Mathematics of Deep Learning
Appendix: Mathematics checklist

Bruno Loureiro
Département d'Informatique, École Normale Supérieure - PSL & CNRS, France

Get in touch at: bruno.loureiro@di.ens.fr

Notation

We denote vectors by lower-case bold letters $\boldsymbol{v} \in \mathbb{R}$ and matrices by upper-case bold letters $\boldsymbol{A} \in \mathbb{R}^{n\times d}$. We define the short-hand $[n] = \{1, \dots, n\}$.

1 Linear algebra

1.1 Important notions

Definition 1 (Column and row space)**.** Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ denote a real-valued rectangular matrix with entries $a_{ij} \in \mathbb{R}$. Define the families of vectors $\boldsymbol{a}_i \in \mathbb{R}^d$, $i \in [n]$ and $\boldsymbol{A}_j \in \mathbb{R}^n$, $j \in [d]$ given by the rows and columns of $\boldsymbol{A}$, respectively. We define the *row* and *column* spaces of $\boldsymbol{A}$ as the vector spaces spanned by these families:

$$\begin{aligned} \mathrm{row}(\boldsymbol{A}) &= \mathrm{span}(\boldsymbol{a}_1, \dots, \boldsymbol{a}_n) \subset \mathbb{R}^d \\ \mathrm{col}(\boldsymbol{A}) &= \mathrm{span}(\boldsymbol{A}_1, \dots, \boldsymbol{A}_d) \subset \mathbb{R}^n \end{aligned} \tag{1.1}$$

Note that seen as a linear transformation $\boldsymbol{A} : \mathbb{R}^d \to \mathbb{R}^n$, the column space is simply its image $\mathrm{col}(\boldsymbol{A}) = \mathrm{Im}(\boldsymbol{A})$.

⚠ For any $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, $\mathrm{col}(\boldsymbol{A}) = \mathrm{row}(\boldsymbol{A}^\top)$.

Definition 2 (Rank)**.** The *rank* of a real-valued rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is the dimension of its column space.

$$\mathrm{rank}(\boldsymbol{A}) = \dim(\mathrm{col}(\boldsymbol{A})) \tag{1.2}$$

In other words, it is the number of linearly independent columns of $\boldsymbol{A}$.

From the definition above, one might wonder why defining the rank as the dimension of the column space and not the row space. Actually, an important result is that these two potential notions are the same.

Theorem 1. For any real-valued rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, the dimension of the column space is the same as the dimension of the row space:

$$\dim(\mathrm{col}(\boldsymbol{A})) = \dim(\mathrm{row}(\boldsymbol{A})) \tag{1.3}$$

Therefore, we have:

$$\mathrm{rank}(\boldsymbol{A}) \leq \min(n, d) \tag{1.4}$$

Definition 3 (Full-rank matrix)**.** A real-valued rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is said to be *full-rank* if:

$$\operatorname{rank}(\boldsymbol{A}) = \min(n, d) \tag{1.5}$$

The most important result in linear algebra for the purpose of this course is the singular-value decomposition.

Theorem 2 (Singular value decomposition)**.** Any real-valued rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ can be decomposed as:

$$\boldsymbol{A} = \sum_{i=1}^{\operatorname{rank}(\boldsymbol{A})} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \tag{1.6}$$

where $\sigma_i \geq 0$ are non-negative real numbers known as the *singular values* and $\boldsymbol{u}_i \in \mathbb{R}^n$, $\boldsymbol{v}_i \in \mathbb{R}^d$ are known as the left and right *singular vectors*. Moreover, the singular vectors form an orthonormal family with respect to the Euclidean scalar product: $\boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{ij}$, $\boldsymbol{v}_i^\top \boldsymbol{v}_j = \delta_{ij}$.

Remark 1. Without loss of generality we can (and will) assume the singular values $\sigma_i(\boldsymbol{A})$ are non-increasing: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ where $r = \operatorname{rank}(\boldsymbol{A})$. Often, the SVD is written as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ are the orthogonal matrices with columns $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ and $\boldsymbol{D} \in \mathbb{R}^{r\times r}$ is a diagonal matrix of singular values $d_{ij} = \sigma_i \delta_{ij}$.

⚠ Sometimes in the literature you will find $\boldsymbol{A} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}^\top$ with $\tilde{\boldsymbol{U}} \in \mathtt{O}(n)$, $\tilde{\boldsymbol{V}} \in \mathtt{O}(d)$ and $\tilde{D} \in \mathbb{R}^{n\times d}$ obtained by completing $\boldsymbol{U}, \boldsymbol{V}$ with an orthonormal basis of $\mathbb{R}^n$ and $\mathbb{R}^d$, respectively. In this case, $\tilde{\boldsymbol{D}} \in \mathbb{R}^{n\times d}$ is a rectangular matrix with a block given by $\boldsymbol{D}$ and zero elsewhere.

1.2 Important classes of matrices

There are a few classes of real valued square matrices which will often appear in the lectures. Here we review the most important ones.

- A square matrix $\boldsymbol{O} \in \mathbb{R}^{n\times n}$ is said to be **orthogonal** if:

  $$\boldsymbol{O}^\top \boldsymbol{O} = \boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_n \tag{1.7}$$

  Note that orthogonal matrices are always invertible $\boldsymbol{O}^\top = \boldsymbol{O}^{-1}$, and as linear transformations they define isometries, i.e. they preserve the Euclidean norm of vectors:

  $$||\boldsymbol{O}\boldsymbol{v}||_2 = ||\boldsymbol{v}||_2 \tag{1.8}$$

  for any $\boldsymbol{v} \in \mathbb{R}^n$. The set of orthogonal matrices define a group, known as the orthogonal group $\boldsymbol{O}(n) = \{\boldsymbol{O} \in \mathbb{R}^{n\times n} : \boldsymbol{O}^\top \boldsymbol{O} = \boldsymbol{I}_n\}$. From eq. (1.7), it is immediate to show that $\det(\boldsymbol{O}) \in \{-1, +1\}$. Orthogonal matrices such that $\det(\boldsymbol{O}) = +1$ are also known as *rotations*, while orthogonal matrices with $\det(\boldsymbol{O}) = -1$ are known as *reflections*. The set of rotations $\mathtt{SO}(n) = \{\boldsymbol{O} \in \mathbb{R}^{n\times n} : \boldsymbol{O}^\top \boldsymbol{O} = \boldsymbol{I}_n \text{ and } \det \boldsymbol{O} = 1\} \subset \mathtt{O}(n)$ defines a subgroup of $\mathtt{O}(n)$ known as the *special orthogonal group*. Orthogonal matrices have complex eigenvalues $\lambda_i \in \mathbb{C}$ with modulus $|\lambda_i| = 1$.

- A square matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is said to be **symmetric** if:

  $$\boldsymbol{M}^\top = \boldsymbol{M} \tag{1.9}$$

  Every real symmetric matrix can be diagonalised over the real numbers:

  $$\boldsymbol{M} = \boldsymbol{O}\boldsymbol{D}\boldsymbol{O}^\top \tag{1.10}$$

where $\boldsymbol{O} \in \mathtt{O}(n)$ is an orthogonal matrix and $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with entries $d_{ij} = \lambda_i(\boldsymbol{M})\delta_{ij}$. Note that the rows (or columns) of $\boldsymbol{O}$ are precisely the normalised eigenvectors of $\boldsymbol{M}$. A symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is said to be **positive semi-definite** $\boldsymbol{M} \succeq 0$ if its spectrum is non-negative $\lambda_i(\boldsymbol{M}) \geq 0$ for all $i \in [n]$, and is said to be **positive-definite** $\boldsymbol{M} \succ 0$ if its spectrum is positive $\lambda_i(\boldsymbol{M}) > 0$ for all $i \in [n]$.

⚠ A symmetric matrix can have zero eigenvalues, so it might not be invertible. However, a positive-definite symmetric matrix $\boldsymbol{M} \succ 0$ is always invertible.

Example 1. For any real valued rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, the square matrices $\boldsymbol{A}^\top\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{n\times n}$ are symmetric positive semi-definite matrices.

- A square matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is said to be a **projection** if it is idempotent:

$$\boldsymbol{P}^2 = \boldsymbol{P} \tag{1.11}$$

From this, it follows that a projection matrix can only have eigenvalues 0 or 1: $\lambda_i(\boldsymbol{P}) \in \{0, 1\}$. Therefore, a projection matrix can always be written as:

$$\boldsymbol{P} = \sum_{i=1}^{\mathrm{rank}(\boldsymbol{P})} \boldsymbol{v}_i\boldsymbol{v}_i^\top \tag{1.12}$$

As the name suggests, projection matrices $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ geometrically define projections into a linear subspace of $\mathrm{Im}(\boldsymbol{P}) \subset \mathbb{R}^n$ of dimension $\mathrm{rank}(\boldsymbol{P})$. More explicitly, this subspace is precisely the span of the eigenvectors corresponding to the non-zero eigenvalues $V = \mathrm{span}(\boldsymbol{v}_i)$. An **orthogonal projection** $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is a projection which is also orthogonal $\boldsymbol{P} \in \mathtt{SO}(n)$, and correspond to the case where the eigenvalues $\boldsymbol{v}_i$ are orthonormal vectors. Finally, every orthogonal projection defines an orthogonal decomposition $\mathbb{R}^n = \mathrm{Im}(\boldsymbol{P}) \oplus \mathrm{Ker}(\boldsymbol{P})$, for which we can associate another orthogonal projection matrix $\boldsymbol{P}_\perp \in \mathbb{R}^{n\times n}$, which is the projection on its orthogonal complement $\mathrm{Ker}(\boldsymbol{P})$.

⚠ With the exception of the identity $\boldsymbol{I}_n$, a projection matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is never invertible.

Example 2. Let $\boldsymbol{v} \in \mathbb{R}^n$ denote a unit-norm vector $||\boldsymbol{v}||_2 = 1$. Then:

$$\boldsymbol{P} = \boldsymbol{v}\boldsymbol{v}^\top, \qquad \boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{v}\boldsymbol{v}^\top \tag{1.13}$$

define a orthogonal projection in the line $L = \{\alpha\boldsymbol{v} \in \mathbb{R}^n : \alpha \in \mathbb{R}\}$ and its orthogonal complement.

1.3 Matrix norms

Just as for vectors, there are different useful notions of norm for matrices. Here we discuss the most relevant for the lectures. Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ denote a real-valued rectangular matrix with singular values $(\sigma_j(\boldsymbol{A}))_{j\in[r]}$, where $r := \mathrm{rank}(\boldsymbol{A})$. Without loss of generality, we assume $\sigma_j(\boldsymbol{A})) \geq 0$ are non-decreasing. We define the following matrix norms:

- The **Frobenius norm** of $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is defined as:

$$||\boldsymbol{A}||_{\mathtt{F}} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} A_{ij}^2} = \sqrt{\mathrm{Tr}\,\boldsymbol{A}^\top\boldsymbol{A}} = \sqrt{\sum_{i=1}^{r}\sigma_i(\boldsymbol{A})^2} \tag{1.14}$$

- The **operator norm** of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is defined as:

$$||\boldsymbol{A}||_{\mathtt{op}} = \sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}} ||\boldsymbol{A}\boldsymbol{v}||_2 = \sigma_1(\boldsymbol{A}) \tag{1.15}$$

where we recall $\sigma_1(\boldsymbol{A})$ is the top singular value of $\boldsymbol{A}$.

- The **nuclear norm** of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is defined as:

$$||A||_* = \mathrm{Tr}\left(\sqrt{\boldsymbol{A}\boldsymbol{A}^\top}\right) = \sum_{i=1}^{r} \sigma_i(\boldsymbol{A}) \tag{1.16}$$

Remark 2. All the norms above are a particular case of a more general class of norms known as Schatten norms:

$$||\boldsymbol{A}||_p = \left(\sum_{i=1}^{r} \sigma_i(\boldsymbol{A})^p\right)^{1/p}. \tag{1.17}$$

More precisely, they correspond to the case $p = 1, 2, \infty$.

Lemma 1. For any real valued matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we have:

$$||\boldsymbol{A}||_{\mathtt{op}} \leq ||\boldsymbol{A}||_{\mathtt{F}} \leq ||\boldsymbol{A}||_* \tag{1.18}$$

Proof. Since the norms are positive, it is equivalent to show:

$$||\boldsymbol{A}||^2_{\mathtt{op}} \leq ||\boldsymbol{A}||^2_{\mathtt{F}} \leq ||\boldsymbol{A}||^2_* \tag{1.19}$$

The first inequality is immediate: since $\sigma_i(\boldsymbol{A}) \geq 0$, the sum can only be larger than any of the terms:

$$||\boldsymbol{A}||^2_{\mathtt{F}} = \sum_{i=1}^{r} \sigma_i(\boldsymbol{A})^2 \geq \sigma_i(\boldsymbol{A}) \text{ for all } i \in [r]. \tag{1.20}$$

The second inequality follow from noting that:

$$\begin{aligned} ||\boldsymbol{A}||^2_* = \left(\sum_{i=1}^{r} \sigma_i\right)^2 &= \sum_{i,j=1}^{r} \sigma_i\sigma_j = \sum_{i=1}^{r} \sigma_i^2 + \sum_{i\neq j} \sigma_i\sigma_j \\ &= ||\boldsymbol{A}||^2_{\mathtt{F}} + \sum_{i\neq j} \sigma_i\sigma_j \geq ||\boldsymbol{A}||^2_{\mathtt{F}} \end{aligned} \tag{1.21}$$

since $\sigma_i(\boldsymbol{A}) \geq 0$. □

Lemma 2. All the norms above are equivalent since:

- $||\boldsymbol{A}||_{\mathtt{F}} \leq ||\boldsymbol{A}||_* \leq \sqrt{r}||\boldsymbol{A}||_{\mathtt{F}}$
- $||\boldsymbol{A}||_{\mathtt{op}} \leq ||\boldsymbol{A}||_* \leq r||\boldsymbol{A}||_{\mathtt{op}}$
- $||\boldsymbol{A}||_{\mathtt{op}} \leq ||\boldsymbol{A}||_{\mathtt{F}} \leq \sqrt{r}||\boldsymbol{A}||_{\mathtt{op}}$

1.4 Matrix identities

Let $\boldsymbol{U} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times n}$ be two rectangular matrices. We have the following useful identities:

- The traces of the resolvent and co-resolvent are related as:

$$\mathrm{Tr}(\boldsymbol{U}\boldsymbol{V} - z\boldsymbol{I}_n)^{-1} = \mathrm{Tr}(\boldsymbol{V}\boldsymbol{U} - z\boldsymbol{I}_d)^{-1} - \frac{n-d}{z} \tag{1.22}$$

  Taking the derivative with respect to z on both sides, this also implies:

$$\mathrm{Tr}(\boldsymbol{U}\boldsymbol{V} - z\boldsymbol{I}_n)^{-2} = \mathrm{Tr}(\boldsymbol{V}\boldsymbol{U} - z\boldsymbol{I}_d)^{-2} - \frac{n-d}{z^2} \tag{1.23}$$

- Push-through identity:

$$(\boldsymbol{U}\boldsymbol{V} - z\boldsymbol{I}_n)^{-1}\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{V}\boldsymbol{U} - z\boldsymbol{I}_d)^{-1} \tag{1.24}$$

- Block inversion formula:

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & -(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix} \tag{1.25}$$

  where $\boldsymbol{A}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, $\boldsymbol{C} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{D} \in \mathbb{R}^{m\times m}$.

- Sherman-Morrison lemma: Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ denote an invertible matrix and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{d\times d}$ two vectors such that $\boldsymbol{v}^\top \boldsymbol{A}^{-1}\boldsymbol{u} \neq -1$. Then:

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v})^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}} \tag{1.26}$$

2 Probability

A few good references to catch up:

- Roman Vershynin's book "*High-dimensional probability: an introduction with applications in data science*", freely available online.
- Chapter 1 of Philippe Rigollet and Jan-Christian Hütte lecture notes on "*High-Dimensional Statistics*", freely available online.

2.1 Geometry of random variables

Definition 4 (L^p norm of a r.v.)**.** Let X denote a random variable. The L^p norm of X is given by:

$$||X||_{L^p} = (\mathbb{E}[|X^p|])^{1/p}, p \in [1, \infty) \tag{2.1}$$

This can be extended to $p = \infty$ by defining:

$$||X||_{L^\infty} = \text{ess sup}|X| \tag{2.2}$$

It can be shown that indeed this defines a norm, and therefore the linear space:

$$L^p = \{X : ||X||_{L^p} \leq \infty\} \tag{2.3}$$

defines a Banach space.

Remark 3. Definition 4 still makes sense for $p \in (0, 1)$. However, in this case $|| \cdot ||_{L^p}$ is not a norm.

The space L^2 is also a Hilbert space, with inner product defined as:

$$\langle X, Y\rangle_{L^2} = \mathbb{E}[|XY|] \tag{2.4}$$

Note that $||X||_{L^p}$ is an increasing function of p. This implies the following inclusion of L^p spaces:

$$L^\infty \subset \cdots \subset L^2 \subset L^1 \tag{2.5}$$

This is quite intuitive: a bounded random variable has all moments, a random variable with p moments has all $p - 1$ moments, and so on. However, having finite p moments for all p does not imply X is almost surely bounded. The Gaussian distribution is an example.

Lemma 3. Let $G \sim \mathcal{N}(0,1)$, then for all $p \in [1,\infty)$:

$$||G||_{L^p} \leq \sqrt{p} \tag{2.6}$$

and $||G||_{L^\infty} = \infty$ since Gaussian variables are unbounded.

L^p spaces are a particular case of a more general geometry of random variables, known as *Orlicz spaces.*

Definition 5 (Orlicz spaces)**.** Let ψ denote a convex increasing function such that:

$$\lim_{x\to 0} \psi(x) = 0, \qquad \lim_{x\to\infty} \psi(x) = \infty. \tag{2.7}$$

For any random variable X, we define the *Orlicz norm* of X as:

$$||X||_\psi = \inf\{k > 0 : \mathbb{E}[\psi(|X|/k)] \leq 1\} \tag{2.8}$$

Further, we define the *Orlicz space* associated to ψ as:

$$L_\psi = \{X : ||X||_\psi < \infty\} \tag{2.9}$$

Note that for $\psi(x) = x^p$, we retrieve $L_\psi = L^p$. However, Orlicz spaces allow us to defined a more refined geometry of random variables, that allow us to distinguish different classes of random variables that have all moments but are not necessarily bounded. For instance, $\psi_p(x) = e^{x^p} - 1$ defines a family of Orlicz spaces L_{ψ_p} that sit exactly in between L^p and L^∞. For instance, note that $L_{\psi_1} \subset L^p$ since exponentials grow faster than polynomials. However, $L^\infty \subset L_{\psi_1}$ since the expectation of the exponential of a bounded random variable is finite. Therefore, $L^\infty \subset L_{\psi_1} \subset L^p$ for any $p > 1$. More generally, L_{ψ_p} define a hierarchy of Orlicz spaces based on the tails of the distributions, with tails which are lighter as p increases. Two important examples are sub-exponential and sub-Gaussian random variables.

Definition 6 (Sub-Gaussian r.v.)**.** A random variable X is sub-Gaussian if:

$$||X||_{\psi_2} = \inf\left\{C > 0 : \mathbb{E}\left[\exp\left(\frac{X^2}{C^2}\right)\right] \leq 2\right\} \leq \infty \tag{2.10}$$

In other words, it is the Orlicz space L_{ψ_2} with $\psi_2(x) = e^{x^2} - 1$. The following are equivalent characterisations:

- **Gaussian tails:** $\exists c_1$ such that for all $t > 0$:

$$\mathbb{P}(|X| \geq t) \leq 2e^{-\frac{t^2}{c_1^2}} \tag{2.11}$$

- **Moments:** $\exists c_2$ such that for all $p > 1$:

$$||X||_{L^p} \leq c_2\sqrt{p} \tag{2.12}$$

- **Moment generating function:** $\exists c_3$ such that if $\mathbb{E}[X] = 0$:

$$\mathbb{E}[e^{tX}] \leq e^{c_3^2 t^2}, \qquad t \in \mathbb{R} \tag{2.13}$$

Example 3. Some popular examples of sub-Gaussian random variables are:

- Gaussian random variables are sub-Gaussian. In particular, if we have $G \sim \mathcal{N}(0,\sigma^2)$, then:

$$||G||_{\psi_2} \leq C\sigma \tag{2.14}$$

- Bernouilli random variables $X \sim \text{Ber}({}^{1}\!/_{2})$ are sub-Gaussian random variables. In particular, we have:

$$||X||_{\psi_2} \leq \frac{1}{\sqrt{\log 2}} \tag{2.15}$$

- Bounded random variables are sub-Gaussian random variables. In particular, we have:

$$||X||_{\psi_2} \leq \frac{||X||_{L^\infty}}{\sqrt{\log 2}} \tag{2.16}$$

Intuitively, sub-Gaussian variables are variables that have the same tail as Gaussian random variables. We can define sub-exponential random variables similarly.

Definition 7 (Sub-Exponential r.v.)**.** A random variable X is *sub-exponential* if:

$$||X||_{\psi_1} = \inf\left\{C > 0 : \mathbb{E}\left[\exp\left(\frac{X}{C}\right)\right] \leq 2\right\} \leq \infty \tag{2.17}$$

In other words, it is the Orlicz space L_{ψ_2} with $\psi_2(x) = e^x - 1$. The following are equivalent characterisations:

- **Exponential tails:** $\exists c_1$ such that for all $t > 0$:

$$\mathbb{P}(|X| \geq t) \leq 2e^{-\frac{t}{c_1^2}} \tag{2.18}$$

- **Moments:** $\exists c_2$ such that for all $p > 1$:

$$||X||_{L^p} \leq c_2 p \tag{2.19}$$

- **Moment generating function:** $\exists c_3$ such that if $\mathbb{E}[X] = 0$:

$$\mathbb{E}[e^{tX}] \leq e^{c_3^2 t^2}, \qquad |t| \leq {}^{1}\!/_{c_3} \tag{2.20}$$

Remark 4. Note that from the perspective of the MGF, the only difference between sub-exponential and sub-Gaussian random variables is that the former holds for t bounded. Therefore, we can informally view sub-Gaussian random variables as the class sub-exponential random variables with $c_3 \to 0$.

Proposition 1. A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover:

$$||X^2||_{\psi_1} = ||X||_{\psi_2}^2 \tag{2.21}$$

2.2 Classical inequalities

In this section, we review some classical inequalities in probability.

Proposition 2 (Jensen's inequality)**.** Let X denote a real-valued random variable. Then, for any convex function φ:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)] \tag{2.22}$$

⚠ It is a common mistake to inverse the direction of Jensen's inequality.

Proposition 3 (Holder's inequality)**.** For any random variables $X \in L^p$ and $Y \in L^q$ where $p, q \in [1, \infty]$ are conjugate variables ${}^{1}\!/_{p} + {}^{1}\!/_{q} = 1$:

$$|\mathbb{E}[XY]| \leq ||X||_p ||Y||_q \tag{2.23}$$

Remark 5. The case $p = q = 2$ is known as the Cauchy-Schwarz inequality:

$$|\mathbb{E}[XY]| \leq ||X||_{L^2} ||Y||_{L^2} \tag{2.24}$$

Proposition 4 (Minkowski's inequality)**.** For any random variables $X, Y \in L^p$ and $p \in [1, \infty]$:

$$||X + Y||_p \leq ||X||_p + ||Y||_p \tag{2.25}$$

2.3 Tail inequalities

Proposition 5 (Markov's inequality)**.** Let X denote a non-negative random variable. Then for all $t > 0$:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \tag{2.26}$$

i.e. the probability that X is at least t is at most the expectation divided by t. Note that when $\mathbb{E}[X] > 0$ we can equivalently write:

$$\mathbb{P}(X \geq t\,\mathbb{E}[X]) \leq \frac{1}{t} \tag{2.27}$$

Proposition 6 (Chebyshev's inequality)**.** Let X denote a random variable with mean $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Then, for all $t > 0$

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{2.28}$$

Proposition 7 (Chernoff's inequality)**.** Let X denote a real random variable. Then, for all $a \in \mathbb{R}$ and $t > 0$:

$$\mathbb{P}(X \geq a) \leq \mathbb{E}[e^{t(X-a)}] \tag{2.29}$$

Note that this holds for all $t > 0$, it is also common to take the infimum over t:

$$\mathbb{P}(X \geq a) \leq \inf_{t \geq 0} \mathbb{E}[e^{t(X-a)}] \tag{2.30}$$

2.4 Concentration inequalities for the sum of random variables

Proposition 8 (Hoeffding's inequality)**.** Let $X_1, \ldots, X_n$ denote independent, zero mean sub-Gaussian random variables. Then, for every $t > 0$, we have:

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum\limits_{i=1}^{n} ||X_i||_{\psi_2}^2}\right) \tag{2.31}$$

where $||\cdot||_{\psi_2}$ is the sub-Gaussian norm from definition 6.

Remark 6 (Particular cases of Hoeffding's inequality)**.** The following particular cases of Hoeffding's inequality are useful.

- **Bernouilli:** Let $X_1, \ldots, X_n$ denote independent symmetric Bernouilli random variables, i.e. $X_i \in \{-1, +1\}$ with:

  $$\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = +1) = \frac{1}{2}. \tag{2.32}$$

  In this case, applying proposition 8 leads to the following tail bound:

  $$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2e^{-\frac{t^2}{2n}} \tag{2.33}$$

- **Bounded:** Let $X_1, \dots, X_n$ denote independent random variables which are bounded almost surely, i.e. $X_i \in [a_i, b_i]$ a.s. Then, for all $t > 0$ their sum $S_n = \sum_{i=1}^{n} X_n$ satisfy:

$$\mathbb{P}\left(S_n - \mathbb{E}S_n \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{2.34}$$

Proposition 9 (Bernstein's inequality)**.** Let $X_1, \dots, X_n$ note independent sub-exponential random variables with $\mathbb{E}[X_i] = 0$. Then, there exists a constant c such that for all $t > 0$ the sum $S_n = \sum_{i=1}^{n} X_i$ satisfy:

$$\mathbb{P}(|S_n| \geq t) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sigma^2}, \frac{t}{k}\right)\right) \tag{2.35}$$

where $c > 0$ is an absolute constant and:

$$\sigma^2 = \sum_{i=1}^{n} ||X_i||_{\psi_1}, \qquad k = \max_{i\in[n]} ||X_i||_{\psi_1} \tag{2.36}$$

where $|| \cdot ||_{\psi_1}$ is the sub-exponential norm from definition 7.

2.5 Convergence of random variables

In this appendix, we review the different notions of convergence for random variables. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space. We start with one of the strongest forms of convergence:

Definition 8 (Almost sure convergence)**.** We say a sequence of random variables $(X_n)_{n\geq 1}$ converges *almost surely* (a.s.) to a random variable X and denote $X_n \overset{a.s.}{\to} X$ if there exists a measurable set $\Omega' \in \mathcal{F}$ such that:

- $\mathbb{P}(\Omega') = 1$.
- For all $\omega \in \Omega'$, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$

Intuitively, almost sure convergence means that $X_n \to X$ just as for deterministic variables, excepts perhaps for exceptional events that have probability zero as $n \to \infty$ (hence the "almost").

Definition 9 (Convergence in probability)**.** We say a sequence of random variables $(X_n)_{n\geq 1}$ converges *in probability* to a random variable X and denote $X_n \overset{P}{\to} X$ if for every $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0 \tag{2.37}$$

While almost sure convergence is a statement about the convergence of values taken by a random variable, convergence in probability is a statement about the convergence of probabilities. A particularly intuitive case is when the limiting random variable X is deterministic, i.e. $X = x$ with probability 1. In this case, we can visualise the convergence in probability as the distribution of X_n getting more and more peaked around $X = x$ as $n \to \infty$.

Example 4. Let $X_n \sim \text{Unif}([-\frac{1}{n}, \frac{1}{n}])$ denote a sequence of uniform random variables. We have $X_n \overset{P}{\to} 0$.

Almost sure convergence implies convergence in probability (see Grimmett and Stirzaker (2020) for a proof), but the converse is not true. A standard example is the following:

Example 5. Consider a sequence of binary random variables $X_n \in \{0, 1\}$ such that:

$$\mathbb{P}(X_n = 1) = \frac{1}{n}, \qquad \mathbb{P}(X_n = 0) = 1 - \frac{1}{n} \tag{2.38}$$

Then, we have $X_n \xrightarrow{P} 0$ since:

$$\lim \mathbb{P}(X_n = 1) = 0, \qquad \lim \mathbb{P}(X_n = 0) = 1 \tag{2.39}$$

However, X_n does not converge almost surely to 0. To see this, consider the event that X_n takes the value 1: $E_n = \{X_n = 1\}$. We have:

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty \tag{2.40}$$

By the Borel-Cantelli lemma, a sequence of independent events with probability that sum to ∞ must happen infinitely often.

Definition 10 (Convergence in L^p)**.** We say a sequence of random variables $(X_n)_{n\geq 1}$ converges *in* L^p (or p-th mean) to a random variable X and denote $X_n \xrightarrow{L^p} X$ if:

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0 \tag{2.41}$$

Note that this is equivalent to convergence in $L^p(\mathbb{P})$ norm. Convergence in L^p implies convergence in probability, but the converse if not true. Note that convergence in L^p does not implied and does not imply almost sure convergence: in general these are unrelated.

Example 6. Let $U \sim \text{Unif}([0, 1])$, and define:

$$X_n = \sqrt{n}\, \mathbf{1}_{(0,1/n)}(U) = \begin{cases} \sqrt{n} & \text{if } U \in (0, 1/n) \\ 0 & \text{otherwise} \end{cases} \tag{2.42}$$

Then, X_n convergences in probability to 0 since for all $0 < \epsilon < 1$:

$$\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}\left(\sqrt{n}\mathbf{1}_{(0,1/n)}(U) > \epsilon\right) = \mathbb{P}\left(0 \leq U \leq \frac{1}{n}\right) = \frac{1}{n} \tag{2.43}$$

which goes to zero as $n \to \infty$. However, X_n does not converge to zero in L^2 since:

$$\mathbb{E}[X_n^2] = n \int_0^{n/2} \mathrm{d}t = 1 \tag{2.44}$$

Note that all the notions so far easily generalise to random vectors or matrices by simply taking an adapted norm. Finally, the last common notion of convergence is convergence in distribution, which we first define for real valued variables:

Definition 11 (Convergence in distribution)**.** We say a sequence of random variables $(X_n)_{n\geq 1}$ converges *in distribution* to a random variable X and denote $X_n \xrightarrow{d} X$ if:

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq t) = \mathbb{P}(X \leq t) \tag{2.45}$$

for all t for which the c.d.f. $\mathbb{P}(X \leq t)$ is continuous.

Convergence in distribution is the weakest form of convergence discussed here. Indeed, it is implied by convergence in probability (and hence by both almost sure and L^p convergence). Note that the condition "for all t for which the c.d.f. $\mathbb{P}(X \leq t)$ is continuous" is important, as highlighted by the following example:

Example 7. Consider a sequence of Gaussian random variables with decreasing variance $X_n \sim \mathcal{N}(0, 1/n)$. We have:

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \lim_{n\to\infty} \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2n}}\right)\right] = \begin{cases} 1 & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{2.46}$$

Therefore, $X_n \xrightarrow{d} X$ with $\mathbb{P}(X = 0) = 1$ which has c.d.f. $\mathbb{P}(X \leq x) = \Theta(x)$, since the discontinuity point $x = 0$ can be ignored.

Since this definition of convergence in distribution relies on the c.d.f., if does not straightforwardly generalise to random vectors. A more adapted and equivalent notion is known in this context as *weak convergence*:

Definition 12 (Weak convergence)**.** We say a sequence of random vectors $(X_n)_{n\geq 1}$ in $\mathbb{R}^d$ *weakly converges* to a random vector X and denote $X_n \xrightarrow{d} X$ if for any bounded continuous function $f : \mathbb{R}^d \to \mathbb{R}$:

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \tag{2.47}$$

This definition can be easily extended to any metric space. Note that "bounded continuous" f can also be exchanged for "bounded Lipschitz". Several equivalent characterisations of weak convergence are given by the Portmanteau lemma.

Summary

We can summarise the discussion in this Appendix in Figure 1. Note that several converse results under stronger assumptions exist. We refer the reader to Chapter 7 of Grimmett and Stirzaker (2020) for a full discussion. Finally, we state the following result which is useful in the context of statistics:

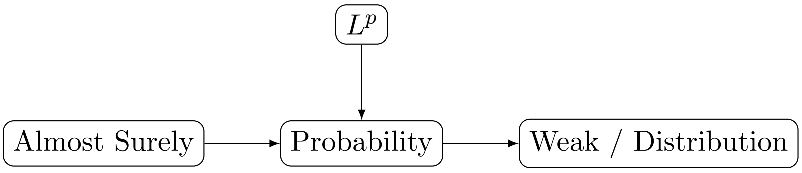

Figure 1: Different notions of convergence for random variables, with the respective implications.

Lemma 4. Let X_n be an unbiased estimate of $\alpha \in \mathbb{R}$. Then, if $\operatorname{Var}(X_n) \to 0$ as $n \to \infty$, $X_n \xrightarrow{L^2} \alpha$ (and hence also in probability).

Proof. By definition, we have $\mathbb{E}[X_n] = \alpha$. Therefore:

$$\mathbb{E}[|X_n - \alpha|^2] = \mathbb{E}[|X_n - \mathbb{E}[X_n]|^2] = \operatorname{Var}(X_n) \to 0 \text{ as } n \to \infty \tag{2.48}$$

which implies convergence in squared-mean. □

2.6 Limit theorems

Theorem 3 (Strong law of large numbers)**.** Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$, and consider the empirical mean:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.49}$$

Then, as $n \to \infty$:

$$\bar{X}_n \xrightarrow{a.s.} \mu \tag{2.50}$$

Theorem 4 (Lindeberg-Lévy central limit theorem)**.** Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $\operatorname{Var}(X_i) = \sigma^2 < \infty$, and consider the empirical mean:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.51}$$

Then, as $n \to \infty$:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{2.52}$$

In other words, letting $Z_n = {}^{\sqrt{n}}/_{\sigma}(\bar{X}_n - \mu)$, for any $t \in \mathbb{R}$:

$$\mathbb{P}(|Z_n| \geq t) \to \mathbb{P}(|G| \geq t) = \int_t^{\infty} \frac{\mathrm{d}x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2.53}$$

point-wise as $n \to \infty$.

Corollary 1. Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ finite variance. Then:

$$\mathbb{E}[|\bar{X}_n - \mu|] = O({}^1/_{\sqrt{d}}), \qquad \text{as } n \to \infty \tag{2.54}$$

3 Analysis

3.1 Lipschitz functions

In the course, we will often need to control the regularity of function. A particularly useful notion of regularity is how the slope/derivative of the function changes point-wise. Functions which have a "gentle" change of the slope are more regular than "spiky" functions for which the slope can vary abruptly. This notion is formalised by Lipschitz function.

Definition 13 (Lipschitz function)**.** Let (X, d_X) and (Y, d_Y) denote metric spaces. A function $f : X \to Y$ is called L-Lipschitz if there exists $L \in \mathbb{R}$ such that for all $x, y \in X$:

$$d_Y(f(x), f(y)) \leq L \cdot d_X(x, y) \tag{3.1}$$

The constant L is known as the Lipschitz constant of f, and the infimum over all L defines a norm, known as the Lipschitz norm of f:

$$||f||_{\text{Lip}} = \inf \ \{L \in \mathbb{R} : d_Y(f(x), f(y)) \leq L \cdot d_X(x, y) \text{ for all } x, y \in X\} \tag{3.2}$$

Lipschitz functions with $||f||_{\text{Lip}} < 1$ are also known as *contractions.*

A particular example of interest for the lectures is the case of functions in a normed vector space, where $X = \mathbb{R}^n$ and $Y = \mathbb{R}$ and eq. (3.1) reads:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L||\boldsymbol{x} - \boldsymbol{y}|| \tag{3.3}$$

Proposition 10 (Properties of Lipschitz functions)**.** Lipschitz functions satisfy the following properties:

1. Every Lipschitz function is uniformly continuous.
2. Every differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz, and:
$$||f||_{\text{Lip}} \leq \sup_{x \in \mathbb{R}^n} ||\nabla f(\boldsymbol{x})||_2 \tag{3.4}$$
3. The composition of two Lipschitz maps is Lipschitz, with:
$$||f \circ g||_{\text{Lip}} = ||f||_{\text{Lip}}||g||_{\text{Lip}} \tag{3.5}$$

⚠ The converse is not true: there are functions which are not everywhere differentiable but are still Lipschitz, for example $f(x) = |x|$ is a 1-Lipschitz function since $||x| - |y|| \leq |x - y|$ for all $x, y \in \mathbb{R}$.

Example 8. Some useful examples of Lipschitz functions.

- For a fixed vector $\boldsymbol{\theta} \in \mathbb{R}^n$, the inner product:
$$f(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle \tag{3.6}$$
is a Lipschitz function on $\mathbb{R}^n$ with:
$$||f||_{\text{Lip}} = ||\boldsymbol{\theta}||_2 \tag{3.7}$$
- More generally, any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ acting as a linear operator:
$$\begin{aligned} \boldsymbol{A} : \mathbb{R}^d &\to \mathbb{R}^n \\ \boldsymbol{x} &\mapsto \boldsymbol{A}\boldsymbol{x} \end{aligned}$$
Is a Lipschitz function with:
$$||\boldsymbol{A}||_{\text{Lip}} = ||\boldsymbol{A}||_{\texttt{op}} \tag{3.8}$$
- Any norm $f(\boldsymbol{x}) = ||\boldsymbol{x}||$ on $\mathbb{R}^n$ is a Lipschitz function. The Lipschitz norm of f is the smallest f that satisfies:
$$||\boldsymbol{x}|| \leq L||\boldsymbol{x}||_2, \text{ for all } \boldsymbol{x} \in \mathbb{R}^n \tag{3.9}$$
For example, the L^1 norm:
$$f(\boldsymbol{x}) = ||\boldsymbol{x}||_1 = \sum_{i=1}^{n} |x_i| \tag{3.10}$$
is a Lipschitz function with Lipschitz constant $L = \sqrt{n}$. More generally, the L^p norms have Lipschitz constant $L = n^{\max(0, 1/2 - 1/p)}$.
- The rectified linear unit $f(x) = \max(0, x)$ is a 1-Lipschitz function.

- The Logistic loss $\ell(x) = \log(1 + e^{-x})$ is a 1-Lipschitz function.
- The Hinge loss $\ell(x) = \max(0, 1 - x)$ is a 1-Lipschitz function.

It is also useful to have in mind examples of functions which are not Lipschitz (and why). The most common features of non-Lipschitz functions are: (a) unbounded derivative/slope; (b) Discontinuities; (c) Infinite oscillations. In some cases, a Lipschitz function can be defined by restricting the domain of non-Lipschitz functions to exclude the singularities. Below, we give a few useful examples:

Example 9. The following functions are not Lipschitz everywhere in their domain.

- The logarithm $f(x) = \log x$ is not a Lipschitz function in $\mathbb{R}_+$ since $f'(x) = \frac{1}{x}$. However, it is a Lipschitz function in any domain $[a, \infty)$ with $a > 0$, with Lipschitz constant $L = {}^1\!/_a$.
- The quadratic function $f(x) = x^2$ is not Lipschitz in $\mathbb{R}$ since its derivative $f'(x) = x$ is unbounded. However, the truncated quadratic $f(x) = \min(1, x^2)$ is a Lipschitz function with constant $L = 2$.
- The square root function $f(x) = \sqrt{x}$ is not a Lipschitz function in $\mathbb{R}_+$ since $f'(x) = {}^1\!/_{2\sqrt{x}}$ grows unbounded as $x \to 0^+$. However, it is a Lipschitz function in any interval $[a, \infty)$ with $a > 0$ with Lipschitz constant $L = {}^1\!/_{2\sqrt{x}}$.
- The exponential function $f(x) = e^x$ is not Lipschitz.
- The Heavyside step function:

$$\Theta(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{3.11}$$

  is not Lipschitz because of the discontinuity at $x = 0$.
- The function $f(x) = \sin({}^1\!/_x)$ is not Lipschitz on $(0, \infty)$, due to the fast oscillations close to 0^+. One can define a Lipschitz function by restricting it to an interval $[a, \infty)$ with $a > 0$, with Lipschitz constant $L = {}^1\!/_{a^2}$

4 Big-O notation

In these notes, we often employ the so-called *Big-O* notation, a handy way of comparing the order of magnitude or limiting behaviour of functions. In this appendix, we give a formal definition and discuss some intuition.

Definition 14 (Big-O notation)**.** Let $f, g : \mathbb{R} \to \mathbb{R}$ denote two real-valued functions. We say:

- "$f(x)$ is big-O of $g(x)$" and write $f(x) = O(g(x))$ as $x \to \infty$ if there exists $M > 0$ and $x_0 \in \mathbb{R}$ such that:

$$|f(x)| < M|g(x)| \text{ for all } x > x_0 \tag{4.1}$$

  Intuitively, $f(x) = O(g(x))$ means $f(x)$ is "at most" $g(x)$, meaning that one can make it $f(x)$ as large as g by multiplying by a constant (with respect to x_0). It is used to denote asymptotic upper bounds. If $g(x)$ is non-zero beyond a certain point, this is equivalent to:

$$\limsup_{x \to \infty} \frac{f(x)}{g(x)} < \infty \tag{4.2}$$

- "$f(x)$ is little-O of $g(x)$" and write $f(x) = o(g(x))$ as $x \to \infty$ if for every $\varepsilon > 0$, there exists constant $x_0 \in \mathbb{R}$ such that:

$$|f(x)| < \varepsilon |g(x)| \text{ for all } x > x_0 \tag{4.3}$$

  Intuitively, $f(x) = o(g(x))$ means that $g(x)$ grows much faster than $f(x)$, or equivalently that $f(x)$ is of lower order than $g(x)$. If $g(x)$ is non-zero beyond a certain point, this is equivalent to:

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0 \tag{4.4}$$

Note that both notions can be easily generalised for other limits than infinity.

Remark 7. Although it is widespread to write $f(x) = O(g(x))$ and $f(x) = o(g(x))$, the use of the equality is an abuse of notation, since this is not a symmetric statement. For instance, $O(x) = O(x^2)$ but $O(x^2) \neq O(x)$. The equality here should be understood in the same sense as we use in English: "*Aristotle is a man, but a man is not necessarily Aristotle*". A more precise notation would be to saw $f(x) < O(g(x))$ or $f(x) \in O(g(x))$, with $O(g(x))$ thought as a class of functions h satisfying eq. (4.6).

Properties 1. The following important properties hold:

- Multiplicative constants are irrelevant: if $f(x) = O(g(x))$, then $100f(x) = O(g(x))$.
- When adding two functions, we only care about the larger one. For example $x^3 + 100x^2 = O(x^3)$.
- For all $a, b > 0$, we have $x^a = O(x^b)$ if and only if $a \leq b$ and $x^a = o(x^b)$ if and only if $a < b$.
- Polynomials are always smaller than exponentials: $x^a = o(e^{x^\epsilon})$ for every $a, \epsilon > 0$, even if ϵ is much smaller than a. For example, $x^{100} = o(e^{\sqrt{x}})$.
- Logarithms are always smaller than polynomials: $(\log x)^a = o(x^\epsilon)$ for all $a, \epsilon > 0$, even if ϵ is much smaller than a. For example, $100x^2 \log x = o(x^3)$.

An useful and related notion is the big-Theta:

Definition 15 (Big-Θ)**.** Let $f, g : \mathbb{R} \to \mathbb{R}$ denote two real-valued functions. We say "f is theta of g" and write $f(x) = \Theta(g(x))$ as $x \to \infty$ if both $f(x) = O(g(x))$ and $g(x) = O(f(x))$ as $x \to \infty$. In order words, there exists constants $m, M > 0$ and $x_0 \in \mathbb{R}$ such that for $x > x_0$:

$$mg(x) < f(x) < Mg(x) \tag{4.5}$$

Intuitively, $f(x) = \Theta(g(x))$ means that f is of the same order as g. It is also common to see the notation $f(x) \asymp g(x)$ and to say f and g are asymptotically equivalent.

A complementary notion, often used in the context of computer science is the big-Ω.

Definition 16 (Big-Ω notation)**.** Let $f, g : \mathbb{R} \to \mathbb{R}$ denote two real-valued functions. We say:

- "$f(x)$ is big-Ω of $g(x)$" and write $f(x) = \Omega(g(x))$ as $x \to \infty$ if $g(x) = O(f(x))$ as $x \to \infty$. More precisely, there exists $M > 0$ and $x_0 \in \mathbb{R}$ such that:

$$|f(x)| > M|g(x)| \text{ for all } x > x_0 \tag{4.6}$$

  Intuitively, $f(x) = O(g(x))$ means $f(x)$ is "at least" $g(x)$. It is used to denote asymptotic lower bounds. If $g(x)$ is non-zero beyond a certain point, this is equivalent to:

$$\liminf_{x \to \infty} \frac{f(x)}{g(x)} > 0 \tag{4.7}$$

- "$f(x)$ is little-ω of $g(x)$" and write $f(x) = \omega(g(x))$ as $x \to \infty$ if $g(x) = o(f(x))$ as $x \to \infty$. More precisely, for every $\varepsilon > 0$, there exists constant $x_0 \in \mathbb{R}$ such that:

$$|f(x)| > \varepsilon |g(x)| \text{ for all } x > x_0 \tag{4.8}$$

  Intuitively, $f(x) = \omega(g(x))$ means that $f(x)$ dominates $g(x)$ asymptotically. If $g(x)$ is non-zero beyond a certain point, this is equivalent to:

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \infty \tag{4.9}$$

Example 10. Some examples with Big-Ω:

- $4x^2 - 3x + 2 = \Omega(x^2)$
- $x^5 = \omega(x^4)$
- $e^x = \omega(x^a)$ for any $a > 0$.

References

Geoffrey Grimmett and David Stirzaker. *Probability and random processes.* Oxford university press, 2020.