M, N \in \mathbb{R}^{m \times n} \), \( \text{rank}(M + N) \leq \text{rank}(M) + \text{rank}(N) \)
Eigenvalue

**Similar Matrices:**
Two matrices $M, N \in \mathbb{R}^{m \times m}$ are similar if $\exists$ a non-singular $S \in \mathbb{R}^{m \times m}$ such that $M = S^{-1}NS$

**Eigenvalues, Eigenvectors:**
Given $M \in \mathbb{R}^{m \times m}$

- An eigen pair $(x, \lambda)$ satisfy $Mx = \lambda x$

**Properties:**
- The characteristic polynomial of $M$ is defined as $\det(\lambda I - M) = 0$
- Eigenvalues are roots of the characteristic polynomial
- Similar matrices have the same eigenvalues
- All eigenvalues of symmetric matrices are real
- If $M \in \mathbb{R}^{m \times m}$ has $m$ distinct eigenvalues, then it is diagonalizable
Eigensystems

Diagonalizable:
- A matrix $M \in \mathbb{R}^{m \times m}$ is diagonalizable if it is similar to a diagonal matrix
- A symmetric real matrix is always diagonalizable!

Matrix Decomposition:
We can decompose a symmetric real matrix as $O^\top \Lambda O$
where $O$ orthogonal and $\Lambda$ diagonal

Orthogonality:
All eigenvectors of symmetric matrices $M$ with different eigenvalues are mutually orthogonal

Proof
For two distinct eigen pairs $(x, \lambda)$ and $(x', \lambda')$

$$\lambda x^\top x' = (M x)^\top x' = x^\top (M^\top x') = x^\top (M x') = \lambda' x^\top x'$$

hence $\lambda' = \lambda$ or $x^\top x' = 0$
Orthogonality

Orthonormal Set:
A set \( \{x_1, x_2, \ldots, x_n\} \) is orthonormal if \( \langle x_i, x_j \rangle = 0 \) if \( i \neq j \)
and \( \|x_i\| = 1 \) for all \( i \)

Orthogonal Matrix:
- An orthogonal matrix \( M \in \mathbb{R}^{m \times m} \) is made up of orthonormal rows and columns
- Not difficult to see that \( MM^\top = 1 \)
- Equivalently \( M^{-1} = M^\top \)

Properties:
- Orthogonal transformations preserve matrix norms
Matrix Invariants

Trace:
- Trace is the sum of the diagonal elements
- For symmetric matrices $\text{tr}(MN) = \text{tr}(NM)$
- Orthogonal matrices preserve trace since $\text{tr}(O^\top MO) = \text{tr}(MOO^\top) = \text{tr} M$
- It can be shown that $\text{tr}(M) = \sum_{i=1}^{m} \lambda_i$

Determinant:
Antisymmetric multi-linear form, i.e. swapping columns or rows changes the sign, adding elements in rows and columns is linear. Useful form

$$\det M = \prod_{i=1}^{m} \lambda_i$$

Invariant under orthogonal transformations
Matrix Norms

Definition:
We call a function $\| \cdot \| : \mathbb{R}^{m \times m} \rightarrow \mathbb{R}^+$ a matrix norm if for all $M, N \in \mathbb{R}^{m \times m}$ we have

1. $\|M\| = 0$ iff $M = 0$
2. $\|cM\| = |c|\|M\|$ for all $c \in \mathbb{R}$
3. $\|M + N\| = \|M\| + \|N\|$ (Cauchy Schwartz?)
4. $\|MN\| = \|M\|\|N\|$

Matrix norms are closely related to the eigenvalues of the matrix (more on this later)

Examples:
1. The $\ell_1$ norm is defined as $\|M\|_1 := \sum_{i,j=1}^{m} |m_{ij}|$
2. The $\ell_2$ norm is defined as $\|M\|_1 := \left( \sum_{i,j=1}^{m} m_{ij}^2 \right)^{\frac{1}{2}}$
**Operator Norm:** Using $M \in \mathbb{R}^{m \times m}$ we have

$$\| M \|_2^2 = \max_{x \in \mathbb{R}^m} \frac{\| M x \|_2^2}{\| x \|_2^2}$$

$$= \max_{x \in \mathbb{R}^m \text{ and } \| x \| = 1} \| M x \|_2^2$$

$$= \max_{x \in \mathbb{R}^m \text{ and } \| x \| = 1} x^\top O \Lambda O^\top O \Lambda O x$$

$$= \max_{x' \in \mathbb{R}^m \text{ and } \| x' \| = 1} x'^\top \Lambda^2 x'$$

$$= \max_{i \in [m]} \lambda_i^2.$$

**Frobenius Norm:**

Likewise we obtain $\| M \|_{\text{Frob}}^2 = \text{tr} O \Lambda O^\top O \Lambda O^\top = \text{tr} \Lambda^2 = \sum_{i=1}^m \lambda_i^2.$
Positive Matrices

Positive Definite Matrix:
A matrix $M \in \mathbb{R}^{m \times m}$ for which for all $x \in \mathbb{R}^m$ we have

$$x^\top M x \geq 0 \text{ if } x \neq 0$$

This matrix has only positive eigenvalues since for all eigenvectors $x$ we have $x^\top M x = \lambda x^\top x = \lambda \| x \|^2 > 0$ and thus $\lambda > 0$.

Induced Norms and Metrics:
Every positive definite matrix induces a norm via

$$\| x \|_M^2 := x^\top M x$$

The triangle inequality can be seen by writing

$$\| x + x' \|_M^2 = (x + x')^\top M^{\frac{1}{2}} M^{\frac{1}{2}} (x + x') = \| M^{\frac{1}{2}} (x + x') \|^2$$

and using the triangle inequality for $M^{\frac{1}{2}} x$ and $M^{\frac{1}{2}} x'$. 
Idea:
Can we find something similar to the eigenvalue / eigenvector decomposition for arbitrary matrices?

Decomposition:
Without loss of generality assume $m \geq n$ For $M \in \mathbb{R}^{m \times n}$ we may write $M$ as $U \Lambda O$ where $U \in \mathbb{R}^{m \times n}, O \in \mathbb{R}^{n \times n}$, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.
Furthermore $O^T O = OO^T = U^T U = 1$.

Useful Trick:
Nonzero eigenvalues of $M^T M$ and $MM^T$ are the same.
This is so since $M^T M x = \lambda x$ and hence $(MM^T) M x = \lambda M x$ or equivalently $(MM^T) x' = \lambda x'$.
Basic Idea:
We have events, denoted by sets $X \subset \mathcal{X}$ in a space of possible outcomes $\mathcal{X}$. Then $P(X)$ tells us how likely is that an event $x$ with $x \in X$ will occur.

Basic Axioms:
- $Pr(X) \in [0, 1]$ for all $X \subseteq \mathcal{X}$
- $Pr(\mathcal{X}) = 1$
- $Pr(\bigcup_i X_i) = \sum_i Pr(X_i)$ if $X_i \cap X_j = \emptyset$ for all $i \neq j$

I am hiding gory details about $\sigma$-algebra on $\mathcal{X}$ here.

Simple Corollary:
$$Pr(X_i \cup X_j) = Pr(X_i) + Pr(X_j) - Pr(X_i \cap X_j)$$
Two Sets:

We can consider the space of events \((x, y) \in X \times Y\) and ask how likely events in the product space are.

Independence:

- If the events \(X \subseteq X\) and \(Y \subseteq Y\) are independent we have \(\Pr(X, Y) = \Pr(X) \cdot \Pr(Y)\).
- Here \(\Pr(X, Y)\) is the probability that any \((x, y)\) with \(x \in X\) and \(y \in Y\) occur.

Conditional Probability:

- Knowing that some event has happened will change our belief about the probability of related events i.e. \(\Pr(Y|X) \Pr(X) = \Pr(Y, X)\).
- This implies \(\Pr(Y, X) \leq \min(\Pr(X), \Pr(Y))\).
Bayes’ Rule

Marginalization:
We can sum out parts of a joint distribution to get the marginal distribution of a subset: \( \Pr(x) = \sum_y \Pr(x, y) \)

Bayes Rule:
- Using conditional probabilities
  \[
  \Pr(X|Y) \Pr(Y) = \Pr(X, Y) = \Pr(Y, X) = \Pr(Y|X) \Pr(X)
  \]
- Bayes’ rule:
  \[
  \Pr(X|Y) = \frac{\Pr(Y|X) \Pr(X)}{\Pr(Y)}
  \]

Application:
- Can infer how likely a hypothesis is, given some experimental evidence
AIDS-Test:

We want to find out likely it is that a patient really has AIDS (event $X$) if the test is positive (event $Y$).

Roughly $0.1\%$ of all Australians are infected ($\Pr(X) = 0.001$)

The probability of a false positive is say $1\%$ ($\Pr(Y|\overline{X}) = 0.01$ and $\Pr(Y|X) = 1$)

By Bayes’ rule

$$
\Pr(X|Y) = \frac{\Pr(Y|X) \Pr(X)}{\Pr(Y|X) \Pr(X) + \Pr(Y|\overline{X}) \Pr(\overline{X})}
$$

$$
= \frac{1 \times 0.001}{1 \times 0.001 + 0.01 \times 0.999} = 0.091
$$

The probability of having AIDS even when the test is positive is just $9.1\%$!
Reliability of Eye-Witness:

- An eye-witness is 90% sure and that there were 20 people at the crime scene.
- What is the probability that the guy identified committed the crime?
- Bayes’ rule again

\[
Pr(X|Y) = \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.1 \times 0.95} = 0.3213 = 32\%
\]

That’s a worry . . .
Computing $\Pr(X)$:

If we deal with continuous valued $X$ we need integrals.

$$\Pr(X) := \int_X d\Pr(x) = \int_X p(x)\,dx$$

Note that the last equality only holds if such a $p(x)$ exists. For the rest of this course we assume that such a $p$ exists ...
Bayes’ Rule for Densities

**Multivariate Densities:**

Densities on product spaces \((X \times Y)\) are given by \(p(x, y)\)

**Conditional Densities:**

For independent variables the densities factorize and we have

\[ p(x, y) = p(x)p(y) \]

For dependent variables (i.e. \(x\) tells us something about \(y\) and vice versa) we obtain

\[ p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x) \]

**Bayes’ Rule:**

Solving for \(p(y \mid x)\) yields

\[ p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \]
Example: $p(x) = 1 + \sin x$

A factorizing distribution
Definition:
If we want to denote the fact that variables $x$ and $y$ are drawn at random from an underlying distribution, we call them random variables.

IID variables:
- Independent and Identically Distributed RV
- Density factorizes into

\[
p(\{x_1, \ldots, x_m\}) = \prod_{i=1}^{m} p(x_i)
\]

Dependent Random Variables:
For prediction purposes we want to estimate $y$ from $x$. In this case we want that $y$ is dependent on $x$. If $p(x, y) = p(x)p(y)$ we could not predict at all!
Marginalization:

Given $p(x, y)$ we can integrate out $y$ to obtain $p(x)$ via

$$p(x) = \int_y p(x, y) \, dy$$

Conditioning:

If we know $y$, we can obtain $p(x \mid y)$ via Bayes rule, i.e.

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} = \frac{p(x, y)}{p(y)}.$$

A similar trick, however, is to note that the dependence of the RHS on $x$ lies only in $p(x, y)$ and therefore we obtain

$$p(x \mid y) = \frac{p(x, y)}{\int_x p(x, y) \, dx}.$$
**Definition:**
The expectation of a function \( f(x) \) with respect to the random variable \( x \) is defined as

\[
E_x[f(x)] := \int_X f(x) \, d\Pr(x) = \int_X f(x) \, p(x) \, dx
\]

The last equation is valid if a density exists.

**Intuition:**
It is the mean value we get by sampling a large number of \( x \) according to \( p(x) \) and evaluating \( f(x) \) on the drawn sample.

**Other Facts:**
- Moments are expectations of higher orders
- Knowledge of all the moments completely determines the distribution
Examples

Uniform Distribution:
Assume the uniform distribution on \([0, 10]\). What is the expected value of \(f(x) = x^2\)?

\[
E_x[f(x)] = \int_{[0,10]} f(x)p(x)\,dx = \int_{[0,10]} x^2 \frac{1}{10} \,dx = 33\frac{1}{3}
\]

Roulette:
What is the expected loss in roulette when we bet on a number, say \(j\) (we win \(36\):\(1\) if the number is hit and \(0\):\(1\) otherwise)?

\[
E_x[f(x)] = \sum_{i=1, i\neq j}^{37} -1\cdot \frac{1}{37} + 35\cdot \frac{1}{37} = -\frac{1}{37}
\]
Mean and Mode

Mean:
- Expected value of the random variable i.e. $\mu := \mathbb{E}_x[x]$.

Mode:
- Largest value of the density $p(x)$.
- Most frequently observed values of $x$.
- Mode and mean do not coincide in general.
Variance

**Definition:**
- Amount of variation in the random variable
- First center and then compute second order moment

\[
\sigma^2 := \mathbb{E}_x \left[ (x - \mathbb{E}_x[x])^2 \right] = \mathbb{E}_x x^2 - (\mathbb{E}_x[x])^2
\]

**Normalization:**
- Rescale data to zero mean and unit variance
- Preprocess data by \( x \rightarrow \frac{x - \mu}{\sigma} \)

**Tails of Distributions:**
- Note that the variance need not always exist
- Tails of distributions give an idea about how sharply concentrated the distribution is around its mean
- Long-tailed distributions can be killers for insurance companies!
Markov’s Inequality

Markov’s Inequality:
If \( x \) takes only non-negative values then
\[
\Pr(x \geq a) \leq \frac{E[x]}{a}
\]

Proof:
We write
\[
E[x] = \int_0^a x p(x) \, dx + \int_a^\infty x p(x) \, dx
\]
\[
\geq \int_a^\infty x p(x) \, dx \quad \text{non-negativity}
\]
\[
\geq a \int_a^\infty p(x) \, dx = a \Pr(x \geq a)
\]

Observation:
Completely independent of the distribution!
Chebyshev’s Inequality

Chebyshev’s Inequality:
For any random variable $x$ we can bound deviations of $x$ from its mean $\mathbb{E}[x]$ by

$$\Pr(|x - \mathbb{E}[x]| > C) \leq \frac{\sigma^2}{C^2}$$

Proof:
Apply Markov’s inequality to $y := (x - \mathbb{E}[x])^2$

Applications:
- Information about some measurement is easy to get
- Easy to estimate the variance too
- Don’t know anything about the distribution :-(
- Still can make statements about probability of deviating from the mean!
**Jensen’s Inequality**

**Jensens’s Inequality:**

If $f$ is a convex function and $X$ is a random variable then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

**Picture:**

Notice how expectation is a linear operator.
Entropy

Definition:
- Measures the *disorder* of a system
- Defined as

\[ H(p) = - \sum_x p(x) \log p(x) \]

Properties:
- \( H(p) > 0 \) unless only one possible outcome
- Maximal value occurs for uniform \( p \)
- Deep connections to information theory exist
Normal Distribution

The Formula:

\[ p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

Mean:

- Notice that \( p(\mu + \xi) = p(\mu - \xi) \) because 
\[ ((\mu + \xi) - \mu)^2 = \xi^2 = ((\mu - \xi) - \mu)^2 \]
- Hence the mean of \( p(x) \) is \( \mu \)

Variance:

The variance of \( p(x) \) is \( \sigma^2 \). We show this by proving that

\[
\text{Var} x = \int_{\mathbb{R}} p(x)(x - \mu)^2 \, dx = \int_{\mathbb{R}} p(\mu + \xi)\xi^2 \, d\xi
= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{\xi^2}{2}} \xi^2 \, d\xi = \sigma^2
\]
Normal Distribution in $\mathbb{R}$: Mean 1, Variance 3

Normal Distribution in $\mathbb{R}^2$: Mean $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, Variance $\begin{bmatrix} 6 & 4 \\ 4 & 4 \end{bmatrix}$
Covariance and Correlation

Covariance:

- For a multivariate distribution
  \[ \text{Cov } \mathbf{x} := \mathbf{E} \left[ (\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top \right] \]

- We now compute a matrix instead of a single number
- In particular
  \[ (\text{Cov } \mathbf{x})_{ij} = \mathbf{E} \left[ (x_i - \mu_i)(x_i - \mu_i) \right] \]

Correlated Variables:

- Measures degree of association between variables
- If positively correlated then
  \[ (\text{Cov } \mathbf{x})_{ij} = \sqrt{(\text{Cov } \mathbf{x})_{ii} (\text{Cov } \mathbf{x})_{jj}} \]
- For uncorrelated variables their covariance vanishes
Multivariate Normal

The Formula:

- \( \Sigma \in \mathbb{R}^{m \times m} \) is positive definite
- The mean \( \mu \in \mathbb{R}^m \)

\[
p(x) = \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} \exp \left( -\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu) \right)
\]

Mean:

Obviously this is \( \mu \) (we can check that by symmetry)

Variance:

- Tedious calculation shows that \( \text{Var}(x) = \Sigma \)
- Hint: decompose \( \Sigma = O^\top \Lambda O \)
- Hint: use \( \det \Sigma = \det O \Sigma O^\top = \prod_i \lambda_i \), where \( \lambda_i \) are the eigenvalues of \( \Sigma \)
Decay of Atoms:
- The probability that a atom decays within 1 sec is $1 - p$
- The probability that it decays in $n$ sec is $1 - p^n$
- In the continuous domain the probability of decay after time $T$ is
  \[
P(\xi \leq T) = 1 - \exp(-\lambda T) = \int_0^T p(t) dt\]

Laplacian Distribution:
- Consequently, $p(t)$ is given by $\lambda \exp(-\lambda T)$
- It is a particularly long-tailed distribution

Mean and Variance:
- Mean is given by $\mu = \frac{1}{\lambda}$
- Variance is given by $\frac{2}{\lambda^2}$
Why Gaussians are good for you: If we have many independent errors, the net effect will be a single error with normal distribution.

Theorem:
Denote by $\xi_i$ random variables with variance $\sigma_i \leq \bar{\sigma}$ for some $\bar{\sigma}$ and with mean $\mu_i \leq \bar{\mu}$ for some $\bar{\mu}$, then the random variable $\xi := \sum_{i=1}^{m} \frac{\xi_i - \mu_i}{\sqrt{\sum_{i=1}^{m} \sigma_i^2}}$ has zero mean and unit variance. Furthermore for $m \to \infty$ the random variable $\xi$ will be normally distributed.
Hoeffding’s Bound

Sum of Random Variables:
- Consider the average of $m$ random variables $\xi_i \in [0, 1]$

$$\xi := \frac{1}{m} \sum_{i=1}^{m} \xi_i$$

Will $\xi$ be concentrated around its mean?

Hoeffding’s Theorem:
- For any $\varepsilon > 0$ the probability of large deviations of $\xi$ from $E[\xi]$ is bounded by

$$\Pr (|\xi - E[\xi]| \geq \varepsilon) \leq 2 \exp (-2\varepsilon^2 m)$$

- things get exponentially better, the more random variables we average over (i.e. more the number of observations)
Questions?