A Refresher on Some Mathematical Subjects
or how to stay fresh all the time?

Mohammad Reza Karimi

ETH Zürich

Spring 2018
Topics Covered

Linear Algebra

Multivariate Analysis

Probability Theory
Outline

Linear Algebra

Multivariate Analysis

Probability Theory
Matrix Multiplication

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

\[ b = \sum_{j=1}^{n} x_j a_j \]

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

\[ uv^\top = [v_1u \ v_2u \ \cdots \ v_mu] \]
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 $Ax = 0$.

\[ \ker(A) = \{0\} \iff A \text{ is injective.} \]

- **Rank** of a matrix $A$ is the dimension of its range. It’s equal to $\dim(\text{col space}) = \dim(\text{row space})$.

\[ \dim \ker(A) + \text{rank}(A) = \#\text{cols of } A \]
A is invertible or nonsingular iff it is square and full rank. Equivalently, having det(A) \neq 0, or ker(A) = \{0\}.

Multiplication by \( A^{-1} \) is a change of basis:
Orthogonality

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

- Two vectors are *orthogonal* if their inner product is zero.

- Let $\{q_1, \ldots, q_n\}$ be a set of pairwise orthogonal vectors in $\mathbb{R}^n$. Then
  $$\forall v \in \mathbb{R}^n : \quad v = \sum_{i=1}^{n} (q_i^\top v)q_i = \sum_{i=1}^{n} (q_iq_i^\top)v$$

Note: $q_iq_i^\top$ is orthogonal projection onto direction $q_i$, which is a rank-one operator.
A matrix $U$ is **unitary** or **orthogonal** if $U^TU = I$, i.e. $\langle u_i, u_j \rangle = \delta_{i,j}$.

If $U$ is unitary, then it preserves angles,

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

and also lengths,

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

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

- A function $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}$, satisfying
  - $\forall x \in \mathbb{R}^n : \| x \| \geq 0$, $\| x \| = 0 \iff x = 0$,
  - $\forall \lambda \in \mathbb{R} : \| \lambda x \| = |\lambda| \| x \|$
  - $\forall x, y \in \mathbb{R}^n : \| x + y \| \leq \| x \| + \| y \|$

- 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 $p \in (0, 1)$
  - $\| x \|_{\infty} = \max_{1 \leq i \leq n} |x_i|$

- Unit balls:

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

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

  $$|\langle x, y \rangle| \leq \| x \|_p \| y \|_q,$$ 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 $A : (\mathbb{R}^n, \| \cdot \|_p) \rightarrow (\mathbb{R}^m, \| \cdot \|_q)$. Then we define

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

- 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)

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

\[ \|A\|_1 = 4 \]

\[ \|A\|_2 \approx 2.92 \]
Singular Values and Singular Vectors

- 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 $\sigma_1, \sigma_2$ 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 $\Sigma$ is a diagonal matrix, i.e.

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

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

$$\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n).$$
Eigenvalues and Eigenvectors

If for some vector $v \neq 0$ we have $Av = \lambda v$ then $v$ is an **eigenvector** of $A$ associated to the **eigenvalue** $\lambda$. In this case we have $(A - \lambda I)v = 0$, 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:

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

which is different from SVD.
Outline

Linear Algebra

Multivariate Analysis

Probability Theory
Notion of the Derivative

- As you may recall, for real differentiable functions \( f : \Omega \subseteq \mathbb{R} \rightarrow \mathbb{R} \), 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 \( f : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m \) and \( x_0 \in \Omega \). We call \( f \) to be differentiable at \( x_0 \) iff there is a linear function \( Df(x_0) : \mathbb{R}^n \rightarrow \mathbb{R}^m \), for which we have

  \[
  \lim_{\|h\| \rightarrow 0} \frac{|f(x_0 + h) - f(x_0) - Df(x_0)h|}{\|h\|} = 0
  \]

- We call \( Df(x_0) \) the derivative or differential of \( f \) at the point \( x_0 \).
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

\[
Df(x_0) = \begin{bmatrix}
\frac{\partial f_i}{\partial x_j}(x_0)
\end{bmatrix}_{1\leq i\leq m, 1\leq j\leq n},
\]

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 $f : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $\nabla f(x_0) = 0$. (The first derivative test)
- So we have the following (first-order) approximation for $x$ sufficiently close to $x_0$:

$$f(x) \approx f(x_0) + \nabla f(x_0)^\top (x - x_0)$$
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 differentiable 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, i.e. 
  \[ \lim_{\|h\| \to 0} \frac{R(h)}{\|h\|} = 0. \] 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 \( \|h\|^2 \).
Let $A$ be an $n \times n$ symmetric matrix. We define the **quadratic form** induced by $A$ to be

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

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 $\nabla f(x) = 2Ax$. 

---

**Multidimensional Quadratic Functions**
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 $x_0$:

$$f(x) \approx f(x_0) + \nabla f(x_0) ^\top (x - x_0) + \frac{1}{2} (x - x_0) ^\top D^2 f(x_0) (x - x_0).$$
In this example, we assume that $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is twice-differentiable, having a local minimum at $x_0$. This picture demonstrates a neighborhood of $x_0$:
Outline

Linear Algebra

Multivariate Analysis

Probability Theory
Basic Notations

- Let $\Omega$ 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),
  - $\Omega_2 = \{\text{Hit, Miss}\}$ (indicator),
  - $\Omega_3 = \{0, 10, 20, \ldots, 100\}$ (score of the throw).

- A family $\mathcal{F}$ of subsets of $\Omega$, called the **events**, are also interesting for us. Rather than asking whether a certain outcome has happen, 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 to be closed under these operations.
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 $i \in \Omega$.
- $\Omega = \mathbb{N}$, $\mathcal{F} = \mathcal{P}(\mathbb{N})$, $\mathbb{P}(\{i\}) = \frac{1}{2i}$ 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 function $X : \Omega \rightarrow \mathbb{R}$ 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 $\mathbb{P}(A) = \text{area}(A)/\text{area}(\mathbb{D}^2)$. 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 $\Omega$ is the inverse image $X^{-1}$.

- We can only ask a question $A$ from $X$ if the inverse image of $A$ is already inside $\mathcal{F}$, i.e. $X^{-1}(A) \in \mathcal{F}$. 

Distributions

- Any random variable induces a probability space 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 \( \lim_{x \to \infty} F(x) = 1 \), \( \lim_{x \to -\infty} F(x) = 0 \).
  - \( \mathbb{P}(a < X \leq b) = F(b) - F(a) \). 

![Graph of cumulative distribution function](image)
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

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

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 $\epsilon$ 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 \( \Omega \) 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

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

- The joint density can also be defined as

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

- 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
  \[ \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \]

- 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 \textit{not} follow that they are independent.
Conditional Probability

- How does information affect our belief?
- “knowing” that an event $B$ has occurred, 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 $\Omega$. 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 $\Omega$. We have

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

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

$$P(A_1, \ldots, A_n) = P(A_1) P(A_2 | A_1) P(A_3 | A_1, A_2) \cdots P(A_n | A_1, \ldots, A_{n-1})$$
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,

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

- 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 $\mu$ to represent its mean.
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

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

- For a distribution, we show its variance by \( \sigma^2 \).
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 each other.

- Up to time $n$, we take the average of what we saw,

$$\bar{X}_n := \frac{X_1 + \cdots + X_n}{n}.$$ 

- WLLN states that for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) = 0$$

- SLLN states that

$$\mathbb{P}\left(\lim_{n \to \infty} \bar{X}_n = \mu\right) = 1$$
Central Limit Theorem(s)

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

$$\sqrt{n}(\bar{X}_n - \mu) \overset{d}{\rightarrow} \mathcal{N}(0, \sigma^2)$$

- $\overset{d}{\rightarrow}$ means convergence in distribution, i.e. pointwise convergence of distribution functions.

- $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean $\mu$ and variance $\sigma^2$.

- 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...
What was not covered!

Please check these on the web! If you had troubles, you can contact me:

- Covariance, Covariance Matrix of $n$ RVs
- Multivariate Normal Distribution

All the best!