Journal of Machine Learning Research 14 (2013) 2837-2856 Submitted 3/13; Revised 6/13; Published 9/13

A Binary-Classification-Based Metric between Time-Series Distributions and Its Use in Statistical and Learning Problems

Daniil Ryabko DANIIL.RYABKO@INRIA.FR
Jérémie Mary JEREMIE.MARY@INRIA.FR
SequeL-INRIA/LIFL-CNRS
Université de Lille, France
40, avenue de Halley
59650 Villeneuve d'Ascq
France

Editor: Léon Bottou

Abstract

A metric between time-series distributions is proposed that can be evaluated using binary classification methods, which were originally developed to work on i.i.d. data. It is shown how this metric can be used for solving statistical problems that are seemingly unrelated to classification and concern highly dependent time series. Specifically, the problems of time-series clustering, homogeneity testing and the three-sample problem are addressed. Universal consistency of the resulting algorithms is proven under most general assumptions. The theoretical results are illustrated with experiments on synthetic and real-world data.

Keywords: time series, reductions, stationary ergodic, clustering, metrics between probability distributions

1. Introduction

Binary classification is one of the most well-understood problems of machine learning and statistics: a wealth of efficient classification algorithms has been developed and applied to a wide range of applications. Perhaps one of the reasons for this is that binary classification is conceptually one of the simplest statistical learning problems. It is thus natural to try and use it as a building block for solving other, more complex, newer or just different problems; in other words, one can try to obtain efficient algorithms for different learning problems by reducing them to binary classification. This approach has been applied to many different problems, starting with multi-class classification, and including regression and ranking (Balcan et al., 2007; Langford et al., 2006), to give just a few examples. However, all of these problems are formulated in terms of independent and identically distributed (i.i.d.) samples. This is also the assumption underlying the theoretical analysis of most of the classification algorithms.

In this work we consider learning problems that concern time-series data for which independence assumptions do not hold. The series can exhibit arbitrary long-range dependence, and different time-series samples may be interdependent as well. Moreover, the learning problems that we consider—the three-sample problem, time-series clustering, and homogeneity testing—at first glance seem completely unrelated to classification.

We show how the considered problems can be reduced to binary classification methods, via a new metric between time-series distributions. The results include asymptotically consistent algorithms, as well as finite-sample analysis. To establish the consistency of the suggested methods, for clustering and the three-sample problem the only assumption that we make on the data is that the distributions generating the samples are stationary ergodic; this is one of the weakest assumptions used in statistics. For homogeneity testing we have to make some mixing assumptions in order to obtain consistency results (this is indeed unavoidable, as shown by Ryabko, 2010b). Mixing conditions are also used to obtain finite-sample performance guarantees for the first two problems.

The proposed approach is based on a new distance between time-series distributions (that is, between probability distributions on the space of infinite sequences), which we call *telescope distance*. This distance can be evaluated using binary classification methods, and its finite-sample estimates are shown to be asymptotically consistent. Three main building blocks are used to construct the telescope distance. The first one is a distance on finite-dimensional marginal distributions. The distance we use for this is the following well-known metric: $d_{\mathcal{H}}(P,Q) := \sup_{h\in\mathcal{H}} |\mathbf{E}_P h - \mathbf{E}_Q h|$ where P,Q are distributions and $\mathcal{H}$ is a set of functions. This distance can be estimated using binary classification methods, and thus can be used to reduce various statistical problems to the classification problem. This distance was previously applied to such statistical problems as homogeneity testing and change-point estimation (Kifer et al., 2004). However, these applications so far have only concerned i.i.d. data, whereas we want to work with highly-dependent time series. Thus, the second building block are the recent results of Adams and Nobel (2012), that show that empirical estimates of $d_{\mathcal{H}}$ are consistent (under certain conditions on $\mathcal{H}$) for arbitrary stationary ergodic distributions. This, however, is not enough: evaluating $d_{\mathcal{H}}$ for (stationary ergodic) time-series distributions means measuring the distance between their finite-dimensional marginals, and not the distributions themselves. Finally, the third step to construct the distance is what we call *telescoping*. It consists in summing the distances for all the (infinitely many) finite-dimensional marginals with decreasing weights. The resulting distance can “automatically” select the marginal distribution of the right order: marginals which cannot distinguish between the distributions give distance estimates that converge to zero, while marginals whose orders are too high to have converged have very small weights. Thus, the estimate is dominated by the marginals which can distinguish between the time-series distributions, or converges to zero if the distributions are the same. It is worth noting that a similar telescoping trick is used in different problems, most notably, in sequence prediction (Solomonoff, 1978; B. Ryabko, 1988; Ryabko, 2011); it is also used in the distributional distance (Gray, 1988), see Section 8 below.

We show that the resulting distance (telescope distance) indeed can be consistently estimated based on sampling, for arbitrary stationary ergodic distributions. Further, we show how this fact can be used to construct consistent algorithms for the considered problems on time series. Thus we can harness binary classification methods to solve statistical learning problems concerning time series. A remarkable feature of the resulting methods is that the performance guarantees obtained do not depend on the approximation error of the binary classification methods used, they only depend on their estimation error.

Moreover, we analyse some other distances between time-series distributions, the possibility of their use for solving the statistical problems considered, and the relation of these distances to the telescope distance introduced in this work.

To illustrate the theoretical results in an experimental setting, we chose the problem of time-series clustering, since it is a difficult unsupervised problem which seems most different from the problem of binary classification. Experiments on both synthetic and real-world data are provided. The real-world setting concerns brain-computer interface (BCI) data, which is a notoriously challenging application, and on which the presented algorithm demonstrates competitive performance.

A related approach to address the problems considered here, as well as some related problems about stationary ergodic time series, is based on (consistent) empirical estimates of the distributional distance, see Ryabko and Ryabko (2010), Ryabko (2010a), Khaleghi et al. (2012), as well as Gray (1988) about the distributional distance. The empirical distance is based on counting frequencies of bins of decreasing sizes and "telescoping." This distance is described in some detail in Section 8 below, where we compare it to the telescope distance. Another related approach to time-series analysis involves a different reduction, namely, that to data compression (B. Ryabko, 2009).

1.1 Organisation

Section 2 is preliminary. In Section 3 we introduce and discuss the telescope distance. Section 4 explains how this distance can be calculated using binary classification methods. Sections 5 and 6 are devoted to the three-sample problem and clustering, respectively. In Section 7, under some mixing conditions, we address the problems of homogeneity testing, clustering with unknown k, and finite-sample performance guarantees. In Section 8 we take a look at other distances between time-series distributions and their relations to the telescope distance. Section 9 presents experimental evaluation.

2. Notation and Definitions

Let $(\mathcal{X},\mathcal{F}_1)$ be a measurable space (the domain), and denote $(\mathcal{X}^k,\mathcal{F}_k)$ and $(\mathcal{X}^{\mathbb{N}},\mathcal{F})$ the product probability space over $\mathcal{X}^k$ and the induced probability space over the one-way infinite sequences taking values in $\mathcal{X}$. Time-series (or process) distributions are probability measures on the space $(\mathcal{X}^{\mathbb{N}},\mathcal{F})$. We use the abbreviation $X_{1..k}$ for $X_1,\dots,X_k$. A set $\mathcal{H}$ of functions is called *separable* if there is a countable set $\mathcal{H}'$ of functions such that any function in $\mathcal{H}$ is a pointwise limit of a sequence of elements of $\mathcal{H}'$.

A distribution ρ is called stationary if $\rho(X_{1..k}\in A)=\rho(X_{n+1..n+k}\in A)$ for all $A\in\mathcal{F}_k$, $k,n\in\mathbb{N}$. A stationary distribution is called (stationary) ergodic if

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1..n-k+1}\mathbb{I}_{X_{i..i+k}\in A}=\rho(A)\ \ \rho-\text{a.s.}$$

for every $A\in\mathcal{F}_k$, $k\in\mathbb{N}$. (This definition, which is more suited for the purposes of this work, is equivalent to the usual one expressed in terms of invariant sets, see, e.g., Gray, 1988.)

3. A Distance between Time-Series Distributions

We start with a distance between distributions on $\mathcal{X}$, and then we extend it to distributions on $\mathcal{X}^{\mathbb{N}}$. For two probability distributions P and Q on $(\mathcal{X},\mathcal{F}_1)$ and a set $\mathcal{H}$ of measurable functions on $\mathcal{X}$, one can define the distance

$$d_{\mathcal{H}}(P,Q):=\sup_{h\in\mathcal{H}}|\mathbf{E}_P h-\mathbf{E}_Q h|. \tag{1}$$

This metric in its general form has been studied at least since the 80's (Zolotarev, 1983); its special cases include Kolmogorov-Smirnov (Kolmogorov, 1933), Kantorovich-Rubinstein (Kantorovich and Rubinstein, 1957) and Fortet-Mourier (Fortet and Mourier, 1953) metrics. Note that the distance function so defined may not be measurable; however, it is measurable under mild conditions which we assume whenever necessary. In particular, separability of $\mathcal{H}$ is a sufficient condition (separability is required in most of the results below).

We are interested in the cases where $d_{\mathcal{H}}(P,Q) = 0$ implies $P = Q$. Note that in this case $d_{\mathcal{H}}$ is a metric (the rest of the properties are easy to see). For reasons that will become apparent shortly (see Remark below), we are mainly interested in the sets $\mathcal{H}$ that consist of indicator functions. In this case we can identify each $f \in \mathcal{H}$ with the indicator set $\{x : f(x) = 1\} \subset \mathcal{X}$ and (by a slight abuse of notation) write $d_{\mathcal{H}}(P,Q) := \sup_{h\in\mathcal{H}} |P(h) - Q(h)|$. In this case it is easy to check that the following statement holds true.

Lemma 1 *$d_{\mathcal{H}}$ is a metric on the space of probability distributions over $\mathcal{X}$ if and only if $\mathcal{H}$ generates $\mathcal{F}_1$.*

The property that $\mathcal{H}$ generates $\mathcal{F}_1$ is often easy to verify directly. First of all, it trivially holds for the case where $\mathcal{H}$ is the set of halfspaces in a Euclidean $\mathcal{X}$. It is also easy to check that it holds if $\mathcal{H}$ is the set of halfspaces in the feature space of most commonly used kernels (provided the feature space is of the same or higher dimension than the input space), such as polynomial and Gaussian kernels.

Based on $d_{\mathcal{H}}$ we can construct a distance between time-series probability distributions. For two time-series distributions ρ_1, ρ_2 we take the $d_{\mathcal{H}}$ between k-dimensional marginal distributions of ρ_1 and ρ_2 for each $k \in \mathbb{N}$, and sum them all up with decreasing weights.

Definition 2 (telescope distance $D_{\mathbf{H}}$) *For two time series distributions ρ_1 and ρ_2 on the space $(\mathcal{X}^{\mathbb{N}}, \mathcal{F})$ and a sequence of sets of functions $\mathbf{H} = (\mathcal{H}_1, \mathcal{H}_2, \dots)$ define the* telescope distance

$$D_{\mathbf{H}}(\rho_1,\rho_2) := \sum_{k=1}^{\infty} w_k \sup_{h\in\mathcal{H}_k} |\mathbf{E}_{\rho_1} h(X_1,\dots,X_k) - \mathbf{E}_{\rho_2} h(Y_1,\dots,Y_k)|, \tag{2}$$

where w_k, $k \in \mathbb{N}$ is a sequence of positive summable real weights (e.g., $w_k = 1/k^2$ or $w_k = 2^{-k}$).

Lemma 3 *$D_{\mathbf{H}}$ is a metric if and only if $d_{\mathcal{H}_k}$ is a metric for every $k \in \mathbb{N}$.*

Proof The statement follows from the fact that two process distributions are the same if and only if all their finite-dimensional marginals coincide. ■

Definition 4 (empirical telescope distance $\hat{D}$) *For a pair of samples $X_{1..n}$ and $Y_{1..m}$ define the* empirical telescope distance *as*

$$\hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) := \sum_{k=1}^{\min\{m,n\}} w_k \sup_{h\in\mathcal{H}_k} \left| \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} h(Y_{i..i+k-1}) \right|. \tag{3}$$

All the methods presented in this work are based on the empirical telescope distance. The key fact is that it is an asymptotically consistent estimate of the telescope distance, that is, the latter can be consistently estimated based on sampling.

Theorem 5 *Let $\mathbf{H} = (\mathcal{H}_k)_{k\in\mathbb{N}}$ be a sequence of separable sets $\mathcal{H}_k$ of indicator functions (over $\mathcal{X}^k$) of finite VC dimension such that $\mathcal{H}_k$ generates $\mathcal{F}_k$. Then for every stationary ergodic time series distributions ρ_X and ρ_Y generating samples $X_{1..n}$ and $Y_{1..m}$ we have*

$$\lim_{n,m\to\infty} \hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) = D_{\mathbf{H}}(\rho_X, \rho_Y) a.s.$$

Note that $\hat{D}_{\mathbf{H}}$ is a biased estimate of $D_{\mathbf{H}}$, and, unlike in the i.i.d. case, the bias may depend on the distributions; however, the bias is $o(n)$.

Remark. The condition that the sets $\mathcal{H}_k$ are sets of indicator function of finite VC dimension comes from the results of Adams and Nobel (2012), who show that for any stationary ergodic distribution ρ, under these conditions, $\sup_{h\in\mathcal{H}_k} \frac{1}{n-k+1}\sum_{i=1}^{n-k+1} h(X_{i..i+k-1})$ is an asymptotically consistent estimate of $\sup_{h\in\mathcal{H}_k} \mathbf{E}_\rho h(X_1,\dots,X_k)$. This fact implies that $d_{\mathcal{H}_k}$ can be consistently estimated, from which the theorem is derived.

Proof [of Theorem 5] As established by Adams and Nobel (2012), under the conditions of the theorem we have

$$\lim_{n\to\infty} \sup_{h\in\mathcal{H}_k} \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) = \sup_{h\in\mathcal{H}_k} \mathbf{E}_{\rho_X} h(X_1,\dots,X_k) \ \rho_X\text{-a.s.} \tag{4}$$

for all $k\in\mathbb{N}$, and likewise for ρ_Y. Fix an $\varepsilon > 0$. We can find a $T\in\mathbb{N}$ such that

$$\sum_{k>T} w_k \le \varepsilon. \tag{5}$$

Note that T depends only on ε. Moreover, as follows from (4), for each $k = 1..T$ we can find an N_k such that

$$\left| \sup_{h\in\mathcal{H}_k} \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \sup_{h\in\mathcal{H}_k} \mathbf{E}_{\rho_X} h(X_{1..k}) \right| \le \varepsilon/T. \tag{6}$$

Let $N_k := \max_{i=1..T} N_i$ and define analogously M for ρ_Y. Thus, for $n \ge N$, $m \ge M$ we have

$$\begin{aligned}
&\hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) \\
&\quad \le \sum_{k=1}^{T} w_k \sup_{h\in\mathcal{H}_k} \left| \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} h(Y_{i..i+k-1}) \right| + \varepsilon \\
&\qquad \le \sum_{k=1}^{T} w_k \sup_{h\in\mathcal{H}_k} \left\{ \left| \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \mathbf{E}_{\rho_1} h(X_{1..k}) \right| \right. \\
&\qquad\qquad + |\mathbf{E}_{\rho_1} h(X_{1..k}) - \mathbf{E}_{\rho_2} h(Y_{1..k})| \\
&\qquad\qquad \left. + \left| \mathbf{E}_{\rho_2} h(Y_{1..k}) - \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} h(Y_{i..i+k-1}) \right| \right\} + \varepsilon \\
&\qquad\qquad\qquad \le 3\varepsilon + D_{\mathbf{H}}(\rho_X, \rho_Y),
\end{aligned}$$

where the first inequality follows from the definition (3) of $\hat{D}_{\mathbf{H}}$ and from (5), and the last inequality follows from (6). Since ε was chosen arbitrary the statement follows. ■

4. Calculating $\hat{D}_\mathbf{H}$ Using Binary-Classification Methods

The methods for solving various statistical problems that we suggest are all based on $\hat{D}_\mathbf{H}$. The main appeal of this approach is that $\hat{D}_\mathbf{H}$ can be calculated using binary classification methods. Here we explain how to do it.

The definition (3) of $D_\mathbf{H}$ involves calculating l summands (where $l := \min\{n,m\}$), that is

$$\sup_{h\in\mathcal{H}_k}\left|\frac{1}{n-k+1}\sum_{i=1}^{n-k+1}h(X_{i..i+k-1})-\frac{1}{m-k+1}\sum_{i=1}^{m-k+1}h(Y_{i..i+k-1})\right| \qquad (7)$$

for each $k = 1..l$. Assuming that $h \in \mathcal{H}_k$ are indicator functions, calculating each of the summands amounts to solving the following k-dimensional binary classification problem. Consider $X_{i..i+k-1}$, $i = 1..n-k+1$ as class-1 examples and $Y_{i..i+k-1}$, $i = 1..m-k+1$ as class-0 examples. The supremum (7) is attained on $h \in \mathcal{H}_k$ that minimizes the empirical risk, with examples weighted with respect to the sample size. Indeed, we can define the weighted empirical risk of any $h \in \mathcal{H}_k$ as

$$\frac{1}{n-k+1}\sum_{i=1}^{n-k+1}(1-h(X_{i..i+k-1}))+\frac{1}{m-k+1}\sum_{i=1}^{m-k+1}h(Y_{i..i+k-1}), \qquad (8)$$

minimising which can be easily seen to be equivalent to (7).

Thus, as long as we have a way to find $h \in \mathcal{H}_k$ that minimizes empirical risk, we have a consistent estimate of $D_\mathcal{H}(\rho_X,\rho_Y)$, under the mild conditions on $\mathbf{H}$ required by Theorem 5. Since the dimension of the resulting classification problems grows with the length of the sequences, one should prefer methods that work in high dimensions, such as soft-margin SVMs (Cortes and Vapnik, 1995).

A particularly remarkable feature is that *the choice of $\mathcal{H}_k$ is much easier* for the problems that we consider *than in the binary classification* problem. Specifically, if (for some fixed k) the classifier that achieves the minimal (Bayes) error for the classification problem is not in $\mathcal{H}_k$, then obviously the error of an empirical risk minimizer will not tend to zero, no matter how much data we have. In contrast, all we need to achieve asymptotically 0 error in estimating $\hat{D}$ (and therefore, in the learning problems considered below) is that the sets $\mathcal{H}_k$ generate $\mathcal{F}_k$ and have a finite VC dimension (for each k). This is the case already for the set of half-spaces in $\mathbb{R}_k$. In other words, the *approximation* error of the binary classification method (the classification error of the best f in $\mathcal{H}_k$) is not important. What is important is the estimation error; for asymptotic consistency results it has to go to 0 (hence the requirement on the VC dimension); for non-asymptotic results, it will appear in the error bounds, see Section 7. Thus, we have the following statement.

Claim 1 *The error $|D_\mathbf{H}(\rho_X,\rho_Y)-\hat{D}_\mathbf{H}(X,Y)|$, and thus the error of the algorithms below, can be much smaller than the error of classification algorithms used to calculate $D_\mathbf{H}(X,Y)$.*

We can conclude that, beyond the requirement that $\mathcal{H}_k$ generate $\mathcal{F}_k$ for each $k \in \mathbb{N}$, the choice of H_k (or, say, of the kernel to use in SVM) is entirely up to the needs and constraints of specific applications.

Remark (number of summands in $\hat{D}_\mathbf{H}$) Finally, we note that while in the definition of the empirical distributional distance (3) the number of summands is l (the length of the shorter of the two samples), it can be replaced with any γ_l such that $\gamma_l \to \infty$, without affecting any asymptotic consistency results. In other words, Theorem 5, as well as all the consistency statements below, holds true for l replaced with any non-decreasing function γ_l that tends to infinity with l. A practically viable choice is $\gamma_l = \log l$; in fact, there is no reason to choose faster growing γ_n since the estimates for higher-order summands will not have enough data to converge. This is also the value we use in the experiments.

Remark (relation to total variation) An illustrative example[1] of the choice of the sets $\mathcal{H}_k$ is the set of indicators of all measurable subsets of $\mathcal{X}^k$. In this case each summand in (2) is the total variation distance between the k-dimensional marginal distributions of ρ_1 and ρ_2. Take, for simplicity, $k = 1$; denoting P and Q the corresponding single-dimensional marginals, the distance becomes $\sup_A |P(A) - Q(A)|$ (cf. (1)). This supremum is reached on the set $A^* := \{x \in \mathcal{X} : f(x) \geq g(x)\}$, where f and g are densities of P and Q with respect to some arbitrary measure that dominates both P and Q (e.g., $1/2(P+Q)$). A binary classifier corresponding to a set A declares P if $x \in A$ and Q otherwise. The optimal classification error is $\inf_A(1 - P(A) + Q(A)) = 1 - \sup_A(P(A) + Q(A)) = 1 - P(A^*) + Q(A^*)$ (cf. (8)). In general, estimating the total variation distance (and finding the best classifier) is not possible, so using smaller sets $\mathcal{H}_k$ can be viewed as a regularization of this problem.

5. The Three-Sample Problem

We start with a conceptually simple problem known in statistics as the three-sample problem (sometimes also called time-series classification). We are given three samples $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_m)$ and $Z = (Z_1, \dots, Z_l)$. It is known that X and Y were generated by different time-series distributions, whereas Z was generated by the same distribution as either X or Y. It is required to find out which one is the case. Both distributions are assumed to be stationary ergodic, but no further assumptions are made about them (no independence, mixing or memory assumptions). The three sample-problem for dependent time series has been addressed by Gutman (1989) for Markov processes and by Ryabko and Ryabko (2010) for stationary ergodic time series. The latter work uses an approach based on the distributional distance.

Indeed, to solve this problem it suffices to have consistent estimates of some distance between time series distributions. Thus, we can use the telescope distance. The following statement is a simple corollary of Theorem 5.

Theorem 6 *Let the samples $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_m)$ and $Z = (Z_1, \dots, Z_l)$ be generated by stationary ergodic distributions ρ_X, ρ_Y and ρ_Z, with $\rho_X \neq \rho_Y$ and either (i) $\rho_Z = \rho_X$ or (ii) $\rho_Z = \rho_Y$. Let the sets $\mathcal{H}_k$, $k \in \mathbb{N}$ be separable sets of indicator functions over $\mathcal{X}^k$. Assume that each set $\mathcal{H}_k$, $k \in \mathbb{N}$ has a finite VC dimension and generates $\mathcal{F}_k$. A test that declares that (i) is true if $\hat{D}_{\mathbf{H}}(Z,X) \leq \hat{D}_{\mathbf{H}}(Z,Y)$ and that (ii) is true otherwise, makes only finitely many errors with probability 1 as $n,m,l \to \infty$.*

It is straightforward to extend this theorem to more than two classes; in other words, instead of X and Y one can have an arbitrary number of samples from different stationary ergodic distributions. A further generalization of this problem is the problem of time-series clustering, considered in the next section.

[1] This example was suggested by an anonymous reviewer.

6. Clustering Time Series

We are given N time-series samples $X^1 = (X^1_1, \dots, X^1_{n_1}), \dots, X^N = (X^N_1, \dots, X^N_{n_N})$, and it is required to cluster them into K groups, where, in different settings, K may be either known or unknown. While there may be many different approaches to define what should be considered a good clustering, and, thus, what it means to have a consistent clustering algorithm, for the problem of clustering time-series samples there is a natural choice, proposed by Ryabko (2010a): Assume that each of the time-series samples $X^1 = (X^1_1, \dots, X^1_{n_1}), \dots, X^N = (X^N_1, \dots, X^N_{n_N})$ was generated by one out of K different time-series distributions $\rho_1, \dots, \rho_K$. These distributions are unknown. The *target clustering* is defined according to whether the samples were generated by the same or different distributions: the samples belong to the same cluster if and only if they were generated by the same distribution. A clustering algorithm is called *asymptotically consistent* if with probability 1 from some n on it outputs the target clustering, where n is the length of the shortest sample $n := \min_{i=1..N} n_i \geq n'$.

Again, to solve this problem it is enough to have a metric between time-series distributions that can be consistently estimated. Our approach here is based on the telescope distance, and thus we use $\hat{D}$.

The clustering problem is relatively simple if the target clustering has what is called the *strict separation property* (Balcan et al., 2008): every two points in the same target cluster are closer to each other than to any point from a different target cluster. The following statement is an easy corollary of Theorem 5.

Theorem 7 *Let the sets $\mathcal{H}_k$, $k \in \mathbb{N}$ be separable sets of indicator functions over $\mathcal{X}^k$. Assume that each set $\mathcal{H}_k$, $k \in \mathbb{N}$ has a finite VC dimension and generates $\mathcal{F}_k$. If the distributions $\rho_1, \dots, \rho_K$ generating the samples $X^1 = (X^1_1, \dots, X^1_{n_1}), \dots, X^N = (X^N_1, \dots, X^N_{n_N})$ are stationary ergodic, then with probability 1 from some $n := \min_{i=1..N} n_i$ on the target clustering has the strict separation property with respect to $\hat{D}_{\mathbf{H}}$.*

With the strict separation property at hand, if the number of clusters K is known, it is easy to find asymptotically consistent algorithms. Here we give some simple examples, but the theorem below can be extended to many other distance-based clustering algorithms.

The *average linkage* algorithm works as follows. The distance between clusters is defined as the average distance between points in these clusters. First, put each point into a separate cluster. Then, merge the two closest clusters; repeat the last step until the total number of clusters is K. The *farthest point* clustering works as follows. Assign $c_1 := X^1$ to the first cluster. For $i = 2..K$, find the point X^j, $j \in \{1..N\}$ that maximizes the distance $\min_{t=1..i} \hat{D}_{\mathbf{H}}(X^j, c_t)$ (to the points already assigned to clusters) and assign $c_i := X^j$ to the cluster i. Then assign each of the remaining points to the nearest cluster. The following statement is a corollary of Theorem 7.

Theorem 8 *Under the conditions of Theorem 7, average linkage and farthest point clusterings are asymptotically consistent, provided the correct number of clusters K is given to the algorithm.*

Note that we do not require the samples to be independent; the joint distributions of the samples may be completely arbitrary, as long as the marginal distribution of each sample is stationary ergodic. These results can be extended to the online setting in the spirit of Khaleghi et al. (2012).

For the case of unknown number of clusters, the situation is different: one has to make stronger assumptions on the distributions generating the samples, since there is no algorithm that is consistent for all stationary ergodic distributions (Ryabko, 2010b); such stronger assumptions are considered in the next section.

7. Speed of Convergence

The results established so far are asymptotic out of necessity: they are established under the assumption that the distributions involved are stationary ergodic, which is too general to allow for any meaningful finite-time performance guarantees. While it is interesting to be able to establish consistency results under such general assumptions, it is also interesting to see what results can be obtained under stronger assumptions. Moreover, since it is usually not known in advance whether the data at hand satisfies given assumptions or not, it appears important to have methods that have *both* asymptotic consistency in the general setting and finite-time performance guarantees under stronger assumptions. It turns out that this is possible: for the methods based on $\hat{D}$ one can establish both the asymptotic performance guarantees for all stationary ergodic distributions and finite-sample performance guarantees under stronger assumptions, namely the uniform mixing conditions introduced below.

Another reason to consider stronger assumptions on the distributions generating the data is that some statistical problems, such as homogeneity testing or clustering when the number of clusters is unknown, are provably impossible to solve under the only assumption of stationary ergodic distributions, as shown by Ryabko (2010b).

Thus, in this section we analyse the speed of convergence of $\hat{D}$ under certain mixing conditions, and use it to construct solutions for the problems of homogeneity and clustering with an unknown number of clusters, as well as to establish finite-time performance guarantees for the methods presented in the previous sections.

A stationary distribution on the space of one-way infinite sequences $(\mathcal{X}^{\mathbb{N}}, \mathcal{F})$ can be uniquely extended to a stationary distribution on the space of two-way infinite sequences $(\mathcal{X}^{\mathbb{Z}}, \mathcal{F}_{\mathbb{Z}})$ of the form $\dots, X_{-1}, X_0, X_1, \dots$.

Definition 9 (β-mixing coefficients) *For a process distribution* ρ *define the mixing coefficients*

$$\beta(\rho,k) := \sup_{\substack{A\in\sigma(X_{-\infty..0}),\\ B\in\sigma(X_{k..\infty})}} |\rho(A\cap B)-\rho(A)\rho(B)|$$

where $\sigma(..)$ *denotes the sigma-algebra of the random variables in brackets.*

When $\beta(\rho,k)\to 0$ the process ρ is called uniformly β-mixing (with coefficients $\beta(\rho,k)$); this condition is much stronger than ergodicity, but is much weaker than the i.i.d. assumption. For more information on mixing see, for example, Bosq (1996).

7.1 Speed of Convergence of $\hat{D}$

Assume that a sample $X_{1..n}$ is generated by a distribution ρ that is uniformly β-mixing with coefficients $\beta(\rho,k)$. Assume further that $\mathcal{H}_k$ is a set of indicator functions with a finite VC dimension d_k, for each $k\in\mathbb{N}$.

Since in this section we are after finite-time bounds, we fix a concrete choice of the weights w_k in the definition of $\hat{D}$ (Definition 2),

$$w_k := 2^{-k}. \tag{9}$$

The general tool that we use to obtain performance guarantees in this section is the following bound that can be obtained from the results of Karandikar and Vidyasagar (2002).

$$q_n(\rho,\mathcal{H}_k,\varepsilon) := \rho\left(\sup_{h\in\mathcal{H}_k}\left|\frac{1}{n-k+1}\sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \mathbf{E}_\rho h(X_{1..k})\right| > \varepsilon\right)$$
$$\le n\beta(\rho, t_n-k) + 8t_n^{d_k+1}e^{-l_n\varepsilon^2/8}, \quad (10)$$

where t_n are any integers in $1..n$ and $l_n = n/t_n$. The parameters t_n should be set according to the values of β in order to optimize the bound.

One can use similar bounds for classes of finite Pollard dimension (Pollard, 1984) or more general bounds expressed in terms of covering numbers, such as those given by Karandikar and Vidyasagar (2002). Here we consider classes of finite VC dimension only for the ease of the exposition and for the sake of continuity with the previous section (where it was necessary).

Furthermore, for the rest of this section we assume geometric β-mixing distributions, that is, $\beta(\rho,t) \le \gamma^t$ for some $\gamma < 1$. Letting $l_n = t_n = \sqrt{n}$ the bound (10) becomes

$$q_n(\rho,\mathcal{H}_k,\varepsilon) \le n\gamma^{\sqrt{n}-k} + 8n^{(d_k+1)/2}e^{-\sqrt{n}\varepsilon^2/8}. \quad (11)$$

Lemma 10 *Let two samples $X_{1..n}$ and $Y_{1..m}$ be generated by stationary distributions ρ_X and ρ_Y whose β-mixing coefficients satisfy $\beta(\rho_\cdot,t) \le \gamma^t$ for some $\gamma < 1$. Let $\mathcal{H}_k$, $k \in \mathbb{N}$ be some sets of indicator functions on $\mathcal{X}^k$ whose VC dimension d_k is finite and non-decreasing with k. Then*

$$P(|\hat{D}_{\mathbf{H}}(X_{1..n},Y_{1..m}) - D_{\mathbf{H}}(\rho_X,\rho_Y)| > \varepsilon) \le 2\Delta(\varepsilon/4,n') \quad (12)$$

where $n' := \min\{n,m\}$, the probability is with respect to $\rho_X \times \rho_Y$ and

$$\Delta(\varepsilon,n) := -\log\varepsilon(n\gamma^{\sqrt{n}+\log(\varepsilon)} + 8n^{(d_{-\log\varepsilon}+1)/2}e^{-\sqrt{n}\varepsilon^2/8}). \quad (13)$$

Proof From (9) we have $\sum_{k=-\log\varepsilon/2}^{\infty} w_k < \varepsilon/2$. Using this and the definitions 2 and 4 of $D_{\mathbf{H}}$ and $\hat{D}_{\mathbf{H}}$ we obtain

$$P(|\hat{D}_{\mathbf{H}}(X_{1..n_1},Y_{1..n_2}) - D_{\mathbf{H}}(\rho_X,\rho_Y)| > \varepsilon) \le \sum_{k=1}^{-\log(\varepsilon/2)} (q_n(\rho_X,\mathcal{H}_k,\varepsilon/4) + q_n(\rho_Y,\mathcal{H}_k,\varepsilon/4)),$$

which, together with (11), implies the statement. ■

7.2 Homogeneity Testing

Given two samples $X_{1..n}$ and $Y_{1..m}$ generated by distributions ρ_X and ρ_Y respectively, the problem of homogeneity testing (or the two-sample problem) consists in deciding whether $\rho_X = \rho_Y$. A test is called (asymptotically) consistent if its probability of error goes to zero as $n' := \min\{m,n\}$ goes to infinity. As mentioned above, in general, for stationary ergodic time series distributions there is no asymptotically consistent test for homogeneity (Ryabko, 2010b) (even for binary-valued time series); thus, stronger assumptions are in order.

Homogeneity testing is one of the classical problems of mathematical statistics, and one of the most studied ones. Vast literature exits on homogeneity testing for i.i.d. data, and for dependent processes as well. We do not attempt to survey this literature here. Our contribution to this line of research is to show that this problem can be reduced (via the telescope distance) to binary classification, in the case of strongly dependent processes satisfying some mixing conditions.

It is easy to see that under the mixing conditions of Lemma 10 a consistent test for homogeneity exists, and finite-sample performance guarantees can be obtained. It is enough to find a sequence $\varepsilon_n \to 0$ such that $\Delta(\varepsilon_n, n) \to 0$ (see (13)). Then the test can be constructed as follows: say that the two sequences $X_{1..n}$ and $Y_{1..m}$ were generated by the same distribution if $\hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) < \varepsilon_{\min\{n,m\}}$; otherwise say that they were generated by different distributions.

Theorem 11 *Under the conditions of Lemma 10 the probability of Type I error (the distributions are the same but the test says they are different) of the described test is upper-bounded by* $2\Delta(\varepsilon/4, n')$. *The probability of Type II error (the distributions are different but the test says they are the same) is upper-bounded by* $2\Delta((\delta-\varepsilon)/4, n')$ *where* $\delta := D_{\mathbf{H}}(\rho_X, \rho_Y)$.

Proof The statement is an immediate consequence of Lemma 10. Indeed, for the Type I error, the two sequences are generated by the same distribution, so the probability of error of the test is given by (12) with $D_{\mathbf{H}}(\rho_X, \rho_Y) = 0$. The probability of Type II error is given by $P(D_{\mathbf{H}}(\rho_X, \rho_Y) - \hat{D}_{\mathbf{H}}(X_{1..n_1}, Y_{1..n_2}) > \delta - \varepsilon)$, which is upper-bounded by $2\Delta((\delta-\varepsilon))/4, n')$ as follows from (12). ■

The optimal choice of ε_n may depend on the speed at which d_k (the VC dimension of $\mathcal{H}_k$) increases; however, for most natural cases (recall that $\mathcal{H}_k$ are also parameters of the algorithm) this growth is polynomial, so the main term to control is $e^{-\sqrt{n}\varepsilon^2/8}$.

For example, if $\mathcal{H}_k$ is the set of halfspaces in $\mathcal{X}^k = \mathbb{R}^k$ then $d_k = k+1$ and one can choose $\varepsilon_n := n^{-1/8}$. The resulting probability of Type I error decreases as $\exp(-n^{1/4})$.

7.3 Clustering with a Known or Unknown Number of Clusters

If the distributions generating the samples satisfy certain mixing conditions, then we can augment Theorems 7 and 8 with finite-sample performance guarantees.

Theorem 12 *Let the distributions* $\rho_1, \dots, \rho_k$ *generating the samples* $X^1 = (X^1_1, \dots, X^1_{n_1}), \dots, X^N = (X^N_1, \dots, X^N_{n_N})$ *satisfy the conditions of Lemma 10. Let* $n := \min_{i=1..N} n_i$ *and* $\delta := \min_{i,j=1..N, i\neq j} D_{\mathbf{H}}(\rho_i, \rho_j)$. *Then with probability at least* $1 - N(N-1)\Delta(\delta/12, n')$ *the target clustering of the samples has the strict separation property. In this case single linkage and farthest point algorithms output the target clustering.*

Proof Note that a sufficient condition for the strict separation property to hold is that for every pair i, j of samples generated by the same distribution we have $\hat{D}_{\mathbf{H}}(X^i, X^j) \le \delta/3$, and for every pair i, j of samples generated by different distributions we have $\hat{D}_{\mathbf{H}}(X^i, X^j) \ge 2\delta/3$. Using Lemma 10, the probability of such an even (for each pair) is upper-bounded by $2\Delta(\delta/12, n')$, which, multiplied by the total number $N(N-1)/2$ of pairs gives the statement. The second statement is obvious. ■

As with homogeneity testing, while in the general case of stationary ergodic distributions it is impossible to have a consistent clustering algorithm when the number of clusters k is unknown, the situation changes if the distributions satisfy certain mixing conditions. In this case a consistent clustering algorithm can be obtained as follows. Assign to the same cluster all samples that are at most ε_n-far from each other, where the threshold ε_n is selected the same way as for homogeneity testing: $\varepsilon_n \to 0$ and $\Delta(\varepsilon_n, n) \to 0$. The optimal choice of this parameter depends on the choice of $\mathcal{H}_k$ through the speed of growth of the VC dimension d_k of these sets.

Theorem 13 *Given N samples generated by k different stationary distributions ρ_i, $i = 1..k$ (unknown k) all satisfying the conditions of Lemma 10, the probability of error (misclustering at least one sample) of the described algorithm is upper-bounded by*

$$N(N-1)\max\{\Delta(\varepsilon/4, n'), \Delta((\delta-\varepsilon)/4, n')\}$$

where $\delta := \min_{i,j=1..k, i\neq j} D_{\mathbf{H}}(\rho_i, \rho_j)$ and $n = \min_{i=1..N} n_i$, with n_i, $i = 1..N$ being lengths of the samples.

Proof The statement follows from Theorem 11. ■

8. Other Metrics for Time-Series Distributions

The previous sections introduce a new metric on the space of time-series distributions, and use its empirical estimates to solve several learning problems. In this section we attempt to put the telescope distance into a more general context, and take a broader look at metrics between time-series distributions.

Introduce the notation μ_k for the k-dimensional marginal distribution of a time-series distribution μ.

8.1 sum **Distances**

Observe that the telescope distance $D_{\mathbf{H}}$ has the form

$$D(\mu, \nu) = \sum_{k\in\mathbb{N}} w_k d_k(\mu_k, \nu_k), \tag{14}$$

where w_k are summable positive real weights.

It is easy to see that distances of this form can be consistently estimated, as long as d_k can be consistently estimated for each $k \in \mathbb{N}$; this is formalized in the following statement.

Proposition 14 (estimating sum-based distances) *Let $\mathcal{C}$ be a set of distributions over $\mathcal{X}^{\mathbb{N}}$. Let $d_k, k \in \mathbb{N}$ be a series of distances on the spaces of distributions over $\mathcal{X}^k$, such that $d_k(\mu_k, \nu_k) \leq a \in \mathbb{R}$ for all $\mu, \nu \in \mathcal{C}$ and such that there exists a series $\hat{d}_k(X_{1..n}, Y_{1..n}), k \in \mathbb{N}$ of their consistent estimates: for each $\mu, \nu \in \mathcal{C}$ we have $\lim_{n\to\infty} \hat{d}_k(X_{1..n}, Y_{1..n}) = d_k(\mu_k, \nu_k)$ a.s., whenever $\mu, \nu \in \mathcal{C}$ are chosen to generate the sequences. Then the distance D given by (14) can be consistently estimated using the estimate $\sum_{k\in\mathbb{N}} w_k \hat{d}_k(X_{1..n}, Y_{1..n})$.*

Proof The proof is an easy generalization of the proof of Theorem 5, with the condition on $\hat{d}_k$ used instead of (4). ■

Clearly, $D_{\mathbf{H}}$ is an example of a distance in the form (14), and it satisfies the conditions of the proposition with $\mathcal{C}$ being the set of all stationary ergodic processes.

Another example of a distance in the form (14) is given by the so-called distributional distance (Gray, 1988; Shields, 1996), whose definition is given below. Empirical estimates of this distance are asymptotically consistent for stationary ergodic time series, and thus can be used (Ryabko and Ryabko, 2010; Ryabko, 2010a; Khaleghi et al., 2012; Khaleghi and Ryabko, 2012; Ryabko, 2012) to solve various statistical problems, including those considered above.

To define the distributional distance, let, for each $k, l \in \mathbb{N}$, the set $B^{k,l}$ be some partition of the set $\mathcal{X}^k$, such that the set $B^k = \cup_{l\in\mathbb{N}} B^{k,l}$ generates $\mathcal{F}_k$. Let also $\mathcal{B} = \cup_{k=1}^{\infty} B^k$. Note that the set $\{B \times \mathcal{X}^{\mathbb{N}} : B \in B^{k,l}, k, l \in \mathbb{N}\}$ generates $\mathcal{F}$.

Definition 15 (distributional distance) *The distributional distance is defined for a pair of processes* ρ_1, ρ_2 *as follows*

$$D_{dd}(\rho_1,\rho_2) := \sum_{m,l=1}^{\infty} w_m w_l \sum_{B\in B^{m,l}} |\rho_1(B) - \rho_2(B)|, \tag{15}$$

where $w_k, k \in \mathbb{N}$ *is a summable sequence of positive real weights (e.g.,* $w_j = 2^{-j}$*).*

Remark. A more general definition, which is not specific to time-series distributions, is to take any sequence $B_j \in \mathcal{F}_1$, $j \in \mathbb{N}$ of events that generate the sigma-algebra $\mathcal{F}$ of a probability space $(\mathcal{X}, \mathcal{F})$, and then define

$$D'_{dd}(\rho_1,\rho_2) := \sum_{j=1}^{\infty} w_j |\rho_1(B_j) - \rho_2(B_j)|; \tag{16}$$

see Gray (1988) for a general treatment. The latter definition is sometimes more convenient for theoretical analysis (Ryabko, 2012), while the distance (15), which makes explicit the marginal distributions on $\mathcal{X}^m$, $m \in \mathbb{N}$ and the level l of discretisation $B^{m,l}$ of each set $\mathcal{X}^m$, is more suited for time-series, and, specifically, for implementing algorithms, see Ryabko and Ryabko (2010), Khaleghi et al. (2012) and Khaleghi and Ryabko (2012).

In general, it is perhaps impossible to tell which distance, specifically, $D_{\mathbf{H}}$ or D_{dd}, should be preferred for which problem. Conceptually, one of the advantages of the telescope distance $D_{\mathbf{H}}$ is that one can use different sets $\mathbf{H}$—the choice that makes it adaptable to applications. Another is that one can reuse readily available classification methods for calculating its empirical estimates. One formal way to compare different metrics is to compare the resulting topologies. This is done in the end of this section.

8.2 sup Distances

A different way to construct a distance between time-series distributions based on their finite-dimensional marginals is to use the supremum instead of summation in (14):

$$d(\mu,\nu) = \sup_{k\in\mathbb{N}} d_k(\mu_k,\nu_k). \tag{17}$$

Some commonly used metrics are defined in the form (17) or have natural interpretations in this form, as the following two examples show.

Definition 16 (total variation) *For time-series distributions* ν, μ *the total variation distance between them is defined as* $D_{tv}(\mu,\nu) := \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|$.

It is easy to see that $D_{tv}(\mu,\nu) = \sup_{k\in\mathbb{N}} \sup_{A\in\mathcal{F}_k} |\mu(A)-\nu(A)|$, so that the total variation distance has the form (17).

However, the total variation distance is not very useful for time-series distributions for the following two reasons. First of all, for stationary ergodic distributions it is degenerate: $D_{tv}(\mu,\nu)=1$ if and only if $\mu\neq\nu$. This follows from the fact that any two different stationary ergodic distributions are singular. Such a distance could still be useful as a formalization of the problem of homogeneity testing. However, the problem of homogeneity testing is impossible to solve based on sampling for stationary ergodic distributions (and even for a smaller family of B processes, see below) (Ryabko, 2010b), so the use of this distance remains limited to more restrictive classes of distributions.

This hints at an intrinsic problem with distances defined in the form (17). The problem is in the difficulties to estimate such metrics based on sampling. At each time step t we observe only a sample of finite length, say n_t, and based on this we want to estimate a quantity that involves k-dimensional marginals for all k, including those with $k > n_t$. Considering a growing (with t) number of marginals for the estimate may be a route to take, but this turns out to be difficult to analyse, especially if no rates of convergence can be established for the set of time-series distributions at hand. This problem is highlighted by the example of the so-called $\bar{d}$ distance, whose definition follows.

Definition 17 ($\bar{d}$ distance) *Assume some distance δ over $\mathcal{X}$ is given. For two time-series distributions μ and ν define*

$$\bar{d}(\mu,\nu) := \sup_{k\in\mathbb{N}} \frac{1}{k} \inf_{p\in P} \sum_{i=1}^{k} \mathbf{E}_p \delta(x_i,y_i),$$

where P is the set of all distributions over $\mathcal{X}^k\times\mathcal{X}^k$ generating a pair of sequences $x_{1..k}, y_{1..k}$ whose marginal distributions are μ_k and ν_k correspondingly.

A process is called a *B-process* (or a Bernoulli process) if it is in the $\bar{d}$-closure of the set of all aperiodic stationary ergodic k-step Markov processes, where $k\in\mathbb{N}$. For more information on $\bar{d}$-distance and B-processes see Gray (1988) and Shields (1996). The set of B-processes is a strict subset of the set of all stationary ergodic time-series distributions. It turns out that $\bar{d}$ distance is impossible to estimate for the latter, while it can be estimated for the former (Ornstein and Weiss, 1990).

Theorem 18 (Ornstein and Weiss, 1990) *There exists an estimator $\hat{\bar{d}}(X_{1..n},Y_{1..n})$ such that, if $X_{1..n},Y_{1..n}$ are generated by B-processes μ and ν then $\hat{\bar{d}}(X_{1..n},Y_{1..n})\to\bar{d}(\mu,\nu)$ a.s. However, for any estimator $\hat{\bar{d}}(X_{1..n},Y_{1..n})$ there is a pair of stationary ergodic processes μ and ν such that $\limsup_{n\to\infty}|\hat{\bar{d}}(X_{1..n},Y_{1..n})-\bar{d}(\mu,\nu)|>1/2$.*

8.3 Comparison with the Distributional Distance

In this section we show that the telescope distance is stronger than the distributional distance in the topological sense. Since in fact both the telescope distance and the distributional distance are families of distances (the telescope distance depends on the sequence $\mathbf{H}$), we will fix a simple natural choice of each of these metrics. In general, different choices of parameters produce topologically non-equivalent metrics; it is easy to check that the analysis in this section extends to many other natural choices.

Thus, for the purpose of this section, let us fix $\mathcal{X} = \mathbb{R}$ and let H_k^0 be the set of halfspaces in $\mathcal{X}^k$. Denote $\mathbf{H}^0 := (\mathcal{H}_k^0 : k \in \mathbb{N})$. Clearly, these $\mathcal{H}_k$ satisfy all the conditions of the theorems of Sections 5 and 6.

For the distributional distance (Definition 15), set $B^{k,l}$ to be the partition of the set $\mathcal{X}^k$ into k-dimensional cubes with volume $h_l^k = (1/l)^k$. Denote D_{dd}^0 the distributional distance D_{dd} with this set of parameters.

Definition 19 *A metric d_1 is said to be* stronger *than a metric d_2 if any sequence that converges in d_1 also converges in d_2. If, in addition, d_2 is not stronger than d_1, then d_1 is called* strictly stronger.

Note that for the distributional distance, if we use the same sets B_k to generate the sigma algebras $\mathcal{X}^k$ then the distance defined by (15) is stronger than the distance defined by (16).

Theorem 20 *$D_{\mathbf{H}^0}$ is strictly stronger than D_{dd}^0.*

Proof Fix any $\varepsilon > 0$ and find a $T \in \mathbb{N}$ such that $\sum_{m,l>T} w_m w_l < \varepsilon$. Let ρ_i, $i \in \mathbb{N}$ be a sequence of process measures that converges in $D_{\mathbf{H}^0}$. Let A^k be the set of all complements to $\mathcal{X}^k$ of cubes with sides of length s, for all $s \in \mathbb{N}$. Note that any cube B in B_k, as well as any set A in A^k, can be obtained by intersecting $2k$ halfspaces. Therefore, we have

$$\sup_{B \in B^k \cup A^k} |\rho_i(B) - \rho_j(B)| \le 2k d_{\mathcal{H}_k}(\rho_i, \rho_j) \le 2k w_k^{-1} D_{\mathbf{H}^0}(\rho_i, \rho_j), \tag{18}$$

where the second inequality follows from the definition of $D_{\mathbf{H}^0}$. Observe that for each $i \in \mathbb{N}$ one can find a set $A_i \in A^k$ such that $\rho_i(A_i) < \varepsilon/2$. From this, (18) and the fact that the sequence ρ_i converges in $D_{\mathbf{H}^0}$, we conclude that there is a set $A \in A^k$ such that

$$\rho_i(A) < \varepsilon$$

for all $i \ge j_k$. For all $k, l \in \mathbb{N}$ one can find $M_{k,l} \in \mathbb{N}$ such that the complement of A (which is a cube in $\mathcal{X}^k$) is contained in the union of $M_{k,l}$ cubes from $B_{k,l}$. Let $M := \max_{k,l \le T} M_{k,l}$ and $J := \max_{i \le T} j_i$. Using (18) and the definition of the partitions $B^{k,l}$ we can derive

$$\sum_{B \in B^{k,l}, B \not\subseteq A_k} |\rho_i(B) - \rho_j(B)| \le 2MTw_T^{-1} D_{\mathbf{H}^0}(\rho_i, \rho_j)$$

for any $i, j \ge J$ and all $k, l \le T$. Increasing J if necessary to have $2MTw_T^{-1} D_{\mathbf{H}^0}(\rho_i, \rho_j) < \varepsilon$ for all $i, j \ge J$, we obtain

$$D_{dd}^0(\rho_i, \rho_j) \le \sum_{m,l=1}^{T} w_m w_l \sum_{B \in B^{m,l}, B \not\subseteq A_m} |\rho_i(B) - \rho_j(B)| + 2\varepsilon \le 3\varepsilon$$

for all $i, j > J$, which means that the sequence $\rho_i, i \in \mathbb{N}$ converges in D_{dd}^0. Thus, $D_{\mathbf{H}^0}$ is stronger than D_{dd}^0.

It remains to show that D_{dd}^0 is not stronger than $D_{\mathbf{H}^0}$. To see this, consider the following sequence of subsets of $\mathcal{X} = \mathbb{R}$. f is the dot $\{0\}$, and f_k is the interval $[0, 1/k]$, for each $k \in \mathbb{N}$. Define the distributions ν_j for $j \in \mathbb{N}$ as uniform on f_j, and let ν be concentrated on f; since we need time-series distributions, extend this i.i.d. for all $n \in \mathbb{N}$. It is easy to check that $\lim_{i \in \mathbb{N}} D_{dd}^0(\nu_i, \nu_0) = 0$ while $D_{\mathbf{H}^0}(\nu_i, \nu_0) = 1$ for all $i > 0$. ■

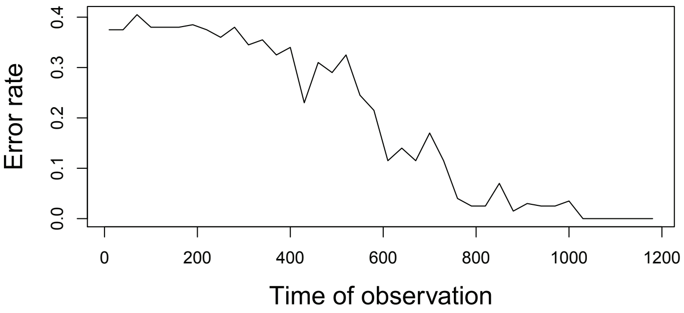

Figure 1: Error of two-class clustering using $\mathrm{TS}_{\mathrm{SVM}}$; 10 time series in each target cluster, averaged over 20 runs.

9. Experimental Evaluation

For experimental evaluation we chose the problem of time-series clustering. The average-linkage clustering is used, with the telescope distance between samples calculated using an SVM, as described in Section 4. In all experiments, SVM is used with radial basis kernel, with default parameters of libsvm (Chang and Lin, 2011). The parameters w_k in the definition of the telescope distance (Definition 2) are set to $w_k := k^{-2}$.

9.1 Synthetic Data

For the artificial setting we chose highly-dependent time-series distributions which have the same single-dimensional marginals and which cannot be well approximated by finite- or countable-state models. Variants of this family of distributions are standard examples in ergodic theory and dynamical systems (see, for example, Billingsley, 1965; Gray, 1988; Shields, 1996). The distributions $\rho(\alpha)$, $\alpha \in (0,1)$, are constructed as follows. Select $r_0 \in [0,1]$ uniformly at random; then, for each $i = 1..n$ obtain r_i by shifting r_{i-1} by α to the right, and removing the integer part. The time series $(X_1, X_2, \dots)$ is then obtained from r_i by drawing a point from a distribution law $\mathcal{N}_1$ if $r_i < 0.5$ and from $\mathcal{N}_2$ otherwise. $\mathcal{N}_1$ is a 3-dimensional Gaussian with mean of 0 and covariance matrix $\mathrm{Id} \times 1/4$. $\mathcal{N}_2$ is the same but with mean 1. If α is irrational[2] then the distribution $\rho(\alpha)$ is stationary ergodic, but does not belong to any simpler natural distribution family; in particular, it is not a B-processes (Shields, 1996). The single-dimensional marginal is the same for all values of α. The latter two properties make all parametric and most non-parametric methods inapplicable to this problem.

In our experiments, we use two process distributions $\rho(\alpha_i), i \in \{1,2\}$, with $\alpha_1 = 0.31...$, $\alpha_2 = 0.35...$,. The dependence of error rate on the length of time series is shown on Figure 1. One clustering experiment on sequences of length 1000 takes about 5 min. on a standard laptop.

9.2 Real Data

To demonstrate the applicability of the proposed methods to realistic scenarios, we chose the brain-computer interface data from BCI competition III (Millán, 2004). The data set consists of (pre-processed) BCI recordings of mental imagery: a person is thinking about one of three subjects

[2] In the experiments we used a `longdouble` with a long mantissa

	s_1	s_2	s_3
TS_{SVM}	**84%**	**81%**	**61%**
DTW	46%	41%	36%
KCpA	79%	74%	61%
SVM	76%	69%	60%

Table 1: Clustering accuracy in the BCI data set. 3 subjects (columns), 4 methods (rows). Our method is TS_{SVM}.

(left foot, right foot, a random letter). Originally, each time series consisted of several consecutive sequences of different classes, and the problem was supervised: three time series for training and one for testing. We split each of the original time series into classes, and then used our clustering algorithm in a completely unsupervised setting. The original problem is 96-dimensional, but we used only the first 3 dimensions (using all 96 gives worse performance). The typical sequence length is 300. The performance is reported in Table 1, labelled TS_{SVM}. All the computation for this experiment takes approximately 6 minutes on a standard laptop.

The following methods were used for comparison. First, we used dynamic time wrapping (DTW) (Sakoe and Chiba, 1978) which is a popular base-line approach for time-series clustering. The other two methods in Table 1 are from the paper of Harchaoui et al. (2008). The comparison is not fully relevant, since the results of Harchaoui et al. (2008) are for different settings; the method KCpA was used in change-point estimation method (a different but also unsupervised setting), and SVM was used in a supervised setting. The latter is of particular interest since the classification method we used in the telescope distance is also SVM, but our setting is unsupervised (clustering). On this data set the telescope distance demonstrates better performance than the comparison methods, which indicates that it can be useful in real-world scenarios.

10. Outlook

We have proposed a binary-classifier-based metric and shown how it can be used to solve several problems concerning highly dependent time series. The consistency results obtained concern the use of the empirical risk minimizer as a binary classifier. For applications this suggests using classifiers that approximate empirical risk minimizers over target sets of (indicator) functions. It is easy to extend the definition of the metric so that any classifier can be used, including such classifiers as nearest-neighbours rules. However, in order to extend the obtained results to such classifiers, one would need to establish the consistency of the empirical estimates of the resulting metric between time-series distributions, which means extending the results concerning the corresponding classifiers from the i.i.d. samples to stationary ergodic time series. Note that, while consistency of the empirical estimates of the time-series metric used is sufficient for the analysis of the learning problems considered in this work, it is not sufficient for some other learning problems concerning dependent time series that rely on a metric between time-series distributions. For example, some change-point problems for stationary ergodic time series can be solved using the distributional distance (Ryabko and Ryabko, 2010; Khaleghi and Ryabko, 2012, 2013). It remains to see whether the same results can be obtained with the telescope distance and its generalizations.

Acknowledgments

This work is an extended version of the NIPS'12 paper (Ryabko and Mary, 2012). The authors are grateful to the anonymous reviewers for the numerous constructive comments that helped us to improve the paper. This research was partially supported by the French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007-2013, ANR project Lampada (ANR-09-EMER-007) and by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 231495 (project CompLACS).

References

T. M. Adams and A. B. Nobel. Uniform approximation of Vapnik-Chervonenkis classes. *Bernoulli*, 18(4):1310–1319, 2012.

M.-F. Balcan, N. Bansal, A. Beygelzimer, D. Coppersmith, J. Langford, and G. Sorkin. Robust reductions from ranking to classification. In Nader Bshouty and Claudio Gentile, editors, *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 604–619. 2007.

M.-F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 671–680. ACM, 2008.

P. Billingsley. *Ergodic Theory and Information*. Wiley, New York, 1965.

D. Bosq. *Nonparametric Statistics for Stochastic Processes*. Estimation and Prediction. Springer, 1996.

Ch.-Ch. Chang and Ch.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

R. Fortet and E. Mourier. Convergence de la répartition empirique vers la répartition théoretique. *Ann. Sci. Ec. Norm. Super., III. Ser*, 70(3):267–285, 1953.

R. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, 1988.

M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35(2):402–408, 1989.

Z. Harchaoui, F. Bach, and E. Moulines. Kernel change-point analysis. In *Advances in Neural Information Processing Systems 21*, pages 609–616, 2008.

L. V. Kantorovich and G. S. Rubinstein. On a function space in certain extremal problems. *Dokl. Akad. Nauk USSR*, 115(6):1058–1061, 1957.

R.L. Karandikar and M. Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statistics and Probability Letters*, 58:297–307, 2002.

A. Khaleghi, D. Ryabko, J. Mary, and P. Preux. Online clustering of processes. In *AISTATS*, JMLR W&CP 22, pages 601–609, 2012.

A. Khaleghi and D. Ryabko. Locating changes in highly dependent data with unknown number of change points. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3095–3103. 2012.

A. Khaleghi and D. Ryabko. Nonparametric multiple change point estimation in highly dependent time series. In *Proc. 24th International Conf. on Algorithmic Learning Theory (ALT'13)*, Singapre, 2013. Springer.

D. Kifer, Sh. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proc. the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB'04, pages 180–191, 2004.

A.N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *G. Inst. Ital. Attuari*, pages 83–91, 1933.

J. Langford, R. Oliveira, and B. Zadrozny. Predicting conditional quantiles via reduction to classification. In *Proc. of the 22th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

J. del R. Millán. On the need for on-line learning in brain-computer interfaces. In *Proc. of the Int. Joint Conf. on Neural Networks*, 2004.

D.S. Ornstein and B. Weiss. How sampling reveals a process. *Annals of Probability*, 18(3):905–930, 1990.

D. Pollard. *Convergence of Stochastic Processes*. Springer, 1984.

B. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24:87–96, 1988.

B. Ryabko. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series. *IEEE Transactions on Information Theory*, 55:4309–4315, 2009.

D. Ryabko. Clustering processes. In *Proc. the 27th International Conference on Machine Learning (ICML 2010)*, pages 919–926, Haifa, Israel, 2010a.

D. Ryabko. Discrimination between B-processes is impossible. *Journal of Theoretical Probability*, 23(2):565–575, 2010b.

D. Ryabko. On the relation between realizable and non-realizable cases of the sequence prediction problem. *Journal of Machine Learning Research*, 12:2161–2180, 2011.

D. Ryabko. Testing composite hypotheses about discrete ergodic processes. *Test*, 21(2):317–329, 2012.

D. Ryabko and B. Ryabko. Nonparametric statistical inference for ergodic processes. *IEEE Transactions on Information Theory*, 56(3):1430–1435, 2010.

D. Ryabko and J. Mary. Reducing statistical time-series problems to binary classification. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2069–2077. 2012.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

P. Shields. *The Ergodic Theory of Discrete Sample Paths*. AMS Bookstore, 1996.

R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Information Theory*, IT-24:422–432, 1978.

V. M. Zolotarev. Probability metrics. *Theory of Probability and Its Applications.*, 28(2):264–287, 1983.