Unsupervised Model-Free Representation Learning

Daniil Ryabko

INRIA Lille, France
daniil@ryabko.net

Abstract. Numerous control and learning problems face the situation where sequences of high-dimensional highly dependent data are available, but no or little feedback is provided to the learner. In such situations it may be useful to find a concise representation of the input signal, that would preserve as much as possible of the relevant information. In this work we are interested in the problems where the relevant information is in the time-series dependence. Thus, the problem can be formalized as follows. Given a series of observations $X_0, \dots, X_n$ coming from a large (high-dimensional) space $\mathcal{X}$, find a representation function f mapping $\mathcal{X}$ to a finite space $\mathcal{Y}$ such that the series $f(X_0), \dots, f(X_n)$ preserve as much information as possible about the original time-series dependence in $X_0, \dots, X_n$. For stationary time series, the function f can be selected as the one maximizing the time-series information $I_\infty(f) = h_0(f(X)) - h_\infty(f(X))$ where $h_0(f(X))$ is the Shannon entropy of $f(X_0)$ and $h_\infty(f(X))$ is the entropy rate of the time series $f(X_0), \dots, f(X_n), \dots$. In this paper we study the functional $I_\infty(f)$ from the learning-theoretic point of view. Specifically, we provide some uniform approximation results, and study the behaviour of $I_\infty(f)$ in the problem of optimal control.

1 Introduction

In many learning and control problems one has to deal with the situation where the input data is high-dimensional and abundant, but the feedback for the learning algorithm is scarce or absent. In such situations, finding the right *representation* of the data can be the key to solving the problem. The focus of this work is on problems in which all or a large significant part of the relevant information is in the time-series dependence of the process. This is the case in many applications, starting with speech or hand-written text recognition, and, more generally, including control and learning problems in which the input is a stream of sensor data of an agent interacting with its environment.

A more formal exposition of the problem follows. First, assume that we are given a stationary sequence $X_0, \dots, X_n, \dots$ where X_i belong to a large (continuous, high-dimensional) space $\mathcal{X}$. For the moment, assume that the problem is non-interactive (the control part is introduced later). We are looking for a compact representation $f(X_0), \dots, f(X_n), \dots$ where $f(X_i)$ belong to a small (for example, finite) space $\mathcal{Y}$.

Let us first consider the following "*ideal*" situation. There exists a function $f : \mathcal{X} \to \mathcal{Y}$ such that each random variable X_i is *independent of the rest of the sample* $X_0, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ *given* $f(X_i)$ (for each $i, n \in \mathbb{N}$). That is, all the time-series dependence is in the sequence $f(X_0), \dots, f(X_N)$, and, given this sequence, the original sequence $X_0, \dots, X_n, \dots$ can be considered as noise, in the sense that X_i are

S. Jain et al. (Eds.): ALT 2013, LNAI 8139, pp. 342–354, 2013.

conditionally independent. In this case we say that $(X_i)_{i\in\mathbb{N}}$ *are conditionally independent given* $(f(X_i))_{i\in\mathbb{N}}$. It is shown in [15] that in this "ideal" situation the function f maximizes the following information criterion

$$I_\infty(f) := h(f(X_0)) - h_\infty(f(X)), \tag{1}$$

where $h(f(X_0))$ is the Shannon entropy of the first element and h_∞ is the entropy rate of the (stationary) time series $f(X_0), \ldots, f(X_n), \ldots$. This means that for any other function $g : \mathcal{X} \to \mathcal{Y}$ we have $I_\infty(f) \geq I_\infty(g)$, with equality if and only $(X_i)_{i\in\mathbb{N}}$ are also conditionally independent given $(g(X_i))_{i\in\mathbb{N}}$.

This allows us to pass to the *non-ideal* situation, in which there is no function f that satisfies the conditional independence criterion. Given a set of functions mapping $\mathcal{X}$ to $\mathcal{Y}$, the function that preserves *the most* of the time-series dependence can be defined as the one that maximizes (1) (this is opposed to the ideal case, in which such a function f preserves all of the time-series dependence).

In this work we show that under certain conditions it is possible to estimate (1) *uniformly* over a set $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$. Importantly, the estimation can be carried out without estimating the distribution of the original time series $(X_i)_{i\in\mathbb{N}}$.

Of particular interest (especially to control problems) is the case where the time series $(X_i)_{i\in\mathbb{N}}$ form a Markov process. In this case, in the "ideal" situation (when $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$) one can show that the process $(f(X_i))_{i\in\mathbb{N}}$ is also Markov, and $I_\infty(f) = I_1(f) := h(f(X_0)) - h(f(X_1)|f(X_0))$. In general, we show that, in the Markov case, to select a function that maximizes $I_\infty(f)$ it is enough to maximize $I_1(f)$.

Next, assume that at each time step i we are allowed to take an *action* A_i, and the next observation X_{i+1} depends not only on $X_0, \ldots, X_n$, but also on the actions $A_1, \ldots, A_n$. Thus, we are considering the *control* problem, and the time series $(X_i)_{i\in\mathbb{N}}$ do not have to be stationary any more. In this situation, the time-series information $I_\infty(f)$ becomes dependent on the policy of the learner (that is, on the way the actions are chosen). However, we can show that in the Markov case, under some mild connectivity conditions, to select the function f that maximizes $I_\infty(f)$, it is enough to consider just one policy that takes all actions with non-zero probability. This means that one can find the representation function f while executing a random policy, without any feedback from the environment (i.e., without rewards). One can then use this representation to solve the target control problem more easily.

Related Work. Learning representations, feature learning, model learning, as well as model and feature selection, are different variants and different names of the same general problem: making the data more amenable to learning. From the vast literature available on these problems we only mention a few that are somehow related to the approach in this work. First, note that in our "ideal" (conditional independence) case, if we further assume that (X_i) form a Markov chain, then we get a special case of Hidden Markov models (HMM) [11], with (unobserved) $f(X_i)$ being hidden states. Indeed, as it was mentioned, in this case the variables $f(X_i)$ form a Markov chain (Section 3.2), and the conditional independence property means that X_i depend on the rest of the sequences only through $f(X_i)$, thus $f(X_i)$ are hidden states. Note that this is a special case of HMM, since the observed variables X_i themselves are also Markovian; however, this

Markov process is over a large (continuous) state space, so it is difficult to use it directly. Thus, the general case (non-ideal situation, X_i are not necessarily Markov) is, in a certain sense, a generalization of HMMs. A related approach to finding representations in HMMs is that of [18] (see also [6]). The setting of [18] can be related to our setting in Sections 3.2, 3.3. Specifically, [18] considers environments generated by HMMs, where the hidden states are deterministic functions of the observed variables. The approach of [18] is then to maximize a penalized likelihood function, where the penalty is for larger state spaces. Consistency results are obtained for the case of finite or countably infinite sets of maps (representation functions) that are given by so-called finite-state machines of bounded memory, one of which is the true environment.

From a different perspective, if X_i are independent and identically distributed and, instead of the time-series dependence (which is absent in this case), we want to preserve as much as possible of the information about another sequence of variables (labels) $Y_1, \dots, Y_n$, then one can arrive at the information bottleneck method [20]. The information bottleneck method can, in turn, be seen as a generalization of the rate-distortion theory of Shannon [17]. Applied to dynamical systems, the information bottleneck method can be formulated [1] as follows: minimize $I(\text{past; representation}) - \beta I(\text{representation; future})$, where β is a parameter. A related idea is that of causal states [16]: two histories belong to the same causal state iff they give the same conditional distribution over futures. What distinguishes the approach of this work from those described, is that we never have to consider the probability distribution of the input time series X_i directly — only through the distribution of the representations $f(X_i)$. Thus, modelling or estimating X_i is not required; this is particularly important for empirical estimates.

For the control problem, to relate the proposed approach to others, first observe that in the case of an MDP, in the "ideal" scenario (there exists a function $f : \mathcal{X} \to \mathcal{Y}$ such that $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$) for any states $x, x' \in \mathcal{X}$ for which $f(x) = f(x')$ all the transition probabilities are the same. In other words, states $x, x' \in \mathcal{X}$ for which $f(x) = f(x')$ are equivalent in a very strong sense, and the function f can be viewed as state aggregation. Generalizations of this equivalence and aggregation (in the presence of rewards or costs) are studied in the bisimulation and homomorphism literature [3, 2, 19, 12]. The main difference of our approach (besides the absence of rewards) is in the treatment of approximate (non-ideal) cases and in the way we propose to find the representation (aggregation) functions. In bisimulation this is approached via a metric on the state space, defined using a distance between the transition (and reward) probability distributions, which then has to be estimated [2, 19]. In our approach, all that has to be estimated concerns the representations $f(X)$, rather than the observations (states) X themselves.

In the context of supervised reinforcement learning (that is, in the presence of rewards), a related problem is that of finding a (concise) representation of the input space, such that the resulting process on representations is Markovian [8, 9].

It should also be noted that the conditional independence property has been previously studied in a different context (classification) in [14]. The latter work shows that if the objects $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given the labels $(Y_i)_{i\in\mathbb{N}}$ then, effectively, one can use classification methods developed to work in the case of i.i.d.

object-label pairs. Combined with the results of this work this means that in the ideal (conditional independence) case one can *decompose* a learning problem into i.i.d. classification and learning the time-series dependence. It is also worth noting that the quantity (1) has been studied in a different context: [13] uses it to construct a statistical test for the hypothesis that a time series consists of independent and identically distributed variables. Furthermore, one can show (see below) that for stationary time series $I_\infty(f)$ equals to the following mutual information $I(X_0; X_{-1}, X_{-2}, \dots)$; this characteristic of time series has been extensively studied [4].

Organization. The rest of the paper is organized as follows. Section 2 introduces some notation and definitions. Section 3 introduces the model and gives the main results concerning representation functions for stationary time series. Section 3.2 considers the special case of (stationary) Markov chains; Section 3.1 presents results on uniform empirical approximation of time-series information. Finally, Section 3.3 extends the model and results to the control problem.

2 Preliminaries

Let $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ be measurable spaces. $\mathcal{X}$ is assumed to be large (e.g., a high-dimensional Euclidean space) and $\mathcal{Y}$ small. For simplicity of exposition, we assume that $\mathcal{Y}$ is finite; however, the results can be extended to infinite (and continuous) spaces $\mathcal{Y}$ as well.

Time-series (or process) distributions are probability measures on the space $(X^\mathbb{N}, \mathcal{F}_\mathbb{N})$ of one-way infinite sequences (where $\mathcal{F}_\mathbb{N}$ is the induced sigma-algebra of $X^\mathbb{N}$). We use the abbreviation $X_{0..k}$ for $X_0, \dots, X_k$. A distribution ρ is stationary if $\rho(X_{0..k} \in A) = \rho(X_{n+1..n+k} \in A)$ for all $A \in \mathcal{F}_{\mathcal{X}^k}$, $k, n \in \mathbb{N}$ (with $\mathcal{F}_{\mathcal{X}^k}$ being the sigma-algebra of $\mathcal{X}^k$).

A stationary distribution on $\mathcal{X}^\mathbb{N}$ can be uniquely extended to a distribution on $\mathcal{X}^\mathbb{Z}$ (that is, to a time series $\dots, X_{-1}, X_0, X_1, \dots$); we will assume such an extension whenever necessary.

For a random variable Z denote $h(Z)$ its entropy. Define $h(f)$ as the entropy of $f(X_0)$

$$h_0(f) := h(f(X_0)), \tag{2}$$

and $h_k(f)$ the k-order entropy of $f(\mathbf{X})$

$$h_k(f) := \mathbb{E}_{X_0,\dots,X_{k-1}} h(f(X_k)|f(X_0), \dots, f(X_{k-1})). \tag{3}$$

For stationary time series $(f(X_i))_{i\in\mathbb{N}}$ the entropy rate is defined as

$$h_\infty(f) := \lim_{k\to\infty} h_k(f).$$

When we speak about conditional distributions the equality of distributions should be understood in the "almost sure" sense.

2.1 Time-Series Information

The "ideal" representation function (which may or may not exist, depending on the distribution) is defined as a function f such that $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$.

Definition 1 (Conditional Independence given Representations). *$(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$, if, for all n, k, and all $i_1, \dots, i_k \neq n$, X_n is independent of $X_{i_1}, \dots, X_{i_k}$ given $f(X_n)$:*

$$P(X_n|f(X_n), X_{i_1}, \dots, X_{i_k}) = P(X_n|f(X_n)) \text{ a.s.} \tag{4}$$

Definition 2. *The time-series information of a series $f(X_0), \dots, f(X_n), \dots$ is defined as*

$$I_\infty(f) := h_0(f) - h_\infty(f). \tag{5}$$

The following theorem established in [15] shows that an "ideal" representation maximizes the time-series information.

Theorem 1 ([15]). *Let $(X_i)_{i\in\mathbb{N}}$ be stationary, and let $f : \mathcal{X} \to \mathcal{Y}$ be such that $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$. Then for any $g : \mathcal{X} \to \mathcal{Y}$ we have $I_\infty(f) \geq I_\infty(g)$, with equality if and only if $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(g(X_i))_{i\in\mathbb{N}}$.*

Define also the k-order time-series information as follows

$$I_k(f) := h_0(f) - h_k(f) = I(f(X_k); f(X_0), \dots, f(X_{k-1})).$$

The following lemma (also from [15]) helps to understand the nature of the quantities $I_\infty(f)$ and $I_k(f)$.

Lemma 1. *If the time series $(X_i)_{i\in\mathbb{Z}}$ is stationary then*

$$I_\infty(f) = I(f(X_0); f(X_{-1}), f(X_{-2}), \dots). \tag{6}$$

Proof. Denote $Y_i := f(X_i)$. We have

$$\begin{aligned} I_\infty(f) &= \lim_{k\to\infty} h(Y_0) - h(Y_0|Y_{-1}, \dots, Y_{-k}) \\ &= \lim_{k\to\infty} I(Y_0; Y_{-1}, \dots, Y_{-k}) = I(Y_0; Y_{-1}, Y_{-2}, \dots), \end{aligned}$$

where the first equality follows from the stationarity of $(X_i)_{i\in\mathbb{Z}}$ and for the last see, e.g., [4, Lemma 5.6.1] □

3 Main Results

Given a set $\mathcal{F}$ of representation functions $f : \mathcal{X} \to \mathcal{Y}$, the function that is "closest" to satisfying the conditional independence property 1 can be defined as the one that maximizes (5). If the set $\mathcal{F}$ is finite and the time series $(X_i)_{i\in\mathbb{N}}$ is stationary, then it is possible to find the function that maximizes (5) given a large enough sample of the time series, without knowing anything about its distribution [15].

The situation is more difficult if the space of representation functions is infinite (possibly uncountable); moreover, we would like to introduce learner's actions into the process, potentially making the time series $(X_i)_{i\in\mathbb{N}}$ non-stationary.

These scenarios are considered in this work.

3.1 Uniform Approximation

Given an infinite (possibly uncountable) set $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$, we want to find a function that maximizes $I_\infty(f)$. Here we first consider the problem of approximating $I_k(f)$, and then based on it proceed with the problem of approximating $I_\infty(f)$.

Since we do not know $I_k(f)$, we can select a function that maximizes the empirical estimate $\hat{I}_k(f)$. The question arises, under what conditions is this procedure consistent? The requirements we impose to obtain consistency of this procedure are of the following two types: first, the set $\mathcal{F}$ should be sufficiently small, and, second, the time series $(X_i)_{i\in\mathbb{N}}$ should be such that uniform (over $\mathcal{F}$) convergence guarantees can be established. Here the first condition is formalized in terms of VC dimension, and the second in terms of mixing times. We show that, under these conditions, the empirical estimator is indeed consistent and learning-theory-style finite-sample performance guarantees can be established.

For a function $f : \mathcal{X} \to \mathcal{Y}$ and a sample $X_1, \dots, X_n$ define the following estimators: $\hat{p}_f(y) := \frac{1}{n}\sum_{k=1}^n \mathbb{I}(f(x) = y)$, and analogously for $\hat{p}_f(y_1, \dots, y_k)$ and multivariate entropies.

Definition 3 (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 absolutely regular; this condition is much stronger than ergodicity, but is much weaker than the i.i.d. assumption.

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 [7]. Let $\mathcal{F}$ be a set of VC dimension d and let ρ be a stationary distribution over $\mathcal{X}^\infty$. Then

$$q_n(\rho, \mathcal{F}, \varepsilon) := \rho\Big(\sup_{g\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_\rho g(X_1)| > \varepsilon\Big) \\ \le n\beta(\rho, t_n) + 8t_n^{d+1}e^{-l_n\varepsilon^2/8}, \quad (7)$$

where t_n is integer 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.

Furthermore, 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 (7) becomes

$$q_n(\rho, \mathcal{F}, \varepsilon) \le n\gamma^{\sqrt{n}} + 8n^{(d+1)/2}e^{-\sqrt{n}\varepsilon^2/8} =: \Delta(d, \varepsilon, n, \gamma). \quad (8)$$

Geometric β-mixing properties can be demonstrated for large classes of (k-order) (PO)MDPs [5], and for many other distributions.

Theorem 2. *Let the time series $(X_i)_{i\in\mathbb{N}}$ be generated by a stationary distribution ρ whose β-mixing coefficients satisfy $\beta(\rho,m)\le\gamma^m$ for some $\gamma<1$. Let $\mathcal{F}$ be a set of functions $f:\mathcal{X}\to\mathcal{Y}$ such that for each $y\in\mathcal{Y}$ the VC dimension of the set $\{\mathbb{I}_{\{x\in\mathcal{X}:g(x)=y\}}:g\in\mathcal{F}\}$ is not greater than d. Furthermore, assume that, for some $k\in\mathbb{N}$, there exist an $\alpha>0$ such that for any $g\in\mathcal{F}$ and any $y_1,\dots,y_k\in\mathcal{Y}$ we have*

$$P(g(X_0)=y_1,\dots,g(X_k)=y_k))\ge\alpha^k. \tag{9}$$

Then

$$P(\sup_{g\in\mathcal{F}}|\hat{I}_k(g)-I_k(g)|>\varepsilon)\le 4|\mathcal{Y}|^{k+1}\Delta(7kd,-k\varepsilon/4|\mathcal{Y}|^{k+1}\log\alpha,n-k,\gamma) \tag{10}$$

for every $\varepsilon<\alpha$.

Proof. First, note that from stationarity of $(X_i)_{i\in\mathbb{N}}$ we get

$$I_k(g)=h(g(X_0))-h(g(X_0),\dots,g(X_k))+h(g(X_1),\dots,g(X_k))$$

so that

$$\begin{aligned}\sup_{g\in\mathcal{F}}|\hat{I}_k(g)-I_k(g)|&\le\sup_{g\in\mathcal{F}}|\hat{h}(g(X_0))-h(g(X_0))|\\&\quad+\sup_{g\in\mathcal{F}}|\hat{h}(g(X_0),\dots,g(X_k))-h(g(X_0),\dots,g(X_k))\\&\quad+\sup_{g\in\mathcal{F}}|\hat{h}(g(X_1),\dots,g(X_k))-h(g(X_1),\dots,g(X_k))|=:T_1+T_2+T_3.\end{aligned}$$

Introduce the shorthand notation $p_g(y):=P(g(X_0)=y)$. From the conditions of the theorem we know that $p_g(y)\ge\alpha$ for any g,y; but we also will need the same to hold for the estimates $\hat{p}$. So, consider the following event

$$B:=\{\inf_{g\in\mathcal{F}}\inf_{y\in\mathcal{Y}}\hat{p}(g(X_0)=y)\le\alpha/2\},$$

and the following simple decomposition

$$P(T_1>\varepsilon)\le P(B)+P(T_1>\varepsilon|\neg B). \tag{11}$$

From (9) and the bound (8) we obtain

$$P(B)\le|\mathcal{Y}|\Delta(d,\alpha/2,n,\gamma). \tag{12}$$

The Taylor expansion of a function u differentiable around t can be expressed as $u(t) = u(c) + (t-c)u'(\theta c + (1-\theta)t)$ for some $\theta \in (0,1)$. Using this for the function $u(p) = p \log p$ we obtain

$$\begin{aligned}
|\hat{h}_0(g) - h_0(g)| &= |\sum_{y\in\mathcal{Y}} (p_g(y)\log p_g(y) - \hat{p}_g(y)\log\hat{p}_g(y)| \\
&= |\sum_{y\in\mathcal{Y}} (p_g(y) - \hat{p}_g(y))(1 + \log(\theta p_g(y) + (1-\theta)\hat{p}_g(y))| \\
&\leq -\log\alpha \sum_{y\in\mathcal{Y}} |p_g(y) - \hat{p}_g(y)| \\
&= -\log\alpha \sum_{y\in\mathcal{Y}} |\mathbb{E}\mathbb{I}_{g(X_0)=y} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{g(X_0)=y}|,
\end{aligned}$$

where the last inequality uses (9) and holds under the assumption $\hat{p}_g(y) > \alpha/2$ for all $y \in \mathcal{Y}$ (that is, conditional on $\neg B$). From this, (12), (11) and (8) we obtain

$$\begin{aligned}
P(T_1 > \varepsilon) &\leq |\mathcal{Y}|(\Delta(d, \alpha/2, n, \gamma) + \Delta(d, -\varepsilon/|\mathcal{Y}|\log\alpha, n, \gamma)) \\
&\leq 2|\mathcal{Y}|\Delta(d, -\varepsilon/|\mathcal{Y}|\log\alpha, n, \gamma),
\end{aligned}$$

where we have used $\varepsilon < \alpha$.

It remains to repeat the same analysis for T_2 and T_3. Clearly, it is enough to consider only T_2 (T_3 is analogous). The difference with T_1 is that instead of the entropy of one variable $h(g(X_0))$ we are have to deal with the entropies of $(k+1)$-tuples $h(g(X_0), \ldots, g(X_k))$. First, observe that, from the definition of mixing, if a process ρ generating $X_0, X_1, X_2, \ldots$ is mixing with coefficients $\beta(\rho, m)$ then the process made of tuples $(X_0, \ldots, X_k), (X_1, \ldots, X_{k+1}), \ldots$ is mixing with coefficients $\beta(\rho, m-k)$. Next, for the VC dimensions, observe that if a set

$$\{\{x : g(x) = y\} : g \in \mathcal{F}\}$$

has VC dimension d (for every $y \in \mathcal{Y}$) then the set

$$\{\{(x_1, \ldots, x_k) : g_i(x_i) = y_i, i = 1..k+1\} : (g_1, \ldots, g_k) \in \mathcal{F}^k\}$$

has VC dimension bounded by $7kd$ (for all $(y_1, \ldots, y_k) \in \mathcal{Y}^k$); see [21], which also gives a more precise bound. Now we can repeat the derivation for T_2, and obtain the resulting bound (10). □

The condition (9) requires that there is sufficient noise in the time series $g(X_i)$ for every $g \in \mathcal{F}$. While this condition is rather mild, we think that it is an artefact of the analysis and can probably be avoided.

We proceed to construct an estimator of $I_\infty(g)$ which is uniformly consistent over a set $\mathcal{F}$ of functions g, provided the time series satisfies mixing conditions. To this end, denote $\delta_k(n)$ the right-hand side of (10). Observe that for each fixed $k \in \mathbb{N}$, $\delta_k(n)$ decreases exponentially fast with n. Therefore, it is possible to find a non-decreasing

sequence $k_n : n \in \mathbb{N}$ such that $\delta_{k_n}(n)$ decreases exponentially fast with n, while $k_n \to \infty$. Define

$$\hat{I}_\infty(g) := \hat{I}_{k_n}(g). \tag{13}$$

Furthermore, observe that, for any stationary time series we have, by definition, $h_\infty(g) = \lim_{k\to\infty} h_k(g)$. For uniform approximation of I_∞ we need this convergence to hold uniformly over the set $\mathcal{F}$. This is akin to the mixing conditions, but, in general, does not follow from them. Thus, we strengthen the mixing conditions by requiring that the following holds

$$\lim_{k\to\infty} \sup_{g\in\mathcal{F}} |h_\infty(g) - h_k(g)| = 0. \tag{14}$$

The following statement is easy to show from Theorem 2, the definition (13) of $\hat{I}_\infty$ and (14).

Theorem 3. *Under the conditions of Theorem 2, if (14) holds true then*

$$\sup_{g\in\mathcal{F}} |\hat{I}_\infty(g) - I_\infty(g)| \to 0 \text{ a.s.}$$

3.2 Time-Series Information for Markov Chains

For the control problem, a special role is played by Markov environments; we first look at the simplifications gained by making this assumption in the stationary case.

If the $(X_i)_{i\in\mathbb{N}}$ form a stationary (k-order) Markov process then the situation simplifies considerably. First, if $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$ then $(f(X_i))_{i\in\mathbb{N}}$ also form a stationary (k-order) Markov chain. Moreover, to find the function that maximizes the time-series information (1) it is enough to find the function that maximizes a simpler quantity $I_k(f) = I(f(X_0); f(X_1), \ldots, f(X_k))$, as the following theorem shows. In the theorem and henceforth, for the sake of simplicity of notation, we only consider the case $k = 1$; the general case is analogous.

Theorem 4. *Suppose that X_i form a stationary Markov process and $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$. Then*

(i) $(f(X_i))_{i\in\mathbb{N}}$ also form a stationary Markov chain;
(ii) $I_\infty(f)$ is the mutual information between $f(X_0)$ and $f(X_1)$:

$$I_\infty(f) = I_1(f) = I(f(X_0), f(X_1)), \tag{15}$$

(iii) for any $g : \mathcal{X} \to \mathcal{Y}$ we have $I_1(f) \geq I_1(g)$ with equality if and only if $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(g(X_i))_{i\in\mathbb{N}}$.

Proof. We use the notation $Y_i := f(X_i)$. First note that from the definition (1) of conditional independence and using the chain rule for entropies, it is easy to show that for any $n, k, i_1, \ldots, i_k \in \mathbb{N}$ we have

$$h(Y_n|Y_{i_1}, X_{i_1}, \ldots, Y_{i_k}, X_{i_k}) = h(Y_n|Y_{i_1}, \ldots, Y_{i_k}). \tag{16}$$

For the first statement of the theorem, observe that

$$h(Y_{n+1}|Y_1 \ldots, Y_n) = h(Y_{n+1}|Y_1, X_1, \ldots, Y_n, X_n) \\ = h(Y_{n+1}|Y_n, X_n) = h(Y_{n+1}|Y_n), \quad (17)$$

where we have used successively (16), the Markov property for $(X_i)_{i\in\mathbb{N}}$ and again (16).

For the second statement, first note that $h_\infty = h_1$ for Markov chains, implying (15). Next, for any $g : \mathcal{X} \to \mathcal{Y}$ the process $g(X_i)$ is stationary, which implies $h_\infty(g(X)) \leq h_1(g(X))$. Thus, using Theorem 1, we obtain

$$I_1(f) = I_\infty(f) \geq I_\infty(g) \geq h_0(g) - h_1(g) = I_1(g).$$

□

3.3 The Active Case: MDPs

In this section we introduce learner's actions into the protocol. The setting is a sequential interaction between the learner and the environment. Given are a space of observations $\mathcal{X}$ and of a space actions $\mathcal{A}$, where $\mathcal{A}$ is assumed finite. At each time step $i \in \mathbb{N}$ the environment provides an observation X_i, the learner takes an action A_i, then the next observation X_{i+1} is provided, and so on. Each next observation X_{i+1} is generated according to some (unknown) probability distribution $P(X_{i+1}|X_0, A_0, \ldots, X_i, A_i)$. Actions are generated by a probability distribution π that is called a *policy*; in general, it has the form $\pi(A_{i+1}|X_0, A_0, \ldots, X_i, A_i, X_{i+1})$.

Note that we do not introduce costs or rewards into consideration. Thus, we are dealing with an unsupervised version of the problem; the goal is just to find a concise representation that preserves the dynamics of the problem.

Definition 4 (Conditional Independence, Active Case). *For a policy π, an environment P and a measurable function f we say that $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$* under the policy π *if*

$$P^\pi(X_n|f(X_n), A_n, X_{i_1}, A_{i_1}, \ldots, X_{i_k}, A_{i_k}) = P^\pi(X_n|f(X_n)) \text{ a.s.} \quad (18)$$

for all $n, k \in \mathbb{N}$, and all $i_1, \ldots, i_k \in \mathbb{N}$ such that $i_j \neq n$, $j = 1..k$, where P^π refers to the joint distribution of X_i and A_i generated according to P and π.

The focus in this section is on time-homogeneous Markov environments, that is, on Markov Decision Processes (MDPs). Thus, we assume that X_{i+1} only depends on X_i and A_i, so that P can be identified with a function from $\mathcal{X} \times \mathcal{A}$ to the space $\mathcal{P}(\mathcal{X})$ of probability distributions on $\mathcal{X}$

$$P(X_{i+1}|X_0, A_0, \ldots, X_{i-1}, A_{i-1}, X_i = x, A_i = a) = P(X_{i+1}|x, a).$$

In this case observations X_i are called *states*.

A policy is called *stationary* if each action only depends on the current state; that is, $\pi(A_{i+1}|X_0, A_0, \ldots, X_i, A_i, X_{i+1} = x) = \pi(A_{i+1}|x)$ where, for each $x \in \mathcal{X}$, $\pi(A|x)$ is a distribution over $\mathcal{A}$.

Call an MDP *admissible* if any stationary policy π has a (unique up to sets of measure 0) stationary distribution P^π over states. The notation $\mathbb{E}^\pi, P^\pi, h^\pi, I_k^\pi$, etc. refers to the stationary distribution of the policy π.

For MDPs we introduce the following policy-independent definition of conditional independence.

Definition 5 (Conditional Independence, MDPs). *For an admissible MDP and a measurable function $f : \mathcal{X} \to \mathcal{Y}$ we say that $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$ if, for every stationary policy π, $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $f(X_i)$ under policy π.*

Call a stationary policy π *stochastic* if $\pi(a|x) \geq \alpha > 0$ for every $x \in \mathcal{X}$ and every $a \in \mathcal{A}$.

Call an admissible MDP *(weakly) connected* if for every *stochastic* policy π and every stationary policy π' we have $P^\pi \gg P^{\pi'}$ (that is, for any measurable $S \subset \mathcal{X} \times \mathcal{A}$ $P^{\pi'}(S) > 0$ implies $P^\pi(S) > 0$). It is easy to see that in this definition one can replace "for every *stochastic* policy π" with "there exists a stationary policy π." Note the difference with a much stronger property that is sometimes called ergodic or recurrent MDP [10]; the latter property would be obtained if we remove the word *"stochastic"* in the definition (allowing, in particular, all deterministic policies π).

It is easy to see that for discrete MDPs this definition coincides with the usual definition of weak connectedness (for any pair of states s_1, s_2 there is a policy that gets from s_1 to s_2 in a finite number of steps with non-zero probability).

Theorem 5. *Fix an admissible weakly connected MDP and a stationary stochastic policy π. Then $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$ if and only if $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$ under π.*

Proof. We only have to prove the "if" part (the other part is obvious). Let π_0 be any stationary policy. Introduce the notation $Y_i := f(X_i)$ and

$$U_0 := (X_{-1}, A_{-1}, A_0, X_1, A_1).$$

We have to establish (18) for P^{π_0}; note that since the process is Markov we can take $k = 2$, $i_1 = 1$, $i_2 = -1$ in (18) w.l.o.g.; thus, we need to demonstrate

$$P^{\pi_0}(X_0|Y_0) = P^{\pi_0}(X_0|Y_0, U_0) \text{ a.s.} \tag{19}$$

Since the policy π is stochastic, the measure P^π dominates P^{π_0}. Therefore, the following probability-one statements are non-vacuous:

$$P^\pi(X_0|Y_0) = P^\pi(X_0|Y_0, U_0) = P^{\pi_0}(X_0|Y_0, U_0) \text{ a.s.}$$

for all $i \in \mathbb{N}$, where the first equality follows from (18), and the second follows from the fact that conditionally on the actions the distributions P^π and P^{π_0} coincide. Moreover,

$$P^{\pi_0}(X_0|Y_0) = \mathbb{E}^{\pi_0}_{U_0} P^{\pi_0}(X_0|Y_0, U_0) = \mathbb{E}^{\pi_0}_{U_0} P^\pi(X_0|Y_0) = P^\pi(X_0|Y_0) \text{ a.s.}$$

Thus, $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$ under π_0; since π_0 was chosen arbitrary, this concludes the proof.

Corollary 1. *Fix an admissible MDP and a stationary stochastic policy π. Assume that for some $f : \mathcal{X} \to \mathcal{Y}$ $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f(X_i))_{i\in\mathbb{N}}$. If $f' = \operatorname{argmax}_g I_1^\pi(g)$ then $(X_i)_{i\in\mathbb{N}}$ are conditionally independent given $(f'(X_i))_{i\in\mathbb{N}}$.*

Proof. The statement follows from Theorems 1 and 5. □

Consider the following scenario. A real-life control problem is given, in which an (average, discounted) cost has to be optimized. In addition, a simulator for this problem is available; running the simulator does not incur any costs, but also does not provide any information about the costs — it only simulates the dynamics of the problem. Given such a simulator, and a set $\mathcal{F}$ of representation functions, one can first execute a random policy to find the best representation function f as the one that maximizes $\hat{I}_1(f)$. Under the conditions given in Section 3.1, the resulting estimator is consistent. One can then use the representation function found to learn the optimal policy in the real problem (with costs).

The problem of solving (efficiently) both problems together— learning the representation and the finding the optimal policy in a control problem— is left for future work.

4 Outlook

This work together with [15] lays some theoretical foundations for building representations functions for time series in an unsupervised, model-free way. Specifically, the results on uniform approximation demonstrate that it is statistically possible to find good approximations to the best representation functions in a large (continuous) sets $\mathcal{F}$ of such functions. This can be done by selecting the function that maximizes empirical time-series information, or its k-order version I_k. The next important step is to develop efficient algorithms for finding such functions for specific sets $\mathcal{F}$. Another interesting question is what results can be obtained if we do not require uniform convergence. In particular, whether it is possible to find, perhaps in some weak asymptotic sense, a function that maximizes I_∞ over the set of all (measurable) functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Our conjecture is that this is possible for stationary ergodic time series.

For the control problem, we have shown that a consistent approach to find representations is just to take random actions and select the best representation for the resulting time series. The resulting representation can then be used to learn the optimal policy for an actual control problem. This may be a reasonable approach if the representation can be found in a simulated scenario. Yet, it is clear that this is not the most efficient way. A natural question is how to find a policy that allows one to find the best representation using as little time (or samples) as possible.

Acknowledgments. This work was supported by FP7/2007-2013 under grant agreements 270327 (CompLACS) and 216886 (PASCAL2), by the French National Research Agency (project Lampada ANR-09-EMER-007), the Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007-2013.

References

[1] Creutzig, F., Globerson, A., Tishby, N.: Past-future information bottleneck in dynamical systems. Phys. Rev. E 79, 041925 (2009)
[2] Ferns, N., Castro, P.S., Precup, D., Panangaden, P.: Methods for computing state similarity in markov decision processes. In: Proceedings of UAI (2006)
[3] Givan, R., Dean, T., Greig, M.: Equivalence notions and model minimization in markov decision processes. Artificial Intelligence 147(1), 163–223 (2003)
[4] Gray, R.: Entropy and information theory. Springer (1990)
[5] Hernández-Lerma, O., Lasserre, J.B.: Markov chains and invariant probabilities. Birkhäuser (2003)
[6] Hutter, M.: Feature reinforcement learning: Part I. Unstructured MDPs. Journal of General Artificial Intelligence 1, 3–24 (2009)
[7] Karandikar, R.L., Vidyasagar, M.: Rates of uniform convergence of empirical means with mixing processes. Statistics and Probability Letters 58, 297–307 (2002)
[8] Maillard, O., Munos, R., Ryabko, D.: Selecting the state-representation in reinforcement learning. In: NIPS, Granada, Spain, pp. 2627–2635 (2011)
[9] Maillard, O., Nguyen, P., Ortner, R., Ryabko, D.: Optimal regret bounds for selecting the state representation in reinforcement learning. In: ICML, Atlanta, USA. JMLR W&CP, vol. 28(1), pp. 543–551 (2013)
[10] Puterman, M.L.: Markov decision processes: discrete stochastic dynamic programming, vol. 414. Wiley-Interscience (2009)
[11] Rabiner, L., Juang, B.: An introduction to hidden Markov models. IEEE ASSP Magazine 3(1), 4–16 (1986)
[12] Ravindran, B., Barto, A.G.: Relativized options: Choosing the right transformation. In: Machine Learning, Proceedings of the Twentieth International Conference, ICML 2003, Washington, DC, USA, August 21-24, vol. 2, pp. 608–615 (2003)
[13] Ryabko, B., Astola, J.: Universal codes as a basis for time series testing. Statistical Methodology 3, 375–397 (2006)
[14] Ryabko, D.: Pattern recognition for conditionally independent data. Journal of Machine Learning Research 7, 645–664 (2006)
[15] Ryabko, D.: Time-series information and learning. In: Proc. 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey. IEEE (2013)
[16] Shalizi, C.R., Crutchfield, J.P.: Computational mechanics: Pattern and prediction, structure and simplicity. Journal of Statistical Physics 104(3-4), 817–879 (2001)
[17] Shannon, C.E.: Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec. 4, 142–163 (1959)
[18] Sunehag, P., Hutter, M.: Consistency of feature markov processes. In: Hutter, M., Stephan, F., Vovk, V., Zeugmann, T. (eds.) ALT 2010, LNCS, vol. 6331, pp. 360–374. Springer, Heidelberg (2010)
[19] Taylor, J., Precup, D., Panangaden, P.: Bounding performance loss in approximate mdp homomorphisms. In: Advances in Neural Information Processing Systems, vol. 21, pp. 1649–1656 (2009)
[20] Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, pp. 368–377 (1999)
[21] van der Vaart, A., Wellner, J.A.: A note on bounds for VC dimensions. Institute of Mathematical Statistics Collections 5, 103 (2009)