Generative kernels

André Martins
atm@priberam.pt

December 9, 2006

Abstract

Kernel methods are a field of intensive research in machine learning. Lately, much attention has been dedicated to the problem of "kernel learning", i.e., choosing the kernel that best suits a particular task. Many discriminative approaches avoid handling this problem directly, ignoring the process of data generation to represent them as vectors in a suitable Euclidean space. By contrast, generative approaches propose devising kernels directly by properly modeling the generation of data; this way prior knowledge about their structure may be introduced. In this framework, objects are modeled as outcomes of random processes, for example HMMs for strings, or multinomials for text documents.

The aim of this report is providing a survey on generative kernels, capturing the essential theoretical aspects on which they settle, and highlighting the main geometrical ideas.

1 Introduction

Although the theoretical background of kernel methods has been established many decades ago, the support vector classifier, which is perhaps the first "conscient" application of kernels in a learning algorithm, was introduced only in the early 90s [Vap00]. After that, kernel methods were adopted for many other learning tasks besides classification, such as regression, principal component analysis, independent component analysis, etc. (see [SS02, STC04]). Their great popularity derives from the fact that a (positive definite) kernel corresponds to an inner product in some feature space. This allows extending the capacity of linear algorithms that depend only on pairwise inner products among data to a nonlinear framework, by replacing each inner product by a kernel evaluation.

Addressing a particular learning problem with a kernel-based approach requires choosing a kernel function that properly captures the similarity among data. The choice of the kernel should reflect our prior knowledge about how data is generated. While classic approaches usually focus on how to obtain easily computable kernels by representing data as vectors in

a suitable Euclidean space, thus ignoring how their generation is governed, this is sometimes a misleading perspective, specially for structured objects that don't naturally "live" in Euclidean spaces, such as strings or text documents. Although there has been some impressive results from nongenerative approaches specially devised to handle this sort of data, like convolutional kernels for discrete objects [Hau99], we concentrate here on generative approaches that prefer to model objects as outcomes of random processes, for example HMMs for strings, or multinomials for text documents. The aim of this report is to give a general overview on recent generative strategies to devise kernels, while clarifying the essential aspects of the background theory on which they settle.

In Section 2 we revise some basic notions of topology and measure theory that form the background for the subsequent sections. We also define some concepts from information theory that are later used, such as the Jensen-Shannon divergence. Section 3 characterizes positive and negative kernels, and highlights important relations between these two classes. We briefly mention semigroup kernels, that are intensively studied in [BCR84] and applied in [CV05, CFV05]. This section culminates with the proof that entropy is a negative definite function and that the Jensen-Shannon divergence is the square of a metric, something that only recently was taking into account by machine learners [Top00, Top02, ES03]. Section 4 aims to introduce an information geometrical perspective, inspired by [AN01, MR93]. We shall see that many generative kernels rely on intuitive reasonings underlying the geometry of probability distributions and measures. Section 5 illustrates all this by mentioning some examples of generative kernels that were recently devised, mainly for structured objects like images, text documents or strings. Finally, Section 6 concludes the paper.

2 Topological spaces and measures

We begin by revising the basic notions of topology and measure theory that are later used. Since we are going to deal with objects in structured domains that are not necessarily Euclidean, it makes sense to work in more general topological spaces. Specifically we are going to work with Radon measures in Hausdorff spaces; these are more general than the Lebesgue-Borel measure traditionally used in Euclidean spaces with the product topology.

2.1 Topological spaces

Let X be a set and $\mathscr{P}(X)$ the collection of its parts. A *topology* on X is a collection $\mathscr{T} \subseteq \mathscr{P}(X)$ that contains both $\varnothing$ and X, and that is closed under finite intersections and arbitrary unions; the members of $\mathscr{T}$ are called *open sets*. A set X together with its topology $\mathscr{T}$ is called a *topological space* and denoted $(X, \mathscr{T})$, or simply X when the underlying topology is clear from the

context. A set $S \subseteq X$ is called *closed* if its complement $S^c \equiv \complement_X S$ is open. Obviously $\varnothing$ and X are simultaneously open and closed for any topology.

Given a topological space $(X, \mathscr{T})$, a *basis* for the topology $\mathscr{T}$ is any family of sets $\{B_i\}_{i\in I}$ that generates $\mathscr{T}$ by taking finite intersections and arbitrary unions of its elements. Any subset $S \subseteq X$ becomes itself a topological space if we endow it with the *induced topology*, where $V \subseteq S$ is declared open if there is some $U \in \mathscr{T}$ such that $V = U \cap S$.

A map f between two topological spaces $(X, \mathscr{T}_X)$ and $(Y, \mathscr{T}_Y)$ is called *continuous* if the inverse image of any open set in Y is open in X, i.e., if $f^{-1}(V) \in \mathscr{T}_X$ for all $V \in \mathscr{T}_Y$. Since $f^{-1}(\complement_Y V) = \complement_X f^{-1}(V)$ we can replace "open" by "closed" in this definition. If $x \in X$, any open set U containing x is called a *neighborhood* of x.

A subset S of X is said to be *compact* if any open covering of S has a finite subcovering, i.e. if for any family of open sets $\{S_i\}_{i\in I}$ such that $S \subseteq \bigcup_{i\in I} S_i$ there exists a finite subfamily $\{S_{i_1}, \ldots, S_{i_n}\}$ such that $S \subseteq S_{i_1} \cup \ldots \cup S_{i_n}$. An important fact is that the image of a compact set under a continuous map is a compact set, i.e., *continuous maps preserve compactness.*

Example 2.1 (Discrete topology.) *Given a set X, declare any of its subsets to be open. This leads to the "discrete topology". The singletons of X form a basis for this topology. In this topology, a set is compact if and only if it is finite.*

Example 2.2 (Ordinary topology in $\mathbb{R}$.) *Let $X = \mathbb{R}$ and define a set S open if any point $x \in S$ belongs to an open interval contained in S (as usual). This leads to the "ordinary topology". Then, a set $C \subseteq X$ is compact if and only if it is closed and bounded.*

Example 2.3 (Banach spaces.) *Let V be a vector space over $\mathbb{C}$ (analogously, over $\mathbb{R}$). A* norm *in V is a function $\|.\| : V \to \mathbb{R}_+$ that satisfies, for all $\alpha \in \mathbb{C}$ and $u, v \in V$:*

- $\|u\| = 0$ *if and only if* $u = 0$,
- $\|\alpha u\| = |\alpha| \cdot \|x\|$,
- $\|u + v\| \leq \|u\| + \|v\|$.

If V is endowed with a norm, we may define a family of open balls *of the form*

$$B_\epsilon(u) = \{v \in V : \|v - u\| < \epsilon\}, \tag{2.1}$$

and this induces a topology, called the ordinary topology *in V: declare a set S open if any point in S is the center of some open ball contained in S.*

If V is complete *with respect to its norm, i.e., if any Cauchy sequence in V converges to an element of V, then it is called a* Banach space.

It turns out that a set $S \subseteq V$ is compact if and only if it is complete and covered by a finite number of balls of fixed radium[1].

Example 2.4 (Hilbert spaces.) *Let V be a vector space over $\mathbb{C}$. An* inner product *in V is a function $\langle .,. \rangle : V \times V \to \mathbb{C}$ satisfying for all $u, v, w \in V$ and all $\alpha, \beta \in \mathbb{C}$:*

- $\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle$,
- $\langle u, v \rangle = \overline{\langle v, u \rangle}$,
- $\langle u, u \rangle \geq 0$ *with equality if and only if* $u = 0$.

The first two properties mean that $\langle .,. \rangle$ is sesquilinear *(linear in the first argument and conjugate linear in the second). The analogous definition for vector spaces over $\mathbb{R}$ skips the conjugate sign, hence implying bilinearity.*

Any inner product induces a norm via $\|x\| \equiv \langle x, x \rangle^{1/2}$. Hence we can also define open balls and obtain the ordinary topology in V.

If V is complete with respect to the induced norm, it is called an Hilbert space*. Hence, Hilbert spaces are particular cases of Banach spaces.*

Example 2.5 (Metric spaces.) *A* metric space *is a set X endowed with a* metric*, i.e., a function $d : X \times X \to \mathbb{R}_+$ that satisfies, for all $x, y, z \in X$:*

- $d(x, y) = 0$ *if and only if* $x = y$.
- $d(x, y) = d(y, x)$ *for all* $x, y \in X$.
- $d(x, z) \leq d(x, y) + d(y, z)$ *for all* $x, y, z \in X$.

We may also define open balls in a metric space, through

$$B_\epsilon(x) = \{y \in X : d(x, y) < \epsilon\} \tag{2.2}$$

and obtain the ordinary topology in the usual way. Defining Cauchy sequences and completeness with respect to the metric allows characterizing compact sets in X analogously.

Any normed vector space is a metric space, defining $d(x, y) \equiv \|y - x\|$. Although metric spaces are more general than normed vector spaces, they do not introduce a great "degree of generality". In fact, every metric space can be embedded in a normed vector space in the following way [Lan93]: define, for each $x \in X$, the function $f_x : X \to \mathbb{R}$ by $f_x(y) = d(x, y)$. Notice that $d(x, y) = \|f_x - f_y\|_S$, where $\|.\|_S$ is the sup-norm, $\|f\|_S \equiv \sup_{x \in X} |f(x)|$. Fix an element $a \in X$ and define $g_x = f_x - f_a$. It turns out that the map $x \mapsto g_x$ is an isometry, i.e., a distance-preserving embedding of X into the normed vector space of bounded functions on X.

[1] An alternative characterization is: S is compact if and only if it has the Bolzano-Weierstrass property (every sequence in S has a point of accumulation in S).

Example 2.6 (Hausdorff spaces.) *An* Hausdorff *space is a topological space $(X, \mathscr{T})$ where any pair of distinct points can be separated by open sets, that is, for any $x, y \in X$ with $x \neq y$ there exist $U, V \in \mathscr{T}$ with $U \cap V = \varnothing$ such that $x \in U$ and $y \in V$ (see Figure 1).*

Hausdorff spaces generalize metric spaces: any metric space under the ordinary topology is Hausdorff. An important fact is that, if X is Hausdorff, then any compact subset $C \subseteq X$ is necessarily closed. In particular, any singleton is closed.

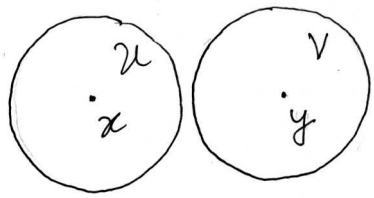


Figure 1: Hausdorff space: two points separated by open sets.

2.2 Measures

Let X be a set. A *σ-algebra* on X is a collection $\mathscr{M} \subseteq \mathscr{P}(X)$ that contains $\varnothing$ and that is closed under taking complements and countable unions (hence it is also closed under taking countable intersections); the members of $\mathscr{M}$ are called *measurable sets*, and $(X, \mathscr{M})$ is called a *measurable space*.

If X is endowed with a topology, a natural σ-algebra is the algebra $\mathscr{B}(X)$ of the Borel subsets of X, i.e., the algebra generated by the open subsets of X. An element of $\mathscr{B}(X)$ is accordingly called *Borel measurable*.

A *positive measure* on a measurable space $(X, \mathscr{M})$ is a map

$$\mu : \mathscr{M} \to [0, \infty] \tag{2.3}$$

(possibly taking infinite values) which is *countably additive*, i.e., such that $\mu(\varnothing) = 0$ and $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ for any sequence of mutually disjoint measurable sets $\{A_i\}_{i \in \mathbb{N}}$. A measurable space together with a measure μ is called a *measured space* and denoted $(X, \mathscr{M}, \mu)$. A positive measure defined on $\mathscr{B}(X)$ is called a *Borel measure*.

To define the *integral* in a measured space $(X, \mathscr{M}, \mu)$, we first consider *step functions* and then proceed to *μ-measurable* functions. A *step function* is a function $\varphi : X \to \mathbb{R}$ that is step with respect to some partition $\{A_1, \ldots, A_r\}$ of some set $A \subseteq X$ of finite measure. The *integral* of φ is then defined as $\int \varphi \, d\mu = \sum_{i=1}^{r} \mu(A_i)\varphi(A_i)$. A function $f : X \to \mathbb{R}$ is called *μ-measurable* if it is the pointwise limit of a sequence of step functions $\{\varphi_n\}_{n \in \mathbb{N}}$ *almost everywhere* (i.e. in any point of $X \setminus Z$ where Z is some set of null measure). In that case, the integral of f is defined as $\int f \, d\mu = \lim \int \varphi_n \, d\mu$.

The case $X = \mathbb{R}^n$ endowed with the Lebesgue-Borel measure corresponds to the Lebesgue integral.

We next proceed by introducing Radon measures on Hausdorff spaces, following [BCR84]. This framework is generic enough for our purposes while avoiding the "pathologies" that occur in the general measure theory in arbitrary sets.

Definition 2.7 (Radon measure.) *Let X be an Hausdorff space. A* Radon measure *on X is a Borel measure satisfying:*

- $\mu(C) < \infty$ *for each compact subset* $C \subseteq X$,
- $\mu(B) = \sup\{\mu(C) : C \subseteq B, C \text{ compact}\}$ *for each* $B \in \mathscr{B}(X)$.

We denote the set of all Radon measures on X by $M_+(X)$.

Example 2.8 (Finite Radon measures.) *A Radon measure is* finite *if $\mu(X) < \infty$; the set of finite Radon measures is denoted $M_+^b(X)$. By considering the second requirement of Definition 2.7 on B^c, we conclude that finite Radon measures satisfy*

$$\mu(B) = \inf\{\mu(U) : B \subseteq U, U \text{ open}\} \tag{2.4}$$

for each $B \in \mathscr{B}(X)$ (this is stated without proof in [BCR84]). In fact, we have that

$$\begin{aligned} \mu(B) &= \mu(X) - \mu(B^c) = \\ &= \mu(X) - \sup\{\mu(C) : C \subseteq B^c, C \text{ compact}\} = \\ &= \inf\{\mu(C^c) : B \subseteq C^c, C \text{ compact}\}. \end{aligned} \tag{2.5}$$

Since X is Hausdorff, the compactness of C implies that C^c is open, and

$$\mu(B) = \inf\{\mu(U) : B \subseteq U, U \text{ open}, U^c \text{ compact}\}. \tag{2.6}$$

Let's see that we may skip the "U^c compact" restriction, i.e., that

$$\bigvee_{\epsilon > 0} \quad \bigexists_{C \text{ compact}} \quad B \subseteq C^c \quad \text{and} \quad \mu(C^c) < \mu(U) + \epsilon. \tag{2.7}$$

In fact, by Def. 2.7 it exists $C \subseteq U^c$ compact such that $\mu(C) > \mu(U^c) - \epsilon$ for an arbitrarily small ϵ. Hence $\mu(C^c) = \mu(X) - \mu(C) < \mu(U) + \epsilon$, and $B \subseteq U \subseteq C^c$ as wanted.

Definition 2.9 (Molecular measures.) *The* support *of a Radon measure μ on X is defined as*

$$\operatorname{supp}(\mu) = \{x \in X : \mu(U) > 0 \text{ for each neighborhood } U \text{ of } x\}. \tag{2.8}$$

Radon measures with a finite support are called molecular measures*; the set of all molecular measures on X is denoted* $\mathrm{Mol}_+(X)$.

Example 2.10 *The Lebesgue measure on the Euclidean space $\mathbb{R}^n$ equipped with the ordinary topology is a Radon measure with support $\mathbb{R}^n$. In fact, although any countable set has zero measure, any neighborhood of a point in $\mathbb{R}^n$ has strictly positive measure.*

Example 2.11 *Let X be an Hausdorff space equipped with the discrete topology, and let $x \in X$. The Dirac measure $\varepsilon_x : \mathscr{B}(X) \to [0, \infty]$ is defined as*

$$\varepsilon_x(A) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A. \end{cases} \tag{2.9}$$

ε_x is a Radon measure and $\operatorname{supp}(\varepsilon_x) = \{x\}$. Thus any Dirac measure is a molecular measure. Conversely, it can be shown that any molecular measure is a finite convex combination of Dirac measures (this is omitted in [BCR84]): let $\mu \in \operatorname{Mol}_+(X)$ have support $\{x_1, \ldots, x_n\}$ for some $n \in \mathbb{N}$. We first prove that any set that does not intersect $\operatorname{supp}(\mu)$ has null measure. From the definition of Radon measure, it suffices to prove this for any compact set C. Since $C \cap \operatorname{supp}(\mu) = \varnothing$, any $y \in C$ must admit a neighborhood U_y with null measure. The family of neighborhoods $(U_y)_{y \in C}$ covers C, hence from the compactness of C there is a finite subcovering $(U_{y_i})_{i=1,\ldots,m}$ implying that $\mu(C) \leq \sum_{i=1}^m \mu(U_{y_i}) = 0$, as we wanted to prove. Now, let $a_i = \mu(\{x_i\})$ for $i = 1, \ldots, n$, which are finite numbers since the sets $\{x_i\}$ are compact. Then, for any set A, we have

$$\begin{aligned} \mu(A) &= \mu(A \cap \operatorname{supp}(\mu)) + \mu(A \cap (\operatorname{supp}(\mu))^c) = \\ &= \mu(A \cap \operatorname{supp}(\mu)) = \\ &= \sum_{i=1}^n a_i \varepsilon_{x_i}(A). \end{aligned} \tag{2.10}$$

A *Radon probability measure* on X is a Radon measure $\mu \in M_+(X)$ such that $\mu(X) = 1$. The set of Radon probability measures on X is denoted $M_+^1(X)$; it can be seen as the set of equivalence classes of finite Radon measures by the equivalence relation in $M_+^b(X)$

$$\mu_1 \equiv \mu_2 \quad \text{if there exists } \lambda \in \mathbb{R} \text{ such that } \mu_1 = \lambda \mu_2. \tag{2.11}$$

which means that any finite Radon measure can be normalized to give rise to a Radon probability measure.

Consider a measured space $(X, \mathscr{M}, \nu)$ where X is Hausdorff, and ν is *σ-finite*, i.e., is such that X can be written as the countable union of sets of finite measure[2]. We say that a measure μ is *ν-absolutely continuous* (denoted $\mu \ll \nu$) if μ vanishes wherever ν does, i.e., if $\nu(A) = 0$ implies $\mu(A) = 0$ for any $A \in \mathscr{M}$. If this happens, the Radon-Nikodym theorem

[2]E.g. the union of closed unit intervals in $\mathbb{R}$ with the Lebesgue-Borel measure.

guarantees the existence (and uniqueness up to equivalence within measure zero) of a *density function* $f : X \to \mathbb{R}_+$ such that, for all $A \in \mathscr{M}$, we have $\mu(A) = \int_A f d\nu$.[3] This density is called the *Radon-Nikodym derivative* and denoted $f = \frac{d\mu}{d\nu}$, since

$$\mu(A) = \int_A d\mu = \int_A f d\nu. \tag{2.12}$$

If $g : X \to \mathbb{R}$ is an arbitrary function, we define its $L^1(\nu)$-norm

$$\|g\|_1 \equiv \int_X |g| \, d\nu = \int_X |g(x)| \, d\nu(x) \tag{2.13}$$

whenever the integral exists. If g is a density, we have $\|g\|_1 = \int_X g d\nu$, and if in particular g is the density of a probability measure, then $\|g\|_1 = 1$.

Suppose now that X is a countable set. In this particular case, the *counting measure* (the measure that assigns to each finite set its number of elements, and ∞ to each infinite set[4]) becomes σ-finite, and may be taken as the dominating measure ν. Then, $\int_X f d\nu = \sum_{x \in X} f(x)$, and we may replace integrals by sums. If instead $X \subseteq \mathbb{R}^n$, we may take ν to be the usual Lebesgue-Borel measure, resulting then $\int_X g d\nu = \int_X g(x) dx$. So, the usage of integrals with respect to a dominating measure provides us with a unified framework for dealing with measures in both the "discrete" and "continuous" scenarios.

2.3 Entropy and divergence measures

Let $(X, \mathscr{M}, \nu)$ be a measured space where X is Hausdorff and ν is a σ-finite Radon measure. Let $M_+^h(X) \subseteq M_+^b(X)$ denote the set of finite Radon ν-absolutely continuous measures whose density $f : X \to \mathbb{R}_+$ satisfies $\|f \cdot \log f\|_1 < \infty$. Denote by $\frac{d}{d\nu} M_+^h(X)$ the set of densities[5] of those measures. The *entropy function* $h : \frac{d}{d\nu} M_+^h(X) \to \mathbb{R}$ is defined by

$$h(f) = -\int_X f \log f d\nu, \tag{2.14}$$

where $0 \log 0 = 0$ by convention.

Remark 2.12 *This definition of entropy generalizes the traditional notions of discrete and differential entropies presented for example in [CT91]. Denote by $M_+^{1,h}(X) = M_+^h(X) \cap M_+^1(X)$ the set of Radon probability measures with finite entropy. If $X \subseteq \mathbb{R}^n$, ν is the Lebesgue-Borel measure, and*

[3] The σ-finiteness of ν is indeed necessary. For a counterexample take ν to be the counting measure on $\mathbb{R}$ and μ the Lebesgue-Borel measure; the existence of a density f would imply $f = 0$ and as consequence $\mu = 0$.

[4] When X is countable it may be seen as the sum of all Dirac measures on X, $\sum_{x \in X} \varepsilon_x$.

[5] More exactly, the set of equivalence classes of densities that are equal almost everywhere.

$P \in M_+^{1,h}(X)$ is a probability measure with density $p = \frac{dP}{d\nu}$, then $h(p)$ reduces to the differential entropy

$$h(p) = -\int_X p(x) \log p(x) dx. \tag{2.15}$$

If, instead, X is a countable set, ν is the counting measure, and $P \in M_+^{1,h}(X)$ is a probability measure with probability mass function $x \mapsto p(x) = P(\{x\})$, then $h(p) \equiv H(p)$ is the discrete entropy

$$H(p) = -\sum_{x \in X} p(x) \log p(x). \tag{2.16}$$

Let f and g be respectively the densities (with respect to the dominating measure ν) of measures μ_f and μ_g in $M_+^h(X)$, such that μ_f is μ_g-absolutely continuous (i.e. $\mu_f \ll \mu_g \ll \nu$). The *Kullback-Leibler divergence* between f and g is defined by

$$\begin{aligned} D(f\|g) &= \int_X f \log \frac{f}{g} d\nu = \\ &= -h(f) - \int_X f \log g d\nu. \end{aligned} \tag{2.17}$$

If f and g are probability densities, the Kullback-Leibler divergence can be seen as a dissimilarity measure between the two distributions. It verifies $D(f\|g) = 0$ if and only if $f = g$ almost everywhere. However, it is not a metric (it is not symmetric and it does not satisfy the triangle inequality).

It is clear that $M_+(X)$ and $M_+^b(X)$ are convex cones, and that $M_+^1(X)$ is a convex set. By linearity of the integral, so are the respective sets of densities. So we can talk about "mixtures of densities". These may be characterized by the following divergence measure:

Definition 2.13 *Let $f_1, \ldots, f_n$ be densities of measures in $M_+^h(X)$, and $f = \alpha_1 f_1 + \ldots + \alpha_n f_n$ a mixture defined by coefficients $\alpha_1, \ldots, \alpha_n \in \mathbb{R}_+$. The* generalized Jensen-Shannon divergence *of $f_1, \ldots, f_n$ with respect to that mixture is defined by*

$$J(f_1, \ldots, f_n; \alpha_1, \ldots, \alpha_n) \equiv h\left(\sum_{i=1}^n \alpha_i f_i\right) - \sum_{i=1}^n \alpha_i h(f_i), \tag{2.18}$$

The restriction of J to probability densities is defined analogously requiring $\sum_{i=1}^n \alpha_i = 1$. The particular case where $n = 2$ and $\alpha_1 = \alpha_2 = \frac{1}{2}$ is simply called Jensen-Shannon divergence *between f and g and denoted $J(f\|g)$:*

$$J(f\|g) \equiv h\left(\frac{f+g}{2}\right) - \frac{h(f) + h(g)}{2}. \tag{2.19}$$

It is straightforward that the Jensen-Shannon divergence relates to the Kullback-Leibler divergence via

$$J(f\|g) = \frac{1}{2}D\left(f\middle\|\frac{f+g}{2}\right) + \frac{1}{2}D\left(g\middle\|\frac{f+g}{2}\right). \tag{2.20}$$

From (2.20), one can see that J is symmetric and inherits from the Kullback-Leibler divergence the property that $J(f\|g) = 0$ if and only if $f = g$ almost everywhere. A remarkable fact that will be shown later is that $\sqrt{J}$ satisfies the triangle inequality, i.e., the Jensen-Shannon divergence is actually the square of a metric.

2.4 The Jensen-Shannon divergence as mutual information

In [GBGC$^+$02] several interpretations of the Jensen-Shannon divergence are given, namely in the fields of statistical physics, information theory and mathematical statistics. We reproduce here the interpretation concerning information theory.

Let $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$ be a finite alphabet and $t_1, \ldots, t_m$ be m strings emitted by different sources, with lengths $|t_1|, \ldots, |t_m|$. Consider the n-length string $s = t_1 \ldots t_m$ formed by concatenating the m strings, where $n = \sum_{i=1}^{m} |t_i|$. Suppose that we choose a random position $1 \leq p \leq n$ with a uniform probability distribution and define the random variables

$$\begin{aligned} \sigma &\equiv \text{“the symbol at position } p\text{”, } \sigma = s[p] \\ t &\equiv \text{“the string (among } t_1, \ldots, t_m\text{) corresponding to position } p\text{”} \end{aligned}$$

Let $p(\sigma_i) = \Pr\{\sigma = \sigma_i\}$ and $p(t_j) = \Pr\{t = t_j\} = \frac{|t_j|}{n}$, and denote the corresponding conditional and joint probabilities by $p(\sigma_i|t_j)$ and $p(\sigma_i, t_j)$. A typical question in information theory is: “How much information I can we obtain from learning the identity of the symbol σ about the identity of the substring t, provided the probability distribution $p(\sigma_i, t_j)$?”. I is called the *mutual information in σ about t* and is defined by

$$I \equiv \sum_{i=1}^{k}\sum_{j=1}^{m} p(\sigma_i, t_j) \log \frac{p(\sigma_i, t_j)}{p(\sigma_i)p(t_j)} \tag{2.21}$$

From Bayes' rule, $p(\sigma_i, t_j) = p(\sigma_i|t_j)p(t_j)$, and (2.21) may be rewritten as

$$\begin{aligned} I &= \sum_{i=1}^{k}\sum_{j=1}^{m} p(\sigma_i|t_j)p(t_j)\log\frac{p(\sigma_i|t_j)}{p(\sigma_i)} = \\ &= \sum_{j=1}^{m} p(t_j)\sum_{i=1}^{k} p(\sigma_i|t_j)\log p(\sigma_i|t_j) - \sum_{i=1}^{k}\left(\sum_{j=1}^{m} p(t_j)p(\sigma_i|t_j)\right)\log p(\sigma_i) = \\ &= -\sum_{j=1}^{m} p(t_j)H(p(.|t_j)) + H\left(\sum_{j=1}^{m} p(t_j)p(\sigma_i|t_j)\right) = \\ &= J\left(p(.|t_1),\ldots,p(.|t_m);p(t_1),\ldots,p(t_m)\right), \end{aligned} \tag{2.22}$$

i.e., the Jensen-Shannon divergence among the probability mass functions of σ conditioned on each substring, with mixture coefficients the substring probabilities $p(t_i) = \frac{|t_i|}{n}$, equals the mutual information in σ about t. If, as an extreme example, all these probability mass functions are equal, $p(.|t^1) = \ldots = p(.|t^m)$, then the Jensen-Shannon divergence vanishes, meaning that knowing the identity of σ does not give us any information about the substring where it was picked from.

3 Positive and negative definite kernels

In the last years, the machine learning community have dedicated great attention to kernel methods [SS02, STC04]. The ability of these methods to *represent* "nonlinearities" as "linearities in a feature space" enables extending the classic linear algorithms for classification, regression, and other learning tasks, to nonlinear scenarios. Perhaps support vector machines [Vap00] is the first historical example (in early 90s) where kernel methods were applied in the field of statistical learning theory. More recently, it was shown that for many learning tasks one could use a larger class of kernels rather than the usual "positive kernels". It happens that there are many results concerning relations between classes of kernels, representation in feature spaces, connections with harmonic analysis, etc., that are known for many time but only recently began to be used in machine learning. We now give a basic review of some of those results, following [BCR84].

3.1 Definition and properties

In order not to loose generality, we consider all functions to be complex-valued, unless otherwise stated. So, in what follows, if z is a complex number, we denote its conjugate by $\overline{z}$. Also, $z \geq 0$ means $\mathrm{Re}(z) \geq 0$ and $\mathrm{Im}(z) = 0$. If A is a matrix, A^* denotes its conjugate transpose.

A *positive semidefinite matrix* is a hermitian matrix K that, for any choice of a column vector x, satisfies $x^*Kx \geq 0$. A widely known fact is that positive semidefinite matrices have real and nonnegative eigenvalues. A squared matrix may be seen as a function defined on $I \times I$, where I is the finite set of indices. The following is a generalization of this concept to functions whose domain $X \times X$ is not necessarily finite.

In what follows, X is a nonempty set.

Definition 3.1 (Positive definite kernel.) *A function $\varphi : X \times X \to \mathbb{C}$ is called a* positive definite kernel *if and only if*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i\overline{c_j}\varphi(x_i, x_j) \geq 0 \tag{3.1}$$

for all $n \in \mathbb{N}$, $\{x_i, \ldots, x_n\} \subseteq X$ and $\{c_i, \ldots, c_n\} \subseteq \mathbb{C}$. If, for any distinct $x_1, \ldots, x_n$, the equality in (3.1) *implies $c_1 = \ldots = c_n = 0$, then the kernel φ is called* strictly *positive definite.*

Definition 3.2 (Negative definite kernel.) *A function $\psi : X \times X \to \mathbb{C}$ is called a* negative definite kernel *if and only if:*

- *ψ is* hermitian, *i.e., $\psi(y, x) = \overline{\psi(x, y)}$ for all $x, y \in X$;*
- *For all $n \in \mathbb{N}$, $\{x_i, \ldots, x_n\} \subseteq X$ and $\{c_i, \ldots, c_n\} \subseteq \mathbb{C}$ with $\sum_{i=1}^{n} c_i = 0$, it holds*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i\overline{c_j}\psi(x_i, x_j) \leq 0 \tag{3.2}$$

If, for any distinct $x_1, \ldots, x_n$, the equality in (3.2) *implies $c_1 = \ldots = c_n = 0$, then the kernel ψ is called* strictly *negative definite. If ψ is (strictly) negative definite, we call $-\psi$ (strictly)* conditionally positive definite.[6]

Remark 3.3 *The analogous of Def. 3.2 for a real-valued ψ is obtained simply by replacing $\{c_i, \ldots, c_n\} \subseteq \mathbb{C}$ by $\{c_i, \ldots, c_n\} \subseteq \mathbb{R}$, removing the conjugate sign and replacing the word "hermitian" by "symmetric".*

In the case of Def 3.1, however, the analogous for real-valued functions has to additionally require φ to be symmetric. In fact, in the general complex-valued case, it follows from the definition that any positive definite kernel is hermitian; however, there are nonsymmetric real-valued functions satisfying (3.1) *for any $\{c_i, \ldots, c_n\} \subseteq \mathbb{R}$ which, of course, are not positive definite kernels.*

[6] [SS02] goes further by mentioning the more general "conditionally positive definite kernels of order k."

Immediate consequences of Defs. 3.1 and 3.2 are that φ is positive (resp. negative) definite if and only if for any finite subset $F \subseteq X$ the restriction $\varphi|_{F\times F}$ is positive (resp. negative) definite. In particular, any positive definite kernel φ is nonnegative on the diagonal $\Delta \equiv \{(x,x) \in X \times X\}$, i.e. $\varphi|_\Delta \geq 0$, as can be seen by setting $F = \{x\}$ for all $x \in X$.

We now list some simple properties of positive and negative definite kernels.

Property 3.4 *If φ is positive definite, then for all $x, y \in X$:*

$$|\varphi(x,y)|^2 \leq \varphi(x,x) \cdot \varphi(y,y) \tag{3.3}$$

Property 3.5 *If ψ is negative definite, then for all $x, y \in X$:*

$$\psi(x,x) + \psi(y,y) \leq 2\,\mathrm{Re}\,\psi(x,y) \tag{3.4}$$

Property 3.6 *Any φ of the form $\varphi(x,y) = f(x) \cdot \overline{f(y)}$, where $f : X \to \mathbb{C}$ is an arbitrary function, is positive definite. In particular, a constant kernel $(x,y) \mapsto c$ is positive definite if and only if $c \geq 0$.*

Property 3.7 *Any ψ of the form $\psi(x,y) = f(x) + \overline{f(y)}$, where $f : X \to \mathbb{C}$ is an arbitrary function, is negative definite. In particular, a constant kernel $(x,y) \mapsto c$ is negative definite if and only if $c \in \mathbb{R}$.*

Proof: The first two properties are easily shown by taking the 2×2 hermitian matrix of the kernel restriction to $\{x, y\}$. The first follows from the nonnegativity of the determinant, and the second from the definition of negative definiteness setting $c_1 = 1$ and $c_2 = -1$. The third follows from

$$\sum_{i,j=1}^{n} c_i \overline{c_j} \varphi(x_i, x_j) = \left| \sum_{i=1}^{n} c_i f(x_i) \right|^2 \geq 0, \tag{3.5}$$

and the fourth from the fact that, if $\sum_{i=1}^{n} c_i = 0$, we have

$$\sum_{i,j=1}^{n} c_i \overline{c_j} \left(f(x_i) + \overline{f(x_j)} \right) = \sum_{i=1}^{n} c_i f(x_i) \sum_{j=1}^{n} \overline{c_j} + \sum_{j=1}^{n} \overline{c_j}\, \overline{f(x_j)} \sum_{i=1}^{n} c_i = 0, \tag{3.6}$$

even with equality. □

In what follows, $\mathcal{K}_+(X)$ and $\mathcal{K}_-(X)$ denote respectively the sets of positive and negative definite kernels defined on $X \times X$. Their strict counterparts will be accordingly denoted $\mathcal{K}_{++}(X)$ and $\mathcal{K}_{--}(X)$.

Property 3.8 *$\mathcal{K}_+(X)$ and $\mathcal{K}_-(X)$ are both convex cones, closed in the topology of pointwise convergence.*

This is easily verifiable from the definition by using the properties of sums. It means that if φ_1 and φ_2 are positive (resp. negative) definite, so is $\lambda_1\varphi_1 + \lambda_2\varphi_2$ for any nonnegative scalars λ_1 and λ_2, and that if $(\varphi_n)_{n\in\mathbb{N}}$ is a sequence of positive (resp. negative) definite kernels converging pointwise to φ, then φ is positive (resp. negative) definite. Regarding integrals as limits of weighted sums, it also implies that $\mathcal{K}_+(X)$ and $\mathcal{K}_-(X)$ are closed under pointwise integration, i.e., if $(\varphi_\theta)_{\theta\in\Theta}$ is a family of positive (resp. negative) definite kernels and μ is a positive measure on Θ such that $\varphi_\theta(x,y)$ is μ-integrable for all $x, y \in X$, then $\varphi : X \times X \to \mathbb{C}$ defined by

$$\varphi(x,y) = \int_\Theta \varphi_\theta(x,y) d\mu(\theta) \tag{3.7}$$

is positive (resp. negative) definite.

The following property goes further by stating that $\mathcal{K}_+(X)$ is closed under taking products.

Property 3.9 *If φ_1 and φ_2 are positive definite, so is $\varphi_1 \cdot \varphi_2$.*

Proof: We follow the proof given in [STC04]. It suffices to show that the Schur product (i.e. the elementwise product) $A \odot B$ of two positive semidefinite $n \times n$ matrices A and B is a positive semidefinite matrix (this was proved by Schur in 1911). Consider first the tensor product $A \otimes B$, which results in a positive semidefinite $n^2 \times n^2$ matrix whose eigenvalues are the products of the pairs of eigenvalues of A and B, containing $A \odot B$ as a principal submatrix (i.e. $[A \odot B]_{ij} = [A \otimes B]_{k_i k_j}$ where $k_1, \ldots, k_n$ is some set of n indices from $1, \ldots, n^2$). Hence for any $c = (c_1, \ldots, c_n)$ there is a $c' = (c'_1, \ldots, c'_{n^2})$ filled with zeros except for the entries $k_1, \ldots, k_n$, where $c'_{k_i} = c_i$. It follows that $c^*(A \odot B)c = c'^*(A \otimes B)c' \geq 0$. □

The two latter results imply that the polynomial and Gaussian kernels (widely used in machine learning) are actually positive definite kernels:

Property 3.10 *Let φ be a positive definite kernel. Any polynomial combination with nonnegative coefficients, $\sum_{i=0}^n \lambda_i \varphi^i$ with each $\lambda_i \geq 0$, is positive definite. Furthermore, if $|\varphi(x,y)| < \rho \leq \infty$, and $f : \mathbb{C} \to \mathbb{C}$ is a holomorphic function in $\{z \in \mathbb{C} : |z| < \rho\}$, $f(z) = \sum_{n=0}^\infty a_n z^n$, where each $a_n \geq 0$, then $f \circ \varphi$ is positive definite. In particular,* $\exp(\varphi)$ *is positive definite.*

3.2 Relations between positive and negative definite kernels

We next list some properties that relate positive and negative definite kernels. To get a picture, let us first say that if V is a vector space over $\mathbb{C}$, it is easy to show that every inner product in V is a positive definite kernel, and any squared distance induced by an inner product is a negative definite kernel. Later we will see that both positive and negative kernels are representable more or less like inner products and squared distances.

A trivial result is that positive definiteness implies conditional positive definiteness, i.e., if $\varphi \in \mathcal{K}_+$ then $-\varphi \in \mathcal{K}_-$. The following properties, whose proofs can be found in [BCR84] and are mainly due to Schoenberg, establish equivalence conditions.

Property 3.11 *Let $\psi : X \times X \to \mathbb{C}$ be an hermitian function, and $x_0 \in X$. Define $\varphi_0, \varphi : X \times X \to \mathbb{C}$ by:*

$$\varphi_0(x, y) = \psi(x, x_0) + \overline{\psi(y, x_0)} - \psi(x, y) \tag{3.8}$$

and

$$\varphi(x, y) = \varphi_0(x, y) - \psi(x_0, x_0) \tag{3.9}$$

Then:

- *φ is positive definite if and only if ψ is negative definite,*
- *If $\psi(x_0, x_0) \geq 0$, then φ_0 is positive definite if and only if ψ is negative definite.*

Property 3.12 *The kernel $\psi : X \times X \to \mathbb{C}$ is negative definite if and only if $\exp(-t\psi)$ is positive definite for all $t > 0$.*

Proof: We first prove the direction '$\Rightarrow$'. Let $\psi \in \mathcal{K}_-(X)$. Since negative definite kernels form a cone, we only need to prove the positive definiteness of $\exp(-t\psi)$ for $t = 1$. Using Property 3.11 and exponentiating both sides of (3.9) we are led to

$$\exp(-\psi(x, y)) = \exp(\varphi(x, y)) \cdot \exp(-\psi(x, x_0)) \cdot \overline{\exp(-\psi(y, x_0))} \cdot \exp(\psi(x_0, x_0))$$

where $\varphi \in \mathcal{K}_+(X)$. Hence, by properties 3.6, 3.9 and 3.10 we conclude that $\exp(-\psi) \in \mathcal{K}_+(X)$. To prove '$\Leftarrow$' we use Property 3.8. If $\exp(-t\psi) \in \mathcal{K}_+(X)$, then $1 - \exp(-t\psi)$ is negative definite, and so is the pointwise limit

$$\psi = \lim_{t \to 0^+} \frac{1}{t}(1 - \exp(-t\psi)). \tag{3.10}$$

□

The following property (whose proof can be found in [BCR84]) uses the result expressed in (3.7), and gives a first clue about how harmonic analysis may be important in the study of positive or negative definiteness. We then give two examples.

Property 3.13 *Let μ be a probability measure on $\mathbb{R}_+$ with positive finite first moment, i.e. such that $0 < \int_0^\infty s d\mu(s) < \infty$, and let $\mathscr{L}\mu$ denote its Laplace transform, i.e.*

$$(\mathscr{L}\mu)(z) = \int_0^\infty \exp(-sz) d\mu(s) \tag{3.11}$$

for $z \in \mathbb{C}_+ = \{z \in \mathbb{C} : \operatorname{Re} z \geq 0\}$ (the complex right half-plane). Then $\psi : X \times X \to \mathbb{C}_+$ is negative definite if and only if $(\mathscr{L}\mu)(t\psi)$ is positive definite for all $t > 0$.

Example 3.14 *The Dirac measure ε_1 has first moment $\int_0^\infty s d\varepsilon_1(s) = 1$. Its Laplace transform is*

$$(\mathscr{L}\varepsilon_1)(z) = \int_0^\infty \exp(-sz) d\varepsilon_1(s) = \exp(-z). \tag{3.12}$$

Hence we have that a kernel $\psi : X \times X \to \mathbb{C}_+$ is negative definite if and only if the composition $(\mathscr{L}\varepsilon_1)(t\psi) = \exp(-t\psi)$ is positive definite for all $t > 0$, which is a weaker version of Property 3.12 for a $\mathbb{C}_+$-valued ψ.

Example 3.15 *Using instead the measure $\mu = \exp(-t)dt$, whose first moment $\int_0^\infty s \exp(-s)ds = 1$ is also positive and finite, we get*

$$\begin{aligned}(\mathscr{L}\mu)(z) &= \int_0^\infty \exp(-sz)\exp(-s)ds = \int_0^\infty \exp(-s(z+1))ds = \\ &= \frac{1}{z+1}.\end{aligned} \tag{3.13}$$

Hence we have that a kernel $\psi : X \times X \to \mathbb{C}_+$ is negative definite if and only if the composition $(\mathscr{L}\mu)(t\psi) = (1 + t\psi)^{-1}$ is positive definite for all $t > 0$. Replacing $t' = \frac{1}{t}$ ($t' > 0 \Leftrightarrow t > 0$) we have that this is equivalent to the positive definiteness of $t' \cdot (t' + \psi)^{-1}$ and hence of $(t' + \psi)^{-1}$. So we may restate this as:

Property 3.16 *The kernel $\psi : X \times X \to \mathbb{C}_+$ is negative definite if and only if $(t + \psi)^{-1}$ is positive definite for all $t > 0$.*

We now present an important class of positive definite kernels, the infinitely divisibles, that we denote by $\mathcal{K}_+^d(X)$.

Definition 3.17 *A positive definite kernel φ is called* infinitely divisible *if for each $n \in \mathbb{N}$ there exists a positive definite kernel φ_n that is a n-th root of φ, i.e., such that $\varphi = (\varphi_n)^n$.*

From Property 3.12 it is clear that if ψ is negative definite then $\varphi = \exp(-\psi)$ is infinitely divisible. The following property goes further and shows that any $\mathbb{R}_{++}$-valued infinitely divisible kernel has this form.

Property 3.18 *Let φ be a positive definite kernel with values in $\mathbb{R}_+$. Then:*

- *$\varphi \in \mathcal{K}_+^d(X)$ if and only if $\varphi^t \in \mathcal{K}_+(X)$ for all $t > 0$.*
- *If φ is $\mathbb{R}_{++}$-valued, then $\varphi \in \mathcal{K}_+^d(X)$ if and only if $-\log \varphi \in \mathcal{K}_-(X)$.*

The following property allows to derive some important negative definite kernels.

Property 3.19 *Let $\psi : X \times X \to \mathbb{C}$ be a negative definite kernel, and let $\mu \in M_+(\mathbb{R}_{++})$. Consider the function $g : D(\mu) \to \mathbb{C}$ defined by*

$$g(z) = \int_0^\infty (1 - \exp(-\lambda z))d\mu(\lambda), \tag{3.14}$$

where $D(\mu)$ is the set of $z \in \mathbb{C}$ for which $\lambda \mapsto (1-\exp(-\lambda z))$ is μ-integrable. Then:

- *If $\psi(X \times X) \subseteq D(\mu)$, then $g \circ \psi$ is negative definite;*
- *A sufficient condition for $\psi(X \times X) \subseteq D(\mu)$ is*

$$\psi|_\Delta \geq 0 \quad \textit{and} \quad \int_0^\infty \lambda(1+\lambda)^{-1} d\mu(\lambda) < \infty. \tag{3.15}$$

Proof: The first statement is a consequence of the negative definiteness of $1 - \exp(-\lambda\psi)$ for any $\lambda > 0$ and the "pointwise integration closure" of $\mathcal{K}_-(X)$ expressed in (3.7). The remaining proof is in [BCR84]. □

Using the formulas

$$z^\alpha = \frac{\alpha}{\Gamma(1-\alpha)} \int_0^\infty (1 - \exp(-\lambda z)) \frac{d\lambda}{\lambda^{\alpha+1}} \tag{3.16}$$

and

$$\log(1+z) = \int_0^\infty (1 - \exp(-\lambda z)) \frac{\exp(-\lambda)}{\lambda} d\lambda \tag{3.17}$$

where $\Gamma : \mathbb{C}_+ \to \mathbb{C}$ denotes the Gamma function

$$\Gamma(s) = \int_0^\infty t^{s-1} \exp(-t) dt \tag{3.18}$$

that are valid for $\operatorname{Re} z \geq 0$ and easily established by deriving both sides of the equations, we have the following very important result as a consequence of Property 3.19.

Property 3.20 *If ψ is negative definite and satisfies $\psi|_\Delta \geq 0$ then so are ψ^α for $0 < \alpha < 1$ and $\log(1+\psi)$.*

Integrating both sides of 3.16 we can also conclude (see [BCR84]) that

Property 3.21 *If $f : X \to \mathbb{C}$ satisfies $\operatorname{Re} f \geq 0$ then for each $\alpha \in [1, 2]$ the kernel $\psi_\alpha(x, y) = -(f(x) + \overline{f(y)})^\alpha$ is negative definite.*

We will see soon how the two latter properties are used to deduce the negative definiteness of the entropy function.

3.3 Hilbert space representation

We are now going to see that any positive or negative definite kernel is representable respectively as an *inner product* or a *squared distance* induced from the inner product in a Hilbert space H (usually called the *feature space* by machine learners) via a *feature map* $\Phi : X \to H$ that maps each data point $x \in X$ to its feature representation $\Phi(x)$. This is the great motivation for using kernel methods in learning, since it allows to generalize algorithms that depend only on distances among data (for example, translation-independent algorithms), or to build nonlinear versions of algorithms that depend uniquely on mutual inner products, by mapping the data (that in general lies in a unstructured input set) to a well-structured Hilbert space. This Hilbert space needs not be unique, and neither do we bother to find the most "nice" Hilbert space in a general situation. The idea is never to perform direct computations in H, which has often a very high dimension, but instead use the kernel function in X to compute inner products or distances in H.

We start with the case of positive definite kernels.

Property 3.22 *A function $\varphi : X \times X$ is a positive definite kernel if and only if there is an Hilbert space H and a mapping $\Phi : X \to H$ such that*

$$\varphi(x, y) = \langle \Phi(x), \Phi(y) \rangle \tag{3.19}$$

for all $x, y \in X$.

Proof: '$\Leftarrow$': It is immediate, since we have

$$\begin{aligned} \sum_i \sum_j c_i \overline{c_j} \varphi(x_i, x_j) &= \sum_i \sum_j c_i \overline{c_j} \langle \Phi(x_i), \Phi(x_j) \rangle = \\ &= \left\langle \sum_i c_i \Phi(x_i), \sum_j c_j \Phi(x_j) \right\rangle \geq 0 \end{aligned} \tag{3.20}$$

from the "nonnegativity on the diagonal" property of the inner product.

'$\Rightarrow$': For each $x \in X$ define the function $\varphi_x : X \to \mathbb{C}$ by $\varphi_x(y) = \varphi(x, y)$. Let H_0 denote the linear subspace of $\mathbb{C}^X$ generated by the functions $\{\varphi_x : x \in X\}$. Let $f, g : X \to \mathbb{C}$ be two elements of H_0; they can be written in the form $f = \sum_i c_i \varphi_{x_i}$ and $g = \sum_j d_j \varphi_{y_j}$ for some, not necessarily unique, $\{c_i\}, \{d_j\} \subseteq \mathbb{C}$ and $\{x_i\}, \{y_j\} \subseteq X$. We are going to see that the function $\langle ., . \rangle : H_0 \times H_0 \to \mathbb{C}$ defined by

$$\begin{aligned} \langle f, g \rangle &= \sum_i \sum_j c_i \overline{d_j} \varphi(x_i, y_j) = \\ &= \sum_j \overline{d_j} f(y_j) = \sum_i c_i g(x_i) \end{aligned} \tag{3.21}$$

is an inner product in H_0. The second line of (3.21) shows that $\langle .,. \rangle$ is indeed well defined (i.e. it does not depend on the choices of the representations of f and g). It also shows that $\langle .,. \rangle$ is sesquilinear. Moreover, $\langle f, f \rangle = \sum_i \sum_j c_i \overline{c_j} \varphi(x_i, x_j) \geq 0$ from the positive definiteness of φ. The sesquilinearity and positivity on the diagonal implies that $\langle .,. \rangle$ is a positive definite kernel on H_0. We can also deduce from (3.21) the *reproducing property*

$$\langle f, \varphi_x \rangle = f(x) \quad \text{for all } f \in H_0 \text{ and } x \in X, \tag{3.22}$$

which implies $\langle \varphi_x, \varphi_y \rangle = \varphi(x, y)$. Since $\langle .,. \rangle$ is a kernel, from Prop. 3.4 we have that $|f(x)|^2 \leq \langle f, f \rangle \cdot \varphi(x, x)$. Hence $\langle f, f \rangle = 0$ if and only if $f = 0$. Therefore $\langle .,. \rangle$ is an inner product, giving H_0 the structure of a pre-Hilbert space. Then $H = \overline{H_0}$ (the completion of H_0 with the norm induced by $\langle .,. \rangle$) is a Hilbert space that we call the *reproducing kernel Hilbert space* (RKHS) associated with φ. Now define Φ to be the mapping $x \mapsto \varphi_x$ and the proof is complete. □

Although a bit more complicated, the negative definite case reduces to the "squared Hilbertian distance" representation in the case of real kernels that vanish on the diagonal (and nowhere else).

Property 3.23 *A function $\psi : X \times X$ is a negative definite kernel if and only if there is a Hilbert space H, a mapping $\Phi : X \to H$, and a function $f : X \to \mathbb{C}$ such that*

$$\psi(x, y) = \|\Phi(x)\|^2 + \|\Phi(y)\|^2 - 2 \langle \Phi(x), \Phi(y) \rangle + f(x) + \overline{f(y)} \tag{3.23}$$

for all $x, y \in X$. Moreover,

- *If there is some $x_0 \in X$ such that $\psi(x, x_0) \in \mathbb{R}$ for all $x \in X$, and if ψ vanishes on the diagonal, $\psi|_\Delta = 0$, then one can choose $f = 0$;*
- *If ψ is real-valued, H may be chosen as a real Hilbert space and equation* (3.23) *becomes*

$$\psi(x, y) = \|\Phi(x) - \Phi(y)\|^2 + f(x) + \overline{f(y)} \tag{3.24}$$

- *If ψ is real-valued and vanishes on the diagonal then in addition $f = 0$, i.e., ψ admits a representation*

$$\psi(x, y) = \|\Phi(x) - \Phi(y)\|^2 \tag{3.25}$$

meaning that $\sqrt{\psi}$ is a semimetric *on X such that Φ is an* isometry. *If furthermore ψ is nonzero outside the diagonal, i.e., $\psi(x, y) = 0$ if and only if $x = y$, then $\sqrt{\psi}$ is a* metric.

Proof: Apply Proposition 3.22 to

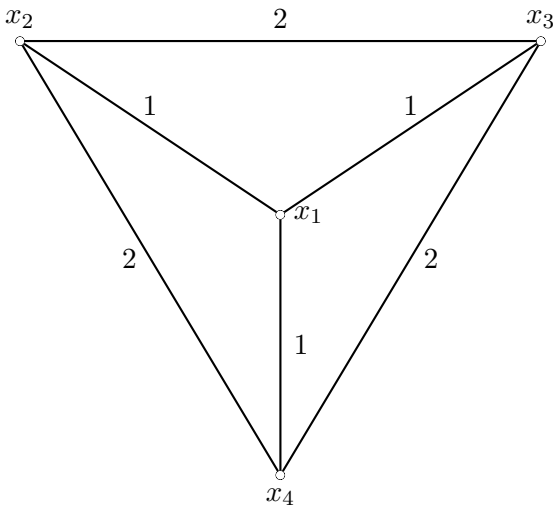
$$\varphi(x,y) = \frac{1}{2}\left(\psi(x,x_0) + \overline{\psi(y,x_0)} - \psi(x,y) - \psi(x_0,x_0)\right) \tag{3.26}$$

which is positive definite if and only if ψ is negative definite by Prop. 3.11, and use the fact that

$$\|u-v\|^2 = \|u\|^2 + \|v\|^2 - 2\,\mathrm{Re}\,\langle u,v\rangle\,. \tag{3.27}$$

□

Remark 3.24 *Notice that it is not true that any squared distance is a negative definite kernel (it is if and only if the distance is Hilbertian). As a counterexample, consider the metric space X formed by the four points x_1, x_2, x_3, x_4 whose distances are represented below:*



The matrix of squared distances,

$$K = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 4 & 4 \\ 1 & 4 & 0 & 4 \\ 1 & 4 & 4 & 0 \end{bmatrix} \tag{3.28}$$

is not negative definite: take for example $c = (-3,1,1,1)$; then $c^T K c = 6 > 0$. In this particular case it is easy to prove that X cannot be isometrically embedded in an Euclidean space. Assume that it could. Then, from

$d(x_1, x_2) + d(x_3, x_1) = d(x_2, x_3) = 2$, *the points* x_1, x_2 *and* x_3 *would be collinear, and analogously for* x_1, x_2 *and* x_4. *But then* $d(x_3, x_4) = 0$, *which is a contradiction.*

According to [BCR84], [SS02], the first known approach to generalize the positive semidefiniteness of matrices to nonfinite sets occurred in the context of integral equations, leading to a famous theorem published by Mercer in 1909. Mercer's theorem considers an alternative representation in a feature space that is not the RKHS of Prop. 3.22:

Theorem 3.25 (Mercer) *Let* $(X, \mathscr{M}, \mu)$ *be a finite measure space and* $\kappa \in L^\infty(X \times X)$ *a symmetric real-valued function such that the integral operator* $T_\kappa : L^2(X) \to L^2(X)$ *defined by*

$$(T_\kappa f)(x) = \int_X \kappa(x, y) f(y) d\mu(y) \tag{3.29}$$

is "Mercer positive definite"[7]*, that is:*

$$\int_{X \times X} \kappa(x, y) f(x) f(y) d\mu(x) d\mu(y) \geq 0 \tag{3.30}$$

for all $f \in L^2(X)$. *Let* $(\psi_j)_{j \in \mathbb{N}}$ *be the* $L^2(X)$*-valued sequence of normalized orthogonal eigenfunctions of* T_κ *and* $(\lambda_j)_{j \in \mathbb{N}}$ *the respective sequence of nonnegative eigenvalues sorted in non-increasing order. Then:*

- $(\lambda_j)_j \in \ell^1$,
- *It holds* $\kappa(x, y) = \sum_{j=1}^\infty \lambda_j \psi_j(x) \psi_j(y)$ *with absolute and uniform convergence almost everywhere in* $X \times X$.

Remark 3.26 *From the second statement it follows that* κ *corresponds almost everywhere to the ordinary inner product in* ℓ^2 *through the mapping* $x \mapsto (\sqrt{\lambda_j}\psi_j(x))_{j \in \mathbb{N}}$. *The feature space associated with this mapping is a space of real-valued* sequences *and not a space of functions in* X *as the RKHS; moreover, the inner product in the Mercer representation is the ordinary inner product for sequences, instead of the more "arbitrary" inner product devised in the RKHS. The uniform convergence property guarantees that one can arbitrarily approximate the kernel function by a representation in a Euclidean feature space (i.e. a feature space of finite dimension), by keeping only the highest eigenvalues. All this would make the Mercer representation very appealing, if not for the fact that most kernel-based algorithms in machine learning merely evaluate the kernel, rather than "working" explicitly in the feature space, and for this it suffices to guarantee the existence of a representation rather than having to choose a particular one.*

[7] It turns out that Mercer's definition of positive definiteness is equivalent to Def. 3.1.

3.4 Semigroup kernels

We now focus on a specific class of positive and negative definite kernels which are called "semigroup kernels" and which are studied with some detail in [BCR84], and see some applications for example in [CFV05]. We will see that semigroup kernels may be devised for probability distributions by considering them as elements in the semigroup of positive measures.

A *semigroup* $(S, \circ)$ (or simply S) is a nonempty set S equipped with an associative composition $\circ$ and a neutral element e. If $\circ$ is also commutative, S is called an *abelian semigroup*. An involution $*$ in a semigroup S is a mapping $* : S \to S$ satisfying $(s \circ t)^* = t^* \circ s^*$ for all $s, t \in S$. This implies $e^* = e$. A *group* is a semigroup where for each element $x \in X$ there exists an inverse $x^{-1} \in X$ verifying $x \circ x^{-1} = x^{-1} \circ x = e$.

The motivation for studying kernels on semigroups in statistical learning theory has to do with the fact that many interesting sets in this framework lack the structure of group: this happens for example with the positive measures $M_+(X)$, that form an abelian semigroup under pointwise addition, but not a group since the only invertible is the null measure.[8]

Though soon we will focus on abelian semigroups with the identical involution, in what follows we consider the general scenario where $(S, \circ, *)$ is a semigroup (possibly a group) with involution.

Definition 3.27 (Positive and negative definite functions.) *Let S be a semigroup with involution. A function $\varphi : S \to \mathbb{C}$ is called* positive definite *if $(s, t) \mapsto \varphi(s^* \circ t)$ is a positive definite kernel. Likewise, it is called* negative definite *if $(s, t) \mapsto \varphi(s^* \circ t)$ is a negative definite kernel. These kernels are accordingly called* semigroup kernels.

We denote the set of positive and negative definite functions in S respectively by $\mathcal{F}_+(S)$ and $\mathcal{F}_-(S)$. The entanglement of the semigroup structure with that of kernels induces some strong properties on such functions. For example, if φ is a positive or negative definite function, the fact that $(s, t) \mapsto \varphi(s^* \circ t)$ is hermitian implies that $\varphi(s^*) = \overline{\varphi(s)}$ for all $s \in S$.

Proposition 3.28 *Let $\varphi \in \mathcal{F}_+(S)$ and $\psi \in \mathcal{F}_-(S)$. The basic properties of positive and negative definite kernels imply, for all $s, t \in S$:*

- $\varphi(s^*) = \overline{\varphi(s)}$*, and analogously for ψ. So if we use the identical involution in S any positive or negative function must be real-valued.*
- $\varphi(s^* \circ s) \geq 0$*. In particular,* $\varphi(e) \geq 0$*.*
- $|\varphi(s^* \circ t)|^2 \leq \varphi(s^* \circ s) \cdot \varphi(t^* \circ t)$*. In particular,* $|\varphi(s)|^2 \leq \varphi(e) \cdot \varphi(s^* \circ s)$*.*

[8] Another example of semigroup that may be of particular interest to us is the semigroup of strings under concatenation, which however is not abelian.

- $2\operatorname{Re}\psi(s^* \circ t) \geq \psi(s^* \circ s) + \psi(t^* \circ t)$. *In particular,* $2\operatorname{Re}\psi(t) \geq \psi(e) + \psi(t^* \circ t)$.
- $\mathcal{F}_+(S)$ *and* $\mathcal{F}_-(S)$ *are convex cones in* $\mathbb{C}^S$*, and are closed in the topology of pointwise convergence.* $\mathcal{F}_+(S)$ *is stable under pointwise products.*

[BCR84] studies in detail positive and negative definite functions. Some interesting results have to do with the ability of representing a wide class of these functions as integrals of semicharacters. This goes beyond our scope, so we conclude this section by deriving the negative definiteness of the entropy function, and introducing the Jensen-Shannon kernel.

Example 3.29 (Entropy as a negative definite function.) *We are going to show that the entropy function defined in* (2.14) *is negative definite. Due to the pointwise integration closure expressed in* (3.7)*, it suffices to show that the function* $\psi : \mathbb{R}_+ \to \mathbb{R}$ *defined by* $\psi(y) = -y \log y$ *is negative definite. We reproduce here the proof given in [BCR84] (Example 6.5.16).*

Note first that

$$\psi(y) = -y \log y = \lim_{\alpha \to 1} \frac{-y^\alpha + y}{\alpha - 1} \tag{3.31}$$

as can be seen by L'Hôpital's rule. Then it suffices to prove that $y \mapsto \frac{-y^\alpha + y}{\alpha - 1}$ *is negative definite for any* $\alpha \in]1 - \epsilon, 1 + \epsilon[\setminus \{1\}$*, with* $\epsilon > 0$ *sufficiently small. We are going to actually prove it for* $\alpha \in]0, 2[\setminus \{1\}$*. The problem clearly reduces to prove that* $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}$ *defined by*

$$\psi_\alpha(y) = \begin{cases} y^\alpha, & 0 < \alpha \leq 1, \\ -y^\alpha, & 1 < \alpha \leq 2 \end{cases} \tag{3.32}$$

is negative definite for any $\alpha \in]0, 2[\setminus \{1\}$*, and this is ensured by Props. 3.20 and 3.21 from the negative definiteness of* ψ_1*,* $-\psi_1$ *and* ψ_2*.*

Example 3.30 (Jensen-Shannon divergence as a squared metric.) *The above result may be used to show that the Jensen-Shannon divergence, as defined in* (2.19)*, is the square of a metric. As pointed out in [Top02] this was only recently taken into account by information theoricists, with several independent proofs being published (for example, in [ES03] a pure analytical proof is given, without the insights provided by kernel theory).*

By the previous example h *is a negative definite function on the semigroup* $\left(\frac{d}{d\nu} M_+^h(X), +\right)$*, so it follows that* $(f, g) \mapsto h(f + g)$ *is a semigroup negative definite kernel on* $\frac{d}{d\nu} M_+^h(X) \times \frac{d}{d\nu} M_+^h(X)$*, and so is the function* $\tilde{J} : \frac{d}{d\nu} M_+^h(X) \times \frac{d}{d\nu} M_+^h(X) \to \mathbb{R}$ *defined by*

$$\tilde{J}(f, g) = h\left(\frac{f + g}{2}\right) \tag{3.33}$$

since $h\left(\frac{f+g}{2}\right) = h\left(\Phi(f) + \Phi(g)\right)$ *with* $f \mapsto \Phi(f) = \frac{f}{2}$. *Since in addition* $(f,g) \mapsto -\frac{h(f)+h(g)}{2}$ *is trivially negative definite (see Prop. 3.7), we have that* J *is a negative definite kernel, for being the sum of negative definite kernels. Since* $J(f,g) = 0$ *if and only if* $f = g$, *we have by property 3.23 that* $\sqrt{J}$ *is a metric.*

Example 3.31 (Jensen-Shannon kernel.) *Following [CV05, CFV05] we now derive a positive definite kernel between measures. Divergence-based kernels have also been devised in [HB04, MHV03].*

First, notice that by Prop. 3.18 the kernel $\kappa_{\tilde{J}} = \exp(-t\tilde{J})$ *is, for any* $t > 0$, *positive definite and, in particular, infinitely divisible. Hence, its normalized version*

$$\begin{aligned} \kappa_J(f,g) &= \frac{\kappa_{\tilde{J}}(f,g)}{\sqrt{\kappa_{\tilde{J}}(f,f)\kappa_{\tilde{J}}(g,g)}} = \\ &= \exp\left(-th\left(\frac{f+g}{2}\right) + t\cdot\frac{h(f)+h(g)}{2}\right) = \\ &= \exp\left(-tJ(f,g)\right) \end{aligned} \tag{3.34}$$

is also, for any $t > 0$, *an infinitely divisible positive definite kernel*[9], *defined on* $\frac{d}{d\nu}M_+^h(X) \times \frac{d}{d\nu}M_+^h(X)$, *and satisfying* $\kappa_J|_\Delta = 1$. *Notice however that* κ_J *is not a semigroup kernel, unlike* $\kappa_{\tilde{J}}$, *although it may be regarded as the normalization of a semigroup kernel.*

4 Information geometry

We now describe some concepts from differential geometry that are useful to characterize certain properties of parametric and non-parametric statistical models. There actually have been many approaches to reinterpret the classical statistical and information theoretic methods through a differential geometric perspective, leading to recent advances in the so-called field of "information geometry" [AN01, MR93]. We start by briefly reviewing some basic concepts and terminology used in the field of differential geometry, and then proceed to show how they apply in the geometry of "statistical manifolds".

4.1 Basic concepts of differential geometry

Intuitively, a manifold M is a set that "behaves" locally like an Euclidean space. We endow M successively with three layers of structure: a *topological* layer, which allows to handle notions as continuity, a *differentiable* layer,

[9] It is easy to show that normalizing a $\mathbb{R}_{++}$-valued kernel preserves infinitely divisible positive definiteness.

which introduces differentiability and smoothness, and a *Riemannian* layer, which allows defining lengths, angles, and curvatures.

4.1.1 Topological manifolds

A *homeomorphism* between two topological spaces X and Y is a map $\varphi : X \to Y$ that is continuous, bijective, and whose inverse φ^{-1} is also continuous (this means that each open subset of X is mapped to an open subset of Y and vice-versa).

Let M be a Hausdorff second countable set, i.e., a Hausdorff space with a countable basis for its topology[10]. We say that M is a n-dimensional *topological manifold* if it is *locally homeomorphic* to an Euclidean space $\mathbb{R}^n$, i.e., if each $p \in M$ has a neighborhood $U \subseteq M$ for which there is a homeomorphism φ from U to an open subset of $\mathbb{R}^n$. These homeomorphisms are called the *charts* or *local coordinate systems* of the manifold. A set of local coordinate systems whose domains cover M is called an *atlas*.

Not all manifolds admit a *global* coordinate system: consider for example the sphere or the torus, for which we can only define local coordinate systems. However, the most addressed in the framework of "information geometry" either admit one, or we are only worried about local behaviors that can be studied using a single coordinate system.

If $p \in M$ is a point in the manifold, and $\varphi : U \subseteq M \to \mathbb{R}^n$ is a local coordinate system such that $p \in U$, we call $\varphi(p) = [\xi^1(p), \dots, \xi^n(p)] \equiv [\xi^1, \dots, \xi^n]$ the local coordinates of p, and the n functions $\xi^i : U \to \mathbb{R}$ the *coordinate functions* (cf. Fig. 2). We denote $\varphi = [\xi^i]$. If $\psi : V \subseteq M \to \mathbb{R}^n$, $\psi = [\rho^i]$ is another local coordinate system such that $p \in V$, then p has coordinates $[\xi^i] \in \mathbb{R}^n$ with respect to φ and $[\rho^i] \in \mathbb{R}^n$ with respect to ψ. The latter may be obtained from the former by applying the *coordinate transformation* or *transition function* $\psi \circ \varphi^{-1} : \varphi(U \cap V) \subseteq \mathbb{R}^n \to \mathbb{R}^n$.

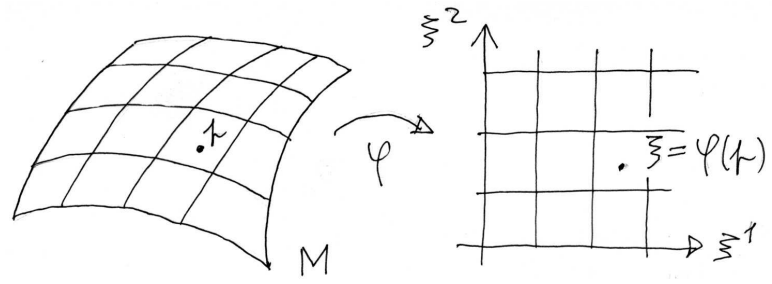


Figure 2: Coordinate functions of a topological manifold.

[10]This assumption excludes some pathological cases. It is not too strong for our purposes: any separable (i.e. that has a dense countable subset) metric space under the ordinary topology is Hausdorff second countable. This includes all the finite Hilbert spaces, which are isomorphic to $\mathbb{R}^n$.

By requiring these coordinate transformations to be *smooth*, we are led to the notion of differentiable manifold.

4.1.2 Differentiable manifolds

Let $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ be open sets. A map $f : U \to V$ is said to be *smooth* or *infinitely differentiable* (denoted C^∞) if it has continuous partial derivatives of all orders; the set of such maps is denoted $C^\infty(U, V)$. If $f \in C^\infty(U, V)$ is bijective and $f^{-1} \in C^\infty(V, U)$, we say that f is a C^∞-*diffeomorphism*.

Let M be a n-dimensional topological manifold with atlas A. We say that M is a *differentiable manifold* if for any pair of charts $\varphi, \psi \in A$ with domains respectively $U \subseteq M$ and $V \subseteq M$, the transition function $\psi \circ \varphi^{-1}$, defined in $\varphi(U \cap V)$, is a C^∞-*diffeomorphism* with respect to its image.

The above notion of smoothness may be extended to maps between manifolds. If M, N are differentiable manifolds, we say that $f : M \to N$ is C^∞ if for any pair of charts $\varphi : U \subseteq M \to \mathbb{R}^m$ and $\psi : V \subseteq N \to \mathbb{R}^n$ such that $f(U) \subseteq V$ we have $\psi \circ f \circ \varphi^{-1} \in C^\infty(\varphi(U), \mathbb{R}^n)$. Notice that the latter is simply the coordinated version of f. A C^∞-*diffeomorphism* between two manifolds M, N is a bijection $f : M \to N$ such that $f \in C^\infty(M, N)$ and $f^{-1} \in C^\infty(N, M)$.

In particular, a function $f : M \to \mathbb{R}$ is C^∞ if $f \circ \varphi^{-1} \in C^\infty(\varphi(U), \mathbb{R})$ for any local chart φ. The set of $C^\infty(M, \mathbb{R})$ functions is denoted $\mathcal{F}(M)$. It is straightforward that f is $C^\infty(M, N)$ if and only if $g \circ f \in \mathcal{F}(M)$ for all $g \in \mathcal{F}(N)$.

4.1.3 Tangent spaces

When M is a n-dimensional surface in $\mathbb{R}^m$ parameterized as $M = \{x \in \mathbb{R}^m : f(x) = 0\}$, where $f : \mathbb{R}^m \to \mathbb{R}^n$ is a differentiable map with component functions $f_1, \ldots, f_n$, the *tangent space* at $p \in M$, denoted T_pM, is simply the vector space orthogonal to $\text{span}\{\nabla_x f_1|_p, \ldots, \nabla_x f_n|_p\}$. *Tangent vectors* are elements of this space; they correspond to "local derivatives of functions along a path".

The definition of tangent space for a general manifold must be intrinsic, and captures the intuitive notion of the tangent space T_pM as the vector space obtained by "locally linearizing" M around p (see Fig. 3). The elements of T_pM, the tangent vectors, are defined as "equivalence classes of curves that have the same velocity vectors at p." Let $\gamma : I \subseteq \mathbb{R} \to M$ be the parameterization of a curve such that $\gamma(a) = p$. If we take a smooth function $f \in \mathcal{F}(M)$, we may take the composition $f \circ \gamma : I \to \mathbb{R}$ and consider its derivative, which is called the directional derivative of f along the curve γ. Introducing local coordinates $[\xi^1, \ldots, \xi^n]$ and defining $\gamma^i(t) \equiv \xi^i(\gamma(t))$ for

$t \in I$, we may write this derivative as:

$$\begin{aligned} \frac{d}{dt} f(\gamma(t)) &= \sum_{i=1}^{n} \left(\frac{\partial f}{\partial \xi^i} \right)_{\gamma(t)} \frac{d\gamma^i(t)}{dt} = \\ &= \left(\frac{\partial f}{\partial \xi^i} \right)_{\gamma(t)} \frac{d\gamma^i(t)}{dt}. \end{aligned} \tag{4.1}$$

where in the last line we used the Einstein notation. The tangent vector $\dot{\gamma}(a)$ at $p = \gamma(a)$ is defined as the operator that maps $f \in \mathcal{F}(M)$ to its directional derivative at $t = a$. The tangent vector at point p of the i-th coordinate curve, $\left(\frac{\partial}{\partial \xi^i}\right)_p : \mathcal{F}(M) \to \mathbb{R}$ maps $f \mapsto \left(\frac{\partial f}{\partial \xi^i}\right)_p$. Considering the set of all curves that pass through p, the set of all tangent vectors corresponding to these curves forms a linear space, that we define to be the tangent space T_pM. Given a coordinate system $[\xi^1, \ldots, \xi^n]$ in a neighborhood of p, the vectors $\left(\frac{\partial}{\partial \xi^1}\right)_p, \ldots, \left(\frac{\partial}{\partial \xi^n}\right)_p$ form a basis for T_pM, which is called the *natural basis* with respect to the coordinate system $[\xi^1, \ldots, \xi^n]$.

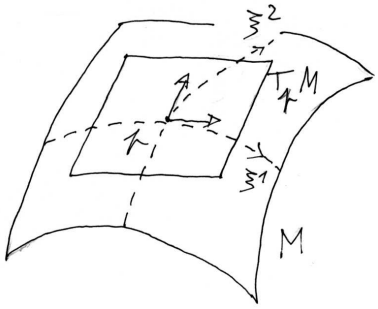


Figure 3: Tangent space.

4.1.4 Submanifolds

Our first description of a tangent space considered a manifold as a surface in $\mathbb{R}^n$. Actually, we often represent a m-dimensional manifold M in an ambient space as a *submanifold* of a n-dimensional manifold N, with $n \geq m$. Technically, we say that M is a submanifold of N if:

- M is endowed with the induced topology of N,
- the embedding $\iota : M \to N$ defined by $\iota(p) = p$ is C^∞, and
- for each $p \in M$ and each coordinate systems $[\theta^j]$ of M and $[\xi^i]$ of N, the matrix $\left(\frac{\partial \xi^i}{\partial \theta^j}\right)_p$ has full rank.

An important property is that, if M is a submanifold of N, then for any point $p \in M$, the tangent space T_pM is a linear subspace of T_pN. We next see some examples (to be later used) of submanifolds of Euclidean spaces.

Example 4.1 (The n-sphere as submanifold of $\mathbb{R}^{n+1}$) *The n-sphere $\mathbb{S}^n$ may be represented in the ambient space $\mathbb{R}^{n+1}$ as*

$$\mathbb{S}_n \equiv \left\{ x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = 1 \right\}. \tag{4.2}$$

A point $p \in \mathbb{S}^n$ may be represented in the Cartesian coordinate system $[\xi^i]$ of $\mathbb{R}^{n+1}$ as $p \equiv (x_1, \ldots, x_{n+1})$. The tangent space $T_p\mathbb{S}_n$ is a linear subspace of $T_p\mathbb{R}^{n+1}$, and can be represented in the natural basis $\left\{ \left(\frac{\partial}{\partial \xi^i} \right)_p \right\}_{i=1}^{n+1}$ of the embedding tangent space $T_p\mathbb{R}^{n+1}$ as follows:

$$T_p\mathbb{S}_n \equiv \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i x_i = 0 \right\}. \tag{4.3}$$

This results from the fact that $T_p\mathbb{S}_n = (\text{span}\{\nabla_x f|_p\})^\perp$, with $f : \mathbb{R}^{n+1} \to \mathbb{R}$ defined as $f(x) = \sum_{i=1}^{n+1} x_i^2 - 1$.

Example 4.2 (The open n-simplex as submanifold of $\mathbb{R}^{n+1}$) *Also the n-dimensional open simplex $\mathbb{P}_n$ is a submanifold of $\mathbb{R}^{n+1}$ and can be represented as*

$$\mathbb{P}^n \equiv \left\{ x \in \mathbb{R}^{n+1} : x_i > 0, \sum_{i=1}^{n+1} x_i = 1 \right\}. \tag{4.4}$$

Analogously, $T_p\mathbb{P}_n$ can be represented in the natural basis of $T_p\mathbb{R}^{n+1}$ as:

$$T_p\mathbb{P}_n \equiv \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i = 0 \right\}. \tag{4.5}$$

4.1.5 Tangent bundle, vector fields and tensor fields

The *tangent bundle* of M is the disjoint union of the tangent spaces of each point in the manifold,

$$TM = \bigsqcup_{p \in M} T_pM. \tag{4.6}$$

Its *position map* is the surjective map $\pi : TM \to M$ that maps $v_p \mapsto p$ for any $v_p \in T_pM$, i.e., $\pi^{-1}(p) = T_pM$. A *vector field* is a map $X : M \to TM$ that associates to each point p a tangent vector in T_pM, i.e., such that $\pi \circ X = I_M$.

From now on we assume that M has a global coordinate system, that we denote by $[\xi^i]$. In this case we may define n vector fields formed by the natural basis, through the mappings $\partial_i \equiv \frac{\partial}{\partial \xi_i} : p \mapsto \left(\frac{\partial}{\partial \xi^i}\right)_p$, for $i = 1, \ldots, n$. The value of any vector field X at a point p may be written as a linear combination $X_p = X_p^i \partial_i$ with $X_p^i \in \mathbb{R}$. Doing this for all $p \in M$ forms n functions $X^i : M \to \mathbb{R}$ that are the components of the vector field X with respect to $[\xi^i]$, so we may write $X = X^i \partial_i$. If each $X^i \in \mathcal{F}(M)$, we say that X is a C^∞-vector field. This definition does not depend on the choice of the coordinate system: if we choose a different global coordinate system $[\rho^j]$, the same vector field is expressible as $X = \tilde{X}^j \tilde{\partial}_j$, where $\tilde{\partial}_j \equiv \frac{\partial}{\partial \rho^j}$, and it holds:

$$\tilde{X}^j = X^i \frac{\partial \rho^j}{\partial \xi^i}. \tag{4.7}$$

We denote by $\mathcal{X}(M)$ the set of C^∞-vector fields in M.

We now introduce the notion of tensor field. For each point p, let $[T_p]_r^0$ and $[T_p]_r^1$ denote respectively the families of multilinear mappings of the form $\underbrace{T_pM \times \ldots \times T_pM}_{r \text{ times}} \to \mathbb{R}$ and of the form $\underbrace{T_pM \times \ldots \times T_pM}_{r \text{ times}} \to T_pM$. A map $A : p \mapsto A_p$ which associates at each point p a multilinear mapping $A_p \in [T_p]_r^q$ is called a *tensor field* of type (q, r), for $q = 0, 1$, where r is called the covariant degree and q the contravariant degree. If for any fixed r vector fields $X_1, \ldots, X_r$ the mapping $A(X_1, \ldots, X_r)$ is C^∞, i.e. it is in $\mathcal{F}$ (resp. in $\mathcal{X}$), we call A a *C^∞-tensor field*. A (C^∞-)vector field is a (C^∞-)tensor field of type $(1, 0)$. We will next introduce a C^∞-tensor field of type $(0, 2)$, the *Riemannian metric*.

4.1.6 Riemannian manifolds

Suppose that for each point $p \in M$, a local inner product is defined on the tangent space T_pM, i.e., a real-valued function $\langle , \rangle_p \to \mathbb{R}$ that satisfies the axioms on Example 2.4. From the bilinearity of the inner product, we have that the map $g : p \mapsto \langle , \rangle_p$ is a tensor field of type $(0, 2)$. If it is in addition a C^∞-tensor field, we say that g is a *Riemannian metric* on M. A manifold M endowed with a Riemannian metric g is called a *Riemannian manifold*, and denoted (M, g). It is important to note that such a metric is not naturally determined by the structure of M, i.e., it is possible to consider an infinite number of Riemannian metrics on M. If we fix a coordinate system $[\xi^i]$, we may represent g by its component functions with respect to $[\xi^i]$, namely $g_{ij} = \langle \partial_i, \partial_j \rangle$. From the bilinearity of the inner product, these component functions completely determine g, i.e., given any two tangent vectors $D, D' \in T_pM$ written in terms of their coordinates as $D = D^i (\partial_i)_p$ and $D' = D'^i (\partial_i)_p$, their inner product is given by

$$\left\langle D, D' \right\rangle_p = g_{ij}(p) D^i D'^j. \tag{4.8}$$

The matrix $G(p) \equiv [g_{ij}(p)]_{1\leq i,j\leq n}$ is obviously positive semidefinite for each $p \in M$. If we choose a different coordinate system $[\rho^k]$ the new components $\tilde{g}_{kl}$ with respect to $[\rho^k]$ relate with g_{ij} via

$$\tilde{g}_{kl} = g_{ij}\left(\frac{\partial \xi^i}{\partial \rho^k}\right)\left(\frac{\partial \xi^j}{\partial \rho^l}\right). \tag{4.9}$$

The Riemannian metric allows us to measure lengths and angles in the manifold. The length $\|\gamma\|$ of a curve $\gamma : [a, b] \to M$ is defined as

$$\|\gamma\| = \int_a^b \|\dot{\gamma}\|\, dt = \int_a^b \sqrt{g_{ij}\dot{\gamma}^i\dot{\gamma}^j}\, dt \tag{4.10}$$

where $\dot{\gamma}^i \equiv \frac{d}{dt}\left(\xi^i \circ \gamma\right)$. This allows defining the geodesic distance $d_g(x, y)$ between two points in the manifold as

$$d_g(x, y) = \inf_{\gamma\in\Gamma(x,y)} \|\gamma\|\,, \tag{4.11}$$

where $\Gamma(x, y)$ denotes the set of piecewise differentiable curves connecting x and y.

4.1.7 Tangent map, pull-back metric, and isometries

Consider now a diffeomorphism $f : M \to N$ between two diferentiable manifolds. In order to map velocity vectors of curves γ in M to velocity vectors of curves $f(\gamma)$ in N, we define the *tangent map*

$$\begin{array}{rccc} f_* : T_pM & \to & T_{f(p)}N \\ v & \mapsto & f_*v \end{array} \tag{4.12}$$

with $r \mapsto (f_*v) \cdot r = v \cdot (r \circ f)$ for all $r \in \mathcal{F}(N)$. The coordinated version of f_* is simply the Jacobian matrix associated with the coordinated version of f [Leb06]. Fig. 4 gives a geometric interpretation.

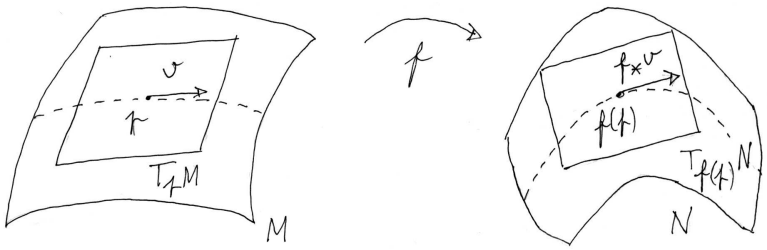


Figure 4: Tangent map.

Notice that a diffeomorphism as above, $f : M \to N$, when (N, h) is a Riemannian manifold, automatically endows M with a Riemannian structure, since the metric h in N induces a metric in f^*h in M, called the

pull-back metric, via the tangent map f_*, as

$$\begin{array}{rcl}(f^*h)_p : T_pM \times T_pM & \to & \mathbb{R} \\ (u,v) & \mapsto & (f^*h)_p(u,v)\end{array} \tag{4.13}$$

with $(f^*h)_p(u,v) = h_{f(p)}(f_*u, f_*v)$. Intuitively, to calculate the inner product of two velocity vectors in T_pM we map them to velocity vectors in $T_{f(p)}N$ and let the metric in N do the job.

Two Riemannian manifolds (M,g) and (N,h) are said to be *isometric* if there is a diffeomorphism $f : M \to N$ such that $g_p(u,v) = (f^*h)_p(u,v)$ for each $p \in M$, and $u, v \in T_pM$. In that case f is called an *isometry*. Isometries preserve the geodesic distance function, i.e., it holds $d_g(x,y) = d_h(f(x), f(y))$ for all $x, y \in M$.

4.2 The geometry of probability models

We now give an important characterization of the geometry of the space of probability distributions as an affine space, having the exponential families as an affine subspace. For this, we follow [MR93]. Then we consider finite dimensional statistical manifolds and show how the Fisher metric induces a Riemannian atructure.

4.2.1 Affine spaces

An *affine space* is a set that becomes a vector space by selecting a point to be the zero point. More formally, it is a set X together with a vector space V, where each vector $v \in V$ corresponds to a translation function $\tau_v : X \to X$ satisfying:

1. $\tau_v \circ \tau_u = \tau_{u+v}$ for all $u, v \in V$,
2. For any $x, y \in X$ there is a unique $v \in V$ such that $y = \tau_v(x)$.

Let X be an affine space. If we declare a point $x_0 \in X$ as the origin, there is a natural bijection $f : X \to V$ that identifies each point $x \in X$ with the vector $v \in V$ such that $x = \tau_v(x_0)$; in particular $f(x_0) = 0$, i.e., the origin is identified with the zero vector. A simple example of affine space is the Euclidean plane: once we choose an origin and fix a basis of vectors each point receives a set of coordinates. Such a coordinate system is called an *affine coordinate system*. An important fact is that a set X with a collection of coordinate systems is an affine space if and only if any two coordinate systems $\theta, \xi : X \to \mathbb{R}$ are *affinely related*, i.e., if and only if there is a matrix A and a vector b such that $\theta^i(x) = A^i_j \xi^j(x) + b^i$ holds for each $x \in X$. Another fundamental geometric property of affine spaces, as we show below, is its *flatness*.

4.2.2 Positive measures as an affine space via the log-likelihood

We now show how positive measures on an event space X form an affine space. To be able to work with measure densities, we restrain ourselves to a set M of measures that are absolutely continuous with respect to each other. If we choose an origin ν for M, we may identify each measure $\mu \in M$ with its density with respect to ν, the Radon-Nikodym derivative $p = \frac{d\mu}{d\nu}$, which is unique up to equivalence within measure zero and satisfies $p > 0$ almost everywhere. To turn M into an affine space, we have to define a *translation* between measures (or, equivalently, densities). Pointwise multiplication by a positive function f would be a natural first candidate (in fact, any density p_1 may be translated to p_2 by a unique positive function f such that $p_2 = fp_1$), but this must be discarded since positive functions do not form a vector space. This may be overcome by considering the vector space R_X of the ν-measurable functions on X, i.e., the *random variables*, and let translation by f mean $p \mapsto \exp(f)p$, using the fact that $\exp(f) > 0$. In terms of measures, the translation function is

$$\begin{array}{rcl} \tau_f : M & \to & M \\ \mu & \mapsto & \tau_f(\mu) = \exp(f)d\mu, \end{array} \tag{4.14}$$

and it is straightforward that this turns M into an affine space. If we choose an origin ν for M, it becomes a vector space, where each measure $\mu \in M$ is identified with the vector $f \in R_X$ that translates ν to μ; this corresponds to the *log-likelihood* map

$$\begin{array}{rcl} \ell : M & \to & R_X \\ \mu \equiv pd\nu & \mapsto & \ell(\mu) = \log p. \end{array} \tag{4.15}$$

Obviously, $\ell(\nu) = \log 1 = 0$, i.e., the base measure corresponds to the zero function.

Denote by $M^1 = \{\mu \in M : \mu(X) = 1\}$ the space of all probability measures in M. M^1 cannot be seen as an affine subspace inside M, since $\int_X \exp(f)pd\nu = 1$ would require $E_p(\exp(f)) = 1$ and the subset of R_X for which this holds is not a vector space. However, instead of regarding each probability measure as a point in M^1, we may regard it as an equivalence class of finite measures in M up to scale (see Fig. 5). Two finite measures are considered equivalent if they are rescalings of each other; in that case their L^1-normalization reaches a unique probability measure. But "rescaling" in the affine space M corresponds to translation by a constant function, i.e., $\lambda p = \exp(\log \lambda)p = \exp(C)p$; hence, if a measure μ corresponds (via the log-likelihood) to a function f in R_X, then the equivalence class of a measure up to scale, say $\mathbb{R}_{++} \cdot \mu$, corresponds to that of a function up to the addition of a constant, say $f + \mathbb{R} \cdot 1$. The set of functions up to constants, i.e. the quotient space $R_X/\mathbb{R} \cdot 1$, is a vector space, hence it is straightforward that measures up to scale define an affine space. The set of *finite* measures up to scale

is a subset of this affine space; it corresponds to the set M^1 of probability measures, and it is not affine.

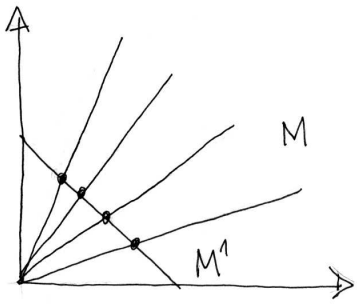


Figure 5: Two ways of regarding a probability distribution: as a point in M^1, or as a "ray" in M.

4.2.3 Exponential families as an affine space

Unless the event space X is finite, the set of positive measures on X is infinite-dimensional, and this poses theoretical problems in treating it as a statistical manifold. Instead, we often consider parametric families of probability distributions

$$S = \{P_\theta \in M_+^1(X) : \theta \in \Theta\}, \tag{4.16}$$

where the open set $\Theta \subseteq \mathbb{R}^n$ is the space of parameters. We assume further that for all $x \in X$ the function $\theta \mapsto p(x;\theta)$, where $p(.;\theta)$ is the density of P_θ with respect to the base measure ν, is a C^∞-diffeomorphism, and that $\mathrm{supp}(P_\theta) = X$ for all θ, which means that $p(x,\theta) > 0$ for all $\theta \in \Theta$ and $x \in X$.

We now show that a family of probability measures which forms a finite-dimensional affine subspace of the set of measures up to scale corresponds to an *exponential family*. Indeed, let S be an r-dimensional affine subspace of the set of measures up to scale. Then S is spanned by linearly independent non-constant random variables $f_1, \ldots, f_r$, and if ν is one of the measures (up to scale) taken as origin, any measure μ in S may be expressed as

$$\mu = \exp(\theta^1 f_1 + \ldots + \theta^r f_r - K)d\nu \tag{4.17}$$

for some coordinates $\theta^1, \ldots, \theta^r \in \mathbb{R}$ and where $K \in \mathbb{R}$ is a constant that scales μ. The probability measures in this subspace correspond to those which are finite, and may be obtained by properly setting $K = K(\theta)$ (called the *cumulant generating function*) to turn μ into a probability measure:

$$K(\theta) = \log \int_X \exp(\theta^1 f_1 + \ldots + \theta^r f_r)d\nu. \tag{4.18}$$

The set $\Theta = \{(\theta^1, \ldots, \theta^r) \equiv \mu : \mu(X) < \infty\}$ corresponding to finite measures up-to-scale (hence, probability measures) is a convex set of $\mathbb{R}^r$, and such families of probability measures are called *exponential families*. A great number of distributions widely used in statistics are exponential families: the normal, gamma, Dirichlet, Bernoulli, multinomial, Poisson, and geometric are some examples.

Example 4.3 (Normal family) *The normal family, usually parameterized as*

$$dP(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx \tag{4.19}$$

where the sample space is $\mathbb{R}$ with the Lebesgue measure dx, and the parameters are $\mu \in \mathbb{R}$ and $\sigma > 0$, is an exponential family. Indeed it can also be parameterized as

$$dP(x; \theta^1, \theta^2) = \exp(f_1(x)\theta^1(x) + f_2(x)\theta^2(x) - K(\theta))dx \tag{4.20}$$

with random variables f_1, f_2 defined by $f_1(x) = x^2$, $f_2(x) = x$, parameters θ^1, θ^2 given by

$$\theta^1 = -\frac{1}{2\sigma^2} \quad \textit{and} \quad \theta^2 = \frac{\mu}{\sigma^2} \tag{4.21}$$

cumulant generating function

$$K(\theta) = \frac{1}{2}\log\left(-\frac{\pi}{\theta^1}\right) - \frac{(\theta^2)^2}{4\theta^1}, \tag{4.22}$$

and the Lebesgue measure as base measure. The parameter $\theta = (\theta^1, \theta^2)$ is called the canonical parameter*, and lies in the open subset of $\mathbb{R}^2$ defined by $\theta^1 < 0$. It follows that the normal family forms an affine subspace of the space of measures up to scale. As an affine space, it holds that any other canonical parameterization (η^1, η^2) is affinely related with (θ^1, θ^2).*

Example 4.4 (Finite nonparametric spaces) *Let X be a finite event space with m events $x_1, \ldots, x_m$ that occur respectively with probabilities $\xi^1, \ldots, \xi^m$. Notice that there are only $m-1$ degrees of freedom in the parameters $\xi^1, \ldots, \xi^m$, since we must have $\sum_{i=1}^m \xi^i = 1$. So, the family of m-sized nonparametric spaces may be parameterized by the open set $\Xi \subseteq \mathbb{R}^{m-1}$*

$$\Xi = \left\{(\xi^1, \ldots, \xi^{m-1}) : \sum_{i=1}^{m-1} \xi^i < 1, \quad \xi^i > 0 \quad \textit{for any } i = 1, \ldots, m\right\}, \tag{4.23}$$

defining $\xi^m = 1 - \sum_{i=1}^{m-1} \xi^i$.

This is also an exponential family. It may be alternatively parameterized with canonical parameters $\theta = (\theta^1, \ldots, \theta^{m-1})$ as

$$dP(x; \theta) = \exp\left(\sum_{i=1}^{m-1} f_i(x)\theta^i(x) - K(\theta)\right) d\nu(x) \tag{4.24}$$

with random variables f_i defined by

$$f_i(x) = \begin{cases} 1, & x = x_i \\ 0, & x \neq x_i, \end{cases} \tag{4.25}$$

parameters θ^i given by

$$\theta^i = \log \frac{\xi^i}{1 - \sum_{i=1}^{m-1} \xi^i} = \log \frac{\xi^i}{\xi^m}, \tag{4.26}$$

cumulant generating function

$$K(\theta) = \log \left(1 + \sum_{i=1}^{m-1} \exp(\theta^i) \right), \tag{4.27}$$

and ν being the counting measure. The parameter space is $\Theta = \mathbb{R}^{m-1}$.

Example 4.5 (Multinomial family) *The multinomial family may be seen as a generalization of the finite nonparametric space and yields a very similar expansion as an exponential family. Multinomials appear frequently in communication theory. Suppose that a source emits symbols $\sigma_1, \ldots, \sigma_m \in \Sigma$ with emission probabilities $\xi^1, \ldots, \xi^m$, and consider Bernoulli sequences of n symbols. Let the random variables $x_1, \ldots, x_m$ mean the number of times each symbol occurs. This gives rise to a multinomial distribution,*

$$dP(x_1, \ldots, x_m; \xi^1, \ldots, \xi^m) = \frac{n!}{x_1! \cdot \ldots \cdot x_m!} \prod_{i=1}^{m} (\xi^i)^{x_i}. \tag{4.28}$$

Again, there are only $m-1$ degrees of freedom, so, the family of m-sized multinomials may be parameterized by the open set $\Xi \subseteq \mathbb{R}^{m-1}$

$$\Xi = \left\{ (\xi^1, \ldots, \xi^{m-1}) : \sum_{i=1}^{m-1} \xi^i < 1, \quad \xi^i > 0 \quad \text{for any } i = 1, \ldots, m \right\}, \tag{4.29}$$

defining $\xi^m = 1 - \sum_{i=1}^{m-1} \xi^i$. Denote $x = (x_1, \ldots, x_m)$ and $\xi = (\xi^1, \ldots, \xi^m)$. The log-likelihood of an element of this family is given by

$$\ell(x; \xi) = \log p(x; \xi) = C(x) + \sum_{i=1}^{m} x_i \log \xi^i, \tag{4.30}$$

where $C(x) = \log(n!) - \sum_{i=1}^{m} \log(x_i!)$ is independent of ξ. The maximum likelihood estimation $\widehat{\xi}$ of the emission probabilities can be obtained by maximizing the previous expression (using a Lagrange multiplier to handle the restriction $\sum_i \xi^i = 1$). As expected, it is given by $\widehat{\xi^i} = x_i/n$, for each $i = 1, \ldots, n$, and it is unbiased, i.e., $E\left(\widehat{\xi^i}\right) = E(x_i/n) = \xi^i$.

The multinomial family is an exponential family; it may be alternatively parameterized with canonical parameters $\theta = (\theta^1, \ldots, \theta^{m-1})$ *as*

$$dP(x;\theta) = \exp\left(\sum_{i=1}^{m-1} f_i(x)\theta^i(x) - K(\theta)\right) d\nu(x) \tag{4.31}$$

with random variables f_i *defined by* $f_i(x) = x_i$*, parameters* θ^i *given by*

$$\theta^i = \log \frac{\xi^i}{1 - \sum_{i=1}^{m-1} \xi^i} = \log \frac{\xi^i}{\xi^m}, \tag{4.32}$$

scale factor

$$K(\theta) = n \log\left(1 + \sum_{i=1}^{m-1} \exp(\theta^i)\right), \tag{4.33}$$

and base measure ν *with density*

$$\frac{n!}{x_1! \cdot \ldots \cdot x_m!} \tag{4.34}$$

with respect to the counting measure. The parameter space is $\Theta = \mathbb{R}^{m-1}$*.*

4.2.4 The Fisher metric

We next endow statistical manifolds with an appropriate Riemannian metric. This is proved to be, through Čencov's theorem [MR93, AN01, Leb05], "the only invariant metric under sufficient statistics transformations".

Let $\ell(x;\theta) = \log p(x;\theta)$ be the log-likelihood function of p_θ. The way $\ell(x;\theta)$ behaves under small perturbations of the parameters may be studied via the *score map* $s : X \times \Theta \to \mathbb{R}^n$,

$$s(x;\theta) = \nabla_\theta \ell(x;\theta) = \nabla_\theta \log p(x;\theta), \tag{4.35}$$

i.e. $[s^i(x;\theta)] = \left[\frac{\partial}{\partial\theta^i}\ell(x;\theta)\right] \equiv [\partial_i \ell(x;\theta)]$. It is straightforward that the expected value of the score is zero:

$$\begin{aligned} E_\theta[s^i_\theta] &= \int_X p(x;\theta) \frac{\partial}{\partial\theta^i} \log p(x;\theta) d\nu(x) = \\ &= \int_X p(x;\theta) \frac{1}{p(x;\theta)} \frac{\partial}{\partial\theta^i} p(x;\theta) d\nu(x) = \\ &= \frac{\partial}{\partial\theta^i} \int_X p(x;\theta) d\nu(x) = \\ &= \frac{\partial}{\partial\theta^i} 1 = 0. \end{aligned} \tag{4.36}$$

For each $\theta \in \Theta$, we define the *Fisher information matrix* $G(\theta) \equiv [g_{ij}(\theta)]$ as the covariance matrix of the scores,

$$\begin{aligned} g_{ij}(\theta) &= E_\theta[s^i_\theta s^j_\theta] = E_\theta[\partial_i \ell_\theta \partial_j \ell_\theta] = \\ &= \int_X p(x;\theta) \left(\frac{\partial}{\partial \theta^i} \log p(x;\theta)\right) \left(\frac{\partial}{\partial \theta^j} \log p(x;\theta)\right) d\nu(x) = \\ &= -E_\theta[\partial_i \partial_j \ell_\theta], \end{aligned} \tag{4.37}$$

where the last equality may be obtained by adding $0 = \frac{\partial^2}{\partial \theta^i \partial \theta^j} \int_X p(x,\theta) d\nu(x)$, interchanging the order of the integration and differentiation, and using the rule for the derivative of products.

The Fisher information matrix $G(\theta)$ is positive semidefinite for all $\theta \in \Theta$, and positive definite if the scores $\{s^i_\theta\}$ are linearly independent when seen as functions on X. If we further assume that each $g_{ij} : \Theta \to \mathbb{R}$ is C^∞, we may define the inner product of the natural basis of the coordinate system $[\theta^i]$ by $g_{ij} = \langle \partial_i, \partial_j \rangle$, leading to a Riemannian metric $g : p_\theta \mapsto \langle , \rangle_\theta$ in S which is called the *Fisher metric*:

$$\begin{aligned} \langle X, Y \rangle_\theta &= \left\langle X^i \partial_i, Y^j \partial_j \right\rangle_\theta = X^i Y^j \langle \partial_i, \partial_j \rangle_\theta = \\ &= X^i Y^j g_{ij} = X^i Y^j E_\theta[(\partial_i \cdot \ell)(\partial_j \cdot \ell)], \end{aligned} \tag{4.38}$$

defined for any tangent vectors $X, Y \in T_{p_\theta} S$ with coordinates $X = X^i \partial_i$ and $Y = Y^i \partial_i$ with respect to the natural basis. Another way to write the previous equation is:

$$\langle X, Y \rangle_\theta = E_\theta[(X \cdot \ell)(Y \cdot \ell)] \tag{4.39}$$

which emphasizes the characterization of tangent vectors as derivative operators.

4.2.5 The embedding curvature

We now give some conditions for a parametric family of probability measures $S = \{P_\theta\}$ to be an exponential family. Suppose that S is a submanifold of the affine space of positive measures M. As seen above, if we choose an origin ν for M, we may associate each measure to the loglikelihood of its density, and M becomes a vector space. Moreover, for each point $p \in S$ we may identify the tangent space $T_p S$ as an affine subspace of $T_p M$ spanned by the score vectors

$$\ell_i(p) \equiv \frac{\partial \ell}{\partial \theta^i}(p) \tag{4.40}$$

for $i = 1, \ldots, r$ (see [MR93] for more details). Then, a necessary and sufficient condition for S to be affine is that the partial derivatives of the scores (i.e. the second derivatives of the log-likelihood)

$$\ell_{ij}(p) \equiv \frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(p) \tag{4.41}$$

lie in the span of the scores $\ell_1, \ldots, \ell_r$ and the constant random variables. If we consider the family $\tilde{S} = \{\exp(\lambda)p : \lambda \in \mathbb{R}, p \in S\}$ obtainable from S by considering scalings of its measures, we may parametrize $\tilde{S}$ with the same parameters as S, $\theta_1, \ldots, \theta_r$, plus an extra parameter $\theta_{r+1} = \lambda$ to account for the scaling factors. Then the loglikelihoods of the measures in $\tilde{S}$ up-to-scale are the same as the loglikelihoods of the distributions in S, up to the addition of a constant random variable, i.e., $\tilde{\ell}(\exp(\lambda)p) = \ell(p) + \lambda$. Hence $\tilde{\ell}_i = \ell_i$ for $i = 1, \ldots, r$, and $\tilde{\ell}_{r+1} = 1$. So the previous condition may be stated "S is an exponential family if and only if each ℓ_{ij} (or $\tilde{\ell_{ij}}$) lies in the span of the scores $\tilde{\ell}_i$." If we introduce the Fisher metric g in $\tilde{S}$ to define the inner product in $T_p\tilde{S}$, another way to state this result is that "S is an exponential family if and only if the component of each ℓ_{ij} normal to $T_p\tilde{S}$ vanishes;" this normal component is denoted α_{ij} and called the *second fundamental form* [MR93] or *embedding curvature* [AN01]; it is given by

$$\alpha_{ij} = \ell_{ij} - g^{mn}E(\ell_{ij}\ell_m)\ell_n - E(\ell_{ij}), \tag{4.42}$$

so we have that S is an exponential family if and only if $\alpha_{ij} = 0$ for each $i, j = 1, \ldots, r$.

4.3 Finite non-parametric spaces

Let's return to the scenario of the examples 4.4 and 4.5, where we assume that the event space $X = \{x_1 \ldots, x_{n+1}\}$ is a finite set. Then the set of probability distributions over X is the well-known probability simplex

$$\overline{\mathbb{P}^n} = \{p : X \to \mathbb{R}^{n+1} : p(x_i) \equiv \theta_i \geq 0, \sum_{i=1}^{n+1} \theta_i = 1\}. \tag{4.43}$$

This is not a manifold, but merely a *manifold with corners* [Leb05]. We may however consider its interior $\mathbb{P}^n$, whose points are the strictly positive probability distributions over X, and that is called the open probability simplex (cf. Example 4.2). We have seen above that $\mathbb{P}^n$ is actually a differentiable manifold. It can be characterized extrinsically as a submanifold of $\mathbb{R}^{n+1}$ through (4.4), or intrinsically through the parameter space expressed in (4.23). Such a probability model is called the *finite non-parametric model*. It is often considered in nonparametric techniques of density estimation, for example in the method of Parzen windows.

We now consider the extrinsic representation provided in Example 4.2. The loglikelihood and its first and second derivatives are given by

$$\ell(\theta) = \log p(x, \theta) = \sum_{i=1}^{n+1} x_i \log \theta^i, \tag{4.44}$$

$$\partial_i \ell(\theta) = \frac{\partial}{\partial \theta^i} \log p(x, \theta) = \frac{x_i}{\theta^i}, \tag{4.45}$$

$$\partial_{ij}\ell(\theta) = \frac{\partial^2}{\partial\theta^i\partial\theta^j}\log p(x,\theta) = -\frac{x_i}{(\theta^i)^2}\delta_i^j, \tag{4.46}$$

and it follows that the Fisher information metric $\mathcal{J}_\theta$ on $\mathbb{P}^n$ is given by:

$$\begin{aligned}\mathcal{J}_\theta(u,v) = \langle u,v\rangle_\theta &= -\sum_{i=1}^{n+1}\sum_{j=1}^{n+1} u^i v^j E_\theta\left(\frac{\partial^2}{\partial\theta^i\partial\theta^j}\log p(x,\theta)\right) = \\ &= \sum_{i=1}^{n+1} u^i v^i E_\theta\left(\frac{x_i}{(\theta^i)^2}\right) = \\ &= \sum_{i=1}^{n+1}\frac{u^i v^i}{\theta_i},\end{aligned} \tag{4.47}$$

where $u, v \in T_p\mathbb{P}^n$ are represented by their coordinates in the natural basis of the embedding tangent space $T_p\mathbb{R}^{n+1}$, and where we used the fact that $E(x_i) = \theta^i$.

Notice that this parameterization of the open simplex $\mathbb{P}^n$ represents also the multinomial distribution, as shown in Example 4.5. The only difference is that the latter includes an additive term in the loglikelihood that is constant in θ (cf. (4.30)), and hence it does not affect the subsequent derivatives nor the expression for the Fisher metric. So, in what follows, we actually may regard this manifold either as the space of all positive distributions in a finite event space, or as the parametric space of the multinomial distribution.

Consider now the positive portion of the n-sphere of radius 2, whose extrinsic representation as a submanifold of $\mathbb{R}^{n+1}$ (cf. (4.2)) is

$$\mathbb{S}^{n,2}_{++} = \left\{\theta \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1}(\theta^i)^2 = 4, \theta^i > 0\right\}, \tag{4.48}$$

and the C^∞ map between manifolds, $f : \mathbb{P}^n \to \mathbb{S}^{n,2}_{++}$, defined in terms of the Euclidean coordinates of $\mathbb{R}^{n+1}$ as:

$$\overline{f}(\theta^1,\ldots,\theta^{n+1}) = (2\sqrt{\theta^1},\ldots,2\sqrt{\theta^{n+1}}), \tag{4.49}$$

whose inverse is $f^{-1} : \mathbb{S}^{n,2}_{++} \to \mathbb{P}^n$ given by

$$\overline{f^{-1}}(\eta^1,\ldots,\eta^{n+1}) = \left(\frac{(\eta^1)^2}{4},\ldots,\frac{(\eta^{n+1})^2}{4}\right). \tag{4.50}$$

Let p be an arbitrary point in $\mathbb{P}^n$ that is mapped to $f(p)$ in $\mathbb{S}^{n,2}_{++}$. The tangent map of f^{-1} at $f(p)$, $f^{-1}_* : T_{f(p)}\mathbb{S}^{n,2}_{++} \to T_p\mathbb{P}^n$, may be obtained by calculating the Jacobian matrix of the coordinated version of f^{-1} (cf. (4.12)),

considering the two manifolds embedded in $\mathbb{R}^{n+1}$; it yields:

$$\begin{aligned} f_*^{-1}(u) &= Ju = \sum_{j=1}^{n+1} u^j \frac{\partial \overline{f_i^{-1}}}{\partial \eta^j} = \\ &= \sum_{j=1}^{n+1} u^j \frac{\eta^i}{2} \delta_j^i = \\ &= \left(\frac{u^1 \eta^1}{2}, \ldots, \frac{u^{n+1} \eta^{n+1}}{2} \right). \end{aligned} \tag{4.51}$$

We can now pullback the Fisher metric on $\mathbb{P}^n$ to $\mathbb{S}_{++}^{n,2}$ through f^{-1}; it yields:

$$\begin{aligned} \langle u, v \rangle_{f(p)} &= \mathcal{J}_p(f_*^{-1}u, f_*^{-1}v) = \\ &= \mathcal{J}_{\frac{(\theta^i)^2}{4}} \left(\left(\frac{u^i \eta^i}{2} \right)_{1 \leq i \leq n+1}, \left(\frac{v^j \eta^j}{2} \right)_{1 \leq i \leq n+1} \right) = \\ &= \sum_{i=1}^{n+1} \frac{u^i \eta^i}{2} \frac{v^i \eta^i}{2} \frac{4}{(\theta^i)^2} = \\ &= \sum_{i=1}^{n+1} u_i v_i = \delta_{f(p)}, \end{aligned} \tag{4.52}$$

which is the Euclidean metric δ on $\mathbb{S}_{++}^{n,2}$ inherited from the embedding Euclidean space $\mathbb{R}^{n+1}$. This makes $f : (\mathbb{P}^n, \mathcal{J}) \to (\mathbb{S}_{++}^{n,2}, \delta)$ an isometry, i.e.,

Proposition 4.6 *The probability simplex $\mathbb{P}^n$ endowed with the Fisher metric is isometric to the positive orthant of the sphere, $\mathbb{S}_{++}^{n,2}$, endowed with the usual Euclidean metric.*

This means that the geodesics in the Riemannian manifold $(\mathbb{P}^n, \mathcal{J})$ correspond to arcs of great circles in $\mathbb{S}_{++}^{n,2}$ via the mapping $f : \mathbb{P}^n \to \mathbb{S}_{++}^{n,2}$ (see Fig. 6). Hence we are able to calculate the geodesic distance between two finite nonparametric or multinomial distributions with probability vectors $\theta = (\theta^1, \ldots, \theta^{n+1})$ and $\theta' = ((\theta')^1, \ldots, (\theta')^{n+1})$ via:

$$d_{\mathcal{J}}(\theta, \theta') = d_\delta(f(\theta), f(\theta')) = 2 \arccos \left(\sum_{i=1}^{n+1} \sqrt{\theta^i \cdot (\theta')^i} \right). \tag{4.53}$$

We will see in the next section how this geodesic distance induces generative kernels defined in the open simplex.

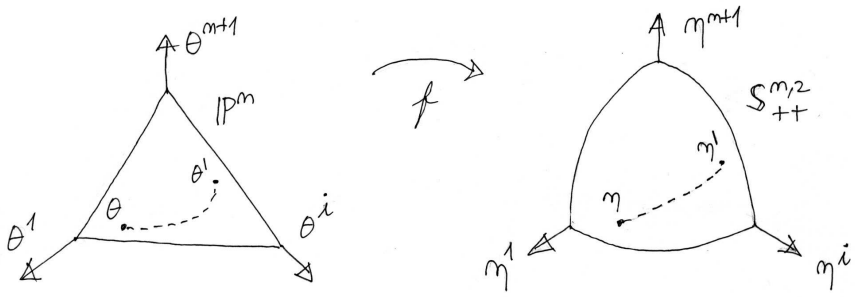


Figure 6: Illustration of Prop. 4.6.

5 Generative kernels

In this section we discuss an important class of kernels that have in common the property of having been designed through some modeling on the data generation; for this reason, they are termed "generative". This is an informal designation that has more to do with the process by which the kernel is devised than with the kernel function itself. In fact, we are going to see that some kernels derived in generative approaches reduce to the classic polynomial or Gaussian kernels that are typically addressed in "nongenerative" approaches.

As a motivation, consider for example a binary classification problem, where the data lies in a input space X and the label set is $Y = \{-1, +1\}$. A pure generative approach starts by using the training data to estimate the conditional densities $p(x|-1)$ and $p(x|+1)$, where $x \in X$, from which it builds a classifier, for example using a "frequentist" point of view, via the maximum likelihood, $\hat{y}_{\mathrm{ML}}(x) = \arg\max_{y \in Y} p(x|y)$, or a "Bayesian" one, via the maximum a posteriori (MAP) $\hat{y}_{\mathrm{MAP}}(x) = \arg\max_{y \in Y} p(x|y)p(y)$, based on some prior belief (e.g. empirical) about the label distribution $p(y)$. By contrast, a kernel-based discriminative approach like a support vector classifier starts by properly *choosing a kernel*, or equivalently a feature map $\Phi : X \to F$, conceptually embeds the data in F and uses the training data to find a linear discriminant in the feature space. Of course, the overall performance will depend strongly on the choice of the kernel: this choice should reflect the prior knowledge about how the data was generated. Lately, there have been many approaches to automatically *learn* the kernel from data (for instance, [LCB+04]). But even in those there is always a prior step where at least a suitable family of kernels must be chosen.

Generative kernels try to combine the advantages of both generative and discriminative approaches by introducing generative models on data and use them to devise a kernel. This can be done in many ways: (i) mapping data to points in a probability space and devising a kernel between probability distributions, (ii) considering a fixed probability distribution and study how

does it "fit" each data point, (iii) assuming that there is some hidden model that governs the data generation and marginalizing with respect to this model, etc. We next present some generative kernels that make use of these various possibilities.

5.1 Marginalization kernels

Marginalization kernels are described in some detail in [STC04], where some applications to hidden Markov models (HMMs) and other graphical models are devised. The idea is to consider a model class M, that we suppose to be discrete (which is the typical scenario where marginalization kernels arise), albeit a generalization for the continuous case is straightforward. Assume that there is a hidden model $m \in M$ governing the generation of data, such that data are conditionally independent given the model, i.e., for $x, y \in X$, $p(x, y|m) = p(x|m)p(y|m)$. If we have a prior $P(m)$ defined on the model class, we can marginalize to obtain:

$$p(x, y) = \sum_{m \in M} p(x|m)p(y|m)P(m) \tag{5.1}$$

and this defines a (sort of naïve Bayes) positive definite kernel on X, since it is a convex combination of kernels of the form $f(x) \cdot f(y)$.

As an example, consider a finite alphabet Σ and let $X = \Sigma^n$ (the set of strings of characters in Σ with length n). If we assume that each string is generated by a HMM with hidden states $h_1, \ldots, h_n \in H$, we may consider $M = H^n$, i.e., each model as a sequence of hidden states. The marginalization is then carried over all possible sequences of hidden states, and the Markov property implies that $P(m) \equiv P(h_1, \ldots, h_n) = P(h_1) \prod_{i=2}^{n} P(h_i|h_{i-1})$, so

$$p(x, y) = \sum_{h \in H^n} \prod_{i=1}^{n} p(x_i|h_i)p(y_i|h_i)P(h_i|h_{i-1}), \tag{5.2}$$

where by convention $P(h_1|h_0) \equiv P(h_1)$. This reasoning may be generalized for other graphical models, so we conclude that in those cases we can treat the joint distribution $p(x, y)$ as a positive definite kernel. This makes applicable a lot of discriminative learning techniques like the support vector machines.

Notice that this kernel was devised from a generative perspective, constrasting to other kernels that are usually applied to discrete data, like the whole class of convolution kernels described in [Hau99].

5.2 The Fisher kernel

The Fisher kernel was one of the earliest purposes of generative kernels, introduced in [JH98]. Consider a parametric family of μ-absolutely continuous

probability distributions on $(X, \mathscr{M}, \mu)$,

$$M_\Theta = \{P_\theta \in M_+^1(X) : \theta \in \Theta\}, \tag{5.3}$$

where $\Theta \subseteq \mathbb{R}^n$, which is supposed to be a differentiable manifold, and can be given a Riemannian structure if we endow it with the Fisher metric $\mathcal{J}$, via the Fisher information matrix (recall (4.37))

$$G(\theta) = [g_{ij}(\theta)] = E_\theta(s_\theta \cdot s_\theta^T), \tag{5.4}$$

where $s_\theta = \nabla_\theta \ell_\theta$ is the score vector, $\ell_\theta = \log p_\theta$ is the loglikelihood and $p_\theta = \frac{dP_\theta}{d\mu}$ is the density associated with the distribution P_θ.

Now fix a parameter setting $\theta \in \Theta$. This is the same as choosing a particular point P_θ in the statistical manifold. We may study how a data point $x \in X$ makes the loglikelihood function $\log p_\theta(x)$ vary in the neighborhood of the parameter setting θ. Specifically, this can be done by computing the score vector at θ,

$$s_\theta(x) = \nabla_\theta \log p_\theta(x), \tag{5.5}$$

which it is the gradient of the loglikelihood $\ell_\theta(x)$ and so corresponds to its steepest ascent direction at θ. Suppose now that we have two data points $x, y \in X$ whose kernel we wish to compute. We can calculate the two score vectors $s_\theta(x)$ and $s_\theta(y)$ and compare them. Geometrically, these are the two vectors in the tangent space $T_{P_\theta} M_\Theta$ that correspond to the steepest ascent directions of the loglikelihood functions $\ell_\theta(x)$ and $\ell_\theta(y)$. However, M_Θ is in general a curved manifold, and so, as pointed out in [Ama98], the natural basis of $T_{P_\theta} M_\Theta$ is not necessarily orthonormed. We should use instead of ∇_θ the *natural gradient* $\tilde{\nabla}_\theta$, that depends on the local Riemannian metric (in this case, given by the Fisher information) in the following way:

$$\tilde{\nabla}_\theta \ell_\theta = G(\theta)^{-1} \nabla_\theta \ell_\theta. \tag{5.6}$$

Notice that under the Euclidean metric, $\tilde{\nabla}_\theta = \nabla_\theta$. Geometrically, the natural gradient corresponds to the direction of steepest ascent *along the manifold*, i.e. the direction that maximizes the loglikelihood function while traversing the minimum distance in the manifold, being the distance given by the Riemannian metric (cf. (4.11)). This is an intrinsic notion, i.e. it does not depend on the chosen coordinate system. Next, to define a positive kernel between x and y, we may consider the natural gradient mapping at θ,

$$\begin{array}{rcl} \Phi : X & \to & T_{P_\theta} M_\Theta \\ x & \mapsto & \Phi(x) = \tilde{\nabla}_\theta \ell_\theta(x), \end{array} \tag{5.7}$$

and, again, use the local inner product in $T_{P_\theta} M_\Theta$ given by Riemannian

metric. This yields the *Fisher kernel*:

$$\begin{aligned}
\kappa(x,y) &= \mathcal{J}_\theta(\Phi(x),\Phi(y)) = \\
&= (\tilde{\nabla}_\theta \ell_\theta(x))^T G(\theta) \tilde{\nabla}_\theta \ell_\theta(y) = \\
&= (\nabla_\theta \ell_\theta(x))^T G(\theta)^{-1} G(\theta) G(\theta)^{-1} \nabla_\theta \ell_\theta(y) = \\
&= (\nabla_\theta \ell_\theta(x))^T G(\theta)^{-1} \nabla_\theta \ell_\theta(y) = \\
&= (s_\theta(x))^T G(\theta)^{-1} s_\theta(y). \qquad (5.8)
\end{aligned}$$

By construction, the Fisher kernel is independent of the choice of the coordinate system. Notice that the matrix $G(\theta)$ depends only on the parameter setting and not on the data points; however, albeit in particular cases it admits a closed form expression, in general it is difficult to compute, and this is a drawback for practical applications. It is immediate, however, that we may define related kernels

$$\kappa(x,y) = (s_\theta(x))^T K s_\theta(y), \qquad (5.9)$$

where K is any positive definite matrix. They are still based in the natural gradient, but unlike the actual Fisher kernel they depend on the parameterization of M_Θ. A wide used version is the "practical Fisher kernel", where K is set to the identity matrix.

The choice of the parameter setting θ is a very important issue in the design of a Fisher kernel[11]. For instance, typical points $x \in X$, i.e., points with high likelihood $\ell_\theta(x)$, will have a small norm $\kappa(x,x)$ since their derivatives are small, while atypical points $y \in X$ may have large derivatives and hence a high norm $\kappa(y,y)$. This effect may be dangerous, since the kernel between similar typical points may be lower than the kernel between very different atypical points. This may be overcome by normalizing the kernel.

Applications of the Fisher kernel to HMMs are given in [STC04].

5.3 Probability product kernels

Although they have a "generative" inspiration, neither the marginalization kernel or the Fisher kernel, described in the previous sections, are directly defined in a probability space. In the case of the Fisher kernel, a particular probability distribution is chosen and kept fixed. In the case of the marginalization kernels, the marginalization in (5.1) is performed making the hidden model vary, but assigning at each time equal versions of the model to both x and y. We now describe a different framework, where data points in X are mapped to probability distributions in a parametric family M_Θ as in (5.3), and a "probability kernel" κ_M is devised in this space (or equivalently, in the space of the corresponding densities).

[11] Why not marginalizing over several values of θ, using a prior $P(\theta)$?

So there are two separate problems: (i) choosing a map $f : X \to M_\Theta$, and (ii) devising a kernel in $M_\Theta \times M_\Theta$. Let's focus on the first problem, i.e., *how to fit a density on an individual datum point*? Notice that this is very different from the usual density estimation problem, where we suppose that many data points are available. Although fitting a density to a single point is not a very interesting approach for estimation purposes, here it is only an intermediate step to devise a kernel. It has the advantage over nongenerative kernels that our choice of the parametric family M_Θ is a good opportunity to reflect our prior knowledge about the data generation. The most obvious choice for the map f is the maximum likelihood estimation

$$x \mapsto p_{\hat{\theta}(x)}, \quad \hat{\theta}(x) \equiv \hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta\in\Theta} p_\theta(x) \tag{5.10}$$

that leads to

$$\kappa(x, y) = \kappa_M \left(p_{\hat{\theta}(x)}, p_{\hat{\theta}(y)} \right). \tag{5.11}$$

If we consider a prior $\pi(\theta)$ on the parameter family, we may use instead the maximum a posteriori estimate

$$\hat{\theta}(x) \equiv \hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta\in\Theta} p_\theta(x)\pi(\theta). \tag{5.12}$$

An alternative "Bayesian-like" strategy is considering the conditional density of the parameters on the datum,

$$p(\theta|x) = \frac{1}{Z(x)} p_\theta(x)\pi(\theta), \tag{5.13}$$

where $Z(x) = \int_\Theta p_\theta(x)\pi(\theta)d\theta$ is a normalizing factor, and taking the true posterior,

$$\begin{aligned} x \mapsto p_x(.) &\equiv \int_\Theta p_\theta(.)p(\theta|x)d\theta = \\ &= \int_\Theta p_\theta(.)\frac{1}{Z(x)}p_\theta(x)\pi(\theta)d\theta = \\ &= \frac{\int_\Theta p_\theta(.)p_\theta(x)\pi(\theta)d\theta}{\int_\Theta p_\theta(x)\pi(\theta)d\theta}. \end{aligned} \tag{5.14}$$

Yet another alternative (also "Bayesian") to devise the kernel in X is to consider the parameters as random variables with the conditional density (5.13) and take the posterior mean:

$$\kappa(x, y) = \int_{\Theta\times\Theta} \kappa_M(p_\theta, p_\xi)p(\theta|x)p(\xi|y)d\theta d\xi. \tag{5.15}$$

This and the subsequent sections explore several choices for the probability kernels $\kappa_M : M_\Theta \times M_\Theta$ themselves.

The simplest way to define a probability kernel is to restrain ourselves to the set of densities that are $L^2(X)$-integrable and consider the standard inner product $\langle p, q\rangle = \int_X p(x)q(x)d\mu(x)$. More generally, we may define a family of *probability product kernels* [JKH04] parameterized by $\alpha > 0$ as kernels in $\frac{d}{d\mu}M_\Theta \cap L^{2\alpha}(X)$ defined by

$$\kappa_\alpha(p, q) = \langle p^\alpha, q^\alpha\rangle = \int_X p(x)^\alpha q(x)^\alpha d\mu(x). \tag{5.16}$$

For $\alpha = \frac{1}{2}$ this is called the *Bhattacharyya kernel*, and for $\alpha = 1$ this is the *expected likelihood kernel*, since it becomes the expectation of one distribution under the other. The Bhattacharyya kernel

$$\kappa_{1/2}(p, q) = \int_X \sqrt{p(x)}\sqrt{q(x)}d\mu(x) \tag{5.17}$$

is known in the statistics literature as the "Bhattacharyya's affinity" between distributions, which relates to the Hellinger's distance

$$H(p, q) = \frac{1}{2}\int_X \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x) \tag{5.18}$$

by $H(p, q) = \sqrt{2 - 2\kappa_{1/2}(p, q)}$; there are also interesting relationships between the Hellinger's distance and other divergence measures, as the Kullback-Leibler or the Jensen-Shannon divergences (see [Top00] for more details). An important property of the Bhattacharyya kernel is that the feature map associated to it maps densities to the unit sphere, i.e., $\kappa_{1/2}|_\Delta = 1$. As pointed out in [JKH04], it turns out that the Bhattacharyya kernel can be computed in closed form for any exponential family (see (4.17)); in fact, if

$$p_\theta(x)d\mu = \exp(\theta^T f(x) - K(\theta))d\nu \tag{5.19}$$

and

$$p_\xi(x)d\mu = \exp(\xi^T f(x) - K(\xi))d\nu \tag{5.20}$$

are two probability distributions belonging to an exponential family, we have that

$$\begin{aligned}
\kappa_{1/2}(p_\theta, p_\xi) &= \int_X \sqrt{p_\theta(x)}\sqrt{p_\xi(x)}d\mu(x) = \\
&= \int_X \exp\left(\left(\frac{\theta+\xi}{2}\right)^T f(x) - \frac{K(\theta)+K(\xi)}{2}\right) d\nu(x) = \\
&= \exp\left(K\left(\frac{\theta+\xi}{2}\right) - \frac{K(\theta)+K(\xi)}{2}\right),
\end{aligned} \tag{5.21}$$

where the last step is due to (4.18).

Example 5.1 (Multivariate normal with variance $\sigma^2 I$.) *For a multivariate normal $N(\mu, \sigma^2 I)$ we have*

$$\theta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right), \tag{5.22}$$

$$f(x) = (x, \|x\|^2), \tag{5.23}$$

and

$$K(\theta) = \log((2\pi)^{k/2}\sigma) + \frac{\|\mu\|^2}{2\sigma^2}. \tag{5.24}$$

So if $p_\theta = N(\mu_1, \sigma^2 I)$ and $p_\xi = N(\mu_2, \sigma^2 I)$, we have

$$\begin{aligned} \kappa_{1/2}(p_\theta, p_\xi) &= \exp\left(K\left(\frac{\theta+\xi}{2}\right) - \frac{K(\theta)+K(\xi)}{2} \right) = \\ &= \exp\left(\frac{\|\mu_1+\mu_2\|^2}{8\sigma^2} - \frac{\|\mu_1\|^2 + \|\mu_2\|^2}{4\sigma^2} \right) = \\ &= \exp\left(-\frac{\|\mu_1-\mu_2\|^2}{8\sigma^2} \right). \end{aligned} \tag{5.25}$$

If we use the maximum likelihood estimation to fit a density of this family to each point $x \in \mathbb{R}^k$, we get the map $x \mapsto N(x, \sigma^2 I)$, and the resulting kernel κ in $\mathbb{R}^k$ obtained from the Bhattacharyya probability kernel is simply the Gaussian kernel with variance $4\sigma^2$, $\kappa(x,y) = \exp\left(-\frac{\|x-y\|^2}{8\sigma^2}\right)$. It is further shown in [JKH04] that also the expected likelihood kernel κ_1 yields, up to a constant factor, a Gaussian kernel, with variance $2\sigma^2$ instead.

Example 5.2 (Multinomial family.) *Recall from* (4.32)-(4.33) *that the canonical parameters $(\theta^i)_{i=1}^{m-1}$ of the multinomial family relate to the symbol probabilities $(\beta^i)_{i=1}^m$ via*

$$\theta^i = \log \frac{\xi^i}{1 - \sum_{i=1}^{m-1} \beta^i} = \log \frac{\beta^i}{\beta^m}, \tag{5.26}$$

and the cumulant generating function is

$$K(\theta) = n \log\left(1 + \sum_{i=1}^{m-1} \exp(\theta^i) \right), \tag{5.27}$$

Hence, the Bhattacharyya kernel between two multinomials p_θ and p_ξ is

$$\begin{aligned}\kappa_{1/2}(p_\theta, p_\xi) &= \exp\left(K\left(\frac{\theta+\xi}{2}\right) - \frac{K(\theta)+K(\xi)}{2}\right) = \\ &= \frac{\left(1+\sum_{i=1}^{m-1}\exp\left(\frac{\theta^i+\xi^i}{2}\right)\right)^n}{\left(1+\sum_{i=1}^{m-1}\exp(\theta^i)\right)^{n/2}\left(1+\sum_{i=1}^{m-1}\exp(\xi^i)\right)^{n/2}} = \\ &= (\beta^m\gamma^m)^{n/2}\left(1+\sum_{i=1}^{m-1}\left(\frac{\beta^i\gamma^i}{\beta^m\gamma^m}\right)^{1/2}\right)^n = \\ &= \left(\sum_{i=1}^{m}\left(\beta^i\gamma^i\right)^{1/2}\right)^n, \end{aligned} \tag{5.28}$$

where $(\beta^i)_{i=1}^m$ and $(\gamma^i)_{i=1}^m$ are the respective symbol probabilities. If, again, we use the MLE to fit densities on the points x and y of $X = \Sigma^n$, we get $\hat{\beta}^i_{ML} = \frac{x_i}{n}$ and $\hat{\gamma}^i_{ML} = \frac{y_i}{n}$; the Bhattacharyya kernel between the multinomials is thus equivalent to the homogeneous polynomial kernel of degree n between the vectors $\left(\sqrt{\frac{x_i}{n}}\right)_{i=1}^n$ and $\left(\sqrt{\frac{y_i}{n}}\right)_{i=1}^n$ of square rooted symbol relative frequencies. When n is not constant, [JKH04] suggests summing over all its possible values, leading to:

$$\kappa(p_\theta, p_\xi) = \sum_{n=0}^{\infty}\left(\sum_{i=1}^{m}\left(\beta^i\gamma^i\right)^{1/2}\right)^n = \left(1-\sum_{i=1}^{m}\left(\beta^i\gamma^i\right)^{1/2}\right)^{-1}. \tag{5.29}$$

5.4 Kullback-Leibler kernel

The Kullback-Leibler kernel [MHV03] is one of the earliest generative kernels to be purposed that operate directly in the probability space. The idea is similar to that of probability product kernels, i.e., fitting to each data point $x \in X$ a μ-absolutely continuous probability distribution with density $p_x \in \frac{d}{d\mu}M_+^1(X)$, for example via the maximum likelihood estimation in a parametric family. The difference is that, instead of considering L^p-inner products of densities, a kernel is devised that uses the more natural Kullback-Leibler divergence $D(.||.)$ (cf. (2.17)). There are however some issues that need to be taken into account. Firstly, $D(.||.)$ is not symmetric. This can be solved by using a symmetrized version $\tilde{D}(.||.)$ defined by $\tilde{D}(p||q) = D(p||q) + D(q||p)$. Secondly, even the symmetrized version fails to be a metric, and no natural way seems to exist that allows devising a positive kernel from it. The approach followed by [MHV03] is simply to define the kernel

$$\kappa(x,y) = \kappa_p(p_x, p_y) = \exp(-\alpha\tilde{D}(p_x||p_y) + \beta) \tag{5.30}$$

restricted to a finite set of data points in X, that by adjustment of the parameters α and β becomes positive definite.

For practical applications this is often harmless since many algorithms work only with a kernel matrix, for example in unsupervised or transductive learning. However, there are some "theoretical" problems when it is necessary to handle unseen data points, for example, in a inductive classifier.

5.5 The heat kernel

In [LL05, Leb05] a new class of kernels on statistical manifolds were introduced, called "information diffusion kernels". This followed the idea of "diffusion kernel" that had already appeared applied to discrete spaces as graphs [KL02]. These kernels have a strong physical interpretation that we sketch here. Again, we start by representing data as points in a statistical manifold $(M_\Theta, \mathcal{J})$, with M_Θ as in (5.3) and $\mathcal{J}$ being the corresponding Fisher metric. Then, a kernel is devised from a particular solution of the *heat diffusion equation* in the manifold. The idea is letting the value of the kernel $\kappa(p_x, p_y)$ express how information flows from p_x to p_y through the manifold. Its construction is based on the notion of *Laplacian* in a Riemannian manifold (M, g).

We first define the (natural) *gradient* as a map that transforms smooth functions into vector fields

$$\operatorname{grad} : \mathcal{F}(M) \to \mathcal{X}(M) \tag{5.31}$$

and satisfies $g_p(\operatorname{grad} f|_p, X_p) = X_p(f)$ for any $f \in \mathcal{F}(M)$ and any $p \in M$. Using local coordinates $[\theta_i]$,

$$(\operatorname{grad} f|_p)_i = \sum_j g^{ij}(p)\partial_j f(p), \tag{5.32}$$

where $g^{ij}(p)$ denotes the (i,j)-entry of the matrix $G^{-1}(p)$ and $\partial_j \equiv \frac{\partial}{\partial \theta_j}$. Next, we define the *divergence* operator as a map that transforms vector fields into smooth functions

$$\operatorname{div} : \mathcal{X}(M) \to \mathcal{F}(M) \tag{5.33}$$

and is given in local coordinates by

$$\operatorname{div} X_p = \frac{1}{\sqrt{\det G(p)}} \sum_i \partial_i \left(\sqrt{\det G(p)} (X_p)_i \right), \tag{5.34}$$

where $\det G(p)$ denotes the determinant of the Gram matrix $G(p)$. These notions of gradient and divergence operators generalize the usual notions in Euclidean spaces, where the gradient may be interpreted as the steepest ascent direction, and the divergence is a measure of outflow minus inflow. So

does the following notion of *Laplacian* operator in the Riemannian manifold (M, g),

$$\Delta : \mathcal{F}(M) \to \mathcal{F}(M), \quad \Delta = \operatorname{div} \circ \operatorname{grad}, \tag{5.35}$$

which can be used to model how heat diffuses through the manifold via the *heat equation*

$$\frac{\partial f}{\partial t} - \Delta f = 0 \tag{5.36}$$

with initial conditions $f(x, 0) = f_0(x)$. Above $f(x, t)$ denotes the flow at point x and time t; the initial conditions are the heat distribution at time zero. The *heat kernel* $\kappa_t(x, y)$ is defined as the solution to the heat equation $f(x, t)$ with initial condition given by Dirac's delta function δ_y. By linearity of the heat equation, we have that the heat kernel generates the solution of the heat equation with arbitrary initial conditions, according to:

$$f(x, t) = \int_M \kappa_t(x, y) f_0(y). \tag{5.37}$$

In the Euclidean case, $(M, g) = (\mathbb{R}, \delta)$, the heat kernel reduces to the Gaussian kernel, $\kappa_t(x, y) = \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{(x-y)^2}{4t}\right)$. In [Leb05] further properties of the heat kernel are given. Unfortunately, for general manifolds there is no closed form solution for the heat kernel; this is case, for example, for the multinomial family that is considered in [Leb05], where the problem of text classification is addressed. The short time behaviour of the solutions can though be studied via the *parametrix expansion*. This yields approximating the heat kernel for the multinomial by

$$\kappa_t(\theta, \theta') = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{t} \arccos^2\left(\sum_{i=1}^{n+1} \sqrt{\theta^i \cdot (\theta')^i}\right)\right) \tag{5.38}$$

which however (since it is a mere approximation) is not guaranteed to be positive definite.

5.6 Negative geodesic distance kernel

In [ZCL05], a kernel that is very similar to (5.38) is proposed. Notice that it is not clear if (5.38) is positive definite or not. From (4.53), we see that it has the form of an exponentiated negative squared distance, where the distance is the geodesic distance $d_{\mathcal{J}}$ on the positive orthant of the sphere. This makes (5.38) a candidate, via Schoenberg's theorem, to be positive definite and even infinitely divisible. However, we have seen from Remark 3.24 that not any squared distance yields a negative definite kernel. A necessary and sufficient condition for (5.38) to be positive definite is the ability to embed the positive orthant of the sphere isometrically in some Hilbert space.

The kernel proposed in [ZCL05] derives from the "negative geodesic distance kernel", which is proved to be conditionally positive definite. This is

the same to say that the geodesic distance $d_{\mathcal{J}}$ itself (and not its square) is negative definite. We refer to [ZCL05] for a proof of this fact. Hence, we can exponentiate the negative geodesic distance and obtain a positive definite kernel, namely

$$\kappa_t(\theta, \theta') = \exp\left(-t \arccos\left(\sum_{i=1}^{n+1} \sqrt{\theta^i \cdot (\theta')^i}\right)\right), \tag{5.39}$$

for $t > 0$. Notice the similarity with the heat kernel (5.38), the only difference being the squared arc-cosine.

5.7 Jensen-Shannon kernel

The *Jensen-Shannon kernel* or entropy kernel between measures was introduced in [CV05, CFV05]. It has been devised above (see (3.34)) when discussing semigroup kernels. There we saw that albeit not being a semigroup kernel, it is semigroup-based, since it is the normalization of a semigroup kernel. The proof that it is positive definite is a consequence of the negative definiteness of the entropy function, and it puts in evidence a big amount of theoretic results on positive and negative kernels.

Unlike some other generative kernels that we have discussed, the Jensen-Shannon kernel is defined between measure densities (with respect to some dominating measure ν), and so its domain is more general than those probability kernels that are defined only in the space of probability densities.

In [CV05, CFV05] the problem of defining a kernel between structured objects (as images, text or sequences) that can be represented as "bags-of-components" was addressed. Suppose first that the set S of basic components is finite (for example, in text objects, S may be the set of words of a closed vocabulary). Assume that, given an object $x \in X$, we can calculate the weight of each component $s \in S$ on x, say $w_s(x)$. This could be, for example, the "number of times" that s appears in x. Then we can map x to a *molecular measure* $\mu \in \mathrm{Mol}_+(S)$ (see Def. 2.9), by making each component $s \in S$ correspond to a Dirac measure centered in s, and give it a weight $w_s(x)$, i.e.

$$\mu = \sum_{s \in S} w_s(x) \varepsilon_s. \tag{5.40}$$

Using the counting measure as base measure, we may associate to each molecular measure its density, which is simply the function $s \mapsto w_s(x)$. This enables us to use the Jensen-Shannon kernel between the molecular measures μ and μ' associated respectively to the objects x and x'. From the definition of Jensen-Shannon divergence, this is done by considering the mixture of μ and μ', which is a sort of "concatenation" of the components of x and x'. This emphasizes the "semigroup" formalism associated with this kernel.

However, if S is an infinite set, this approach is not applicable in general. This happens because the counting measure is not guaranteed to be σ-finite, and hence we cannot associate a density to each absolutely continuous measure. If we use other base measure instead (for example the Lebesgue measure for $S = \mathbb{R}^n$) then molecular measures may not be absolutely continuous with respect to that base measure. In [CFV05] it is suggested to overcome this problem by *smoothing* the molecular measure μ to a measure $\varphi(\mu) \in M_+^h(S)$ (see Sect. 2.3) via a *smoothing kernel* κ. This is done in a similar way as the Parzen window estimation procedure:

$$\begin{array}{rcl} \varphi : \mathrm{Mol}_+(S) & \rightarrow & M_+^h(S) \\ \mu \equiv p\ d\nu & \mapsto & \varphi(\mu) = \sum_{s \in \mathrm{supp}\,\mu} p(s)\kappa(s,.)\ d\nu \end{array} \tag{5.41}$$

For example, if $S = \mathbb{R}^n$ endowed with the Lebesgue measure, and κ is the Gaussian kernel on S, then molecular measures on S are smoothed to mixtures of Gaussians, and the resulting Jensen-Shannon kernel between two objects x and y will compare the entropy of the Gaussian mixtures of the components of x and y with the entropy of each individual Gaussian mixture.

5.8 Multiresolution kernels

Multiresolution kernels were introduced in [CF05]. These kernels are adequate for structured data such as text, images, or sequences, that have smaller components (resp. words, pixel intensity values, or characters). Typically these data are represented as "bags-of-components"; a more powerful representation considers a collection of *nested bags* (see Fig. 7) based on a prior knowledge on the data structure. This allows comparing two objects both in detailed perspectives, stressing local matches between smaller bags, or in a global one, using the entire bag.

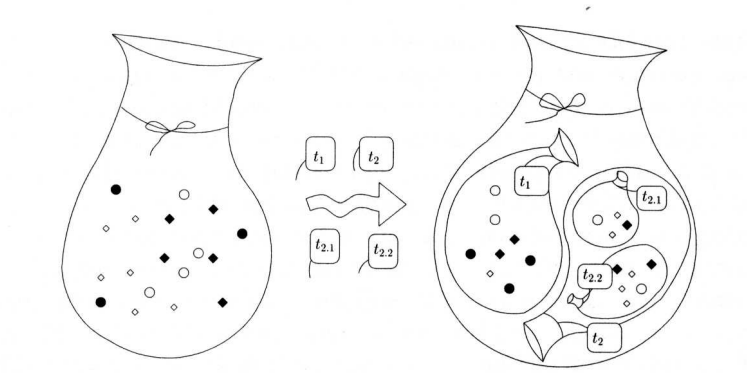


Figure 7: Nested bags of components (extracted from [CF05]).

Suppose, as above, that each point x in the input space X may be mapped to a probability density $p_x \in \frac{d}{d\nu} M_+^1(X)$, and let T be an arbitrary set of conditioning events that can be directly observed in the object (unlike the hidden models considered in the marginalizing kernels). We may decompose p_x as

$$p_x(.) = \sum_{t \in T} p(.|t)p(t) = \sum_{t \in T} \mu_t, \tag{5.42}$$

where $\mu_t \equiv p(.|t)p(t) \in \frac{d}{d\nu} M_+^{\leq 1}(X)$ is the density of a sub-probability measure on X (i.e. a measure whose total mass does not exceed 1). Objects can hence be represented as families of measures of $M_+^{\leq 1}(X)$ indexed by T, i.e., as elements $\mu \equiv (\mu_t)_{t \in T}$ in

$$M_T(X) \equiv \left(M_+^{\leq 1}(X) \right)^T. \tag{5.43}$$

For example, if the objects are images, each event t could correspond to a small region, and T could be a partition of the image in distinct regions. Each image object would then be represented as a family of histograms, one for each region. If, instead, the objects are strings, each event t could be a n_t-gram context, and T could be a context tree. Each string object could then be represented as a family of histograms, one for each context, counting the relative frequencies of the subsequent character.

To achieve the multiresolution kernel, we must first define a family $(\kappa_t)_{t \in T}$ of kernels that measure the object similarity with respect to each event t,

$$\kappa_t(\mu, \mu') \equiv \kappa(\mu_t, \mu'_t), \tag{5.44}$$

where κ is a predefined kernel. If the objects are "structured", it seems reasonable to suppose that some events are similar while others are not. If two events s and t are considered similar (for example two neighboring regions in an image), we may consider a unique event $\{s, t\}$ and a corresponding kernel $\kappa_{\{s,t\}}(\mu, \mu') \equiv \kappa(\mu_s + \mu_t, \mu'_s + \mu'_t)$. More generally, if $T_0 \subseteq T$ is a set of similar events, we may define

$$\kappa_{T_0}(\mu, \mu') \equiv \kappa \left(\sum_{t \in T_0} \mu_t, \sum_{t \in T_0} \mu'_t \right). \tag{5.45}$$

Consider now a finite partition P of T, i.e., a set $P = \{T_1, \ldots, T_n\}$ of disjoint subsets of T that cover T. We define the kernel κ_P induced by the partition P as

$$\kappa_P(\mu, \mu') \equiv \prod_{i=1}^{n} \kappa_{T_i}(\mu, \mu'). \tag{5.46}$$

Such a partition reflects a belief on how the events in T should be associated or dissociated to highlight component similarities or local dissimilarities. Instead of considering the set of all possible partitions of T, which

would be computationally prohibitive, [CF05] suggests obtaining partitions by assuming the existence of a prior hierarchical information on the events in T. This can be made through a "hierarchical" family $H \equiv (P_d)_{d=0}^D$ of partitions

$$P_0 = \{T\}, \ldots, P_D = \{\{t\}_{t \in T}\} \tag{5.47}$$

that has the property that any subset in a partition P_d is included in a (unique by definition of partition) subset of the coarser partition P_{d-1} (see Fig. 8). We assume further that this inclusion is strict. This means that each set $T_i \in P_{d-1}$ is itself "partitioned" into more than one subset in P_d. The multiresolution approach then considers partitions defined by sets in the different hierarchical levels of H. To do this, all the sets in $\bigcup_{d=0}^D P_d$ are taken into account to form the set $\mathscr{P}_H$ of all the partitions that can be built with them. The *multiresolution kernel* is then defined through a prior measure π on $\mathscr{P}_H$ as:

$$\kappa_\pi(\mu, \mu') = \sum_{P \in \mathscr{P}_H} \pi(P) \kappa_P(\mu, \mu'). \tag{5.48}$$

Tracing back to (5.44), we see that κ_π is positive definite whenever the predefined base kernel κ is. The latter may be any kernel that is defined in the space of subprobability measures, for example the Jensen-Shannon kernel (3.34). In [CF05] a suggestion is given to generate the prior measure π through a branching process; further issues concerning the efficient computation of the multiresolution kernel are also provided in this reference.

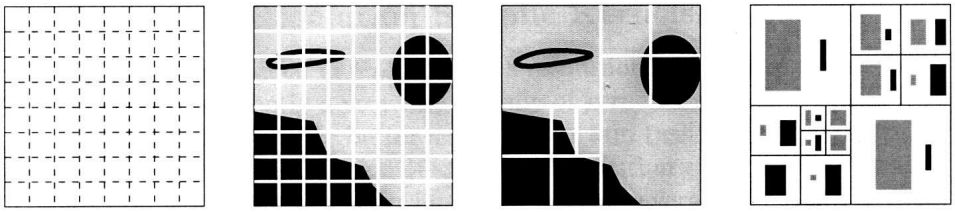


Figure 8: A hierarchy of partitions, and an example of a multiresolution partition (extracted from [CF05]).

5.9 Sequential document representations

We conclude this section by mentioning some recent work that concerns document representation. There has been several approaches that extend the traditional *bag-of-words representation* [SWY75], where each document is represented as a sparse vector in a very large Euclidean space $\mathbb{R}^{|V|}$, with $V = \{w_1, \ldots, w_{|V|}\}$ being a vocabulary of words. A well-known approach for text classification that uses the bag-of-words representation is *latent semantic analysis*. Basically, this method considers a document collection

$C = \{d_1, \ldots, d_n\}$ represented as bags-of-words, and finds a linear subspace[12] of dimension $k \leq n$ where documents may be approximately represented by their projections. This can be done, for example, via a SVD decomposition of the words-by-documents matrix, and keeping only the k largest singular values.

More recent works [Gou99, HH00] generalize this geometrical idea to the manifold of multinomial families: instead of finding a "linear subspace", as in the Euclidean case, they learn a *curved multinomial subfamily* (i.e. a submanifold of the multinomial family, which, unlike the latter, may be no longer affine). Several approaches are considered: for example, if the probability simplex $\mathbb{P}^{|V|-1}$ is represented extrinsically in the ambient space $\mathbb{R}^{|V|}$ we may consider subfamilies formed by intersecting the simplex with an affine subspace of $\mathbb{R}^{|V|}$[13]; alternatively, if we represent the Riemannian manifold of the multinomial family endowed with the Fisher metric as the positive orthant of the $|V|$-sphere with the Euclidean metric, we may consider *spherical subfamilies* [Gou99] which are lower dimensional spheres embedded in the $|V|$-sphere. Otherwise, we may represent $\mathbb{P}^{|V|-1}$ intrinsically and consider *exponential subfamilies*, i.e., subfamilies that are affine in the information geometric sense.

To illustrate this idea, [Gou99] splits a book (Machiavelli's *The Prince*) in several text blocks, its numbered pages, considers each page as a point in the multinomial simplex, learns a 2-dimensional subspace (an extreme example of dimension reduction) and projects each page in this subspace. The result is the representation of the book as a sequential path in $\mathbb{R}^2$, tracking the evolution of the subject matter of the book over the course of its pages (see Fig. 9).

Inspired by this sort of document representations, more recent work [Leb06] suggested a sequential representation of documents by *simplicial curves* (i.e. curves in the probability simplex), that is denoted as the *locally weighted bag-of-words* (lowbow) representation. According to this representation, a length-normalized document is a function $x : [0, 1] \times V \to \mathbb{R}_+$ such that

$$\sum_{w_j \in V} x(t, w_j) = 1, \quad \text{for any } t \in [0, 1]. \tag{5.49}$$

The pure sequential representation of the n-length document $(y_1, \ldots, y_n) \in V^n$ may be written using the above formalism as

$$x(t, w_j) = \varepsilon_{w_j}(y_{\lceil tn \rceil}) = \begin{cases} 1, & \text{if } w_j = y_{\lceil tn \rceil} \\ 0, & \text{if } w_j \neq y_{\lceil tn \rceil}, \end{cases} \tag{5.50}$$

[12] This procedure is known in machine learning as principal component analysis (PCA), and may be generalized to find a nonlinear subspace by using kernel methods (kPCA) [STC04].

[13] These are named "affine subfamilies" in [HH00] albeit this designation is misleading since they are not affine in the information geometrical sense (they are not in general an exponential family).

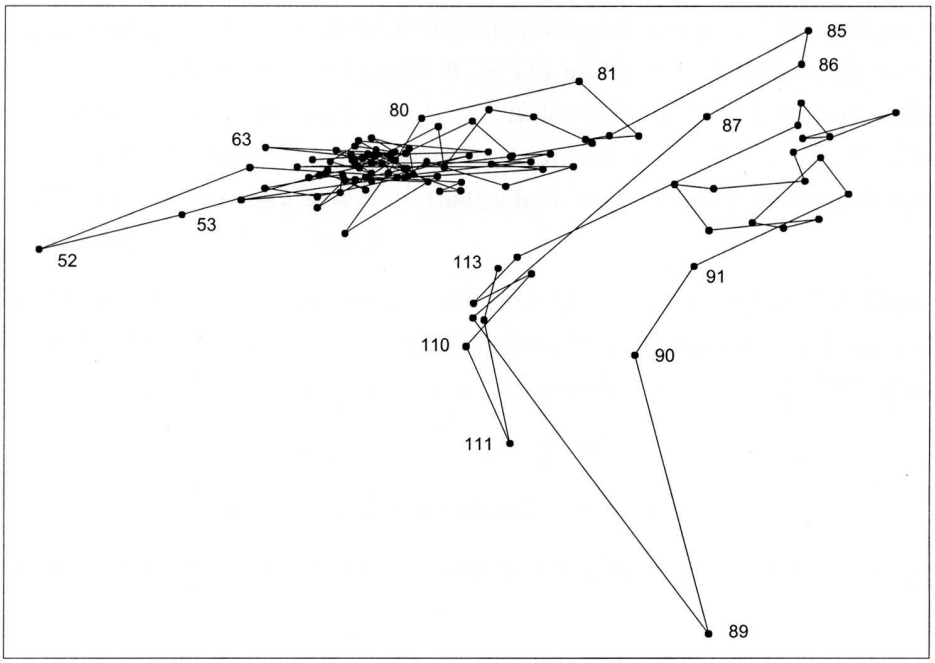


Figure 9: The 113 pages of *The Prince* projected onto a 2-dimensional space (extracted from [Gou99]).

where ε_y denotes the Dirac measure centered at y, and $\lceil a \rceil$ denotes the smallest integer greater than a. The global bag-of-words representation of x corresponds to the point $\rho(x) \in \overline{\mathbb{P}^{|V-1|}}$ parameterized by:

$$\rho^j(x) = \int_0^1 x(t, w_j)dt, \quad j = 1, \ldots, |V|. \tag{5.51}$$

The pure sequential representation in (5.50) may be smoothed via a function $f_{\mu,\sigma} : [0,1] \to \mathbb{R}_{++}$, where $\mu \in [0,1]$ and $\sigma \in \mathbb{R}_{++}$ are respectively a location and a scale parameter. An example of such a smoothing function is the truncated Gaussian defined in $[0,1]$ and normalized. This allows defining the lowbow representation at μ of the n-lenght document $(y_1, \ldots, y_n) \in V^n$ as the function $x : [0,1] \times V \to \mathbb{R}_+$ such that:

$$x(\mu, w_j) = \int_0^1 \varepsilon_{w_j}(y_{\lceil tn \rceil}) f_{\mu,\sigma}(t)dt, \tag{5.52}$$

which is proved [Leb06] to be a continuous and differentiable parameterized curve in the simplex. The scale of the smoothing kernel controls the amount of locality/globality in the document representation (see Fig. 10): when $\sigma \to \infty$ the simplicial curve degenerates to a single point which is the global bow representation (5.51); when $\sigma \to 0$, the curve quickly moves

between the different corners of the simplex approaching the pure sequential representation (5.50). A distance between two documents may defined by integrating pointwise the geodesic distance between the corresponding points in the simplex:

$$d(x, x') = \int_0^1 d_{\mathcal{J}}(x(\mu, .), x'(\mu, .))d\mu, \tag{5.53}$$

with $d_{\mathcal{J}}$ defined in (4.53).

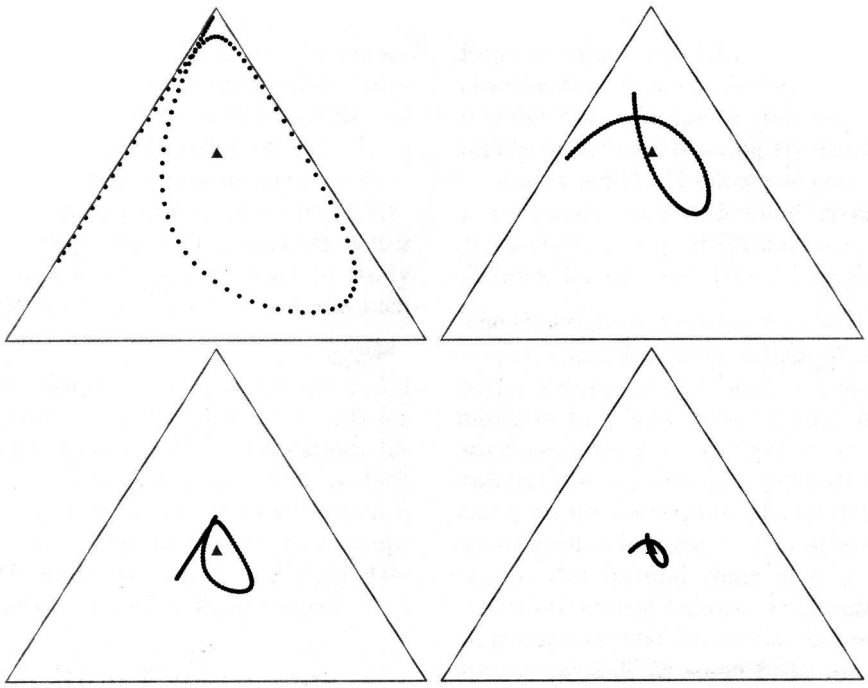


Figure 10: The lowbow representation of a document with $|V| = 3$, for several values of the scale parameter σ (extracted from [Leb06]).

Representing a document as a simplicial curve allows us to use geometric concepts to characterize properties of the document. For example, the tangent vector field along the curve describes sequential "topic trends" and their change; the curvature measures the amount of wigglyness or deviation from a geodesic path. This properties may be useful for tasks like text segmentation or summarization.

6 Conclusions

Throughout this report, a survey was presented about generative techniques to devise kernels on structured objects as strings, text or images. We studied

several different perspectives, some of them making use of recent results on information theory, as is the case with the Jensen-Shannon kernel, and others inspired by the field of information geometry. We emphasized the theoretical aspects and the geometrical insights that support either of these methods. We opted to be as general as possible: for example when studying the theory of positive and negative kernels, we did not choose the most direct path to prove that the Jensen-Shannon divergence is negative definite; instead some intermediate results were presented. Our belief is that while doing so, other ideas may come up.

Future work will concern making some experiments to compare several of these kernels on practical applications, as well as devising other kernels based on these insights, and best suited to particular tasks.

References

[Ama98] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[AN01] Shun-Ichi Amari and H. Nagaoka. *Methods of Information Geometry (Translations of Mathematical Monographs)*. Oxford University Press, 2001.

[BCR84] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, Berlin, 1984.

[CF05] Marco Cuturi and Kenji Fukumizu. Multiresolution kernels, 2005.

[CFV05] Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *J. Mach. Learn. Res.*, 6:1169–1198, 2005.

[CT91] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[CV05] Marco Cuturi and Jean-Philippe Vert. Semigroup kernels on finite sets. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 329–336. MIT Press, Cambridge, MA, 2005.

[ES03] Dominik M. Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

[GBGC+02] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan J. Oliver, and H. E. Stanley. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E*, 65, 2002.

[Gou99] Alan Gous. Spherical subfamily models, November 10 1999.

[Hau99] D. Haussler. Convolution kernels on discrete structures, 1999.

[HB04] Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures, September 28 2004.

[HH00] Keith Hall and Thomas Hofmann. Learning curved multinomial subfamilies for natural language processing and information retrieval. In *Proc. 17th International Conf. on Machine Learning*, pages 351–358. Morgan Kaufmann, San Francisco, CA, 2000.

[JH98] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. Technical report, Dept. of Computer Science, Univ. of California, 1998.

[JKH04] Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[KL02] Risi Imre Kondor and John D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In Claude Sammut and Achim G. Hoffmann, editors, *ICML*, pages 315–322. Morgan Kaufmann, 2002.

[Lan93] Serge Lang. *Real and Functional Analysis*. Springer-Verlag, New York, NY, USA, 1993.

[LCB+04] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[Leb05] G. Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, CMU, January 2005.

[Leb06] G. Lebanon. Sequential document representations and simplicial curves. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

[LL05] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163, 2005.

[MHV03] Pedro J. Moreno, Purdy Ho, and Nuno Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS*. MIT Press, 2003.

[MR93] Michael Murray and John Rice. *Differential Geometry and Statistics*. Number 48 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1st edition, 1993.

[SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. CUP, jun 2004.

[SWY75] G. Salton, A. Wong, and A. C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:229–237, 1975.

[Top00] Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

[Top02] Flemming Topsøe. Inequalities for the Jensen-Shannon divergence (draft version), 2002.

[Vap00] N. Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York., 2000.

[ZCL05] Dell Zhang, Xi Chen, and Wee Sun Lee. Text classification with kernels on the multinomial manifold. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 266–273, New York, NY, USA, 2005. ACM Press.