Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	Instituto Superior Tecnico AVENIDA ROVISCO PAIS LISBOA 1049-001 Portugal **http://www.ist.utl.pt**
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	Compression Kernels
Deliverable	D 2.1
Document type	Report
Dissemination level	Public
Contractual date of delivery	M 12
Project reference number	213250
Status & version	Definitive version
Work package, deliverable responsible	WP 2.2, IST
Author(s)	André T. Martins, Pedro M. Q. Aguiar, Mário A. T. Figueiredo
Additional contributor(s)	-

Contents

Information-Theoretic and Compression Kernels

June 15, 2009

Abstract

This document results from work carried out in the context of the Task 2.2 (named “Compression Kernels”) of Work Package 2 (named “Deriving Similarities for Non-vectorial Data”) of the SIMBAD project.

The central goal of this task is to devise ways to obtain kernels for non-vectorial data, inspired by information and coding theory. More specifically, the idea is to assume that the objects in hands (e.g., to be clustered or classified) were generated by some probabilistic source, and then define kernels between source models. More formally, by defining a family S containing the distributions from which the data points (in the input space X) are assumed to have been generated, and defining a map from X from S (e.g., via maximum likelihood estimation), a distribution in S may be fitted to each datum. Therefore, a kernel defined on $S \times S$ automatically induces a kernel on $X \times X$, through map composition. Some of these kernels have a natural information theoretic interpretation, establishing a bridge between kernel methods and information theory.

The idea in ”compression kernels” is to bypass the need to explicitly form the map from X from S. In principle, it is possible to obtain the entropy of a source by finding the best compression rate achievable for that source, thus bypassing the estimation of an explicit model source. Of course this requires the use of a universal source coding technique (e.g., of the Lempel-Ziv class), and the result is asymptotic. The same idea can be applied to non-parametrically obtain dissimilarities (e.g., the Kullback-Leibler divergence) between sources, without any intermediate estimate of the source model.

1 Kernel Methods

Although the theoretical foundations of kernel methods were established 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 linear algorithms that depend only on pairwise inner products between pairs of objects 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.

Generative kernels (the topic of WP2.1) and information-theoretic kernels are of course intimately related approaches. Both are based on the assumption that the objects of interest where somehow generated by a probabilistic mechanism (a source, in information/coding theoretic terms) and then proceed by defining (dis)similarity measures or kernels between (or among) models of these probabilistic sources. In fact, information-theoretic kernels can be clearly considered as generative kernels, although usually the information-theoretic perspective is more agnostic in the sense that it does not assume that the adopted model reflects the truth.

2 Information Theoretic Kernels

2.1 Kernels

We start by very briefly recalling some basic concepts from kernel theory (for comprehensive accounts, [SS02], [STC04]); in the following, X denotes a nonempty set.

Definition 2.1 *Let $\varphi : X \times X \to \mathbb{R}$ be a symmetric function, that is, a function satisfying $\varphi(y,x) = \varphi(x,y)$, for all $x, y \in X$. φ is called a* positive definite *(pd) kernel if and only if*

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

for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in X$ and $c_1, \ldots, c_n \in \mathbb{R}$.

Definition 2.2 *Let $\psi : X \times X \to \mathbb{R}$ be symmetric. ψ is called a* negative definite *(nd) kernel if and only if*

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

for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in X$ and $c_1, \ldots, c_n \in \mathbb{R}$, satisfying the additional constraint $c_1 + \ldots + c_n = 0$. In this case, $-\psi$ is called conditionally *pd; obviously, positive definiteness implies conditional positive definiteness.*

The sets of positive definite (pd) and negative definite (nd) kernels are both closed under pointwise sums/integrations, the former being also closed under pointwise products; moreover, both sets are closed under pointwise convergence. While pd kernels “correspond” to inner products via embedding in a Hilbert space, nd kernels that vanish on the diagonal and are positive anywhere else, “correspond” to squared Hilbertian distances. These facts, and the following proposition is shown, e.g., in [BCR84].

Proposition 2.3 *The function $\psi : X \times X \to \mathbb{R}$ is a nd kernel if and only if* $\exp(-t\psi)$ *is pd for all $t > 0$.*

2.2 Entropy, the Kullback-Leibler Divergence, and the Jensen-Shannon Divergence

We proceed by presenting mathematically formal definitions of entropy, Kullback-Leibler divergence, and Jensen-Shannon divergence, on arbitrary measured spaces.

Definition 2.4 *Let $(X, \mathscr{M}, \nu)$ be a measured space where X is Hausdorff and ν is a σ-finite Radon measure. Let $M_+^h(X) \subseteq M_+^b(X)$ (where $M_+^b(X)$ is the set of finite Radon measures on 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*[1] *of those measures. The* entropy function *$h : \frac{d}{d\nu} M_+^h(X) \to \mathbb{R}$ is defined by*

$$h(f) = -\int_X f \log f d\nu, \tag{2.1}$$

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

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

This definition of entropy generalizes the traditional notions of discrete and differential Shannon entropies, presented for example in [CT91]. Denote by $M_+^{1,h}(X)$ the set of Radon probability measures with finite entropy. If $X \subseteq \mathbb{R}^n$, ν is the Lebesgue-Borel measure, and $P \in M_+^{1,h}(X)$ is a probability measure with density $p = \frac{dP}{d\nu}$, then $h(p)$ reduces to the standard *differential entropy*

$$h(p) = -\int_X p(x) \log p(x) dx. \tag{2.2}$$

If 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 well known Shannon *discrete entropy*

$$H(p) = -\sum_{x \in X} p(x) \log p(x). \tag{2.3}$$

Definition 2.5 *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*

$$\begin{aligned} D(f\|g) &= \int_X f \log \frac{f}{g} d\nu \\ &= -h(f) - \int_X f \log g \, d\nu. \end{aligned} \tag{2.4}$$

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).

If X is a countable set, ν is the counting measure, and $P, Q \in M_+^{1,h}(X)$ are probability measures with probability mass functions $x \mapsto p(x) = P(\{x\})$ and $x \mapsto q(x) = Q(\{x\})$, then $D(p\|q)$ is the discrete *Kullback-Leibler divergence*

$$D(p\|q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{2.5}$$

$$= -H(p) - \sum_{x \in X} p(x) \log q(x). \tag{2.6}$$

Finally, we review (an extended version of) the definition of the Jensen-Shannon divergence, which may be seen as a smoothed and symmetrized version of the Kullback-Leibler divergence.

Definition 2.6 *Let $f_1, \ldots, f_n$ be probability densities of measures in $M_+^{1,h}(X)$, and $f = \alpha_1 f_1 + \ldots + \alpha_n f_n$ a mixture defined by coefficients $\alpha_1, \ldots, \alpha_n \in \mathbb{R}_+$,* *such that* $\alpha_1 + \ldots + \alpha_n = 1$*. The* generalized Jensen-Shannon divergence *of* $f_1, \ldots, f_n$ *with respect to that mixture is defined by*

$$J(f_1, \ldots, f_n; \alpha_1, \ldots, \alpha_n) \equiv h\left(\sum_{i=1}^{n} \alpha_i f_i\right) - \sum_{i=1}^{n} \alpha_i h(f_i), \tag{2.7}$$

Notice that positivity of the Jensen-Shannon divergence is a direct corollary of the Jensen inequality, since the entropy is a concave function.

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* $JS(f\|g)$*:*

$$JS(f\|g) \equiv h\left(\frac{f+g}{2}\right) - \frac{h(f) + h(g)}{2}. \tag{2.8}$$

In the case of probability mass functions on countable sets, we have H instead of h in (2.8). If, moreover, we restrict ourselves to the case of two distributions, the Jensen-Shannon divergence can also be written as

$$\begin{aligned} JS(p_1, p_2) &= H\left(\frac{p_1 + p_2}{2}\right) - \frac{H(p_1) + h(p_2)}{2} && (2.9) \\ &= \frac{1}{2} D\left(p_1 \middle\| \frac{p_1 + p_2}{2}\right) + \frac{1}{2} D\left(p_2 \middle\| \frac{p_1 + p_2}{2}\right) && (2.10) \end{aligned}$$

as introduced by [Lin91]. It has been shown that $\sqrt{JS}$ satisfies the triangle inequality (hence being a metric) and that, moreover, it is a Hilbertian metric[2] [ES03, Top00], which has motivated its use in kernel-based machine learning [CFV05], [HB05].

2.3 Probability Product Kernels

The Fisher kernel, introduced in the seminal paper [JH98], was arguably the earliest "generative" kernel. In fact, although it has a "generative" inspiration, the Fisher kernel is not defined on a probability space, since a particular distribution is chosen and kept fixed. We now describe a framework where data points in X are mapped to probability distributions in a parametric family M_Θ, and a "probability kernel" κ_M is devised in this space (or equivalently, in the space of the corresponding densities).

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*? Notice that this is very different from the usual density estimation problem, where we suppose that many i.i.d. data points are available. Although fitting a density to a single realization may not seem interesting for estimation purposes, here it is only an

[2] A metric $d : X \times X \to \mathbb{R}$ is Hilbertian if there is some Hilbert space H and an isometry $f : X \to H$ such that $d^2(x, y) = \langle f(x) - f(y), f(x) - f(y)\rangle_H$ holds for any $x, y \in X$ [HB05].

intermediate step towards a kernel and provides an opportunity to embody our prior knowledge about the data generation. The most obvious choice for the map f is maximum likelihood estimation

$$x \mapsto p_{\hat{\theta}(x)}, \quad \hat{\theta}(x) \equiv \hat{\theta}_{\mathrm{ML}}(x) = \arg \max_{\theta \in \Theta} p_{\theta}(x) \tag{2.11}$$

that leads to

$$\kappa(x, y) = \kappa_M \left(p_{\hat{\theta}(x)}, p_{\hat{\theta}(y)} \right). \tag{2.12}$$

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

$$\hat{\theta}(x) \equiv \hat{\theta}_{\mathrm{MAP}}(x) = \arg \max_{\theta \in \Theta} p_{\theta}(x) \pi(\theta). \tag{2.13}$$

A more Bayesian-flavored alternative to obtain a kernel in X is to consider the parameters as random variables and use the posterior mean kernel

$$\kappa(x, y) = \int_{\Theta \times \Theta} \kappa_M(p_\theta, p_\xi) p(\theta|x) p(\xi|y) d\theta d\xi. \tag{2.14}$$

The second problem concerns the choice of the probability kernels $\kappa_M : M_\Theta \times M_\Theta$, which will the topic of the following sections. In particular, we will focus on probability kernels based on information theoretic quantities. Before that, we conclude this subsection by reviewing two other possibilities: the Bhattacharyya and Hellinger distances.

The *Bhattacharyya kernel*, given by

$$\kappa_{1/2}(p, q) = \int_X \sqrt{p(x)} \sqrt{q(x)} d\mu(x), \tag{2.15}$$

is known in the statistics literature as the “Bhattacharyya affinity” between distributions. The Hellinger distance

$$H(p, q) = \frac{1}{2} \int_X \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 d\mu(x) \tag{2.16}$$

is related to Bhattacharyya affinity by $H(p, q) = \sqrt{2 - 2\kappa_{1/2}(p, q)}$. There are also interesting relationships between the Hellinger distance and other divergence measures, as the Kullback-Leibler or the Jensen-Shannon divergences [Top00]. As pointed out in [JKH04], the Bhattacharyya kernel can be computed in closed form for exponential families.

2.4 Kullback-Leibler Kernels

The Kullback-Leibler “kernel” [MHV03] was one of the earliest information-theoretic kernels proposed, operating directly on probability spaces. There are however some issues that need to be taken into account. Firstly, $D(.||.)$

is not even symmetric, thus can't be called a kernel. Asymmetry can be handled by using a symmetrized version $\tilde{D}(.||.)$ defined by $\tilde{D}(p||q) = D(p||q) + D(q||p)$. However, 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

$$\kappa(x, y) = \kappa_p(p_x, p_y) = \exp\left(-(\tilde{D}(p_x||p_y) - \beta)\right) \tag{2.17}$$

restricted to a finite set of data points in X, that by adjustment of the parameter β becomes positive definite. Notice that it is clear from Definition 2.2 and Proposition 2.3 that, if X is a finite set, there exists some β that makes $\tilde{D}(\cdot||\cdot) - \beta$ negative definite, thus $\kappa_p(\cdot, \cdot)$ 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.

2.5 Jensen-Shannon Kernels

The *Jensen-Shannon kernel* or entropy kernel between measures was introduced in [CV05, CFV05]. The proof that it is positive definite is a consequence of the negative definiteness of the entropy function. 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.

The basic result that allows deriving positive definite kernels based on the JS divergence is the fact that the Shannon entropy is a negative definite function. For a detailed proof of this fact, see [BCR84], [Top00] and [CV05]. A more general result, valid for non-extensive generalizations of entropy, was recently presented in [MAF08], [MSX+09].

We are now in a position to present the family of *weighted Jensen-Shannon kernels*, generalizing the JS-based kernels in that they allow using unnormalized measures; equivalently, they allow using different weights for each of the two arguments.

Definition 2.7 (weighted Jensen-Shannon kernels) *The kernel* $\widetilde{k}_{WJS} : (M_+^H(X))^2 \to \mathbb{R}$ *is defined as*

$$\begin{aligned} \widetilde{k}_{WJS}(\mu_1, \mu_2) &= \widetilde{k}_{WJS}(\omega_1 p_1, \omega_2 p_2) \\ &= \left(H(\pi) - J(p_1, p_2; \pi_1, \pi_2)\right)(\omega_1 + \omega_2), \end{aligned}$$

where $p_1 = \mu_1/\omega_1$ *and* $p_2 = \mu_2/\omega_2$ *are the normalized counterpart of* μ_1 *and* μ_2*, and*

$$\pi = (\omega_1/(\omega_1 + \omega_2), \omega_2/(\omega_1 + \omega_2))\,.$$

Analogously, the kernel $k_{WJS} : \left(M_+^H(X) \setminus \{0\}\right)^2 \to \mathbb{R}$ *is simply*

$$k_{WJS}(\mu_1, \mu_2) = k_{WJS}(\omega_1 p_1, \omega_2 p_2) = H(\pi) - J^\pi(p_1, p_2).$$

Proof that the weighted Jensen-Shannon kernels $\widetilde{k}_{\mathrm{WJS}}$ and k_{WJS} are pd can be found in [MSX+09].

The following family of *weighted exponentiated JS kernels*, generalize the so-called *exponentiated JS* kernel, that has been used, and shown to be pd, by [CV05].

Definition 2.8 (Exponentiated JS kernel) *The kernel* $k_{EJS} : M_+^1(X) \times M_+^1(X) \to \mathbb{R}$ *is defined, for* $t > 0$*, as*

$$k_{EJS}(p_1, p_2) = \exp\left[-t\, JS\,(p_1, p_2)\right].$$

Definition 2.9 (Weighted exponentiated JS kernels) *The kernel* $k_{WEJS} : M_+^H(X) \times M_+^H(X) \to \mathbb{R}$ *is defined, for* $t > 0$*, as*

$$\begin{aligned} k_{WEJS}(\mu_1, \mu_2) &= \exp[t\, k_{WJS}(\mu_1, \mu_2)] \\ &= \exp(t\, H(\pi)) \exp\left[-t J^\pi(p_1, p_2)\right]. \qquad (2.18) \end{aligned}$$

Proof that the weighted exponentiated Jensen-Shannon kernels $\widetilde{k}_{\mathrm{WEJS}}$ are pd can be found in [MSX+09].

Another possibility for defining a pd kernel from the Jensen-Shannon divergence is the following.

Definition 2.10 (Jensen-Shannon kernel) *The kernel* $k_{JS} : M_+^1(X) \times M_+^1(X) \to \mathbb{R}$ *is defined as*

$$k_{JS}(p_1, p_2) = \ln 2 - JS(p_1, p_2).$$

Again, proof that this Jensen-Shannon kernel is pd can be found in [MSX+09].

In conclusion, the Jensen-Shannon divergence allows building several positive-definite kernels between pairs of probability distributions.

A interesting avenue for future research is the definition of multi-kernels (i.e., defined not between pairs of objects, but among groups of more than two objects). This may be done by using the generalized Jensen-Shannon divergence, as given by (2.7). Whether such multi-kernels are useful for machine learning applications, and what properties they have is a topic for future research.

3 Computing the Kernels

The fundamental issue in applying the information theoretic kernels described in the previous section concerns the computation of the involved quantities: entropies and Kullback-Leibler divergences. Recall that the Jensen-Shannon divergence may be computed using entropies or Kullback-Leibler divergences, as shown in (2.9)–(2.10).

In this document, we will consider only the case where the objects are strings of symbols from some finite alphabet, which is the case of interest for problems such as text categorization or clustering.

3.1 The Classical Approach: Bags of Words

Arguably the most standard way to compute kernels between strings is to ignore their string nature and consider them as realization of some memoryless source. This is the case o the *bag of words* (BoW) model, where the source is assumed to generate words in a memoryless fashion. There is a large literature on how to preprocess the text and obtain the BoW representation, [Joa02], [BYRN99]. From a probabilistic point of view, the BoW representation corresponds to a memoryless multinomial model; thus text documents can be mapped into multinomial distributions over words via maximum likelihood estimation.

In [MSX$^+$09], we have shown that information theoretic kernels outperform the classical linear kernel applied to the BoW representation, and is competitive with the heat kernel (in fact, the approximation introduced in [LL05]) on the multinomial manifold, specially when instead of the Jensen-Shannon kernel one is allowed to use its non-extensive counterpart, the Jensen-Tsallis kernel.

3.2 String Kernels versus Kernels on Stochastic Processes

Several string kernels (i.e., kernels operating on the space of strings) have been proposed in the literature [Hau99, LSST$^+$02, LEN02, VS03, STC04]. String kernels are defined on $A^* \times A^*$, where A^* is the Kleene closure of a finite alphabet A (i.e., the set of all finite strings formed by characters in A together with the empty string ϵ). The *p-spectrum kernel* (PSK) [LEN02] is associated with a feature space indexed by A^p (the set of length-p strings). The feature representation of a string s, $\Phi^p(s) \triangleq (\phi_u^p(s))_{u \in A^p}$, counts the number of times each $u \in A^p$ occurs as a substring of s,

$$\phi_u^p(s) = |\{(v_1, v_2) : s = v_1 u v_2\}|.$$

The p-spectrum kernel is then defined as the standard inner product in $\mathbb{R}^{|A|^p}$

$$k_{\text{SK}}^p(s,t) = \langle \Phi^p(s), \Phi^p(t) \rangle \,. \tag{3.1}$$

A more general kernel is the *weighted all-substrings kernel* (WASK) [VS03], which takes into account the contribution of all the substrings weighted by their length. This kernel can be viewed as a conic combination of p-spectrum kernels and can be written as

$$k_{\mathrm{WASK}}(s,t) = \sum_{p=1}^{\infty} \alpha_p \, k_{\mathrm{SK}}^{p}(s,t), \tag{3.2}$$

where α_p is often chosen to decay exponentially with p and truncated; for example, $\alpha_p = \lambda^p$, if $p_{\min} \le p \le p_{\max}$, and $\alpha_p = 0$, otherwise, where $0 < \lambda < 1$ is the decaying factor.

Both k_{SK}^{p} and k_{WASK} are trivially positive definite, the former by construction and the latter because it is a conic combination of positive definite kernels. A remarkable fact is that both kernels may be computed in $O(|s| + |t|)$ time (i.e., with cost that is linear in the length of the strings), as shown by [VS03], by using data structures such as suffix trees or suffix arrays [Gus97]. Moreover, with s fixed, any kernel $k(s,t)$ may be computed in time $O(|t|)$, which is particularly useful for classification applications.

As an alternative to these classical string kernels, information theoretic kernels on strings may be obtained by building upon the Jensen-Shannon kernels introduced in the previous section. Notice that joint Jensen-Shannon divergences are simply Jensen-Shannon divergences defined on a product space of the form $X = X_1 \times X_2 \times \cdots \times X_l$, thus they still yield positive definite kernels, as shown in the previous section. This observation suggests a simple way to define Jensen-Shannon kernels between stochastic processes, by applying them to l-th order joint probability functions, as explained next.

Definition 3.1 (l-th order weighted JS kernels) *Let $\mathscr{S}(A)$ be the set of stationary and ergodic stochastic processes that take values on the alphabet A. For $l \in \mathbb{N}$, let the kernel $\widetilde{k}_l : \mathscr{S}(A)^2 \to \mathbb{R}$ be defined as*

$$\widetilde{k}_l((\omega_1, s_1), (\omega_2, s_2)) \triangleq \widetilde{k}_{WJS}(\omega_1 p_{s_1,l}, \omega_2 p_{s_2,l}) \tag{3.3}$$

where $p_{s_1,l}$ and $p_{s_2,l}$ are the l-th order joint probability functions associated with the stochastic sources s_1 and s_2, $\pi = (\omega_1/(\omega_1+\omega_2), \omega_2/(\omega_1+\omega_2))$, and $\widetilde{k}_{WJS}$ was defined in the previous section.

Let the kernel $k_l : \mathscr{S}(A)^2 \to \mathbb{R}$ be defined as

$$k_l((\omega_1, s_1), (\omega_2, s_2)) \triangleq k_{WJS}(\omega_1 p_{s_1,l}, \omega_2 p_{s_2,l}), \tag{3.4}$$

where k_{WJS} was defined in the previous section.

Positive definiteness of these string kernels is a corollary of that of $\widetilde{k}_{\mathrm{WJS}}$ and k_{WJS}, as was shown in [MSX+09]. Of course unweighted versions of these kernels are obtained by setting $\omega_1 = \omega_2$.

In [MSX$^+$09], we have shown that these information theoretic kernels outperform the PSK and the WASK, specially when instead of the classical Jensen-Shannon kernel between stochastic processes one is allowed to use its non-extensive counterpart, the Jensen-Tsallis kernel.

3.3 Compression Kernels

The idea in "compression kernels" is to bypass the need to explicitly form the map $f : X \to M_\Theta$ (introduced in Section 2.3) and compute the information theoretic kernel (e.g., the Jensen-Shannon kernel) directly from a pair of objects x and y (e.g., a pair of strings).

As mentioned in the introduction, it is possible to obtain the entropy of a source by finding the best compression rate achievable for that source, thus bypassing the explicit estimation of a model source. Naturally, this assumes the use of a universal (e.g. Lempel-Ziv) source code; recall that a universal code is one that is able to asymptotically achieve the entropy rate lower bound without prior knowledge of a model of the source [CT91]. The same idea can be applied to non-parametrically obtain dissimilarities (e.g., the Kullback-Leibler divergence) between sources, without any intermediate estimate of the source model.

In the following subsections, we detail these ideas.

3.4 Relationship Between Entropy and Lempel-Ziv Coding

Assume a random sequence $x = (x_1, x_2, ..., x_n)$ that was produced by an unknown (but finite) order stationary Markovian source, with a finite alphabet. Consider the goal estimating the n-th order entropy, or equivalently the logarithm of the joint probability function

$$-(1/n) \log_2 p(x_1, x_2, ..., x_n)$$

(from which the entropy can be obtained). A direct approach is computationally prohibitive for large n. However, an alternative route can be taken using the following fact (see [CT91], [ZL78]): the Lempel-Ziv (LZ) code length for x, divided by n, is a computationally efficient and reliable estimate of the entropy. More formally, let $c(x)$ denote the number of phrases in x resulting from the LZ incremental parsing of x into distinct phrases, such that each phrase is the shortest sequence which is not a previously parsed phrase. Then, the LZ code length for x is approximately

$$c(x) \log_2 c(x); \tag{3.5}$$

moreover, it was shown in [ZM93] that

$$c(x) \log_2 c(x) \xrightarrow{a.s.} -(1/n) \log_2 p(x_1, x_2, ..., x_n),$$

as $n \to \infty$. This fact suggests using the output of an LZ encoder to estimate the entropy of an unknown source without explicitly estimating its model parameters.

3.5 Estimating Kullback-Leibler Divergences

The method proposed in [ZM93] for measuring relative entropy is also based on two Lempel-Ziv-type parsing algorithms:

- The incremental LZ parsing algorithm [ZL78], which is a self parsing procedure of a sequence into $c(z)$ distinct phrases such that each phrase is the shortest sequence that is not a previously parsed phrase. For example, let $n = 11$ and $z = (01111000110)$, then the self incremental parsing yields $(0, 1, 11, 10, 00, 110)$, namely, $c(z) = 6$.

- A variation of the LZ parsing algorithm described in [ZM93], which is a sequential parsing of a sequence z with respect to another sequence x (cross parsing). Let $c(z|x)$ denote the number of phrases in z with respect to x. For example, let z as before and $x = (10010100110)$; then, parsing z with respect to x yields $(011, 110, 00110)$, that is $c(z|x) = 3$.

It was shown in [ZM93] that for two arbitrary (but finite) order Markovian sequences of length n, the quantity

$$\Delta(z||x) = \frac{1}{n} \left[c(z|x) \log_2 n - c(z) \log_2 c(z) \right] \tag{3.6}$$

converges, as $n \to \infty$, to the relative entropy between the two sources that emitted the sequences z and x. Roughly speaking, we can observe (see (3.5)) that $c(z) \log_2 c(z)$ is the measure of the complexity of the sequence z obtained by self-parsing, thus providing an estimate of its entropy, while $(1/n)\, c(z|x) \log_2 n$ can be seen as an estimate of the code-length obtained when coding z using a model for x. From now on we will refer to $\Delta(z||x)$ as the ZM divergence.

Our implementation of the ZM divergence [PF05] uses the LZ78 algorithm to make the self parsing procedure. To perform the cross parsing, we designed a modified LZ77-based algorithm where the dictionary is static and only the lookahead buffer slides over the input sequence, as shown in Figure 1.

Two important parameters of the algorithm are the dictionary size and the maximum length of a matching sequence found in the look-ahead buffer (LAB); both influence the parsing results and determine the compressor efficiency.

Alternative methods to estimate entropies and Kullback-Leibler divergences were recently proposed in [CKV04] and [CKV06]. Those techniques (herein called CKV) is based on the Burrow-Wheeler transform (BWT) for

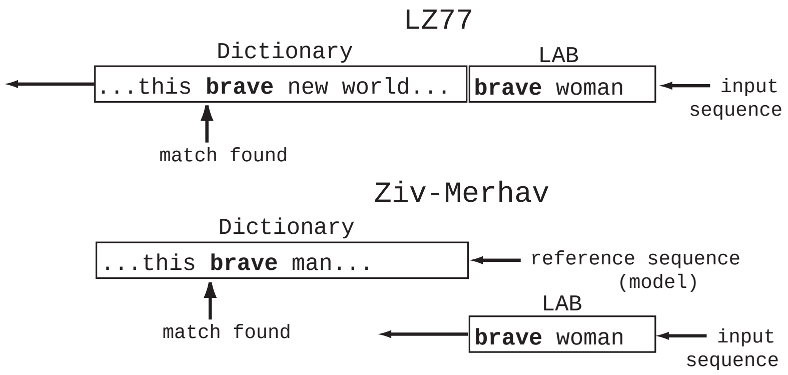

Figure 1: The original LZ77 algorithm uses a sliding window over the input sequence to get the dictionary updated, whereas in the Ziv-Merhav cross parsing procedure the dictionary is static and only the *lookahead buffer* (LAB) slides over the input sequence.

text compression [BW94]. The BWT is a reversible block-sorting method that operates on a sequence of symbols as follows: it produces all cyclic shifts of the original sequence, sorts them lexicographically, and outputs the last column of the sorted table. For finite-memory sources, performing the BWT on a reversed data sequence groups together symbols in the same state (i.e., with the same context). Using the BWT followed by segmentation is the basic idea behind the entropy estimation in [CKV04]. This idea was extended to divergence estimation in [CKV06], by using the joint BWT of two sequences as illustrated in Figure 2.

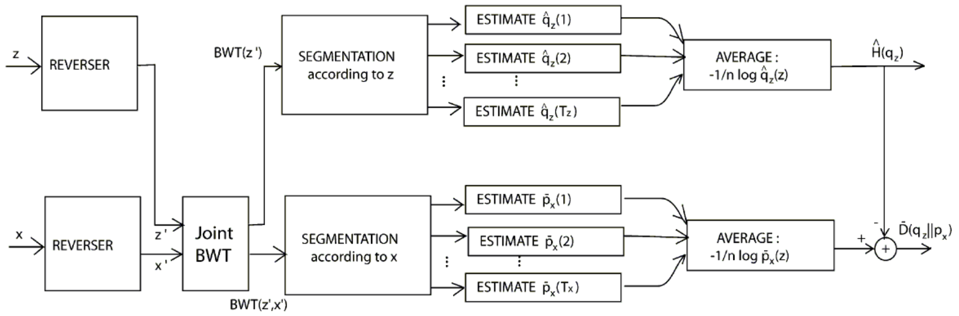

Figure 2: Block diagram of the divergence estimator via the BWT.

In [PF08], we have conducted experiments to compare the theoretical values of the KL divergence with the estimates produced by the ZM and the CKV methods, on pairs of binary sequences with 100, 1000 and 10000 symbols. The sequences were randomly generated from simulated sources using both memoryless and order-1 Markov models. Results for these experiments using 10000 symbols are shown in Figure 3. Each plot compares the true KL divergence with the ZM and CKV estimates, over a varying range of source symbol probabilities. The results show that, for this type of source, the CKV method provides a more accurate KL divergence estimate than the ZM technique (which may even return negative values when the sequences are very similar).

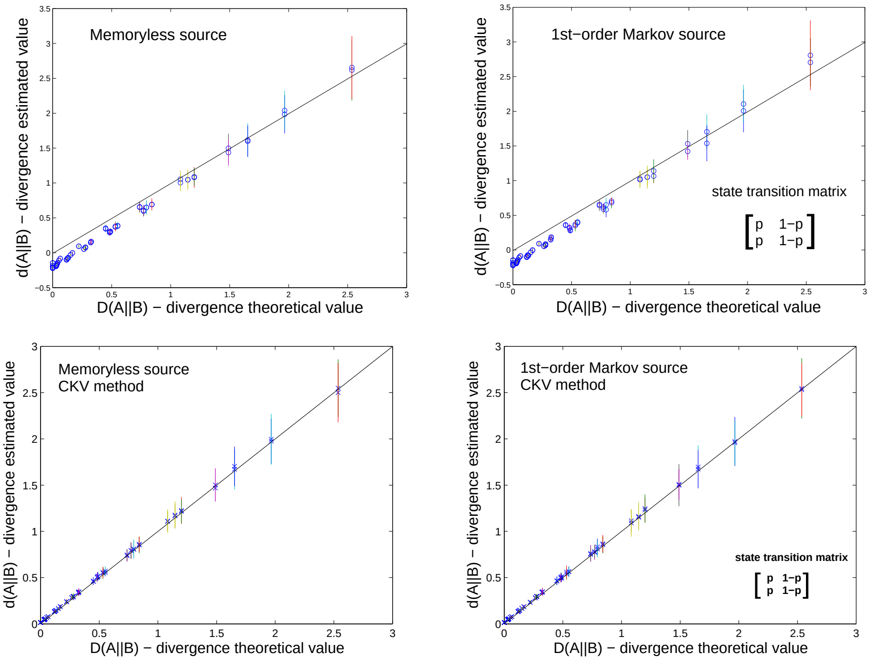

Figure 3: KL divergence estimates obtained by the ZM and CKV methods, versus the theoretical values. Each circle is the sample mean value and the vertical bars the sample standard deviation values, evaluated over 100 pairs of sequences (of length 10000). For the 1st-order Markov source we use the state transition matrix shown and consider a range of values of $p \in [0, 1]$.

We have also run experiments to assess the performance of the ZM and CKV estimators in a text classification task; more specifically, in an authorship attribution problem. We have used the same text corpus that was used in [BCL02]. This corpus contains a set of 86 files of several Italian authors, and can be downloaded from `www.liberliber.it`. In this experiment, each text is classified as belonging to the author of the closest text in the remaining set. In other words, the results reported can be seen as a full *leave-one-out cross-validation* (LOO-CV) performance measure of a nearest-neighbor classifier built using the considered divergence functions.

Author	**No. of texts**	**BCL**	**ZM**	**CKV**
Alighieri	8	7	7	7
Deledda	15	15	15	0
Fogazzaro	5	3	5	4
Guicciardini	6	6	5	0
Macchiavelli	12	11	11	5
Manzoni	4	4	3	4
Pirandello	11	9	11	3
Salgari	11	11	11	8
Svevo	5	5	5	1
Verga	9	7	9	1
Total	**86**	**78**	**82**	**33**

Table 1: Classification of Italian authors: for each author, we report the number of texts considered and three values of classification success rate, obtained using the method of Benedetto, Caglioti and Loreto [BCL02], the Ziv-Merhav method (ZM) and the CKV method.

4 Towards Compression Kernels: Ongoing and Future Work

As shown in the previous subsections, we have LZ-compression-based tools that serve to obtain estimates of the Kullback-Leibler divergence as well as entropy estimates. These tools can now be used to define Jensen-Shannon divergences (using either (2.9) or (2.10) and thus Jensen-Shannon kernels. The last question that needs to be answered in order to implement (2.9) or (2.10) is the following: how to generate a sample from the mixture distribution $(p_1 + p_2)/2$, given a string generated by source p_1 and another by source p_2? It turns out that this question was answered more than a decade ago in [EYFT97], where an algorithm for that purpose was proposed. In fact, the use of the ZM approach for estimating the Kullback-Leibler divergence was also proposed in [EYFT97]. We believe (but have not shown yet) that computing the Jensen-Shannon divergence in the form (2.9) may be a better alternative, since it only involves estimating entropies, rather than Kullback-Leibler divergences. Moreover, the form (2.9) is directly generalizable to more than a pair of arguments.

As far as we know, these compression-based estimates of the Jensen-Shannon (or Kullback-Leibler) divergences has not yet been used to build the corresponding kernels, and used in kernel based classifiers (e.g., support vector machines). This is the work topic we have currently in hands, in the context of the work package WP2 of the SIMBAD project.

References

[BCL02] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Physical Review Letters, 88:4*, 2002.

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

[BW94] M. Burrows and D. Wheeler. A block-sorting lossless data compression algorithm. *Tech. Rep. 124, Digital Systems Research Center*, 1994.

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval.* ACM Press, New York, 1999.

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

[CKV04] H. Cai, S. Kulkarni, and S. Verdu. Universal estimation of entropy via block sorting. *IEEE Transactions on Information Theory*, 50:1551–1561, 2004.

[CKV06] H. Cai, S. Kulkarni, and S. Verdu. Universal divergence estimation for finite-alphabet sources. *IEEE Transactions on Information Theory*, 52:3456–3475, 2006.

[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.

[EYFT97] R. El-Yaniv, S. Fine, and N. Tishby. Agnostic classification of markovian sequences. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 465–471. MIT Press, 1997.

[Gus97] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press, 1997.

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

[HB05] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005.

[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.

[Joa02] T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers, 2002.

[LEN02] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Proceedings of the Pacific Symposium on Biocomputing 2002 (PSB 2002)*, pages 564–575, 2002.

[Lin91] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

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

[LSST$^+$02] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

[MAF08] A. Martins, P. Aguiar, and M. Figueiredo. Tsallis kernels on measures. In *IEEE Information Theory Workshop–ITW'2008*, 2008.

[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.

[MSX$^+$09] A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

[PF05] D. Pereira Coutinho and M. Figueiredo. Information theoretic text classification using the Ziv-Merhav method. *2nd Iberian Conference on Pattern Recognition and Image Analysis – IbPRIA'2005*, 2005.

[PF08] D. Pereira Coutinho and M. Figueiredo. Information theoretic text classification: Experimental evaluation. In *8th International Workshop on Pattern Recognition in Information Systems*, Barcelona, Spain, 2008.

[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.

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

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

[VS03] S.V.N. Vishwanathan and A. J. Smola. Fast kernels for string and tree matching. In K. Tsuda, B. Schölkopf, and J.P. Vert, editors, *Kernels and Bioinformatics*, Cambridge, MA, 2003. MIT Press.

[ZL78] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

[ZM93] J. Ziv and N. Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39:1270–1279, 1993.