A Recap on Some Mathematical Subjects

or "How to stay fresh all the time!"

Mohammad Reza Karimi*

Learning & Adaptive Systems

Spring 2019

*mkarimi@ethz.ch

Topics Covered

Linear Algebra

Multivariate Analysis

Probability Theory

Outline

Linear Algebra

Multivariate Analysis

Probability Theory

Matrix Multiplication

- Let $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$.
  $Ax = b \iff b$ is a linear combination of columns of A.

$$b = \sum_{j=1}^{n} x_j \, a_j$$

Matrix Multiplication

- Let $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$.
  $Ax = b \iff b$ is a linear combination of columns of A.

$$b = \sum_{j=1}^{n} x_j \, a_j$$

- **Outer product.** Let $u \in \mathbb{R}^n, v \in \mathbb{R}^m$. We call $uv^\top$ the outer product of u and v:

$$uv^\top = \begin{bmatrix} v_1 u & v_2 u & \cdots & v_m u \end{bmatrix}$$

Range, Kernel and Rank

- **Range** of a matrix A is the span of its columns.

Range, Kernel and Rank

- **Range** of a matrix A is the span of its columns.
- **Kernel** or *Null Space* of a matrix A is the space of all x such that $A x=\mathbf{0}$.

$$\operatorname{null}(A)=\{\mathbf{0}\} \iff A \text{ is injective.}$$

Range, Kernel and Rank

- **Range** of a matrix A is the span of its columns.
- **Kernel** or *Null Space* of a matrix A is the space of all x such that $Ax = \mathbf{0}$.

$$\mathrm{null}(A) = \{\mathbf{0}\} \iff A \text{ is injective.}$$

- **Rank** of a matrix A is the dimension of its range. It's equal to dim(col space) = dim(row space).

$$\dim \mathrm{null}(A) + \mathrm{rank}(A) = \#\text{cols of A}$$

Inverse

- A is *invertible* or *nonsingular* iff it is square and full rank. Equivalently, having $\det(A) \neq 0$, or $\ker(A) = \{\mathbf{0}\}$.

Inverse

- A is *invertible* or *nonsingular* iff it is square and full rank. Equivalently, having $\det(A) \neq 0$, or $\ker(A) = \{\mathbf{0}\}$.
- Multiplication by A^{-1} is a change of basis:

$$\begin{tikzcd}
\boxed{\begin{array}{c}\text{Coefficients of}\\ \text{expansion of } b\\ \text{in } \{e_1, \ldots, e_n\}\end{array}} \arrow[rr, bend left, "\times A^{-1}"] & & \boxed{\begin{array}{c}\text{Coefficients of}\\ \text{expansion of } b\\ \text{in } \{a_1, \ldots, a_n\}\end{array}} \arrow[ll, bend left, "\times A"]
\end{tikzcd}$$

Orthogonality

- The usual **inner product** of two vectors x and y in $\mathbb{R}^n$ is defined as $\langle x, y \rangle = x^\top y = \sum x_i y_i$.

Orthogonality

- The usual **inner product** of two vectors x and y in $\mathbb{R}^n$ is defined as $\langle x, y \rangle = x^\top y = \sum x_i y_i$.
- Two vectors are *orthogonal* if their inner product is zero.

Orthogonality

- The usual **inner product** of two vectors x and y in $\mathbb{R}^n$ is defined as $\langle x, y \rangle = x^\top y = \sum x_i y_i$.
- Two vectors are *orthogonal* if their inner product is zero.
- Let $\{q_1, \ldots, q_n\}$ be a set of pairwise orthogonal unit vectors in $\mathbb{R}^n$. Then

$$\forall v \in \mathbb{R}^n : \quad v = \sum_{i=1}^{n} (q_i^\top v) q_i = \sum_{i=1}^{n} (q_i q_i^\top) v$$

  Note: $q_i q_i^\top$ is orthogonal projection onto direction q_i, which is a rank-one operator.

Unitary Matrices

- A real square matrix U is **unitary** or *orthogonal* if $U^\top U = UU^\top = I$, *i.e.*, $\langle u_i, u_j \rangle = \delta_{i,j}$.

Unitary Matrices

- A real square matrix U is **unitary** or *orthogonal* if $U^\top U = UU^\top = I$, *i.e.*, $\langle u_i, u_j \rangle = \delta_{i,j}$.
- If U is unitary, then it preserves angles,

  $$\langle Ux, Uy \rangle = \langle x, y \rangle \, ,$$

  as well as lengths,

  $$\|Ux\|_2 = \|x\|_2.$$

  If $\det(U) = 1$, then U is a rigid rotation, and if $\det(U) = -1$, then U includes a reflection.

Problem

By considering what space is spanned by the first n columns of R, show that if R is a nonsingular $m \times m$ upper-triangular matrix, then R^{-1} is also upper-triangular. (The analogous result also holds for lower-triangular matrices.)

Problem

By considering what space is spanned by the first n columns of R, show that if R is a nonsingular $m \times m$ upper-triangular matrix, then R^{-1} is also upper-triangular. (The analogous result also holds for lower-triangular matrices.)

Problem

Show that if a matrix is both triangular and unitary, then it is diagonal.

Norms

- The class of p-norms:

Norms

- The class of p-norms:
  - $\|x\|_1 = \sum_{i=1}^{n} |x_i|$

Norms

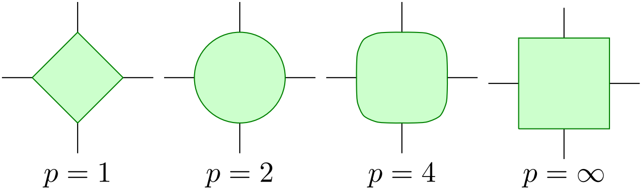
- The class of p-norms:
  - $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
  - $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ for $p \in (1, \infty)$

Norms

- The class of p-norms:
  - $\|x\|_1 = \sum_{i=1}^n |x_i|$
  - $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $p \in (1, \infty)$
  - $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$

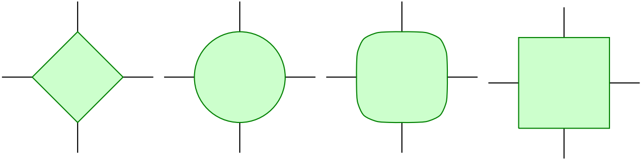
Norms

- The class of p-norms:
  - $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
  - $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ for $p \in (1, \infty)$
  - $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$
- Unit balls:



Norms

- The class of p-norms:
  - $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
  - $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ for $p \in (1, \infty)$
  - $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$
- Unit balls:



$p = 1$ $\quad p = 2$ $\quad p = 4$ $\quad p = \infty$

- The *Hölder inequality*. (Case $p = q = 2$ is known as *Cauchy-Schwartz* inequality)

$$\langle x, y \rangle \le \|x\|_p \|y\|_q, \text{ for } 1/p + 1/q = 1.$$

Matrix Norms

- We can view a matrix as a *linear operator*, and we can define norms on the space of linear operators. A famous norm is the **operator norm** of a matrix A. Let $A : (\mathbb{R}^n, \|\cdot\|_p) \to (\mathbb{R}^m, \|\cdot\|_q)$. Then we define

$$\|A\|_{(p,q)} := \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_q}{\|x\|_p} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|_p = 1}} \|Ax\|_q$$

Matrix Norms

- We can view a matrix as a *linear operator*, and we can define norms on the space of linear operators. A famous norm is the **operator norm** of a matrix A. Let $A : (\mathbb{R}^n, \|\cdot\|_p) \to (\mathbb{R}^m, \|\cdot\|_q)$. Then we define

$$\|A\|_{(p,q)} := \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_q}{\|x\|_p} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|_p = 1}} \|Ax\|_q$$

- Defines the maximum *stretch* of the unit ball.

Matrix Norms

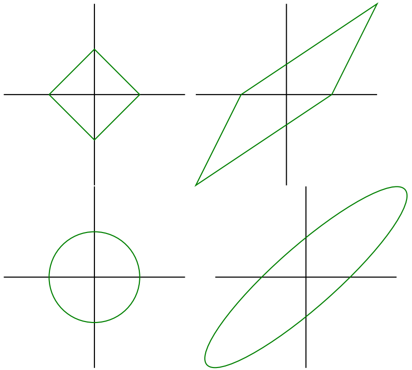
- We can view a matrix as a *linear operator*, and we can define norms on the space of linear operators. A famous norm is the **operator norm** of a matrix A. Let $A : (\mathbb{R}^n, \|\cdot\|_p) \to (\mathbb{R}^m, \|\cdot\|_q)$. Then we define

$$\|A\|_{(p,q)} := \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_q}{\|x\|_p} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|_p = 1}} \|Ax\|_q$$

- Defines the maximum *stretch* of the unit ball.
- When $p = q$ we just write $\|A\|_p$.
  e.g. $\|A\|_2$ is the largest *singular value* of A.

Matrix Norms (Example)

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix}$$



$$\|A\|_1 = 4$$

$$\|A\|_2 \approx 2.92$$

Norms

- Other norms can be defined for matrices. A famous example is the **Frobenious Norm**:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

Norms

- Other norms can be defined for matrices. A famous example is the **Frobenious Norm**:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

- All norms in finite-dimensional vector spaces are *equivalent*, that is, there exists positive constants C_1, C_2, such that

$$C_1 \cdot \|x\|_2 \leq \|x\|_1 \leq C_2 \cdot \|x\|_2$$

  Or, one can squeeze ones unit ball inside the other one.

Notes on Inner Products and Norms

One may introduce different inner products on $\mathbb{R}^n$. Formally, any bilinear[†] symmetric function which satisfies $\langle x, x\rangle \geq 0$ (and $\langle x, x\rangle = 0$ if and only if $x = 0$) could be an inner product. In $\mathbb{R}^n$, for example, it is enough to define the inner product of all $\langle e_i, e_j\rangle := a_{i,j}$. By bilinearity, it will be defined over all $\mathbb{R}^n$. Also, it is assumed that the inner product has values in $\mathbb{R}$. The more general case is when it takes values in $\mathbb{C}$. In that case, we should have $\langle x, y\rangle = \overline{\langle y, x\rangle}$.

Problem

The only p-norm that is induced by some inner product is only the 2-norm: (a) Prove that if the norm $\|\cdot\|$ comes from an inner product, then it should satisfy the parallelogram law:

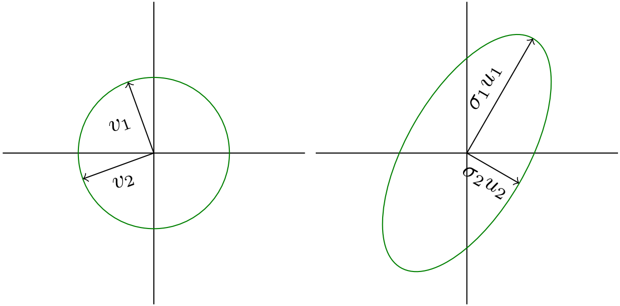
$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

(b) Show that this could happen if and only if $p = 2$.

[†] = linear in both arguments

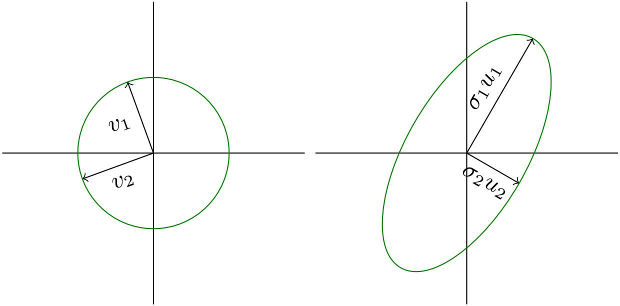
Singular Values and Singular Vectors

- **Theorem.** The image of the unit sphere under a linear transform is always a hyperellipse.



Singular Values and Singular Vectors

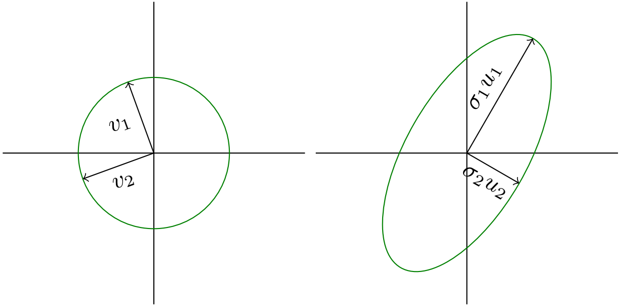
- **Theorem.** The image of the unit sphere under a linear transform is always a hyperellipse.



- We have $Av_i = \sigma_i u_i$.

Singular Values and Singular Vectors

- **Theorem.** The image of the unit sphere under a linear transform is always a hyperellipse.



- We have $Av_i = \sigma_i u_i$.
- v_1, v_2 are called *right singular vectors* and u_1, u_2 the *left singular vectors*. Also σ_1, σ_2 are *singular values*.

Singular Value Decomposition (SVD)

- We can decompose any matrix A in the form

$$A = U\Sigma V^\top,$$

where U and V are unitary and Σ is a diagonal matrix, *i.e.*

$$U = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix},$$

$$V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix},$$

$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n).$$

Singular Value Decomposition (SVD)

- We can decompose any matrix A in the form

$$A = U \Sigma V^\top,$$

where U and V are unitary and Σ is a diagonal matrix, *i.e.*

$$U = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix},$$

$$V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix},$$

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n).$$

- This is the **Full SVD**. There is also a **Reduced SVD**…

Eigenvalues and Eigenvectors

- If for some vector $v \neq \mathbf{0}$ we have $Av = \lambda v$ then v is an **eigenvector** of A associated to the **eigenvalue** λ. In this case we have $(A - \lambda I)v = \mathbf{0}$, and this can only happen when $\det(A - \lambda I) = 0$.

Eigenvalues and Eigenvectors

- If for some vector $v \neq \mathbf{0}$ we have $Av = \lambda v$ then v is an **eigenvector** of A associated to the **eigenvalue** λ. In this case we have $(A - \lambda I)v = \mathbf{0}$, and this can only happen when $\det(A - \lambda I) = 0$.
- If A is real and symmetric, then all eigenvalues are real and eigenvectors can be chosen to be orthogonal to each other.

Eigenvalues and Eigenvectors

- If for some vector $v \neq \mathbf{0}$ we have $Av = \lambda v$ then v is an **eigenvector** of A associated to the **eigenvalue** λ. In this case we have $(A - \lambda I)v = \mathbf{0}$, and this can only happen when $\det(A - \lambda I) = 0$.
- If A is real and symmetric, then all eigenvalues are real and eigenvectors can be chosen to be orthogonal to each other.
- There is also an **Eigenvalue Decomposition** for a diagonalizable square matrix:

$$A = X \Lambda X^{-1},$$

which is different from SVD.

Eigenvalues and Eigenvectors

- If for some vector $v \neq \mathbf{0}$ we have $Av = \lambda v$ then v is an **eigenvector** of A associated to the **eigenvalue** λ. In this case we have $(A - \lambda I)v = \mathbf{0}$, and this can only happen when $\det(A - \lambda I) = 0$.
- If A is real and symmetric, then all eigenvalues are real and eigenvectors can be chosen to be orthogonal to each other.
- There is also an **Eigenvalue Decomposition** for a diagonalizable square matrix:

  $$A = X \Lambda X^{-1},$$

  which is different from SVD.
- $\sigma_i^2(A) = \lambda_i(A^\top A) = \lambda_i(AA^\top)$.

Problem

Let $A \in \mathbb{C}^{m \times m}$ *be hermitian.*[‡] *(a) Prove that all eigenvalues of* A *are real. (b) Prove that if* x *and* y *are eigenvectors corresponding to distinct eigenvalues, then* x *and* y *are orthogonal.*

[‡] The *adjoint* of an $m \times n$ matrix A, written A^*, is the $n \times m$ matrix whose i, j entry is the complex conjugate of j, i entry of A. If $A = A^*$, A is *hermitian*.

Problem

Let $A \in \mathbb{C}^{m\times m}$ be hermitian.[‡] *(a) Prove that all eigenvalues of A are real. (b) Prove that if x and y are eigenvectors corresponding to distinct eigenvalues, then x and y are orthogonal.*

Problem

If u and v are m-vectors, the matrix $A = I + uv^*$ is known as a rank-one perturbation of the identity. Show that if A is nonsingular, then its inverse has the form $A^{-1} = I + \alpha uv^*$ for some scalar α, and give an expression for α. For what u and v is A singular? If it is singular, what is $\mathrm{null}(A)$*?*

[‡] The *adjoint* of an $m \times n$ matrix A, written A^*, is the $n \times m$ matrix whose i, j entry is the complex conjugate of j, i entry of A. If $A = A^*$, A is *hermitian*.

Outline

Notion of the Derivative

- As you may recall, for real differentiable functions $f : \Omega \subseteq \mathbb{R} \to \mathbb{R}$, the derivative $f'(x) = \frac{df}{dx}(x)$ is the *slope* of the **tangent** line at the point x. This notion is geometrically plausible, but unfortunately hard to generalize.

Notion of the Derivative

- As you may recall, for real differentiable functions $f : \Omega \subseteq \mathbb{R} \to \mathbb{R}$, the derivative $f'(x) = \frac{df}{dx}(x)$ is the *slope* of the **tangent** line at the point x. This notion is geometrically plausible, but unfortunately hard to generalize.
- A better notion for derivative would be *the best linear approximation* of a function near a point x.

Notion of the Derivative

- As you may recall, for real differentiable functions $f : \Omega \subseteq \mathbb{R} \to \mathbb{R}$, the derivative $f'(x) = \frac{df}{dx}(x)$ is the *slope* of the **tangent** line at the point x. This notion is geometrically plausible, but unfortunately hard to generalize.
- A better notion for derivative would be *the best linear approximation* of a function near a point x.
- Formally, let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $x_0 \in \Omega$. We call f to be differentiable at x_0 iff there is a *linear* function $Df(x_0) : \mathbb{R}^n \to \mathbb{R}^m$, for which we have

$$\lim_{\|h\| \to 0} \frac{|f(x_0 + h) - f(x_0) - Df(x_0)(h)|}{\|h\|} = 0$$

Notion of the Derivative

- As you may recall, for real differentiable functions $f : \Omega \subseteq \mathbb{R} \to \mathbb{R}$, the derivative $f'(x) = \frac{df}{dx}(x)$ is the *slope* of the **tangent** line at the point x. This notion is geometrically plausible, but unfortunately hard to generalize.
- A better notion for derivative would be *the best linear approximation* of a function near a point x.
- Formally, let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $x_0 \in \Omega$. We call f to be differentiable at x_0 iff there is a *linear* function $Df(x_0) : \mathbb{R}^n \to \mathbb{R}^m$, for which we have

$$\lim_{\|h\| \to 0} \frac{|f(x_0 + h) - f(x_0) - Df(x_0)(h)|}{\|h\|} = 0$$

- We call $Df(x_0)$ the derivative of f at the point x_0.

Derivative

- Take $f(x) = (f_1(x), \ldots, f_m(x))$. We call f_i the **components** of f. If the derivative exists, then we have

$$Df(x_0) = \left[\frac{\partial f_i}{\partial x_j}(x_0)\right]_{1 \le i \le m, 1 \le j \le n},$$

where $\frac{\partial f_i}{\partial x_j}(x_0)$ is the **partial derivative** of f_i w.r.t. x_j at the point x_0, namely

$$\frac{\partial f_i}{\partial x_j}(x_0) = \lim_{\epsilon \to 0} \frac{f_i(x_0 + \epsilon e_j) - f_i(x_0)}{\epsilon}.$$

Derivative

- Take $f(x) = (f_1(x), \ldots, f_m(x))$. We call f_i the **components** of f. If the derivative exists, then we have

$$Df(x_0) = \left[\tfrac{\partial f_i}{\partial x_j}(x_0)\right]_{1 \le i \le m, 1 \le j \le n},$$

  where $\frac{\partial f_i}{\partial x_j}(x_0)$ is the **partial derivative** of f_i w.r.t. x_j at the point x_0, namely

$$\frac{\partial f_i}{\partial x_j}(x_0) = \lim_{\epsilon \to 0} \frac{f_i(x_0 + \epsilon e_j) - f_i(x_0)}{\epsilon}.$$

- If for a function f, all partial derivatives exist and *are continuous* at the point x_0 then f is continuously differentiable at x_0 and its derivative would be the matrix $Df(x_0)$ above.

Notion of the Gradient

- Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be a real-valued differentiable function. Then we have

$$Df(x_0) = \left[\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0)\right] =: \nabla f(x_0)^\top$$

Notion of the Gradient

- Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be a real-valued differentiable function. Then we have

$$Df(x_0) = \left[\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0)\right] =: \nabla f(x_0)^\top$$

- The **gradient** of f has the following properties:

Notion of the Gradient

- Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be a real-valued differentiable function. Then we have

$$Df(x_0) = \left[\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0)\right] =: \nabla f(x_0)^\top$$

- The **gradient** of f has the following properties:
  - It points to the direction in which f has the maximum rate of increase. (likewise, $-\nabla f$ points to the direction of maximum decrease)

Notion of the Gradient

- Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be a real-valued differentiable function. Then we have

$$Df(x_0) = \left[\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0)\right] =: \nabla f(x_0)^\top$$

- The **gradient** of f has the following properties:
  - It points to the direction in which f has the maximum rate of increase. (likewise, $-\nabla f$ points to the direction of maximum decrease)
  - It is always orthogonal to the contour line $\{x \ : \ f(x) = f(x_0)\}$.

Notion of the Gradient

- Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be a real-valued differentiable function. Then we have

$$Df(x_0) = \left[\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0)\right] =: \nabla f(x_0)^\top$$

- The **gradient** of f has the following properties:
  - It points to the direction in which f has the maximum rate of increase. (likewise, $-\nabla f$ points to the direction of maximum decrease)
  - It is always orthogonal to the contour line $\{x \;:\; f(x) = f(x_0)\}$.
  - If f attains a local minimum (or maximum) at some point x_0, then $\nabla f(x_0) = \mathbf{0}$. (The **first-order condition**)

Notion of the Gradient

- Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be a real-valued differentiable function. Then we have

$$Df(x_0) = \left[\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0)\right] =: \nabla f(x_0)^\top$$

- The **gradient** of f has the following properties:
  - It points to the direction in which f has the maximum rate of increase. (likewise, $-\nabla f$ points to the direction of maximum decrease)
  - It is always orthogonal to the contour line $\{x \ : \ f(x) = f(x_0)\}$.
  - If f attains a local minimum (or maximum) at some point x_0, then $\nabla f(x_0) = \mathbf{0}$. (The **first-order condition**)
  - We have the following (first-order) approximation for x sufficiently close to x_0:

$$f(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + o(\|x - x_0\|)$$

Chain Rule

Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $g : \Omega' \subseteq \mathbb{R}^p \to \mathbb{R}^n$. Assume g is differentiable at x_0 and $g(x_0) \in \Omega$ and f is differentiable at $g(x_0)$. Then $f \circ g : \Omega' \to \mathbb{R}^m$ is differntiable at x_0 and we have

$$D(f \circ g)(x_0) = Df(g(x_0)) \circ Dg(x_0)$$

A good example is the directional derivatives. Assume $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$. Let $u \in \mathbb{R}^n$. We want to find the rate of change of f in the direction of u, *i.e.* $\frac{d}{dt} f(x_0 + tu)$ for $t = 0$. Define $g(t) = x_0 + tu$. We have

$$D(f \circ g)(0) = Df(g(0)) \circ Dg(0) = \nabla f(x_0)^\top u.$$

Second-Order Approximation

- The error of the first-order approximation is sub-linear.
  Can we do better?

Second-Order Approximation

- The error of the first-order approximation is sub-linear. Can we do better?
- In the single-variable regime, we can use quadratic polynomials to approximate a function in some neighborhood.

Second-Order Approximation

- The error of the first-order approximation is sub-linear. Can we do better?
- In the single-variable regime, we can use quadratic polynomials to approximate a function in some neighborhood.
- We need to understand what are quadratic functions in multi-dimensional case and try to approximate our function.

Second-Order Approximation

- The error of the first-order approximation is sub-linear. Can we do better?
- In the single-variable regime, we can use quadratic polynomials to approximate a function in some neighborhood.
- We need to understand what are quadratic functions in multi-dimensional case and try to approximate our function.
- We expect our new approximation's error has a faster convergence to 0 than $\|x - x_0\|^2$.

Multidimensional Quadratic Functions

- Let A be an $n \times n$ symmetric matrix. We define the **quadratic form** induced by A to be

$$f \ : \ x \in \mathbb{R}^n \mapsto x^\top A x.$$

Multidimensional Quadratic Functions

- Let A be an $n \times n$ symmetric matrix. We define the **quadratic form** induced by A to be

$$f \ : \ x \in \mathbb{R}^n \mapsto x^\top A x.$$

- Note that the quadratic form is a weighted sum of all possible second degree terms, *e.g.* $x_i x_j$ or x_i^2.

Multidimensional Quadratic Functions

- Let A be an $n \times n$ symmetric matrix. We define the **quadratic form** induced by A to be

$$f \ : \ x \in \mathbb{R}^n \mapsto x^\top A x.$$

- Note that the quadratic form is a weighted sum of all possible second degree terms, *e.g.* $x_i x_j$ or x_i^2.
- We call A a **positive (negative) definite** matrix, iff all eigenvalues of A are positive (negative). We say A is a **positive semi-definite** or p.s.d., if all eigenvalues are nonnegative.

Multidimensional Quadratic Functions

- Let A be an $n \times n$ symmetric matrix. We define the **quadratic form** induced by A to be

$$f \; : \; x \in \mathbb{R}^n \mapsto x^\top A x.$$

- Note that the quadratic form is a weighted sum of all possible second degree terms, *e.g.* $x_i x_j$ or x_i^2.
- We call A a **positive (negative) definite** matrix, iff all eigenvalues of A are positive (negative). We say A is a **positive semi-definite** or p.s.d., if all eigenvalues are nonnegative.
- If A is p.s.d., then the contour levels of f are concentric ellipsoids.

Multidimensional Quadratic Functions

- Let A be an $n \times n$ symmetric matrix. We define the **quadratic form** induced by A to be

$$f \ : \ x \in \mathbb{R}^n \mapsto x^\top A x.$$

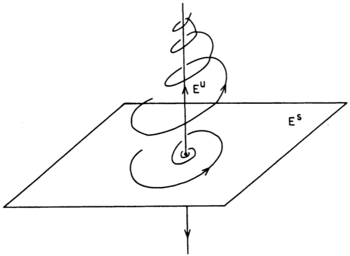
- Note that the quadratic form is a weighted sum of all possible second degree terms, *e.g.* $x_i x_j$ or x_i^2.
- We call A a **positive (negative) definite** matrix, iff all eigenvalues of A are positive (negative). We say A is a **positive semi-definite** or p.s.d., if all eigenvalues are nonnegative.
- If A is p.s.d., then the contour levels of f are concentric ellipsoids.
- One can prove that $\nabla f(x) = 2Ax$.

A Reminder from ODE Theory*

Consider the linear ODE $\dot{x} = Ax$, with A having real nonzero eigenvalues $\lambda_1, \dots, \lambda_n$ and eigenvectors $v_1, \dots, v_n$. Define

$$E^s = \mathrm{Span}\{v_j \mid \lambda_j < 0\},$$
$$E^u = \mathrm{Span}\{v_j \mid \lambda_j > 0\},$$

which are *stable*, and *unstable* subspaces.



Explore. Consider the ODE $\dot{x} = -\nabla f(x)$, when f is a quadratic form. Try to understand the shape of the function using the stable and unstable subspaces.

Positive Definiteness

- Here are some equivalent conditions. A is a real symmetric matrix.

Positive Definiteness

- Here are some equivalent conditions. A is a real symmetric matrix.
  - A is p.s.d. $\iff$ $x^\top Ax \geq 0$ for all x.

Positive Definiteness

- Here are some equivalent conditions. A is a real symmetric matrix.
  - A is p.s.d. $\iff x^\top Ax \geq 0$ for all x.
  - A is p.d. $\iff x^\top Ax > 0$ for all $x \neq 0$.

Positive Definiteness

- Here are some equivalent conditions. A is a real symmetric matrix.
  - A is p.s.d. $\iff x^\top Ax \geq 0$ for all x.
  - A is p.d. $\iff x^\top Ax > 0$ for all $x \neq 0$.
- **Cholskey Decomposition.** A can be decomposed as

$$A = LL^\top,$$

  where L is a lower triangular matrix with positive diagonal entries (if A is p.s.d., nonnegative entries).

Example

Let $A \in \mathbb{R}^{d \times d}$ be a symmetric p.d. matrix, and set $c \in \mathbb{R}^d$. How does the set

$$E = \{x \in \mathbb{R}^d \mid (x - c)^\top A^{-1} (x - c) \leq 1\}$$

look like?

Example

Let $A \in \mathbb{R}^{d\times d}$ be a symmetric p.d. matrix, and set $c \in \mathbb{R}^d$. How does the set

$$E = \{x \in \mathbb{R}^d \mid (x-c)^\top A^{-1}(x-c) \le 1\}$$

look like?

Answer.

Let $A = LL^\top$ be the Cholskey decomposition of A. Thus $A^{-1} = L^{-\top}L^{-1}$. By a change of variable $y = L^{-1}(x-c)$, we get

$$E = \{Ly + c \mid \|y\|_2^2 = y^\top y \le 1\}.$$

Thus, E is the result of an affine transformation applied to the unit Euclidean ball, which is an **ellipsoid**. $\square$

The Hessian

- Assume $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ is twice-differentiable at x_0. Then there exists some symmetric matrix $D^2 f(x_0)$ which we call the **Hessian** of f at x_0, with

$$D^2 f(x_0) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0)\right]_{1 \leq i,j \leq n},$$

where $\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(x_0)$.

The Hessian

- Assume $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ is twice-differentiable at x_0. Then there exists some symmetric matrix $D^2 f(x_0)$ which we call the **Hessian** of f at x_0, with

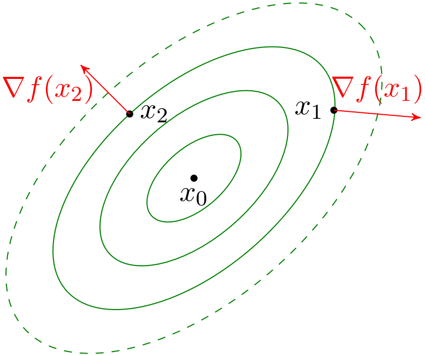
$$D^2 f(x_0) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0) \right]_{1 \leq i,j \leq n},$$

  where $\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(x_0)$.
- We have the following (second-order) approximation for x sufficiently close to x_0:

$$\begin{aligned} f(x) = f(x_0) &+ \nabla f(x_0)^\top (x - x_0) \\ &+ \frac{1}{2}(x - x_0)^\top D^2 f(x_0)(x - x_0) + o(\|x - x_0\|^2) \end{aligned}$$

All-in-one Picture

In this example, we assume that $f : \mathbb{R}^2 \to \mathbb{R}$ is twice-differentiable, having a local minimum at x_0. This picture demonstrates a (sufficiently small) neighborhood of x_0:



Note that around the local minimum, the Hessian of f is positive semi-definite, thus the elliptic contour lines.

KKT Optimality Conditions

Let $f, g_i, h_j : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable convex functions for $i = 1, \ldots, k$ and $j = 1, \ldots, l$. Consider the optimization problem

$$\begin{aligned} & \text{minimize}\, f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \\ & h_j(x) = 0. \end{aligned}$$

Assuming the Slater's Condition, if $x^\star$ is a local minimum of the optimization problem, then there exists constants λ_i and μ_j such that

- $\nabla f(x^\star) + \sum_i \lambda_i \nabla g_i(x^\star) + \sum_j \mu_j \nabla h_j(x^\star) = 0$, (Stationarity)

KKT Optimality Conditions

Let $f, g_i, h_j : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable convex functions for $i = 1, \ldots, k$ and $j = 1, \ldots, l$. Consider the optimization problem

$$\begin{aligned} &\text{minimize}\, f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \\ & h_j(x) = 0. \end{aligned}$$

Assuming the Slater's Condition, if $x^\star$ is a local minimum of the optimization problem, then there exists constants λ_i and μ_j such that

- $\nabla f(x^\star) + \sum_i \lambda_i \nabla g_i(x^\star) + \sum_j \mu_j \nabla h_j(x^\star) = 0$, (Stationarity)
- $g_i(x^\star) \leq 0$ and $h_j(x^\star) = 0$, (Primal Feasibility)

KKT Optimality Conditions

Let $f, g_i, h_j \colon \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable convex functions for $i = 1, \ldots, k$ and $j = 1, \ldots, l$. Consider the optimization problem

$$\begin{aligned} &\text{minimize}\, f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \\ & h_j(x) = 0. \end{aligned}$$

Assuming the Slater's Condition, if $x^\star$ is a local minimum of the optimization problem, then there exists constants λ_i and μ_j such that

- $\nabla f(x^\star) + \sum_i \lambda_i \nabla g_i(x^\star) + \sum_j \mu_j \nabla h_j(x^\star) = 0$, (Stationarity)
- $g_i(x^\star) \leq 0$ and $h_j(x^\star) = 0$, (Primal Feasibility)
- $\lambda_i \geq 0$ for $i = 1, \ldots, k$, (Dual Feasibility)

KKT Optimality Conditions

Let $f, g_i, h_j \colon \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable convex functions for $i = 1, \ldots, k$ and $j = 1, \ldots, l$. Consider the optimization problem

$$\begin{aligned} &\text{minimize}\, f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \\ & h_j(x) = 0. \end{aligned}$$

Assuming the Slater's Condition, if $x^\star$ is a local minimum of the optimization problem, then there exists constants λ_i and μ_j such that

- $\nabla f(x^\star) + \sum_i \lambda_i \nabla g_i(x^\star) + \sum_j \mu_j \nabla h_j(x^\star) = 0$, (Stationarity)
- $g_i(x^\star) \leq 0$ and $h_j(x^\star) = 0$, (Primal Feasibility)
- $\lambda_i \geq 0$ for $i = 1, \ldots, k$, (Dual Feasibility)
- $\lambda_i \cdot g_i(x^\star) = 0$ for $i = 1, \ldots, k$. (Complementary Slackness)

Outline

Linear Algebra

Multivariate Analysis

Probability Theory

Basic Notations

- Let Ω be a set, which we call the **sample space**. This set contains all possible outcomes of an experiment. Note that this set depends on how we model the problem, *e.g.* in the problem of a single dart throw to a circular dartboard, we have the following possibilities:

Basic Notations

- Let Ω be a set, which we call the **sample space**. This set contains all possible outcomes of an experiment. Note that this set depends on how we model the problem, *e.g.* in the problem of a single dart throw to a circular dartboard, we have the following possibilities:
  - $\Omega_1 = \mathbb{R}^2$ (exact place of landing),

Basic Notations

- Let Ω be a set, which we call the **sample space**. This set contains all possible outcomes of an experiment. Note that this set depends on how we model the problem, *e.g.* in the problem of a single dart throw to a circular dartboard, we have the following possibilities:
  - $\Omega_1 = \mathbb{R}^2$ (exact place of landing),
  - $\Omega_2 = \{\text{Hit}, \text{Miss}\}$ (indicator),

Basic Notations

- Let Ω be a set, which we call the **sample space**. This set contains all possible outcomes of an experiment. Note that this set depends on how we model the problem, *e.g.* in the problem of a single dart throw to a circular dartboard, we have the following possibilities:
  - $\Omega_1 = \mathbb{R}^2$ (exact place of landing),
  - $\Omega_2 = \{\text{Hit}, \text{Miss}\}$ (indicator),
  - $\Omega_3 = \{0, 10, 20, \ldots, 100\}$ (score of the throw).

Basic Notations

- Let Ω be a set, which we call the **sample space**. This set contains all possible outcomes of an experiment. Note that this set depends on how we model the problem, *e.g.* in the problem of a single dart throw to a circular dartboard, we have the following possibilities:
  - $\Omega_1 = \mathbb{R}^2$ (exact place of landing),
  - $\Omega_2 = \{\text{Hit}, \text{Miss}\}$ (indicator),
  - $\Omega_3 = \{0, 10, 20, \ldots, 100\}$ (score of the throw).
- A family $\mathcal{F}$ of subsets of Ω, called the **events**, are also interesting for us. Rather than asking whether a certain outcome has happened, we want to ask harder questions. For example, if we want to ask whether the score is higher than 60 or not, we are asking about the event $\{80, 100\}$.

Basic Notations

- Let Ω be a set, which we call the **sample space**. This set contains all possible outcomes of an experiment. Note that this set depends on how we model the problem, *e.g.* in the problem of a single dart throw to a circular dartboard, we have the following possibilities:
  - $\Omega_1 = \mathbb{R}^2$ (exact place of landing),
  - $\Omega_2 = \{\text{Hit}, \text{Miss}\}$ (indicator),
  - $\Omega_3 = \{0, 10, 20, \dots, 100\}$ (score of the throw).
- A family $\mathcal{F}$ of subsets of Ω, called the **events**, are also interesting for us. Rather than asking whether a certain outcome has happened, we want to ask harder questions. For example, if we want to ask whether the score is higher than 60 or not, we are asking about the event $\{80, 100\}$.
- We also want to apply rules of logic. Taking "and" is translated to intersection of events, "or" is union, and "not" is complements. So we desire our family of events $\mathcal{F}$ to be closed under these operations.

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,
  - If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,
  - If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
    - If $A_1, A_2, \ldots$ are pairwise disjoint, $\mathbb{P}(\bigcup A_i) = \sum \mathbb{P}(A_i)$.

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,
  - If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
    - If $A_1, A_2, \ldots$ are pairwise disjoint, $\mathbb{P}(\bigcup A_i) = \sum \mathbb{P}(A_i)$.
- We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,
  - If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
    - If $A_1, A_2, \ldots$ are pairwise disjoint, $\mathbb{P}(\bigcup A_i) = \sum \mathbb{P}(A_i)$.
- We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.
- Examples:

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,
  - If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
    - If $A_1, A_2, \ldots$ are pairwise disjoint, $\mathbb{P}(\bigcup A_i) = \sum \mathbb{P}(A_i)$.
- We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.
- Examples:
  - $\Omega = \{1, \ldots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, and $\mathbb{P}(\{i\}) = \frac{1}{6}$ for all $i \in \Omega$.

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,
  - If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
    - If $A_1, A_2, \dots$ are pairwise disjoint, $\mathbb{P}(\bigcup A_i) = \sum \mathbb{P}(A_i)$.
- We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.
- Examples:
  - $\Omega = \{1, \dots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, and $\mathbb{P}(\{i\}) = \frac{1}{6}$ for all $i \in \Omega$.
  - $\Omega = \mathbb{N}$, $\mathcal{F} = \mathcal{P}(\mathbb{N})$, $\mathbb{P}(\{i\}) = \frac{1}{2^i}$ for all $i \in \mathbb{N}$.

Basic Notations

- We also assign a belief (which we may obtain through experiments, or just arbitrary) to each of these events. However, this assignment should be consistent. It is agreed that the following rules suffice to model our philosophy about beliefs and probabilities. If we assign to each event A, a probability $\mathbb{P}(A)$, we should have:
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
  - $\mathbb{P}(\Omega) = 1$,
  - If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
    - If $A_1, A_2, \dots$ are pairwise disjoint, $\mathbb{P}(\bigcup A_i) = \sum \mathbb{P}(A_i)$.
- We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.
- Examples:
  - $\Omega = \{1, \dots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, and $\mathbb{P}(\{i\}) = \frac{1}{6}$ for all $i \in \Omega$.
  - $\Omega = \mathbb{N}$, $\mathcal{F} = \mathcal{P}(\mathbb{N})$, $\mathbb{P}(\{i\}) = \frac{1}{2^i}$ for all $i \in \mathbb{N}$.
  - $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, $\mathbb{P}((a, b]) = b - a$.

Random Variables

- RVs are different points-of-view to the same probability space. We call a “measurable” function $X : \Omega \to \mathbb{R}$ a **random variable**. We now ask our questions through the lens of X.

Random Variables

- RVs are different points-of-view to the same probability space. We call a "measurable" function $X : \Omega \to \mathbb{R}$ a **random variable**. We now ask our questions through the lens of X.
- In the dart throwing problem, let's suppose $\Omega = \mathbb{D}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{D}^2)$ and $\mathbb{P}$ be the uniform measure, meaning that for $S \in \mathcal{F}$, $\mathbb{P}(S) = \text{area}(S)/\text{area}(\mathbb{D}^2)$. Take X to be

$$X(\omega) = \text{distance of } \omega \text{ to the center of dartboard.}$$

  Now we can ask, whether the throw was further than 0.3 cm, via asking about the event $\{\omega : X(\omega) \geq 0.3\}$. For brevity we write this event as $\{X \geq 0.3\}$.

Random Variables

- RVs are different points-of-view to the same probability space. We call a "measurable" function $X : \Omega \to \mathbb{R}$ a **random variable**. We now ask our questions through the lens of X.
- In the dart throwing problem, let's suppose $\Omega = \mathbb{D}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{D}^2)$ and $\mathbb{P}$ be the uniform measure, meaning that for $S \in \mathcal{F}$, $\mathbb{P}(S) = \text{area}(S)/\text{area}(\mathbb{D}^2)$. Take X to be

  $$X(\omega) = \text{distance of } \omega \text{ to the center of dartboard.}$$

  Now we can ask, whether the throw was further than 0.3 cm, via asking about the event $\{\omega : X(\omega) \geq 0.3\}$. For brevity we write this event as $\{X \geq 0.3\}$.
- Link between X and Ω is the **inverse image** X^{-1}.

Random Variables

- RVs are different points-of-view to the same probability space. We call a "measurable" function $X : \Omega \to \mathbb{R}$ a **random variable**. We now ask our questions through the lens of X.
- In the dart throwing problem, let's suppose $\Omega = \mathbb{D}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{D}^2)$ and $\mathbb{P}$ be the uniform measure, meaning that for $S \in \mathcal{F}$, $\mathbb{P}(S) = \text{area}(S)/\text{area}(\mathbb{D}^2)$. Take X to be

  $$X(\omega) = \text{distance of } \omega \text{ to the center of dartboard.}$$

  Now we can ask, whether the throw was further than 0.3 cm, via asking about the event $\{\omega : X(\omega) \geq 0.3\}$. For brevity we write this event as $\{X \geq 0.3\}$.
- Link between X and Ω is the **inverse image** X^{-1}.
- We can only ask a question $A \in \mathcal{B}(\mathbb{R})$ from X if the inverse image of A is already inside $\mathcal{F}$, *i.e.* $X^{-1}(A) \in \mathcal{F}$.

On Measurability*

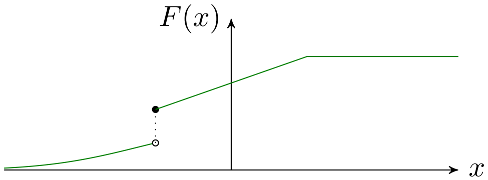
Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a function $X : \Omega \to \mathbb{R}$ is called *measurable* if it respects the structure of $\mathcal{F}$. That is, for all $t \in \mathbb{R}$ one should have

$$\{X \le t\} = X^{-1}((-\infty, t]) \in \mathcal{F}.$$

An example of a non-measurable function is as follows: Take $\Omega = \{\epsdice{1}, \epsdice{2}, \dots, \epsdice{6}\}$ to be the different faces of a die. Define $\mathcal{F} = \{\emptyset, \Omega, \{\epsdice{1}, \epsdice{2}, \epsdice{3}\}, \{\epsdice{4}, \epsdice{5}, \epsdice{6}\}\}$ to be the *information available about the experiment*. Then, the function $X = \mathbf{1}_{\{\epsdice{1}, \epsdice{2}\}}$ is *not* a random variable, as the inverse image $X^{-1}(\{1\}) = \{\epsdice{1}, \epsdice{2}\}$ is not a member of $\mathcal{F}$.
If you are interested, see more about Borel sets, Lebesgue measure, and Measurable functions.

Distributions

- Any random variable induces a probability measure on $\mathbb{R}$, *i.e.* for each interval I we assign the probability $\mathbb{P}(X^{-1}(I))$.

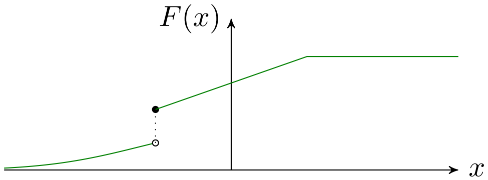


Distributions

- Any random variable induces a probability measure on $\mathbb{R}$, *i.e.* for each interval I we assign the probability $\mathbb{P}(X^{-1}(I))$.
- This defines a right-continuous nondecreasing function $F : \mathbb{R} \to [0, 1]$,

$$F(x) := \mathbb{P}(\{X \leq x\}) = \mathbb{P}(X \leq x) = \mathbb{P}(X^{-1}((-\infty, x]))),$$

  which we call the **cumulative distribution function** (or CDF).

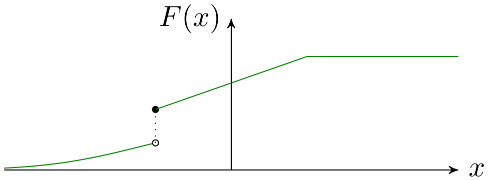


Distributions

- Any random variable induces a probability measure on $\mathbb{R}$, *i.e.* for each interval I we assign the probability $\mathbb{P}(X^{-1}(I))$.
- This defines a right-continuous nondecreasing function $F : \mathbb{R} \to [0, 1]$,

$$F(x) := \mathbb{P}(\{X \leq x\}) = \mathbb{P}(X \leq x) = \mathbb{P}(X^{-1}((-\infty, x]))),$$

  which we call the **cumulative distribution function** (or CDF).
  - We have $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$.

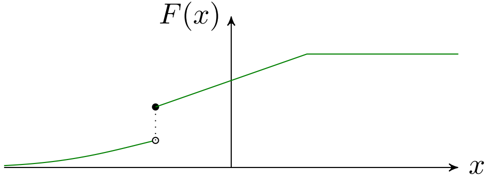


Distributions

- Any random variable induces a probability measure on $\mathbb{R}$, *i.e.* for each interval I we assign the probability $\mathbb{P}(X^{-1}(I))$.
- This defines a right-continuous nondecreasing function $F : \mathbb{R} \to [0, 1]$,

$$F(x) := \mathbb{P}(\{X \leq x\}) = \mathbb{P}(X \leq x) = \mathbb{P}(X^{-1}((-\infty, x]))),$$

  which we call the **cumulative distribution function** (or CDF).
    - We have $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$.
    - $\mathbb{P}(a < X \leq b) = F(b) - F(a)$.



Densities

- Exactly like physical concept of density, we can define density for a random variable (if it is regular enough). For a random variable X we define

$$f(x) := \lim_{\substack{|I| \to 0 \\ x \in I}} \frac{\mathbb{P}(X \in I)}{|I|},$$

where $\mathbb{P}(\cdots)$ replaces "mass" and $|I|$ is in place of "volume".

Densities

- Exactly like physical concept of density, we can define density for a random variable (if it is regular enough). For a random variable X we define

$$f(x) := \lim_{\substack{|I| \to 0 \\ x \in I}} \frac{\mathbb{P}(X \in I)}{|I|},$$

  where $\mathbb{P}(\cdots)$ replaces "mass" and $|I|$ is in place of "volume".
- If F is differentiable at x, then $f(x) = F'(x)$ and by FTC

$$F(x) = \int_{-\infty}^{x} f(y) dy.$$

Densities

- Exactly like physical concept of density, we can define density for a random variable (if it is regular enough). For a random variable X we define

$$f(x) := \lim_{\substack{|I|\to 0\\ x\in I}} \frac{\mathbb{P}(X \in I)}{|I|},$$

  where $\mathbb{P}(\cdots)$ replaces "mass" and $|I|$ is in place of "volume".
- If F is differentiable at x, then $f(x) = F'(x)$ and by FTC

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

- The value of $f(x)$ can be used to estimate probabilities, *e.g.* if $f(x) = 2$, then for a small interval I of size ϵ around x, we know that $\mathbb{P}(X \in I) \approx 2\epsilon$.

Motto!

With random variables and their densities (or distributions) we can even forget about Ω and just look at the probability space that is defined on $\mathbb{R}$ via X. So the following holds:

> *If $(\Omega, \mathcal{F}, \mathbb{P})$ and X are known, we can find the distribution of X (and its density, if it exists).*

Motto!

With random variables and their densities (or distributions) we can even forget about Ω and just look at the probability space that is defined on $\mathbb{R}$ via X. So the following holds:

> *If* $(\Omega, \mathcal{F}, \mathbb{P})$ *and* X *are known, we can find the distribution of* X *(and its density, if it exists).*
>
> *If we know* $F(x)$ *or* $f(x)$*, we can build a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and a random variable* X*, such that the distribution of* X *is exactly* $F(x)$*.*

Joint Distribution and Marginals

- Let X, Y be two random variables over the same probability space. Then we can define the **joint distribution** as

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

Joint Distribution and Marginals

- Let X, Y be two random variables over the same probability space. Then we can define the **joint distribution** as

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

- The **joint density** can also be defined as

$$f_{X,Y}(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y).$$

Joint Distribution and Marginals

- Let X, Y be two random variables over the same probability space. Then we can define the **joint distribution** as

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y).$$

- The **joint density** can also be defined as

$$f_{X,Y}(x,y) = \frac{\partial^2 F}{\partial x \partial y}(x,y).$$

- Given the joint distribution, one can find the distribution of each of variables by **marginalizing**:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad F_X(x) = F_{X,Y}(x,\infty)$$

Independence

- Two "events" A and B are said to be independent if we have
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Independence

- Two "events" A and B are said to be independent if we have
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$
- Two "RVs" X and Y are independent if their joint distribution function factorizes, *i.e.*
$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Independence

- Two "events" A and B are said to be independent if we have
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$
- Two "RVs" X and Y are independent if their joint distribution function factorizes, *i.e.*
$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$
- A sequence of n RVs $X_1, \ldots, X_n$ are said to be independent, iff their joint distribution factorizes. Note that if X_i are pairwise independent, it does *not* follow that they are independent.

Conditional Probability

- How does information affect our belief?

Conditional Probability

- How does information affect our belief?
- "knowing" that an event B has occured, what is the probability of A happening?

Conditional Probability

- How does information affect our belief?
- "knowing" that an event B has occured, what is the probability of A happening?
- Denote by $\mathbb{P}(A|B)$ by conditional probability of A given B.

Conditional Probability

- How does information affect our belief?
- "knowing" that an event B has occured, what is the probability of A happening?
- Denote by $\mathbb{P}(A|B)$ by conditional probability of A given B.
- This can be defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad (\mathbb{P}(B) \neq 0)$$

Conditional Probability

- How does information affect our belief?
- "knowing" that an event B has occured, what is the probability of A happening?
- Denote by $\mathbb{P}(A|B)$ by conditional probability of A given B.
- This can be defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad (\mathbb{P}(B) \neq 0)$$

- **Law of total probability.** Let $A_1, \ldots, A_n$ be a partition of Ω. We have

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Conditional Probability

- How does information affect our belief?
- "knowing" that an event B has occured, what is the probability of A happening?
- Denote by $\mathbb{P}(A|B)$ by conditional probability of A given B.
- This can be defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad (\mathbb{P}(B) \neq 0)$$

- **Law of total probability.** Let $A_1, \ldots, A_n$ be a partition of Ω. We have

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

- **Bayes Rule.**

$$\mathbb{P}(A|B) = \mathbb{P}(B|A)\frac{\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Bayes Rule and the Chain Rule

- Let $A_1, \ldots, A_n$ be a partition of Ω. We have

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

Bayes Rule and the Chain Rule

- Let $A_1, \ldots, A_n$ be a partition of Ω. We have

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^n \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

- **Chain Rule.** Let $A_1, \ldots, A_n$ be arbitrary events. We have

$$\begin{aligned}\mathbb{P}(A_1, \ldots, A_n) = &\mathbb{P}(A_1)\times \\ &\mathbb{P}(A_2|A_1)\times \\ &\mathbb{P}(A_3|A_1, A_2) \times \cdots \times \mathbb{P}(A_n|A_1, \ldots, A_{n-1})\end{aligned}$$

Expected Value

- If I do an experiment multiple times and look at my RV's value, what does the average look like?

Expected Value

- If I do an experiment multiple times and look at my RV's value, what does the average look like?
- This average converges (as I make more experiments) to a certain number, called **expected value** or **mean** of X.

Expected Value

- If I do an experiment multiple times and look at my RV's value, what does the average look like?
- This average converges (as I make more experiments) to a certain number, called **expected value** or **mean** of X.
- By definition,

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\mathbb{R} xf(x)dx$$

Expected Value

- If I do an experiment multiple times and look at my RV's value, what does the average look like?
- This average converges (as I make more experiments) to a certain number, called **expected value** or **mean** of X.
- By definition,

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\mathbb{R} xf(x)dx$$

- Expected value is linear! $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, even when X and Y are not independent. For a distribution, we usually use μ to represent its mean.

Variance

- Variance is a measure of scattering.

Variance

- Variance is a measure of scattering.
- "10^6 pockets, that only one of them has a golden coin with value 10^6" vs. "1 pocket with a coin of value 1". Which one do you choose?

Variance

- Variance is a measure of scattering.
- "10^6 pockets, that only one of them has a golden coin with value 10^6" vs. "1 pocket with a coin of value 1". Which one do you choose?
- Expected value is the same, the second one has lower variance...

Variance

- Variance is a measure of scattering.
- "10^6 pockets, that only one of them has a golden coin with value 10^6" vs. "1 pocket with a coin of value 1". Which one do you choose?
- Expected value is the same, the second one has lower variance...
- Can be defined as

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$$

Variance

- Variance is a measure of scattering.
- "10^6 pockets, that only one of them has a golden coin with value 10^6" vs. "1 pocket with a coin of value 1". Which one do you choose?
- Expected value is the same, the second one has lower variance...
- Can be defined as

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$$

- For a distribution, we show its variance by σ^2.

Law of Large Numbers

- Let's say, you did an experiment infinitely many times. The outcomes are listed as $X_1, X_2, \ldots$. We assume that each time we did the experiment *fresh*! Meaning that X_i does not depend on each other. Usually we say X_i are **iid RVs**; meaning that they have the same distribution and are independent of eachother.

Law of Large Numbers

- Let's say, you did an experiment infinitely many times. The outcomes are listed as $X_1, X_2, \ldots$. We assume that each time we did the experiment *fresh*! Meaning that X_i does not depend on each other. Usually we say X_i are **iid RVs**; meaning that they have the same distribution and are independent of eachother.
- Define the running sum $S_n = X_1 + \cdots + X_n$.

Law of Large Numbers

- Let's say, you did an experiment infinitely many times. The outcomes are listed as $X_1, X_2, \ldots$. We assume that each time we did the experiment *fresh*! Meaning that X_i does not depend on each other. Usually we say X_i are **iid RVs**; meaning that they have the same distribution and are independent of eachother.
- Define the running sum $S_n = X_1 + \cdots + X_n$.
- WLLN states that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\tfrac{S_n}{n} - \mu| > \varepsilon) = 0$$

Law of Large Numbers

- Let's say, you did an experiment infinitely many times. The outcomes are listed as $X_1, X_2, \ldots$. We assume that each time we did the experiment *fresh*! Meaning that X_i does not depend on each other. Usually we say X_i are **iid RVs**; meaning that they have the same distribution and are independent of eachother.
- Define the running sum $S_n = X_1 + \cdots + X_n$.
- WLLN states that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\tfrac{S_n}{n} - \mu| > \varepsilon) = 0$$

- SLLN states that

$$\mathbb{P}\left(\lim_{n\to\infty} \tfrac{S_n}{n} = \mu\right) = 1$$

Central Limit Theorem(s)

- If $X_1, X_2, \ldots$ is an iid seq. of RVs, having mean μ and variance σ^2, we have the following:

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Central Limit Theorem(s)

- If $X_1, X_2, \ldots$ is an iid seq. of RVs, having mean μ and variance σ^2, we have the following:

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- $\xrightarrow{d}$ means **convergence in distribution**, *i.e.* pointwise convergence of distribution functions at their continuity points.

Central Limit Theorem(s)

- If $X_1, X_2, \ldots$ is an iid seq. of RVs, having mean μ and variance σ^2, we have the following:

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- $\xrightarrow{d}$ means **convergence in distribution**, *i.e.* pointwise convergence of distribution functions at their continuity points.
- $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

Central Limit Theorem(s)

- If $X_1, X_2, \ldots$ is an iid seq. of RVs, having mean μ and variance σ^2, we have the following:

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- $\xrightarrow{d}$ means **convergence in distribution**, *i.e.* pointwise convergence of distribution functions at their continuity points.
- $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.
- *Basically*, for large n, every scaled distribution is essentially normal distribution!

Central Limit Theorem(s)

- If $X_1, X_2, \ldots$ is an iid seq. of RVs, having mean μ and variance σ^2, we have the following:

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- $\xrightarrow{d}$ means **convergence in distribution**, *i.e.* pointwise convergence of distribution functions at their continuity points.
- $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.
- *Basically*, for large n, every scaled distribution is essentially normal distribution!
- Good for creating *approximate confidence intervals*

Central Limit Theorem(s)

- If $X_1, X_2, \ldots$ is an iid seq. of RVs, having mean μ and variance σ^2, we have the following:

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- $\xrightarrow{d}$ means **convergence in distribution**, *i.e.* pointwise convergence of distribution functions at their continuity points.
- $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.
- *Basically*, for large n, every scaled distribution is essentially normal distribution!
- Good for creating *approximate confidence intervals*
- Caveat! Speed of convergence, Uniform convergence, regularity conditions…

Normal Distribution

- The density of the multivariate normal distribution is

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

written as $\mathcal{N}(x \mid \mu, \Sigma)$.

Normal Distribution

- The density of the multivariate normal distribution is

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{ -\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu) \right\}$$

  written as $\mathcal{N}(x \mid \mu, \Sigma)$.
- We have for $v \in \mathbb{R}^d$,

$$\mathrm{Var}(v^\top Y) = v^\top \Sigma v.$$

Normal Distribution

- The density of the multivariate normal distribution is

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{ -\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) \right\}$$

  written as $\mathcal{N}(x \mid \mu, \Sigma)$.
- We have for $v \in \mathbb{R}^d$,

$$\mathrm{Var}(v^\top Y) = v^\top \Sigma v.$$

- If one marginalizes a normal random vector, the result would also have a normal distribution.

Normal Distribution

- The density of the multivariate normal distribution is

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{ -\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu) \right\}$$

  written as $\mathcal{N}(x \mid \mu, \Sigma)$.
- We have for $v \in \mathbb{R}^d$,

$$\mathrm{Var}(v^\top Y) = v^\top \Sigma v.$$

- If one marginalizes a normal random vector, the result would also have a normal distribution.
- Any linear combination of normal random vectors, and also the image of a normal random vector under a linear transformation would be a normal random vector (with maybe different mean and covariance matrix).

Much more things to say, but no time!

Hope You Enjoy the Course!