1 Metrics and norms

Let $X$ be a non-empty set, on which we want to have a notion of distance. This notion is formalized by the concept of metric.

Definition 1.1 A metric on $X$ is a function $d : X \times X \to [0, \infty)$ with the following properties.

(a) Reflexivity: $d(x, y) = 0$ iff $x = y$.

(b) Symmetry: $d(x, y) = d(y, x)$ for all $x, y \in X$.

(c) Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$.

The pair $(X, d)$ is called a metric space.

Sketch a triangle with sides of lengths $x, y, z$ to illustrate the triangle inequality, which makes you understand the terminology as well.

For a given $X$ there are many metrics possible. Suppose one chooses a metric $d$, then $d'(x, y) := pd(x, y)$ defines another metric for any $p > 0$ as is easily seen. But also $d''(x, y) := \frac{d(x, y)}{1 + d(x, y)}$ defines a metric (less easy to see).

On $X = \mathbb{R}$, one usually takes $d(x, y) = |x - y|$, the Euclidean metric. On $X = \mathbb{R}^k$ with $k$ an integer greater than 1, there are more than one popular choices. Points $x$ in $\mathbb{R}^k$ have coordinates $x_i$, $i = 1, \ldots, k$. A favourite choice of a metric is $d(x, y) = (\sum_{i=1}^{k}(x_i - y_i)^2)^{1/2}$, called the Euclidean metric on $\mathbb{R}^k$. Think of Pythagoras’ theorem in $\mathbb{R}^2$ for an illustration.

Another metric on $\mathbb{R}^k$ is $d'(x, y) = \sum_{i=1}^{k}|x_i - y_i|$, and yet another one is $d''(x, y) = \max\{|x_i - y_i|, i = 1, \ldots, k\}$. These three metrics are equivalent in the following sense, there exist positive finite constants $C_1, C_2, C_3$ such that for all $x, y \in \mathbb{R}^k$ it holds that $d(x, y) \leq C_1d'(x, y) \leq C_2d''(x, y) \leq C_3d(x, y)$.

There also exist metrics on infinite dimensional spaces, some of these will be discussed below.

Related to the concept of metric is that of a norm. For that one needs that $X$ is a (real) vector space, in which case we have the following definition.

Definition 1.2 A norm on $X$ is a function $\| \cdot \| : X \to [0, \infty)$ with the following properties.

(a) Reflexivity: $\|x\| = 0$ iff $x = 0$.

(b) Homogeneity: $\|ax\| = a\|x\|$ for all $x \in X$ and $a \geq 0$.

(c) Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$.

The pair $(X, \| \cdot \|)$ is called a normed space.
If $X$ is endowed with a norm $\| \cdot \|$, then there is an obvious choice for a metric $d$, namely $d(x, y) = \| x - y \|$. Many of the examples of metrics above are derived from a norm, you check which ones and what the norms there are.

Let $X$ be the set of functions $f : [0, 1] \to \mathbb{R}$. A possible norm on $X$ is $\| f \| = \sup \{|f(x)| : x \in [0, 1]\}$. If $X$ is the space of continuous (in the usual sense) functions on $[0, 1]$ another often used norm is $\| f \|_1 = \int_0^1 |f(x)| \, dx$. In the course we will also use the norm $\| f \|_2 = (\int_0^1 f(x)^2 \, dx)^{1/2}$. You check that all these are indeed norms.

Other examples are the spaces $L^p(S, \Sigma, \mu)$ for $p \in [1, \infty]$ with the $p$-norms $\| f \|_p = (\mu(|f|^p))^{1/p}$ for $p \in [1, \infty)$ and the ‘sup-norm’ $\| f \|_\infty$. Care must be taken here with reflexivity, if $\| f \|_p = 0$ then one can only conclude that $f = 0 \mu$-a.e., which is not the same as $f(x) = 0$ for all $x \in S$. Still, one can call $\| \cdot \|$ a norm with abuse of terminology, which often happens. Another, more fundamental way out is to consider the quotient spaces $L^p(S, \Sigma, \mu)$, whose elements are classes of functions that coincide a.e. We omit a further treatment.

For random variables we consider the spaces $L^p(\Omega, \mathcal{F}, \mathbb{P})$ instead of $L^p(S, \Sigma, \mu)$.

We will often look at convergent sequences $(x_n)$ with limit $x$ in a metric space $(X, d)$. By this we mean sequences satisfying $d(x_n, x) \to 0$ when $n \to \infty$. The concept convergence depends thus on the metric on $X$! And it may happen that some sequence $(x_n)$ in $X$ converges in a metric $d$, but not in a metric $d'$. One has to be careful with the term convergent. Here is an example. Let $f_n(x) = n^{1/2}e^{-nx^2}1_{[0, \infty)}(x)$. Then $\| f_n \|_1 = \frac{1}{\sqrt{n}} \to 0$, whereas $\| f_n \|_2$ is constant.

So $f_n \not\to 0$ (convergence of the $f_n$ to the zero function in the $\| \cdot \|_1$-norm, but the $f_n$ don’t converge (to the zero function) in the $\| \cdot \|_2$-norm.

Convergence in the metrics above on $\mathbb{R}^k$ takes place simultaneously, one has $d(x_n, x) \to 0$ (in the Euclidean metric) iff $d'(x_n, x) \to 0$ iff $d''(x_n, x) \to 0$.

Finally a remark on product spaces. Suppose $(X, d_X)$ and $(Y, d_Y)$ are metric spaces and consider the product space $X \times Y$. There are various ways to define a metric on this product and a convenient is the ‘sum’ of the metrics. For any $(x_1, y_1)$ and $(x_2, y_2)$ in $X \times Y$ we define $d((x_1, y_1), (x_2, y_2)) := d_X(x_1, x_2) + d_Y(y_1, y_2)$. Verify that this $d$ is indeed a metric on $X \times Y$. If $(x_n)$ is a sequence in $X$ with limit $x$ and $(y_n)$ is a sequence in $Y$ with limit $y$, then $(x_n, y_n) \to (x, y)$, when we use the appropriate limits.

### 2 Helly’s lemma

First some notation. For a function $F$ defined on $\mathbb{R}$ we denote by $C_F$ the set of $x \in \mathbb{R}$ where $F$ is continuous.

**Lemma 2.1** Let $(F_n)$ be a sequence of distribution functions. Then there exists a, possibly defective, distribution function $F$ and a subsequence $(F_{n_k})$ such that $F_{n_k}(x) \to F(x)$, for all $x \in C_F$. 


Proof The proof’s main ingredients are an infinite repetition of the Bolzano-Weierstraß theorem combined with a Cantor diagonalization. First we restrict ourselves to working on \( \mathbb{Q} \), instead of \( \mathbb{R} \), and exploit the countability of \( \mathbb{Q} \). Write \( \mathbb{Q} = \{ q_1, q_2, \ldots \} \) and consider the \( F_n \) restricted to \( \mathbb{Q} \). Then the sequence \( (F_{n_k}(q_1)) \) is bounded and along some subsequence \( (n_k^1) \) it has a limit, \( \ell(q_1) \) say. Look then at the sequence \( F_{n_2^1}(q_2) \). Again, along some subsequence of \( (n_k^1) \), call it \( (n_k^2) \), we have a limit, \( \ell(q_2) \) say. Note that along the thinned subsequence, we still have the limit \( \lim_{k \to \infty} F_{n_2^1}(q_1) = \ell(q_1) \). Continue like this to construct a nested sequence of subsequences \( (n_k^j) \) for which we have that \( \lim_{k \to \infty} F_{n_k^j}(q_i) = \ell(q_i) \) holds for every \( i \leq j \). Put \( n_k = n_k^k \), then \( (n_k) \) is a subsequence of \( (n_k^i) \) for every \( i \leq k \). Hence for any fixed \( i \), eventually \( n_k \in (n_k^i) \). It follows that for arbitrary \( i \) one has \( \lim_{k \to \infty} F_{n_k}(q_i) = \ell(q_i) \). In this way we have constructed a function \( \ell : \mathbb{Q} \to [0, 1] \) and by the monotonicity of the \( F_n \) this function is increasing.

In the next step we extend this function to a function \( F \) on \( \mathbb{R} \) that is right-continuous, and still increasing. We put

\[
F(x) = \inf \{ \ell(q) : q \in \mathbb{Q}, q > x \}.
\]

Note that in general \( F(q) \) is not equal to \( \ell(q) \) for \( q \in \mathbb{Q} \), but the inequality \( F(q) \geq \ell(q) \) always holds true. Obviously, \( F \) is an increasing function and by construction it is right-continuous. An explicit verification of the latter property is as follows. Let \( x \in \mathbb{R} \) and \( \varepsilon > 0 \). There is \( q \in \mathbb{Q} \) with \( q > x \) such that \( \ell(q) < F(x) + \varepsilon \). Pick \( y \in (x, q) \). Then \( F(y) < \ell(q) \) and we have \( F(y) - F(x) < \varepsilon \). Note that it may happen that for instance \( \lim_{x \to \infty} F(x) < 1 \), \( F \) can be defective.

The function \( F \) is of course the one we are aiming at. Having verified that \( F \) is a (possibly defective) distribution function, we show that \( F_{n_k}(x) \to F(x) \) if \( x \in C_F \). Take such an \( x \) and let \( \varepsilon > 0 \) and \( q \) as above. By left-continuity of \( F \) at \( x \), there is \( y < x \) such that \( F(x) < F(y) + \varepsilon \). Take now \( r \in (y, x) \cap \mathbb{Q} \), then \( F(y) \leq \ell(r) \), hence \( F(x) < \ell(r) + \varepsilon \). So we have the inequalities

\[
\ell(q) - \varepsilon < F(x) < \ell(r) + \varepsilon.
\]

Then \( \limsup F_{n_k}(x) \leq \lim F_{n_k}(q) = \ell(q) < F(x) + \varepsilon \) and \( \liminf F_{n_k}(x) \geq \liminf F_{n_k}(r) = \ell(r) > F(x) - \varepsilon.\) The result follows since \( \varepsilon \) is arbitrary. \( \square \)

Here is an example for which the limit is not a true distribution function. Let \( \mu_n \) be the Dirac measure concentrated on \( \{ n \} \). Then its distribution function is given by \( F_n(x) = 1_{[n,\infty)}(x) \) and hence \( \lim_{n \to \infty} F_n(x) = 0 \). Hence any limit function \( F \) in Lemma 2.1 has to be the zero function, which is clearly defective.

3 Inverse function theorem (IFT)

The formulation of the theorem is taken from wikipedia, [https://en.wikipedia.org/wiki/Inverse_function_theorem](https://en.wikipedia.org/wiki/Inverse_function_theorem). For functions of more than one variable, the IFT states that if \( F \) is a continuously differentiable function from an
open set of $\mathbb{R}^n$ into $\mathbb{R}^n$, and the total derivative is invertible at a point $p$ (i.e., the Jacobian determinant of $F$ at $p$ is non-zero), then $F$ is invertible near $p$: an inverse function to $F$ is defined on some neighborhood of $q = F(p)$.

Writing $F = (F_1, \ldots, F_n)$, this means that the system of $n$ equations $y = F(x)$, explicitly written as $y_i = F_i(x_1, \ldots, x_n)$ with $i = 1, \ldots, n$, has a unique solution for $x_1, \ldots, x_n$ in terms of $y_1, \ldots, y_n$, provided that we restrict $x$ and $y$ to small enough neighborhoods of $p$ and $q$, respectively.

Finally, the theorem says that the inverse function $F^{-1}$ is continuously differentiable, and its Jacobian derivative at $q = F(p)$ is the matrix inverse of the Jacobian of $F$ at $p$: $J_{F^{-1}}(q) = [J_F(p)]^{-1}$.

To get some intuition, one can argue as follows. Taylor’s theorem says that approximately, in a neighborhood of $p$ and with $q = F(p)$, $y = F(x)$, $A = [J_F(p)]$, $F(x) \approx F(p) + A(x - p)$, leading to 

$$y \approx q + A(x - p),$$

so

$$Ax \approx y - q + Ap.$$ 

Assuming that $A$ is an invertible matrix, one gets

$$x \approx A^{-1}(y - q) + p.$$ 

If $F$ is an affine function, $F(x) = Ax + b$, then the above heuristics is completely correct, and one gets exactly $x = A^{-1}(y - b)$.

Invertibility of $[J_F(p)]$ is a sufficient condition, not a necessary one. This can already be seen when $n = 1$, when $[J_F(p)] = F'(p)$. Let $F(x) = x^3$, $x \in \mathbb{R}$. Then $F$ is everywhere (‘globally’) invertible and $F^{-1}(x) = x^{1/3}$. But at $p = 0$, $F'(p) = 0$.

A well known example for $n = 1$ illustrates the theorem. Let $F(x) = x^2$, then $F$ is not globally invertible (since $F(-x) \neq F(x)$ for all $x$), and then also not ‘locally’ in a neighborhood of $x = 0$. But $F$ is locally invertible in a neighborhood of any $p \neq 0$, since then $F'(p) = 2p \neq 0$. Indeed, if $p > 0$, then $y = x^2$ has a unique solution $x = \sqrt{y}$ if $y$ is (sufficiently) near $q = p^2$, and if $p < 0$, then $y = x^2$ has a unique solution $x = -\sqrt{y}$ if $y$ is near $q = p^2$. Note that (in both last cases), $F^{-1}(q) = \pm\frac{1}{2}\sqrt{q} = \frac{1}{F'(p)}$ as $\sqrt{p^2} = |p|$.

4 **On the proof of Lemma 4.9 in vdV**

Here are some more detailed arguments used in that proof.
If $A_n, B_n$ are events such that $\mathbb{P}(A_n) \to 1$ and $\mathbb{P}(B_n) \to 1$, then also $\mathbb{P}(A_n \cap B_n) \to 1$. Reason as follows, $(A_n \cap B_n)^c = A_n^c \cup B_n^c$ and hence $\mathbb{P}(A_n \cap B_n)^c \leq \mathbb{P}(A_n^c) + \mathbb{P}(B_n^c) \to 0$. This is used the final statement of the first paragraph.

The rule $A = (A \cap B) \cup (A \cap B^c) \subset (A \cap B) \cup B^c$ is used to get the second display. Take $A = \{ \Psi_n(\theta_0 - \varepsilon) < -\eta \} \cap \{ \Psi_n(\theta_0 + \varepsilon) > \eta \}$ and $B = \{ \Psi_n(\hat{\theta}_n) \in [-\eta, \eta] \}$. Then $A \cap B \subset \{ \theta_0 - \varepsilon < \hat{\theta}_n < \theta_0 + \varepsilon \}$.

Here is some extra information on the text below the second display. 
$\Psi_n(\theta_0 - \varepsilon) \to \Psi(\theta_0 - \varepsilon)$ means $\mathbb{P}(|\Psi_n(\theta_0 - \varepsilon) - \Psi(\theta_0 - \varepsilon)| < \delta) \to 1$ for every $\delta > 0$. But $\mathbb{P}(|\Psi_n(\theta_0 - \varepsilon) - \Psi(\theta_0 - \varepsilon)| < \delta) \geq \mathbb{P}(|\Psi_n(\theta_0 - \varepsilon) - \Psi(\theta_0 - \varepsilon)| < \delta)$ and hence also $\mathbb{P}(\Psi_n(\theta_0 - \varepsilon) - \Psi(\theta_0 - \varepsilon) < \delta) \to 1$. Next we develop with $\eta < -\frac{1}{2}\Psi(\theta_0 - \varepsilon)$ (which is positive!),

\[
\mathbb{P}(\Psi_n(\theta_0 - \varepsilon) - \Psi(\theta_0 - \varepsilon) < \delta) \\
= \mathbb{P}(\Psi_n(\theta_0 - \varepsilon) < \Psi(\theta_0 - \varepsilon) + \delta) \\
= \mathbb{P}(\Psi_n(\theta_0 - \varepsilon) < -\eta + \Psi(\theta_0 - \varepsilon) + \delta + \eta) \\
\leq \mathbb{P}(\Psi_n(\theta_0 - \varepsilon) < -\eta + \Psi(\theta_0 - \varepsilon) + \delta - \frac{1}{2}\Psi(\theta_0 - \varepsilon)) \\
= \mathbb{P}(\Psi_n(\theta_0 - \varepsilon) < -\eta + \frac{1}{2}\Psi(\theta_0 - \varepsilon) + \delta) \\
= \mathbb{P}(\Psi_n(\theta_0 - \varepsilon) < -\eta),
\]

if we choose, which we do, $\delta = -\frac{1}{2}\Psi(\theta_0 - \varepsilon) > 0$. It follows from the assumption that $\mathbb{P}(\Psi_n(\theta_0 - \varepsilon) < -\eta) \to 1$.

With similar reasoning one sees $\mathbb{P}(\Psi_n(\theta_0 + \varepsilon) > \eta) \to 1$ and hence $\mathbb{P}(\Psi_n(\theta_0 - \varepsilon) < -\eta, \Psi_n(\theta_0 + \varepsilon) > \eta)$ tends to 1.

5 On Example 4.10 in vdV

Let $\Psi(\theta) = \mathbb{P}(X > \theta) - \mathbb{P}(X < \theta)$ and note that $\Psi$ is nonincreasing. If $X$ has a density $f$ w.r.t. Lebesgue measure, both probabilities here are continuous in $\theta$ and hence there must be a $\theta_0$ such that $\Psi(\theta_0) = 0$, which is then equivalent to $\mathbb{P}(X < \theta_0) = \frac{1}{2}$. One further has

\[
\Psi(\theta_0 - \varepsilon) = 1 - 2\mathbb{P}(X < \theta_0 - \varepsilon) = 2 \int_{\theta_0-\varepsilon}^{\theta_0} f(x) \, dx,
\]

which is strictly positive if $f$ is strictly positive on the interval $[\theta_0 - \varepsilon, \theta_0]$. One similarly shows $\Psi(\theta_0 + \varepsilon) < 0$.

The more general condition $\mathbb{P}(X < \theta_0 - \varepsilon) < \frac{1}{2} < \mathbb{P}(X < \theta_0 + \varepsilon)$ for all positive $\varepsilon$ gives first (let $\varepsilon \to 0$) $\mathbb{P}(X < \theta_0) \leq \frac{1}{2} \leq \mathbb{P}(X \leq \theta_0)$, using right-
continuity of a distribution function. Then
\[
\Psi(\theta_0 - \varepsilon) = \mathbb{P}(X > \theta_0 - \varepsilon) - \mathbb{P}(X < \theta_0 - \varepsilon) \\
> \mathbb{P}(X > \theta_0 - \varepsilon) - \frac{1}{2} \\
\geq \mathbb{P}(X \geq \theta_0) - \frac{1}{2} \\
\geq 0.
\]
It follows that $\Psi(\theta_0 - \varepsilon) > 0$. The inequality $\Psi(\theta_0 + \varepsilon) < 0$ is shown by similar arguments (you try!).

6 On the second display of page 47

The display reads
\[
P\left( |\hat{\Psi}_n(\hat{\theta}_n) | > M \right) \leq P\left( \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_n(X_i) > M \right) + P\left( A_n^c \right).
\]

To prove this, one needs the information in and above the previous display, which is \textit{valid on the event $A_n = \{ \hat{\theta}_n \in B \}$} (this follows from the assumptions in Theorem 4.11):

On $A_n$: $|\hat{\Psi}_n(\hat{\theta}_n) | \leq \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_n(X_i)$.

Let $C = \{ |\hat{\Psi}_n(\hat{\theta}_n) | > M \}$ and $C' = \{ \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_n(X_i) > M \}$, and observe that it now follows

$C \cap A_n \subset C' \cap A_n$.

Use next the disjoint union $C = (B \cap A_n) \cup (C \cap A_n^c)$ which is contained in $(C \cap A_n) \cup A_n^c$, from which it follows that $P(C) \leq P(C \cap A_n) + P(A_n^c)$. Then
\[
P\left( |\hat{\Psi}_n(\hat{\theta}_n) | > M \right) = P(C) \\
\leq P(C \cap A_n) + P(A_n^c) \\
\leq P(C' \cap A_n) + P(A_n^c) \\
\leq P(C') + P(A_n^c) \\
\leq P\left( \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_n(X_i) > M \right) + P(A_n^c),
\]
and we arrive where we wished to be.