Mathematics

Imahn Shekhzadeh*
imahn.shekhzadeh@unige.ch

November 25, 2023

1 Measure Theory

Definition 1.1 (Generated σ-algebra [**generated-sigma-algebras**])**.** Let X be a set and $\mathcal{E} \subset \mathcal{P}(X)$ a non-empty collection of subsets of X. The *smallest* σ-algebra containing all the sets of $\mathcal{E}$ is denoted by $\sigma(\mathcal{E})$.

Corollary 1.2. Let $\mathcal{E}_1$, $\mathcal{E}_2 \subset \mathcal{P}(X)$ be such that $\mathcal{E}_1 \subset \mathcal{E}_2$. Then $\sigma(\mathcal{E}_1) \subset \sigma(\mathcal{E}_2)$.

Definition 1.3 (Measurable function)**.** Let $(X, \mathcal{E})$ and $(Y, \mathcal{F})$ be measurable spaces. A map $f : X \to Y$ is said to be $\mathcal{E}$*-measurable* if

$$f^{-1}(\mathcal{F}) := \left\{ f^{-1}(A) | A \in \mathcal{F} \right\} := \left\{ \{x \in X | f(x) \in A\} \, | A \in \mathcal{F} \right\} \subset \mathcal{E}. \tag{1.1}$$

Theorem 1.4 (Generator and measurable function [**measurable-functions**])**.** Let $(X, \mathcal{E})$ and $(Y, \mathcal{F})$ be measurable spaces and $\mathcal{F} = \sigma(\mathcal{G})$, i.e. $\mathcal{F}$ is the σ-algebra generated by a family $\mathcal{G} \subset \mathcal{P}(Y)$, where $\mathcal{P}(Y)$ denotes the power set of Y. Then $f : X \to Y$ is measurable if and only if

$$f^{-1}(G) \in \mathcal{E} \ \forall G \in \mathcal{G}. \tag{1.2}$$

Proof. "⇒" Since $\mathcal{F} = \sigma(\mathcal{G})$, it obviously holds that $\mathcal{G} \subset \mathcal{F}$ and therefore $f^{-1}(G) \in \mathcal{E} \ \forall G \in \mathcal{G}$ is ensured by f being measurable.

"⇐" Define the set $\mathcal{M} := \{B \subset Y \mid f^{-1}(B) \in \mathcal{A}\}$. First we want to convince ourselves that $\mathcal{M}$ is a σ-algebra on Y.

1. $\emptyset \in \mathcal{M}$, since $f^{-1}(\emptyset) = \{x \in X \mid f(x) \in \emptyset\} = \emptyset$.
2. Let $B \in \mathcal{M}$, then also $Y \backslash B \in \mathcal{M}$, since $f^{-1}(Y \backslash B) = f^{-1}(Y) \backslash f^{-1}(B)$, as can be easily shown by using the definition of the complement of a set. Since $f^{-1}(Y) = X$, it follows that $f^{-1}(Y \backslash B) = X \backslash f^{-1}(B)$. Since by assumption $B \in \mathcal{M}$ (and therefore $f^{-1}(B) \in \mathcal{A}$) and $\mathcal{A}$ itself is a σ-algebra, it follows that $X \backslash f^{-1}(B) \in \mathcal{A}$.
3. Let $B_i \in \mathcal{M}$ for $i \in \mathbb{N}$, then also $\cup_{i \in \mathbb{N}} B_i \in \mathcal{M}$, since

$$f^{-1}\left(\bigcup_{i \in \mathbb{N}} B_i \right) = \bigcup_{i \in \mathbb{N}} f^{-1}(B_i).$$

Since $\mathcal{M}$ is a σ-algebra and since $\mathcal{G} \subset \mathcal{M} \Rightarrow \mathcal{F} = \sigma(\mathcal{G}) \subset \mathcal{M} = \sigma(\mathcal{M})$, it follows that f is measurable. ■

*Computer Science Department, University of Geneva, Route de Drize 7, Carouge, Switzerland

Lemma 1.5 (Push-forward measure)**.** Let $(X, \mathcal{E})$ and $(Y, \mathcal{F})$ be measurable spaces. Given a measurable map $f : X \to Y$ and a measure μ on $\mathcal{E}$, let $f_{\#}\mu$ be defined by

$$f_{\#}\mu(A) := \mu(f^{-1}(A)) \quad \forall A \in \mathcal{F}. \tag{1.3}$$

$f_{\#}\mu$ is a measure on $\mathcal{F}$ and called the *push-forward* of μ under f.

Proof. Obviously, $f_{\#}\mu(\emptyset) = \mu(f^{-1}(\emptyset)) \overset{(1.1)}{=} \mu(\emptyset) = 0$. Also, no matter what kind of set $A \in \mathcal{F}$ we take, since $\mu(A) \geq 0$, the same holds for $f_{\#}\mu(A)$. Finally, let $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ be a sequence of mutually disjoint sets, then:

$$f_{\#}\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) \tag{1.4}$$

$$= \mu\left(f^{-1}\left(\bigcup_{n\in\mathbb{N}} A_n\right)\right) \tag{1.5}$$

$$\overset{(1.1)}{=} \mu\left(\left\{x \in X \middle| f(x) \in \bigcup_{n\in\mathbb{N}} A_n\right\}\right) \tag{1.6}$$

$$= \mu\left(\bigcup_{n\in\mathbb{N}} \{x \in X | f(x) \in A_n\}\right) \tag{1.7}$$

$$= \mu\left(\bigcup_{n\in\mathbb{N}} f^{-1}(A_n)\right) \tag{1.8}$$

$$= \sum_{n\in\mathbb{N}} \mu\left(f^{-1}(A_n)\right) \tag{1.9}$$

$$= \sum_{n\in\mathbb{N}} f_{\#}\mu(A_n) \tag{1.10}$$

■

Corollary 1.6 (Push-forward of a probability measure)**.** Let $(X, \mathcal{E})$ and $(Y, \mathcal{F})$ be measurable spaces. Given a measurable map $f : X \to Y$ and a probability measure on $\mathcal{E}$, the push-forward of μ under f, denoted by $f_{\#}\mu$, is also a probability measure.

Proof. Since μ is in particular a measure and thus $f_{\#}\mu$ is also a measure, we only need to show that

$$f_{\#}\mu(Y) = \mu\left(f^{-1}(\mu)\right) = \mu\left(\{x \in X \mid f(x) \in Y\}\right) = \mu(X) = 1. \tag{1.11}$$

■

Definition 1.7 (σ-finite measure)**.** Let $(X, \mathcal{A})$ be a measurable space and μ a measure on it. If there are sets $A_1, A_2, \ldots \in \mathcal{A}$ with $\mu(A_n) < \infty \; \forall n \in \mathbb{N}$ that satisfy

$$\bigcup_{n\in\mathbb{N}} A_n = X \tag{1.12}$$

then we say μ is *σ-finite*.

Remark 1.8. Obviously, every finite measure is σ-finite; however, the converse does not necessarily hold [**sigma-finite**].

Definition 1.9 (Absolutely continuous measures.)**.** Let μ and ν be two measures on a σ-algebra $\mathcal{A}$. ν is called *absolutely continuous* w.r.t. μ, written as

$$\nu \ll \mu, \tag{1.13}$$

if for each $A \in \mathcal{A}$, $\mu(A) = 0$ implies $\nu(A) = 0$. If μ and ν are both absolutely continuous w.r.t each other μ and ν are called *equivalent*.

Theorem 1.10 (Radon-Nikodym Theorem [**measure-integration**])**.** Let μ be a σ-finite measure on a measurable space $(S, \mathcal{A})$. Then it is equivalent:

1. $\nu \ll \mu$,
2. $d\nu = hd\mu$ for some measurable function $h : S \to \mathbb{R}_+$.

The density h then is μ-a.e. finite and μ-a.e. unique.

Lemma 1.11 (Frechét Inception Distance)**.** For two multivariate Gaussian distributions $\mathcal{G}(\mu_x, \Sigma_x)$, $\mathcal{G}(\mu_y, \Sigma_y)$, the *Frechét Inception Distance* (FID) is defined as [`https://arxiv.org/pdf/1706.08500.pdf`]:

$$d\left(\mathcal{G}(\mu_x, \Sigma_x), \mathcal{G}(\mu_y, \Sigma_y)\right) := \sqrt{||\mu_x - \mu_y||_2^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_y - 2\left(\Sigma_x\Sigma_y\right)^{\frac{1}{2}}\right)}. \tag{1.14}$$

It is a metric.

Proof. Clearly, $d\left(\mathcal{G}(\mu_x, \Sigma_x), \mathcal{G}(\mu_x, \Sigma_x)\right) = 0$. Also, $d\left(\mathcal{G}(\mu_x, \Sigma_x), \mathcal{G}(\mu_y, \Sigma_y)\right) = d\left(\mathcal{G}(\mu_y, \Sigma_y), \mathcal{G}(\mu_x, \Sigma_x)\right)$, which holds because $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for any matrices A and B. To see that $d\left(\mathcal{G}(\mu_x, \Sigma_x), \mathcal{G}(\mu_y, \Sigma_y)\right) \geq 0$, note that $\mathrm{Tr}(\Sigma_x + \Sigma_y - 2\left(\Sigma_x\Sigma_y\right)^{1/2}) = \mathrm{Tr}\left(\left(\Sigma_x^{1/2} - \Sigma_y^{1/2}\right)^2\right) \geq 0$, since the covariance matrices contain the variances on the diagonal, which are obviously non-negative. It now remains to be shown that also the triangle inequality is fulfilled. For this, note that

$$\begin{aligned}
d\left(\mathcal{G}(\mu_x, \Sigma_x), \mathcal{G}(\mu_y, \Sigma_y)\right) &= \sqrt{||\mu_x - \mu_y||_2^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_y - 2\left(\Sigma_x\Sigma_y\right)^{\frac{1}{2}}\right)} && (1.15)\\
&= \sqrt{||\mu_x - \mu_y||_2^2 + \mathrm{Tr}\left(\left(\Sigma_x^{1/2} - \Sigma_y^{1/2}\right)^2\right)} && (1.16)\\
&= \sqrt{||\mu_x - \mu_y||_2^2 + ||\sigma_x - \sigma_y||_2^2}, && (1.17)
\end{aligned}$$

where σ_x and σ_y denote vectors containing the standard deviations of the two Gaussian distributions. Clearly,

$$\begin{aligned}
d\left(\mathcal{G}(\mu_x, \Sigma_x), \mathcal{G}(\mu_z, \Sigma_z)\right) &= \sqrt{||\mu_x - \mu_z||_2^2 + ||\sigma_x - \sigma_z||_2^2} && (1.18)\\
&\leq \sqrt{||\mu_x - \mu_y||_2^2 + ||\mu_y - \mu_z||_2^2 + ||\sigma_x - \sigma_y||_2^2 + ||\sigma_y - \sigma_z||_2^2} && (1.19)\\
&\leq \sqrt{||\mu_x - \mu_y||_2^2 + ||\sigma_x - \sigma_y||_2^2} \;+\; \sqrt{||\mu_y - \mu_z||_2^2 + ||\sigma_y - \sigma_z||_2^2} && (1.20)\\
&= d\left(\mathcal{G}(\mu_x, \Sigma_x), \mathcal{G}(\mu_y, \Sigma_y)\right) + d\left(\mathcal{G}(\mu_y, \Sigma_y), \mathcal{G}(\mu_z, \Sigma_z)\right), && (1.21)
\end{aligned}$$

since $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, as one can directly show by squaring for non-negative x, $y \in \mathbb{R}$. ■

Definition 1.12 (Convergence in probability)**.** Assume we have a sequence of random variables $(X_n)_{n\in\mathbb{N}}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say this sequence converges to another random variable X if

$$\forall \epsilon > 0 : \lim_{n\to\infty} \mathbb{P}\left\{|X_n - X| > \epsilon\right\} = 0. \tag{1.22}$$

2 Functional Analysis

Definition 2.1 (Normed Space)**.** content...

Theorem 2.2. Let $(X, ||\cdot||)$ be a normed space. Then the "second triangle inequality" holds:

$$|||\varphi|| - ||\psi||| \leq ||\varphi - \psi|| \quad \forall \varphi, \psi \in X. \tag{2.1}$$

Proof: For φ, $\psi \in X$ we have

$$||\varphi|| = ||\varphi - \psi + \psi|| \leq ||\varphi - \psi|| + ||\psi|| \Leftrightarrow ||\varphi|| - ||\psi|| \leq ||\varphi - \psi|| \,. \tag{2.2}$$

By exchanging the roles of φ and ψ we obtain

$$||\psi|| - ||\varphi|| \leq ||\varphi - \psi|| \tag{2.3}$$

and thus

$$|||\varphi|| - ||\psi||| \leq ||\varphi - \psi|| \,. \tag{2.4}$$

■

Theorem 2.3. Let $(X, ||\cdot||)$ be a normed space. Then the addition, scalar multiplication and the norm itself are continuous.

Proof:

- ad continuity of the addition: Let $(\varphi_n)_{n\in\mathbb{N}}$ and $(\psi_n)_{n\in\mathbb{N}}$ be convergent sequences in X with limit elements φ, $\psi \in X$, i.e. $\varphi_n \longrightarrow \varphi$ and $\psi_n \longrightarrow \psi$ for $n \to \infty$. Thus

$$0 \leq ||(\varphi_n + \psi_n) - (\varphi + \psi)|| \leq ||\varphi_n - \varphi|| + ||\psi_n - \psi|| \longrightarrow 0 \quad \text{for } n \to \infty \tag{2.5}$$

  and hence $\varphi_n + \psi_n \to \varphi + \psi$ for $n \to \infty$.

- ad continuity of the scalar multiplication: Let $(\alpha_n)_{n\in\mathbb{N}} \in \mathbb{K}$ converge to $\alpha \in \mathbb{K}$ and $(\varphi_n)_{n\in\mathbb{N}} \in X \to \varphi \in X$ for $n \to \infty$. Then

$$0 \leq ||\alpha_n\varphi_n - \alpha\varphi|| = ||\alpha_n(\varphi_n - \varphi) + (\alpha_n - \alpha)\varphi|| \leq ||\alpha_n(\varphi_n - \varphi)|| + ||(\alpha_n - \alpha)\varphi|| \tag{2.6}$$

$$\leq |\alpha_n| \, ||\varphi_n - \varphi|| + |\alpha_n - \alpha| \, ||\varphi|| \overset{n\to\infty}{\longrightarrow} 0. \tag{2.7}$$

  This implies $\alpha_n\varphi_n \to \alpha\varphi$ for $n \to \infty$.

- ad continuity of the norm: Let $\varphi_n \to \varphi$. With Theorem 2.2 we have:

$$0 \leq | \, ||\varphi_n|| - ||\varphi|| \, | \leq ||\varphi_n - \varphi|| \overset{n\to\infty}{\longrightarrow} \tag{2.8}$$

  and hence $||\varphi_n|| \to ||\varphi||$ for $n \to \infty$.

■

Definition 2.4. Two norms $||\cdot||_a$ and $||\cdot||_b$ on a linear space X are said to be equivalent if and only if there exist positive constants $0 < c \leq C < \infty$ such that

$$c \, ||\varphi||_b \leq ||\varphi||_a \leq C \, ||\varphi||_b \quad \forall \varphi \in X. \tag{2.9}$$

(It is also possible to write this as $\tilde{c} \, ||\varphi||_a \leq ||\varphi||_b \leq \tilde{C} \, ||\varphi||_a$ with $\tilde{c} := C^{-1}$ and $\tilde{C} := c^{-1}$, where $0 < \tilde{c} \leq \tilde{C} < \infty$.)

Lemma 2.5. Let X be a linear space and the pairs $(||\cdot||_a, ||\cdot||_c)$ and $(||\cdot||_b, ||\cdot||_c)$ be equivalent. Then also the pair $(||\cdot||_a, ||\cdot||_b)$ is equivalent.

Proof: By assumption, we know that

$$c\,||\varphi||_c \leq ||\varphi||_a \leq C\,||\varphi||_c \quad \forall \varphi \in X \tag{2.10}$$

and

$$d\,||\varphi||_c \leq ||\varphi||_b \leq D\,||\varphi||_c \quad \forall \varphi \in X \tag{2.11}$$

$$\Leftrightarrow ||\varphi||_c \leq \frac{1}{d}\,||\varphi||_b \leq \frac{D}{d}\,||\varphi||_c\,. \tag{2.12}$$

$$\overset{(2.10)}{\Leftrightarrow} \frac{1}{C}\,||\varphi||_a \leq ||\varphi||_c \leq \frac{1}{d}\,||\varphi||_b \leq \frac{D}{d}\,||\varphi||_c \leq \frac{D}{d\cdot c}\,||\varphi||_a \tag{2.13}$$

$$\Leftrightarrow \frac{1}{C}\,||\varphi||_a \leq \frac{1}{d}\,||\varphi||_b \leq \frac{D}{d\cdot c}\,||\varphi||_a \tag{2.14}$$

$$\Leftrightarrow \frac{d}{C}\,||\varphi||_a \leq ||\varphi||_b \leq \frac{D}{c}\,||\varphi||_a\,. \tag{2.15}$$

It is clear that $0 < dC^{-1} \leq Dc^{-1} < \infty$ holds. ■

Theorem 2.6. On a *finite-dimensional* space X over a field $\mathbb{K}$ all norms are equivalent.

*Proof [***werner-fa***]*: Let $\dim(X) = n$, $\{e_1, \ldots, e_n\}$ be a basis of X and $||\cdot||$ a norm on X. We can now show that $||\cdot||$ is equivalent to the Euclidean norm $||\sum_{i=1}^n \alpha_i e_i||_2 = \left(\sum_{i=1}^n |\alpha_i|^2\right)^{1/2}$ as follows: Set $K := \max\{||e_1||, \ldots, ||e_n||\} > 0$. Then from the triangle inequality for $||\cdot||$ we have:

$$||x|| = \left|\left|\sum_{i=1}^n \alpha_i e_i\right|\right| \leq \sum_{i=1}^n ||\alpha_i e_i|| = \sum_{i=1}^n |\alpha_i|\,||e_i|| \tag{2.16}$$

Since $(|\alpha_1|, \ldots, |\alpha_n|)^T$, $(||e_1||, \ldots, ||e_n||)^T \in \mathbb{R}^n$ and

$$\Big\langle (|\alpha_1|, \ldots, |\alpha_n|), (||e_1||, \ldots, ||e_n||) \Big\rangle = \sum_{i=1}^n |\alpha_i|\,||e_i|| \tag{2.17}$$

we can use the Cauchy-Schwarz inequality:

$$\langle (|\alpha_1|, \ldots, |\alpha_n|), (||e_1||, \ldots, ||e_n||) \rangle \leq \left|\left|\sum_{i=1}^n \alpha_i e_i\right|\right|_2 \cdot \left|\left|\sum_{i=1}^n e_i\right|\right|_2 = \sqrt{\sum_{i=1}^n |\alpha_i|^2} \cdot \sqrt{\sum_{i=1}^n ||e_i||^2} \tag{2.18}$$

$$= ||x||_2 \cdot \sqrt{\sum_{i=1}^n K^2} = K\sqrt{n}\,||x||_2 \quad \forall x \in X, \tag{2.19}$$

where in the last line we used that $K = \max\{||e_1||, \ldots, ||e_n||\}$ and $x = \sum_{i=1}^n \alpha_i e_i$. Putting the last Eq. into Eq. (2.16), we have:

$$||x|| \leq \sum_{i=1}^n |\alpha_i|\,||e_i|| \leq K\sqrt{n}\,||x||_2 \quad \forall x \in X. \tag{2.20}$$

Now define the set

$$S := \{x \in X \mid ||x||_2 = 1\}. \tag{2.21}$$

This set is closed since it is the preimage of the closed set $\{1\} \subset \mathbb{R}$ under the continuous function $||\cdot||_2$, cf. Theorem 2.3, 3.12. S is also closed since $S \subset B_r(0) = \{\psi \in X \mid ||\psi||_2 < r\}$ for $r > 0$ (here, we take into account that every norm induces a metric). Thus, S is compact according to Heine-Borel (which applies to every finite-dimensional normed vector space). Since every continuous function takes its minimum on a compact set, we know that $||\cdot||$ has a minimum $m > 0$ on S. Since $x \cdot ||x||_2^{-1} \in S$ for $x \neq 0$, we have (m is the minimum of the function $||\cdot||$):

$$m\,||x||_2 \leq ||x|| \quad \forall x \in X. \tag{2.22}$$

All in all, we proved:

$$m\,||x||_2 \leq ||x|| \leq K\sqrt{n}\,||x||_2 \quad \forall x \in X. \tag{2.23}$$

■

Definition 2.7. Let X be a linear space equipped with two metrics d and d'. Then the metrics are *strongly equivalent* if and only if there exist positive constants $0 < \alpha \leq \beta < \infty$ such that

$$\alpha d(x,y) \leq d'(x,y) \leq \beta d(x,y). \tag{2.24}$$

[**equivalence-metrics**]

Remark 2.8. Obviously, Eq. (2.9) can also be written as

$$c\,||\varphi - \psi||_b \leq ||\varphi - \psi||_a \leq C\,||\varphi - \psi||_b \quad \forall \varphi, \psi \in X. \tag{2.25}$$

Thus, also Theorem 2.6 holds for metrics as well: In a finite-dimensional space X, all metrics are strongly equivalent.

3 Topology

Definition 3.1 (Metric Space [**fa2019**])**.** Let X be a non-empty set. Then the map

$$d : X \times X \to \mathbb{R}$$

is called a *metric* on X if for all φ, ψ, $\chi \in X$ the following properties are given:

- $d(\varphi, \psi) \geq 0$,
- $d(\varphi, \psi) = d(\psi, \varphi)$,
- $d(\varphi, \varphi) = 0$,
- $d(\varphi, \psi) \leq d(\varphi, \chi) + d(\chi, \psi)$.

Definition 3.2 (Topological Space [**topology-singh**])**.** A topology on a set X is a collection τ of subsets of X such that

- the intersection of two members of τ is in τ,
- the union of any collection of members of τ is in τ,
- the empty set $\{\}$ and X itself are in τ.

A set X endowed with a topological structure τ on it is called a *topological space*. The elements of X are called *points*, and the members of τ are called the *open sets*.

Remark 3.3 (Discrete Metric)**.** Let X be an arbitrary set and $d(\varphi, \psi) = 1$ $\forall \varphi, \psi \in X$ and $d(\varphi, \varphi) = 0$. Then d is a metric, called the *discrete metric* on X.

Definition 3.4 (Open Ball)**.** Let (X, d) be a metric space, $\varphi \in X$ and $r > 0$. Then

$$B_r(\varphi) := \{\psi \in X | d(\varphi, \psi) < r\} \subset X \tag{3.1}$$

open ball in X with middle point φ and radius r.

Definition 3.5 (Open Set)**.** Let (X, d) be a metric space. Then a subset $U \subset X$ is called *open* in X if for every $\varphi \in U$ there is an open ball $B_r(\varphi)$ that is contained in U, i.e. $B_r(\varphi) \subset U$.

Definition 3.6 (Closed Set)**.** Let (X, d) be a metric space. Then a subset $A \subset X$ is called *closed* in X if the complement $X \backslash A$ is open according to Definition 3.5.

Theorem 3.7. Open balls are open.

Proof: An illustration is shown in Figure. Let $f \in X$ and for $r > 0$ consider the open ball $U := B_r(f) \subset X$. By definition, for every $\varphi \in U$ it holds that $d(\varphi, f) < r$. Now we show that

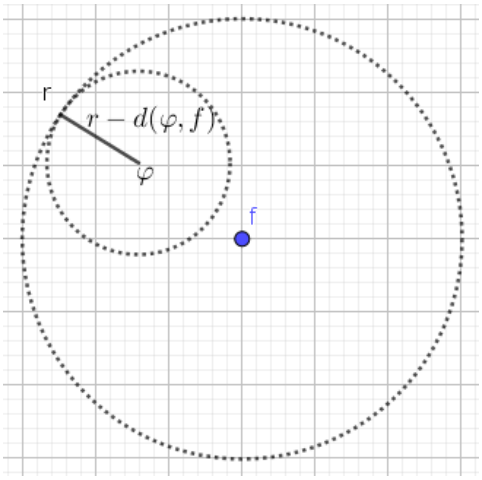
$$B_{r-d(\varphi,f)}(\varphi) \subset U. \tag{3.2}$$

For this consider $\psi \in B_{r-d(\varphi,f)}(\varphi)$, i.e. $d(\psi, \varphi) < r - d(\varphi, f)$. With the triangle inequality we obtain:

$$d(f, \psi) \leq d(f, \varphi) + d(\varphi, \psi) < d(f, \varphi) + r - d(\varphi, f) = r, \tag{3.3}$$

i.e. $d(f, \psi) < r$ and thus $\psi \in B_r(f)$. ■

Theorem 3.8. Let (X, d) be a metric space. The collection of open sets as in Definition 3.5 gives a topology. Thus, every metric space is a topological space.



Definition 3.9 (Bounded Set)**.** Let (X, d) be a metric space. Then $A \subset X$ is called *bounded* if there exists $r > 0$ and $\varphi \in X$ such that $A \subset B_r(\varphi) \subset X$. [**MfPII**]

Definition 3.10 (ϵ-δ definition of continuity)**.** Let D be a subset of a metric space (X, d) and let $p \in D$. Let (Y, d') be another metric space. A function $f : D \to Y$ is called *continuous* at p if for all $\epsilon > 0$ there exists a $\delta > 0$ s.t.

$$d'\left(f\left(z\right), f\left(p\right)\right) < \epsilon \; \forall z \in D \quad \text{with } d\left(z, p\right) < \delta. \tag{3.4}$$

[**MfPI**]

Lemma 3.11. Consider the map $f : \mathbb{R}^m \to \mathbb{R}^n$. Then f is continuous if all the $f_i : \mathbb{R}^m \to \mathbb{R}$, $i \in \{1, \ldots, n\}$ are continuous.

Proof. Since all metrics are strongly equivalent, it does not matter which metric we equip $\mathbb{R}$, $\mathbb{R}^m$ and $\mathbb{R}^n$ with, cf. Remark 2.8; for the following, consider the Euclidean metric. Suppose each f_i is continuous for all $i \in \{1, \ldots, n\}$, i.e.:

$$\forall p \in \mathbb{R}^m : \forall \epsilon_i > 0 \; \exists \delta_i > 0 : d_{\mathbb{R}}(f_i(p), f_i(z)) < \frac{\epsilon_i}{\sqrt{n}} \; \forall z \in D \quad \text{with } d_{\mathbb{R}^m}(p, z) < \delta_i. \tag{3.5}$$

Now define $\epsilon := \max\{\epsilon_1, \ldots, \epsilon_n\}$. Thus:

$$d_{\mathbb{R}^n}(f(p), f(z)) = \sqrt{\sum_{i=1}^{n} \left(f_i(p) - f_i(z)\right)^2} = \sqrt{\sum_{i=1}^{n} d_{\mathbb{R}}^2\left(f_i(p), f_i(z)\right)} < \sqrt{\sum_{i=1}^{n} \left(\frac{\epsilon}{\sqrt{n}}\right)^2} = \epsilon. \tag{3.6}$$

■

Theorem 3.12. For a map $f : X \to Y$ between two metric spaces (X, d_X) and (Y, d_Y) the following statements are equivalent:

(i) f is continuous,

(ii) preimages $f^{-1}(V) := \{x \in X \mid f\left(x\right) \in V\}$ of <u>all</u> open sets $V \subset Y$ are open,

(iii) preimages $f^{-1}(A)$ of <u>all</u> closed sets $A \subset Y$ are closed.

Proof:
(i) $\Rightarrow$ (ii) Assume that f is continuous and that $V \subset Y$ is open and let $a \in f^{-1}\left(V\right)$. Since V is an open set, $\exists \epsilon > 0$ s.t. $B_\epsilon\left(f\left(a\right)\right) \subset V$, cf. Def. 3.5. By assumption, f is continuous at $a \in f^{-1}\left(V\right) \subset X$ and therefore $\forall \epsilon > 0 \; \exists \delta > 0$ s.t. $d_X\left(x, a\right) < \delta \Rightarrow d_Y\left(f\left(x\right), f\left(a\right)\right) < \epsilon \; \forall x \in X$. Put differently, $x \in B_\delta\left(a\right)$ implies $f\left(x\right) \in B_\epsilon\left(f\left(a\right)\right) \subset V$. Thus, $B_\delta(a) \subset f^{-1}(V) \Rightarrow f^{-1}(V) \subset X$ is an open set, cf.

Def. 3.5.

(ii) $\Rightarrow$ (i) Assume that $f^{-1}(Y) \subset X$ is open for $V \subset Y$ open and let $a \in X$, $\epsilon > 0$. From Theorem 3.7 we know that $B_\epsilon(f(a)) \subset Y$ is open. Thus, by assumption, $f^{-1}(B_\epsilon(f(a))) = \{x \in X \mid f(x) \in B_\epsilon(f(a))\} = \{x \in X \mid f(x) \in \{y \in Y \mid d_Y(y, f(a)) < \epsilon\}\}$ is open as well. Clearly, $a \in f^{-1}(B_\epsilon(f(a)))$. Therefore, it follows from Def. 3.5 that $\exists \delta > 0$ s.t. $B_\delta(a) \subset f^{-1}(B_\epsilon(f(a)))$. Thus, $\forall x \in B_\delta(a) : d_X(x, a) < \delta \Rightarrow d_Y(f(x), f(a)) < \epsilon$. This proves that $f : X \to Y$ is continuous in every point $a \in X$. [**cont-functions-open-sets**]

(ii) $\Rightarrow$ (iii) Assume that the preimages $f^{-1}(V)$ of all open sets $V \subset Y$ are open. Since $f^{-1}(Y \backslash V) = f^{-1}(Y) \backslash f^{-1}(V)$, as can be easily shown by using the definition of the complement of a set, we have for all open sets $V \subset Y$:

$$f^{-1}(Y \backslash V) = f^{-1}(Y) \backslash f^{-1}(V) = X \backslash f^{-1}(V). \tag{3.7}$$

Since $f^{-1}(V)$ is open by assumption, $f^{-1}(Y \backslash V) = X \backslash f^{-1}(V)$ is closed.

(iii) $\Rightarrow$ (ii) Assume that the preimages $f^{-1}(A)$ of all closed sets $A \subset Y$ are closed. Then:

$$f^{-1}(Y \backslash A) = f^{-1}(Y) \backslash f^{-1}(A) = X \backslash f^{-1}(A). \tag{3.8}$$

Since $f^{-1}(A)$ is closed by assumption, $f^{-1}(Y \backslash A) = X \backslash f^{-1}(A)$ is open. [**preimage-of-closed-sets**] ■

Corollary 3.13. Let (X, τ_1) and (Y, τ_2) be topological spaces coming from metric spaces. Then $f : X \to Y$ is continuous if and only if $f^{-1}(G)$ for every open set $G \in Y$.

Definition 3.14 (Basis)**.** A *basis* of a topology (M, τ) is a collection of open sets $\mathcal{B}$ such that for all $U \in \tau$ there exists an index set I and corresponding $B_i \in \mathcal{B}$ s. t.

$$\bigcup_{i \in I} B_i = U. \tag{3.9}$$

Example 3.15. Let (X, d) be a metric space, where the metric is the discrete metric from Remark 3.3. Then the collection of all singletons $\{\varphi \in X\}$ is basis of X.

Example 3.16. In a metric space X, the collection

$$\{B_r(x) \mid x \in X, r > 0\}$$

of open balls is a basis for the topology given by the metric on X.

Definition 3.17 (Hausdorff)**.** A topology (M, τ) is called **Hausdorff** if $\forall p \neq q \in M$ $\exists U, V \in \tau$ open with $p \in U$, $q \in V$ such that $U \cap V = \emptyset$.

Lemma 3.18. All metric spaces are Hausdorff spaces.

Proof: A visualization is shown in Fig. 1. To prove this more rigorously, define $U := B_r(p)$ and $V := B_r(q)$ with radius $r := \frac{d(p,q)}{2}$, where from Theorem 3.7 we know that U and V are open. Suppose $U \cap V = \emptyset$ would not hold. Then there exists a $z \in U \cap V$ with

$$d(p, z) < \frac{d(p, q)}{2} \tag{3.10}$$

and

$$d(q, z) < \frac{d(p, q)}{2}. \tag{3.11}$$

Therefore, by the triangle inequality for metric spaces, we have:

$$d(p, q) \leq d(p, z) + d(q, z) < d(p, q), \tag{3.12}$$

which is clearly a contradiction. ■

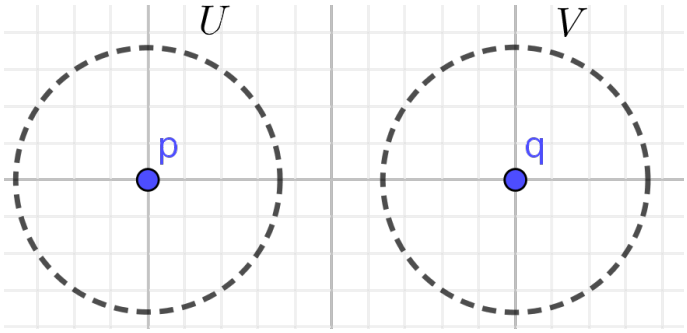


Figure 1: Visualization for why a metric space is a Hausdorff space.

Definition 3.19. Let (X, τ) be a topological space. Then it is said to be *second countable* if τ has a countable basis.

Example 3.20. For any $n > 0$, the topological spaces $\mathbb{R}^n$ are second countable.

Remark 3.21. Metric spaces are not automatically Hausdorff spaces. For example, take an uncountable set X, endow it with the discrete metric and because of Example 3.15, 3.16, the topological space X is not second countable, since X is uncountable.

Definition 3.22 (Homeomorphism)**.** A *homeomorphism* between two topological spaces X and Y is an invertible function $f : X \to Y$ such that f and f^{-1} are continuous [**topology-singh**].

Example 3.23. The Euclidean space $\mathbb{R}^n$, equipped with the usual topology, is homeomorphic to the open ball $B_r(\varphi) = \{x \in \mathbb{R}^n \mid ||x - \varphi|| < r\} \subset \mathbb{R}^n$ (consider the open ball as a metric space with the *induced metric* from the whole space of $\mathbb{R}^n$ and then equip it with the usual topology on a metric space given by the open sets).

Proof: Consider the map

$$f : \mathbb{R}^n \to B_r(\varphi),\ x \mapsto \frac{r \cdot (x - \varphi)}{1 + ||x - \varphi||}.$$

Obviously, f is continuous with inverse

$$f^{-1} : B_r(\varphi) \to \mathbb{R}^n,\ x \mapsto \frac{x}{r - ||x||} + \varphi,$$

which is also continuous. One can easily show that f and f^{-1} are inverses to each other. Thus, f describes a homeomorphism. ■

Example 3.24 (Stereographic projection)**.** The unit sphere S^n, embedded in $\mathbb{R}^{n+1}$, without the north pole, i.e. $S^n \backslash \{p\} \subset \mathbb{R}^{n+1}$, where $S^n := \{x \in \mathbb{R}^{n+1} \mid ||x||_2 = 1\}$ and $p := \{x \in \mathbb{R}^{n+1} \mid x_i = 0\ \forall i \in [1, n], x_{n+1} = 1\}$, is homeomorphic to $\mathbb{R}^n$ (consider the unit sphere as a metric space with the *induced metric* from the whole space of $\mathbb{R}^{n+1}$ and then equip it with the usual topology on a metric space given by the open sets).

Proof. Consider the map

$$f : S^n \backslash \{p\} \to \mathbb{R}^n,\ x \mapsto \frac{1}{1 - x_{n+1}} \left(x_1, \ldots, x_n\right)^T. \tag{3.13}$$

Obviously, f is continuous. Its inverse is given by

$$f^{-1} : \mathbb{R}^n \to S^n \backslash \{p\},\ x \mapsto \begin{pmatrix} 2x_1/\left(1 + ||x||_2^2\right) \\ \vdots \\ 2x_n/\left(1 + ||x||_2^2\right) \\ 1 - 2/\left(1 + ||x||_2^2\right) \end{pmatrix}. \tag{3.14}$$

It is trivial to show that f and f^{-1} are inverses to each other. One can also easily show that $||f^{-1}(x)||_2 = 1$; thus, f^{-1} does indeed bring us to the unit ball S^n. To see that f^{-1} is continuous, note that all components are continuous and thus, according to Lemma 3.11, the function itself is continuous. ■

4 Set Theory

Definition 4.1 (Binary Relation [**binary˙relations**])**.** A *binary relation* over a set X is some relation R where
$\forall x,\, y \in X$ the statement xRy is either true or false.

Definition 4.2 (Equivalence Relation and Class [**equivalence˙relation**])**.** An *equivalence relation* on a set X is a binary relation $\sim$ with the following properties $\forall x,\, y,\, z \in X$:

- **reflexivity**: $x \sim x$,
- **symmetry**: $x \sim y \Leftrightarrow y \sim x$,
- **transitivity**: $(x \sim y) \wedge (y \sim z) \Rightarrow x \sim z$.

The *equivalence class* of an element $x \in X$ is defined as

$$[x] := \{y \in X \mid x \sim y\} = \{y \in X \mid y \sim x\}\,. \tag{4.1}$$

Definition 4.3 (Partially Ordered Set [**kuratowski˙zorn˙lemma**])**.** A *partially ordered set* $(X, \leq)$ is a set X, equipped with a binary relation $\leq$, that satisfies the following properties $\forall x,\, y,\, z \in X$:

- **reflexivity**: $x \leq x$,
- **antisymmetry**: $(x \leq y) \wedge (y \leq x) \Rightarrow x = y$,
- **transitivity**: $(x \leq y) \wedge (y \leq z) \Rightarrow x \leq z$.

Definition 4.4 (Incomparability)**.** In Definition 4.3, the phrasing "partially ordered" is used to emphasize that there might exist elements $x,\, y \in X$ s.t. both $x \leq y$ and $y \leq x$ are wrong. These pairs are called *incomparable*. If either $x \leq y$ or $y \leq x$ is true, then we say that the pair is *comparable*.

Example 4.5. Consider $X := \{\{1\}, \{2\}, \{1, 2\}\}$ with $\subset$ as partial ordering. Obviously, the elements $\{1\}$ and $\{2\}$ are incomparable.

Definition 4.6 (Chain, Upper Bound, Maximal Element)**.** For preparing the Kuratowski-Zorn lemma, the following definitions come in handy:

a) A *chain* C is a partially ordered set where every pair of elements in C is comparable. One might also say that C is a *totally ordered set*.

b) An *upper bound* (if existent) of a subset $S \subset X$, where X is a partially ordered set, is an element $u \in X$ such that

$$s \leq u \;\forall s \in S. \tag{4.2}$$

Since $S \subset X$, S itself is a partially ordered set.

c) A *maximal element* (if existent) of a partially ordered set X is an element $m \in X$ such that

$$\text{if } m \leq x \text{ for some } x \in X, \text{ then } x = m. \tag{4.3}$$

This is equivalent to saying that there is no $x \in X$ such that $m \leq x$ and $x \neq m$.

Remark 4.7. For an arbitrary partially ordered set X, a maximal element (if existent) does not have to be unique. For example, consider $X := \{\{1\}, \{2\}, \{3\}, \{1, 2\}\}$ with $\subset$ as partial ordering. Both $\{3\}$ and $\{1, 2\}$ are maximal elements. However, if we consider chains, then maximal elements are indeed unique by definition.

Theorem 4.8 (Kuratowski-Zorn Lemma)**.** Let $(M, \leq)$ be a non-empty partially ordered set. If every chain $C \subset M$ has an upper bound, then M has a maximal element.

Remark 4.9. The upper bound of every chain $C \subset M$ need not be in C, by definition of a chain, but it must be in M.

5 Differential Geometry

Definition 5.1 (Smooth Atlas [**Lindemann-lec1**])**.** Let M be a second countable Hausdorff topological space. An *n-dimensional smooth atlas on M* is a collection of maps

$$\mathcal{A} = \{(\varphi_i, U_i) \mid i \in I\}, \quad \varphi_i : U_i \to \varphi_i(U_i) \subset \mathbb{R}^n$$

such that all $U_i \subset M$ are open, all φ_i are homeomorphisms, I is an index set and

- $\{U_i \mid i \in I\}$ is an open covering of M,
- $\varphi_i \circ \varphi_j^{-1} : \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)$ are smooth $\forall i, j \in I$.

The tuples (φ_i, U_i), $i \in I$, are so-called *charts* on M, the maps $\varphi_i \circ \varphi_j^{-1}$ are called *transition maps* or *changes of coordinates* and n is the *dimension* of M. ○

Remark 5.2. To see why the domain of the transition maps $\varphi_i \circ \varphi_j^{-1}$ is $\varphi_j(U_i \cap U_j)$, note that the expression $\varphi_i\left(\varphi_j^{-1}(x)\right)$ only makes sense if

$$(x \in \varphi_j(U_j)) \wedge \left(\varphi_j^{-1}(x) \in U_i\right) \Rightarrow (x \in \varphi_j(U_j)) \wedge (x \in \varphi_j(U_i)) \Rightarrow x \in \varphi_j(U_j \cap U_i).$$

Similarly, we can convince ourselves that the codomain of the transition maps $\varphi_i \circ \varphi_j^{-1}$ is given by $\varphi_i(U_i \cap U_j)$. Since $x \in \varphi_j(U_j)$, it follows that $\varphi_j^{-1}(x) \in U_j$. In addition, due to the domain of the homeomorphism φ_i, it must hold that $\varphi_j^{-1}(x) \in U_i$. Thus:

$$\begin{aligned} &\left(\varphi_j^{-1}(x) \in U_j\right) \wedge \left(\varphi_j^{-1}(x) \in U_i\right) \\ \Rightarrow &\left(\varphi_i(\varphi_j^{-1}(x)) \in \varphi_i(U_j)\right) \wedge \left(\varphi_i(\varphi_j^{-1}(x)) \in \varphi_i(U_i)\right) \\ \Rightarrow &\varphi_i\left(\varphi_j^{-1}(x)\right) \in \varphi_i\left(U_i \cap U_j\right). \end{aligned}$$

■

Definition 5.3 (Equivalence of Atlases)**.** Let M be a second countable Hausdorff topological space. Two atlases $\mathcal{A}$ and $\mathcal{B}$ on M are called *equivalent* if $\mathcal{A} \cup \mathcal{B}$ is an atlas on M.

Remark 5.4. To see that not all atlases are equivalent to each other, consider $M = \mathbb{R}$ (which is a second countable Hausdorff topological space). Consider the atlases $\mathcal{A} = \{(\varphi, M)\}$ with $\varphi : M \to M$, $x \mapsto x$ and $\mathcal{B} = \{(\psi, M)\}$ with $\psi : M \to M$, $x \mapsto x^3$. The atlases are not equivalent, since $\varphi \circ \psi^{-1} : M \to M$, $x \mapsto \sqrt[3]{x}$ is not smooth (the derivative is not continuous).

6 Normalizing Flows

Definition 6.1. Let M, N be two manifolds, $g : M \to N$ a differentiable map. If g is a bijection and its inverse $g^{-1} : N \to M$ is differentiable as well, then we call f a *diffeomorphism*. We talk of a C^k *diffeomorphism* if both g and g^{-1} are k-times continuously differentiable.

Theorem 6.2 (Change of variables [**MfPIII**])**.** Let U, $V \subset \mathbb{R}^n$ be open subsets and $T : U \to V$ a diffeomorphism, cf. Def. 6.1. Then the function $f : V \to \mathbb{C} \cup \{\infty\}$ is integrable over V if and only if the function

$$(f \circ T) \cdot \left|\det\left(\frac{\partial T_\mu}{\partial x_\nu}\right)_{\mu\nu}\right| \tag{6.1}$$

is integrable over U. In this case, it holds that

$$\int_U (f \circ T)(x) \cdot \left|\det\left(\frac{\partial T_\mu}{\partial x_\nu}(x)\right)_{\mu,\nu}\right| dx = \int_V f(y) dy. \tag{6.2}$$

Remark 6.3. If T is a diffeomorphism, then also T^{-1} is a diffeomorphism, thus we could also have chosen T^{-1} in the formulation of Theorem 6.2.

Theorem 6.4 (Inverse function theorem [**IFT**])**.** Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable on some open set $V \subset \mathbb{R}^n$ containing $\mathbf{a}$ and suppose $\det Jg(\mathbf{a}) \neq 0$, where J shall be the Jacobi matrix of g. Then there is some open set containing $\mathbf{a}$ and an open set $W \subset \mathbb{R}^n$ containing $g(\mathbf{a})$ such that $g : V \to W$ has a continuous inverse $g^{-1} : W \to V$ which is differentiable for all $\mathbf{y} \in W$.

As matrices, we can write this as

$$J(g^{-1})(\mathbf{y}) = \left[(Jg)\left(g^{-1}(\mathbf{y})\right)\right]^{-1} \tag{6.3}$$

Remark 6.5. An example for a function that is invertible and continuously differentiable but not a diffeomorphism is $g : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^3$. Its inverse is obviously $g^{-1} : \mathbb{R} \to \mathbb{R}, x \mapsto \sqrt[3]{x}$, cf. [**cube-root**] for a nice plot. However, $\frac{dg^{-1}}{dx}|_{x=0}$ does not exist. The reason is that $\det Jg(0) = 0$ and hence the inverse function theorem does not apply.

- Let $\mathbf{U}$ be a random variable and let $p(\mathbf{U})$ describe the probability distribution of it, e.g. a uniform distribution between 0 and 1. We now make a simple transformation and obtain a new random variable $\mathbf{X}$, where we again denote by $p(\mathbf{X})$ the probability distribution of $\mathbf{X}$. We obtain $\mathbf{X}$ in the following way:

  $$p(\mathbf{X}) = p(\mathbf{U}) \left| \det \left(\frac{\partial \mathbf{f}}{\partial \mathbf{U}} \right) \right|^{-1}, \tag{6.4}$$

  where $\mathbf{f}$ denotes an invertible (and hence bijective) mapping.

- Without proof, it holds that

  $$\left| \det \left(\frac{\partial \mathbf{f}}{\partial \mathbf{U}} \right) \right|^{-1} = \left| \det \left(\frac{\partial \mathbf{f}^{-1}}{\partial \mathbf{U}} \right) \right| \tag{6.5}$$

  and thus we can rewrite Eq. (6.4) as

  $$p(\mathbf{U}) = p(\mathbf{X}) \left| \det \left(\frac{\partial \mathbf{f}^{\mathbf{-1}}}{\partial \mathbf{U}} \right) \right|^{-1}. \tag{6.6}$$

  Since we assumed $\mathbf{f}$ to be invertible, $\mathbf{f^{-1}}$ is well-defined.

- In practice, we will want $\mathbf{f}$ to be both invertible and to have a **tractable** Jacobian, i.e. a Jacobian that we can easily calculate. For $\mathbf{f}$ to have a Jacobian at all, each of its first-order partial derivatives must exist [**jacobi-matrix**]. So-called *autoregressive flows* have the property that their Jacobian is an upper triangular matrix. For an upper triangular matrix, it holds that its determinant is given by the product of its diagonal elements [**triangular-matrices**].

Definition 6.6 (Determinant)**.** Let D be an $n \times n$ matrix and let S_n denote the symmetric group over n. Then the determinant of D is defined as:

$$\det(D) := \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)} \right), \tag{6.7}$$

cf. [**leibniz-formula**].

Lemma 6.7. Let A be a $k \times k$, 0 an $k \times n$, C an $n \times k$ and D an $n \times n$ matrix; then

$$\det\left(\begin{pmatrix} A & 0 \\ C & D \end{pmatrix} \right) = \det(A) \det(D). \tag{6.8}$$

Proof. Define

$$B := \begin{pmatrix} A & 0 \\ C & D \end{pmatrix}. \tag{6.9}$$

Clearly,

$$b_{i,j} = \begin{cases} a_{i,j} & i,j \leq k, \\ 0 & i \leq k, j \geq k+1, \\ c_{i-k,j} & i \geq k+1, j \leq k, \\ d_{i-k,j-k} & i,j \geq k+1. \end{cases} \tag{6.10}$$

We can write the determinant of B as

$$\det(B) = \sum_{\sigma \in S_{n+k}} \operatorname{sgn}(\sigma) \prod_{i=1}^{n+k} b_{i,\sigma(i)}. \tag{6.11}$$

From Eq. (6.10) we know that all summands of the form $\sigma(i) = j$ with $i \leq k$, $j \geq k+1$ are 0. Therefore, we can consider all permutations of the form $\sigma(i) = j$ with $i, j \leq k$ or $\sigma(i) = j$ with $i \geq k+1$, $j \leq k$. We can also write this in the form $\sigma(i) = \pi(i)$ for $i \leq k$ and $\sigma(k+i) = k + \tau(i)$ for $1 \leq i \leq n$, where $\pi \in S_k$ and $\tau \in S_n$. Denote the set of all such permutations by $\tilde{S}_{k+n}$. Thus:

$$\begin{aligned}
\det(B) &= \sum_{\sigma \in \tilde{S}_{k+n}} \operatorname{sgn}(\sigma) \prod_{i=1}^{n+k} b_{i,\sigma(i)} && (6.12)\\
&= \sum_{\sigma \in \tilde{S}_{k+n}} \operatorname{sgn}(\sigma) \prod_{i=1}^{k} b_{i,\sigma(i)} \prod_{i=k+1}^{n+k} b_{i,\sigma(i)} && (6.13)\\
&\overset{(6.10)}{=} \sum_{\sigma \in \tilde{S}_{k+n}} \operatorname{sgn}(\sigma) \prod_{i=1}^{k} a_{i,\sigma(i)} \prod_{i=k+1}^{n+k} d_{i-k,\sigma(i)-k} && (6.14)\\
&= \sum_{\sigma \in \tilde{S}_{k+n}} \operatorname{sgn}(\sigma) \prod_{i=1}^{k} a_{i,\sigma(i)} \prod_{i=1}^{n} d_{i,\sigma(i+k)-k} && (6.15)\\
&= \sum_{\pi \in S_k, \tau \in S_n} \operatorname{sgn}(\pi)\operatorname{sgn}(\tau) \prod_{i=1}^{k} a_{i,\pi(i)} \prod_{i=1}^{n} d_{i,\tau(i)} && (6.16)\\
&= \sum_{\pi \in S_k} \operatorname{sgn}(\pi) \prod_{i=1}^{k} a_{i,\pi(i)} \sum_{\tau \in S_n} \operatorname{sgn}(\tau) \prod_{i=1}^{n} d_{i,\tau(i)} && (6.17)\\
&= \det(A)\det(D) && (6.18)
\end{aligned}$$

[**block-triangular-matrix**]

Definition 6.8. Let $h(\bullet\,;\theta) : \mathbb{R} \to \mathbb{R}$ be a bijection parametrized by θ. Then an *autoregressive model* is a function

$$g : \mathbb{R}^D \to \mathbb{R}^D, \begin{pmatrix} x_1 \\ \dots \\ x_D \end{pmatrix} \mapsto \begin{pmatrix} h\left(x_1; \Theta_1\right) \\ h\left(x_2; \Theta_2\left(x_1\right)\right) \\ \dots \\ h\left(x_D; \Theta_D\left(x_1, \dots, x_{D-1}\right)\right) \end{pmatrix} \tag{6.19}$$

The functions Θ_t for $t = 2, \dots, D$ are arbitrary functions whose domain is $\mathbb{R}^{t-1}$, Θ_1 is a constant.

Remark 6.9. The Jacobian matrix of an autoregressive flow is given as follows:

$$Dg = \begin{pmatrix} \partial g_1/\partial x_1 & \partial g_1/\partial x_2 & \dots & \partial g_1/\partial x_D \\ \partial g_2/\partial x_1 & \partial g_2/\partial x_2 & \dots & \partial g_2/\partial x_D \\ \vdots & \vdots & \ddots & \vdots \\ \partial g_D/\partial x_1 & \partial g_D/\partial x_2 & \dots & \partial g_D/\partial x_D \end{pmatrix} \tag{6.20}$$

One can easily convince oneself that Dg is a lower triangular matrix.

Theorem 6.10. Normalizing flows in $\mathbb{R}^D$ come from the push-forward of a measure.

Proof. Let $\mathcal{Y}$, $\mathcal{Z} \subset \mathbb{R}^D$ be open, $\Sigma_{\mathcal{Y}} = \mathcal{B}(\mathcal{Y})$, $\Sigma_{\mathcal{Z}} = \mathcal{B}(\mathcal{Z})$ and $g : \mathcal{Z} \to \mathcal{Y}$ be a diffeomorphism. The function $g : \mathcal{Z} \to \mathcal{Y}$ is measurable if and only if $g^{-1}(G) \in \Sigma_{\mathcal{Z}}$ for every set G that is open in $\mathcal{Y}$, cf. Theorem 1.4 (Borel σ-algebras are generated by the open sets). Since for functions between two topological spaces it holds that they are continuous if and only if the inverse image of an open set is again open, we have that g is indeed measurable.

Now define the probability measure μ on the measurable space $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$ as

$$\mu(I) := \int_I p_{\mathbf{Z}}(z) d\lambda(z) \ \forall I \in \Sigma_{\mathcal{Z}}, \tag{6.21}$$

where $p_{\mathbf{Z}} : \mathcal{Z} \to \mathbb{R}$ shall be a PDF and λ the Lebesgue measure. (The Lebesgue measure is defined on the completion of $\mathcal{B}(\mathbb{R}^D)$, which is "larger" than $\mathcal{B}(\mathbb{R}^D)$.) By considering the push-forward of μ under the measurable map g, we have $\forall J \in \Sigma_{\mathcal{Y}}$:

$$g_\star \mu\,(J) = \mu(g^{-1}(J)) \overset{(6.21)}{=} \int_{g^{-1}(J)} p_{\mathbf{Z}}(z) d\lambda(z) \tag{6.22}$$

Since by assumption $g : \mathcal{Z} \to \mathcal{Y}$ and therefore also the inverse $g^{-1} : \mathcal{Y} \to \mathcal{Z}$ are a diffeomorphism, we can use the change of variables formula from Theorem 6.2, since we assume $p_{\mathbf{Z}}$ to be integrable over $\mathcal{Z}$. We apply Theorem 6.2 to g^{-1} instead of g, cf. Remark 6.3:

$$g_\star \mu(J) = \int_{g^{-1}(J)} p_{\mathbf{Z}}(z) d\lambda(z) = \int_J \underbrace{\left(p_{\mathbf{Z}} \circ g^{-1}\right)(y) \cdot \left|\det Dg^{-1}(y)\right|}_{=p_{\mathbf{Y}}(y)} d\lambda(y) \tag{6.23}$$

■

Definition 6.11. A rational-quadratic function takes the form of a quotient of two quadratic polynomials.

$$\frac{\alpha^{(k)}(\xi)}{\beta^{(k)}(\xi)} = \frac{a_0\xi^2 + a_1\xi + a_2}{b_0\xi^2 + b_1\xi + b_2} \tag{6.24}$$

7 Diffusion-Based Models

Lemma 7.1 (ELBO)**.** Let $q(x_0)$ denote the true (unknown) distribution of a real image x_0, and let $p_\theta(x_0)$ be the model's approximation to $q(x_0)$, then we have the following ELBO-like loss:

$$\mathbb{E}_{q(x_0)}\left[\log p_\theta(x_0)\right] \geq -\mathbb{E}_{q(x_0,\dots,x_T)}\left[\log \frac{q(x_1,\dots,x_T \mid x_0)}{p_\theta(x_0,\dots,x_T)}\right]. \tag{7.1}$$

Proof. [**lilian˙weng**]

$$\log p_\theta(x_0) \geq \log p_\theta(x_0) - D_{\mathrm{KL}}\left[q(x_1, \ldots, x_T \mid x_0) \mid\mid p_\theta(x_1, \ldots, x_T \mid x_0)\right] \tag{7.2}$$

$$= \log p_\theta(x_0) - \mathbb{E}_{q(x_1,\ldots,x_T|x_0)}\left[\log \frac{q(x_1, \ldots, x_T \mid x_0)}{p_\theta(x_1, \ldots, x_T \mid x_0)}\right] \tag{7.3}$$

$$= \log p_\theta(x_0) - \mathbb{E}_{q(x_1,\ldots,x_T|x_0)}\left[\log \frac{q(x_1, \ldots, x_T \mid x_0)}{p_\theta(x_0, \ldots, x_T)/p_\theta(x_0)}\right] \tag{7.4}$$

$$= -\mathbb{E}_{q(x_1,\ldots,x_T|x_0)}\left[\log \frac{q(x_1, \ldots, x_T \mid x_0)}{p_\theta(x_0, \ldots, x_T)}\right] \tag{7.5}$$

$$\Rightarrow \mathbb{E}_{q(x_0)}\left[\log p_\theta(x_0)\right] \geq -\mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_1,\ldots,x_T|x_0)}\left[\log \frac{q(x_1, \ldots, x_T \mid x_0)}{p_\theta(x_0, \ldots, x_T)}\right] \tag{7.6}$$

Assuming that the assumptions of Fubini's theorem hold and from the monotonicity of the expectation, we have:

$$\mathbb{E}_{q(x_0)}\left[\log p_\theta(x_0)\right] \geq \mathbb{E}_{q(x_1,\ldots,x_T|x_0)q(x_0)}\left[\log \frac{q(x_1, \ldots, x_T \mid x_0)}{p_\theta(x_0, \ldots, x_T)}\right] = E_{q(x_0,\ldots,x_T|x_0)}\left[\log \frac{q(x_1, \ldots, x_T \mid x_0)}{p_\theta(x_0, \ldots, x_T)}\right]. \tag{7.7}$$

■

8 SDEs

Definition 8.1 (Wiener process)**.** Let W_t be a real-valued continuous-time stochastic process. It is said to be a *Wiener process* if the following properties hold:

- $W_0 = 0$,
- W has independent increments, i.e. $\forall t > 0$:, the terms $W_{t+u} - W_t$, $u \geq 0$ are independent of past values W_s, $s \leq t$,
- W has Gaussian increments: $W_{t+u} - W_t \sim \mathcal{N}(0, u)$,
- W has continuous paths, i.e. $\forall t$, W_t is continuous in t.

source: `https://en.wikipedia.org/wiki/Wiener_process`

Remark 8.2. If ξ_1, ξ_2, ... be i.i.d random variables with a mean of 0 and standard deviation of 1. For every n, define a continuous time stochastic process

$$W_n(t) := \frac{1}{\sqrt{n}} \sum_{1 \leq k \leq \lfloor nt \rfloor} \xi_k \tag{8.1}$$

This is what makes Wiener processes so powerful (and explains the ubiquity of Brownian motion). According to Donsker's theorem, the above expression becomes a Wiener process.

Definition 8.3 (Globally Lipschitz continuous)**.**

9 Miscellaneous

Lemma 9.1 (Chain rule for KL-divergences)**.** Let $p(x, y)$ and $q(x, y)$ be two arbitrary PDF's. Then the following hods:

$$D_{\mathrm{KL}}(p(x,y)||q(x,y)) = D_{\mathrm{KL}}(p(x)||q(x)) + D_{\mathrm{KL}}(p(y|x)||q(y|x)) \tag{9.1}$$

Proof. Brute-force calculation yields:

$$\begin{aligned}
D_{\mathrm{KL}}(p(x,y)||q(x,y)) &= \int\int p(x,y)\log\frac{p(x,y)}{q(x,y)}d\lambda(x)d\lambda(y) && (9.2)\\
&= \int\int p(x,y)\log\frac{p(x)p(y|x)}{q(x)q(y|x)}d\lambda(x)d\lambda(y) && (9.3)\\
&= \int\int p(x,y)\log\frac{p(x)}{q(x)}d\lambda(x)d\lambda(y) + \int\int p(x,y)\log\frac{p(y|x)}{q(y|x)}d\lambda(x)d\lambda(y) && (9.4)\\
&= D_{\mathrm{KL}}(p(x)||q(x)) + \int p(x)d\lambda(x)\int p(y|x)\log\frac{p(y|x)}{q(y|x)}d\lambda(y) && (9.5)\\
&= D_{\mathrm{KL}}(p(x)||q(x)) + D_{\mathrm{KL}}(p(y|x)||q(y|x)) && (9.6)
\end{aligned}$$

■

Definition 9.2 (Mutual information)**.** Let $(X, Y) \sim P_{(X,Y)} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is the space of all probability measures over the space $\mathcal{X} \times \mathcal{Y}$. The mutual information between the random variables X and Y is now defined as:

$$I(X;Y) := D_{\mathrm{KL}}\left(P_{(X,Y)}||P_X \otimes P_Y\right), \tag{9.7}$$

where P_X and P_Y are the marginal measures of the coupling measure $P_{(X,Y)}$ and $P_X \otimes P_Y$ is the induced **product measure**.

Reference: `https://people.ece.cornell.edu/zivg/ECE_5630_Lectures10.pdf`, Def. 10.1