LEARNABILITY IN PROBLEMS OF SEQUENTIAL INFERENCE

Daniil Ryabko

Habilitation à diriger des recherches
Université Lille 1 - Sciences et Technologies

préparée au sein du centre de recherche INRIA Lille - Nord Europe

Specialité : Mathématiques

Daniil RYABKO

APPRENABILITÉ DANS LES PROBLÈMES DE L'INFÉRENCE SÉQUENTIELLE

LEARNABILITY IN PROBLEMS OF SEQUENTIAL INFERENCE

Habilitation soutenue à Villneuve d'Ascq le 19 décembre 2011 devant le jury composé de

Pr. Peter AUER	University of Leoben	Rapporteur
Pr. Jan BEIRLANT	K.U.Leuven	Examinateur
Pr. Léon BOTTOU	Microsoft	Rapporteur
Pr. Max DAUCHET	Université Lille 1	Examinateur
Pr. László GYÖRFI	Budapest University	Rapporteur
Pr. Rémi MUNOS	INRIA Lille - Nord Europe	Examinateur
Pr. Philippe PREUX	Université Lille 3	Examinateur

Résumé

Les travaux présentés sont dédiés à la possibilité de faire de l'inférence statistique à partir de données séquentielles. Le problème est le suivant. Étant donnée une suite d'observations x_1,...,x_n,... , on cherche à faire de l'inférence sur le processus aléatoire ayant produit la suite. Plusieurs problèmes, qui d'ailleurs ont des applications multiples dans différents domaines des mathématiques et de l'informatique, peuvent être formulés ainsi. Par exemple, on peut vouloir prédire la probabilité d'apparition de l'observation suivante, x_{n+1} (le problème de prédiction séquentielle); ou répondre à la question de savoir si le processus aléatoire qui produit la suite appartient à un certain ensemble H_0 versus appartient à un ensemble différent H_1 (test d'hypothèse) ; ou encore, effectuer une action avec le but de maximiser une certain fonction d'utilité. Dans chacun de ces problèmes, pour rendre l'inférence possible il faut d'abord faire certaines hypothèses sur le processus aléatoire qui produit les données. La question centrale adressée dans les travaux présentés est la suivante : *sous quelles hypothèses l'inférence est-elle possible ?* Cette question est posée et analysée pour des problèmes d'inférence différents, parmi lesquels se trouvent la prédiction séquentielle, les tests d'hypothèse, la classification et l'apprentissage par renforcement.

Abstract

Given a growing sequence of observations x_1,...,x_n,... , one is required, at each time step n, to make some inference about the stochastic mechanism generating the sequence. Several problems that have numerous applications in different branches of mathematics and computer science can be formulated in this way. For example, one may want to forecast probabilities of the next outcome x_{n+1} (sequence prediction); to make a decision on whether the mechanism generating the sequence belongs to a certain family H_0 versus it belongs to a different family H_1 (hypothesis testing); to take an action in order to maximize some utility function.

In each of these problems, as well as in many others, in order to be able to make inference, one has to make some assumptions on the probabilistic mechanism generating the data. Typical assumptions are that x_i are independent and identically distributed, or that the distribution generating the sequence belongs to a certain parametric family. The central question addressed in this work is: *under which assumptions is inference possible?* This question is considered for several problems of inference, including sequence prediction, hypothesis testing, classification and reinforcement learning.

Contents

Publications on which this manuscript is based

Journal papers

[R1] 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.

[R2] D. Ryabko. Testing composite hypotheses about discrete ergodic processes. *Test*, (in press), 2011.

[R3] D. Ryabko. On finding predictors for arbitrary families of processes. *Journal of Machine Learning Research*, 11:581–602, 2010.

[R4] D. Ryabko and B. Ryabko. Nonparametric statistical inference for ergodic processes. *IEEE Transactions on Information Theory*, 56(3):1430–1435, 2010.

[R5] D. Ryabko. Discrimination between B-processes is impossible. *Journal of Theoretical Probability*, 23(2):565–575, 2010.

[R6] D. Ryabko and J. Schmidhuber. Using data compressors to construct order tests for homogeneity and component independence. *Applied Mathematics Letters*, 22(7):1029–1032, 2009.

[R7] D. Ryabko and M. Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, 405(3):274–284, 2008.

[R8] D. Ryabko and M. Hutter. Predicting non-stationary processes. *Applied Mathematics Letters*, 21(5):477–482, 2008.

[R9] D. Ryabko. On sample complexity for computational classification problems. *Algorithmica*, 49(1):69–77, 2007.

[R10] D. Ryabko. Pattern recognition for conditionally independent data. *Journal of Machine Learning Research*, 7:645–664, 2006.

Selected conference papers

[R11] D. Ryabko. Clustering processes. In *Proc. the 27th International Conference on Machine Learning (ICML 2010)*, pages 919–926, Haifa, Israel, 2010.

[R12] D. Ryabko. Sequence prediction in realizable and non-realizable cases. In *Proc. the 23rd Conference on Learning Theory (COLT 2010)*, pages 119–131, Haifa, Israel, 2010.

[R13] D. Ryabko. Testing composite hypotheses about discrete-valued stationary processes. In *Proc. IEEE Information Theory Workshop (ITW'10)*, pages 291–295, Cairo, Egypt, 2010. IEEE.

[R14] D. Ryabko. An impossibility result for process discrimination. In *Proc. 2009 IEEE International Symposium on Information Theory (ISIT)*, pages 1734–1738, Seoul, South Korea, 2009. IEEE.

[R15] D. Ryabko. Characterizing predictable classes of processes. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*, Montreal, Canada, 2009.

[R16] D. Ryabko. Some sufficient conditions on an arbitrary class of stochastic processes for the existence of a predictor. In *Proc. 19th International Conf. on Algorithmic Learning Theory (ALT'08)*, LNAI 5254, pages 169–182, Budapest, Hungary, 2008. Springer.

[R17] D. Ryabko and B. Ryabko. On hypotheses testing for ergodic processes. In *Proc. 2008 IEEE Information Theory Workshop (ITW)*, pages 281–283, Porto, Portugal, 2008. IEEE.

[R18] D. Ryabko and M. Hutter. On sequence prediction for arbitrary measures. In *Proc. 2007 IEEE International Symposium on Information Theory (ISIT)*, pages 2346–2350, Nice, France, 2007. IEEE.

[R19] D. Ryabko and M. Hutter. Asymptotic learnability of reinforcement problems with arbitrary dependence. In *Proc. 17th International Conf. on Algorithmic Learning Theory (ALT)*, LNAI 4264, pages 334–347, Barcelona, Spain, 2006. Springer.

[R20] D. Ryabko. On computability of pattern recognition problems. In *Proc. 16th International Conf. on Algorithmic Learning Theory (ALT)*, LNAI 3734, pages 148–156, Singapore, 2005. Springer.

[R21] D. Ryabko. Application of classical nonparametric predictors to learning conditionally i.i.d. data. In *Proc. 15th International Conf. on Algorithmic Learning Theory (ALT)*, LNAI 3244, pages 171–180, Padova, Italy, 2004. Springer.

[R22] D. Ryabko. Online learning of conditionally i.i.d. data. In *21st International Conference on Machine Learning (ICML)*, pages 727–734, Banff, Canada, 2004.

Acknowledgements

Thanks for reading.

Chapter 1

Introduction

This manuscript summarizes my work on the problem of learnability in making sequential inference. The problem is as follows. Given a growing sequence of observations x_1,...,x_n,... , one is required, at each time step n, to make some inference about the stochastic mechanism generating the sequence. Several problems that have numerous applications in different branches of mathematics and computer science can be formulated in this way. For example, one may want to forecast probabilities of the next outcome x_{n+1} (sequence prediction); to make a decision on whether the mechanism generating the sequence belongs to a certain family H_0 versus it belongs to a different family H_1 (hypothesis testing); to take an action in order to maximize some utility function. In each of these problems, as well as in many others, in order to be able to make inference, one has to make some assumptions on the probabilistic mechanism generating the data. Typical assumptions are that x_i are independent and identically distributed, or that the distribution generating the sequence belongs to a certain parametric family. The central question addressed in this work is: *under which assumptions is inference possible?* This question is considered for several problems of inference, including sequence prediction, hypothesis testing, classification and reinforcement learning.

The motivation for studying such questions is as follows. There are

numerous applications of the problems of sequential inference considered, including the analysis of data coming from a stock market, biological databases, network traffic, surveillance, etc. Moreover, new applications spring up regularly. Clearly, each application requires its own set of assumptions or its own model: indeed, a model that describes well DNA sequences probably should not be expected to describe a network traffic stream as well. At the same time, the methods developed for sequential data analysis, for example, for sequence prediction, are typically limited to a restricted set of models, the construction of which is driven not so much by applications as by the availability of mathematical and algorithmic tools. For example, much of the non-parametric analysis in sequence prediction and the problems of finding an optimal behaviour in an unknown environment is limited to finite-state models, such as Markov or hidden Markov models. One approach to start addressing this problem is to develop a theory that would allow one to check whether a given model is feasible, in the sense that it allows for the existence of a solution to the target problem of inference. Another question is how to find a solution given a model in a general form. This manuscript summarizes some first steps I have made in solving these general problems.

The most important results presented here are as follows. For the problem of *hypothesis testing*, I have obtained a topological characterization (necessary and sufficient conditions) of those (composite) hypotheses $H_0 \subset \mathcal{E}$ that can be consistently tested against $\mathcal{E} \backslash H_0$, where $\mathcal{E}$ is the set of all stationary ergodic discrete-valued process measures. The developed approach, which is based on empirical estimates of the distributional distance, has allowed me to obtain *consistent procedures for change point estimation, process classification and clustering*, under the only assumption that the data (real-valued, in this case) is generated by stationary ergodic distributions: a setting that is much more general than those in which consistent procedures were known before. I have also demonstrated that a consistent test for homogeneity does not exist for the general case of stationary ergodic (discrete-valued) sequences. For the problem of *sequence prediction*, I have shown that if there

is a consistent predictor for a set of process distributions $\mathcal{C}$, then there is a Bayesian predictor consistent for this set. This is a no-assumption result: the distributions in $\mathcal{C}$ can be arbitrary (non-i.i.d., non-stationary, etc.) and the set itself does not even have to be measurable. I have also obtained several descriptions (sufficient conditions) of those sets $\mathcal{C}$ of process distributions for which consistent predictors exist. For the problem of *selecting an optimal strategy* in a reactive environment (perhaps, the most general inference problem considered) I have identified some sufficient conditions on the environments under which it is possible to find a universal asymptotically optimal strategy.

Organization of this manuscript

Each of the subsequent chapters is devoted to a specific group of inference problems. These chapters are essentially constructed from the papers listed below, which are streamlined and endowed with unified introductions and notation. References to the corresponding papers follow the title of of each chapter or section. While trying to keep the contents concise, I have decided to include all proofs (in appendices to the chapters) to make the manuscript self-contained. The contents of the Section 5.1 is extracted from my Ph.D. thesis; the rest of the work is posterior.

Chapter 2

Sequence prediction [R1, R3, R8]

The problem is sequence prediction in the following setting. A sequence $x_1,...,x_n,...$ of discrete-valued observations is generated according to some unknown probabilistic law (measure) μ. After observing each outcome, it is required to give the conditional probabilities of the next observation. The measure μ is unknown. We are interested in constructing predictors ρ whose conditional probabilities $\rho(\cdot|x_1,...,x_n)$ converge (in some sense) to the "true" μ-conditional probabilities $\mu(\cdot|x_1,...,x_n)$, as the sequence of observations increases ($n \to \infty$). In general, this goal is impossible to achieve if nothing is known about the measure μ generating the sequence. In other words, one cannot have a predictor whose error goes to zero for any measure μ. The problem becomes tractable if we assume that the measure μ generating the data belongs to some known class $\mathcal{C}$. The main general problems considered in this chapter are as follows.

(i) Given a set $\mathcal{C}$ of process measures, what are the conditions on $\mathcal{C}$ under which there exists a measure ρ that predicts every $\mu \in \mathcal{C}$? (Section 2.3)

(ii) Is there a general way to construct such a measure ρ? What form should ρ have? (Section 2.2)

(iii) Given two process measures μ and ρ, what are the conditions on them

under which ρ is a predictor for μ? (Section 2.4)

(iv) Given a set $\mathcal{C}$ of process measures, what are the conditions on $\mathcal{C}$ under which there exists a measure ρ which predicts every measure ν that is predicted by at least one measure $\mu \in \mathcal{C}$? (Section 2.5)

The last question is explained by the following consideration. Since a predictor ρ that we wish to construct is required, on each time step, to give conditional probabilities $\rho(\cdot|x_1,...,x_n)$ of the next outcome given the past, for each possible sequence of past observations $x_1,...,x_n$, the predictor ρ itself defines a measure on the space of one-way infinite sequences. This enables us to pose similar questions about the measures μ generating the data and predictors.

The motivation for studying predictors for arbitrary classes $\mathcal{C}$ of processes is two-fold. First of all, the problem of prediction has numerous applications in such diverse fields as data compression, market analysis, bioinformatics, and many others. It seems clear that prediction methods constructed for one application cannot be expected to be optimal when applied to another. Therefore, an important question is how to develop specific prediction algorithms for each of the domains. Apart from this, sequence prediction is one of the basic ingredients for constructing intelligent systems. Indeed, in order to be able to find optimal behaviour in an unknown environment, an intelligent agent must be able, at the very least, to predict how the environment is going to behave (or, to be more precise, how relevant parts of the environment are going to behave). Since the response of the environment may in general depend on the actions of the agent, this response is necessarily non-stationary for explorative agents. Therefore, one cannot readily use prediction methods developed for stationary environments, but rather has to find predictors for the classes of processes that can appear as a possible response of the environment.

To evaluate the quality of prediction we will mostly use expected (with respect to data) average (over time) Kullback-Leibler divergence, as well as

total variation distance (see Section 2.1 for definitions). Prediction in total variation is a very strong notion of performance; in particular, it is not even possible to predict an arbitrary i.i.d. Bernoulli distribution in this sense. Prediction in expected average KL divergence is much weaker and therefore more practical, but it is also more difficult to study, as is explained below.

Next we briefly describe **some of the answers** to the questions (i)-(iv) posed above that we have obtained. It is well-known [11, 46] (see also Theorem 2.2 below) that a measure ρ predicts a measure μ in total variation if and only if μ is absolutely continuous with respect to ρ. This answers question (ii) for prediction in total variation. Moreover, since in probability theory we know virtually everything about absolute continuity, this fact makes it relatively easy to answer the rest of the questions for this measure of performance. We obtain (Theorem 2.31) two characterizations of those sets $\mathcal{C}$ for which consistent predictors exist (question (i)): one of them is separability (with respect to total variation distance) and the other is an algebraic condition based on the notion of singularity of measures. We also show that question (iv) is equivalent two questions (i) and (ii) for prediction in total variation. Perhaps more importantly, it turns out that *whenever there is a predictor that predicts every measure $\mu\in\mathcal{C}$, there exists a Bayesian predictor with a countably discrete prior (concentrated on $\mathcal{C}$) that predicts every measure $\mu\in\mathcal{C}$ as well.* This provides an answer to question (ii). We show (Theorems 2.7, 2.35) that this property also holds to prediction in expected average KL divergence. In both cases (prediction in total variation and in KL divergence) this is a no-assumption result: the set $\mathcal{C}$ can be completely arbitrary (it does not even have to be measurable). We also obtain sufficient conditions, expressed in terms of separability, that provide answers to questions (i) and (iv) for prediction in expected average KL divergence. To provide an answer to question (ii) we find a suitable generalization of the notion of absolute continuity (a requirement which is stronger than local absolute continuity, but much weaker than absolute continuity proper) under which a measure ρ predicts a measure μ in expected average KL divergence

(Section 2.4), and also use this condition to obtain another characterization that addresses the question (i).

The content of this chapter is organized as follows. Section 2.1 provides notation, definition and auxiliary results. In Section 2.2 we show that if there is a predictor that is consistent (in either KL divergence or total variation) for every μ in $\mathcal{C}$, then there exists a Bayesian predictor with a discrete prior which also has this consistency property. Several sufficient conditions on the set $\mathcal{C}$ for the existence of a consistent predictor are provided in Section 2.3. These conditions include separability of $\mathcal{C}$ (with respect to appropriate topologies). In Section 2.4 we address the question of what are the conditions on a measure ρ under which it is a consistent predictor for a single measure μ in KL divergence. Some sufficient conditions are found, that generalize absolute continuity in a natural way.

Finally, in Section 2.5 we turn to the non-realizable version of the sequence prediction problem (question (iv) above); we show that is different from the realizable version if we consider prediction in KL divergence, and obtain some analogues and generalizations of the results on the realizable problem for the non-realizable version.

2.1 Notation and definitions

Let $\mathbf{X}$ be a finite set. The notation $x_{1..n}$ is used for $x_1,...,x_n$. We consider stochastic processes (probability measures) on $\Omega := (\mathbf{X}^\infty, \mathcal{F})$ where $\mathcal{F}$ is the sigma-field generated by the cylinder sets $[x_{1..n}]$, $x_i \in \mathbf{X}, n \in \mathbb{N}$ and $[x_{1..n}]$ is the set of all infinite sequences that start with $x_{1..n}$. For a finite set A denote $|A|$ its cardinality. We use $\mathbb{E}_\mu$ for expectation with respect to a measure μ.

Next we introduce the criteria of the quality of prediction used in this chapter. For two measures μ and ρ we are interested in how different the μ- and ρ-conditional probabilities are, given a data sample $x_{1..n}$. Introduce the

(conditional) total variation distance

$$v(\mu,\rho,x_{1..n}) := \sup_{A\in\mathcal{F}} |\mu(A|x_{1..n}) - \rho(A|x_{1..n})|.$$

Definition 2.1. *Say that ρ predicts μ in total variation if*

$$v(\mu,\rho,x_{1..n}) \to 0 \ \mu\text{-a.s.}$$

This convergence is rather strong. In particular, it means that ρ-conditional probabilities of arbitrary far-off events converge to μ-conditional probabilities. Recall that μ is *absolutely continuous* with respect to ρ if (by definition) $\mu(A)>0$ implies $\rho(A)>0$ for all $A\in\mathcal{F}$. Moreover, ρ predicts μ in total variation if and only if μ is absolutely continuous with respect to ρ:

Theorem 2.2 ([11, 46])**.** *If ρ, μ are arbitrary probability measures on $(\mathbf{X}^\infty,\mathcal{F})$, then ρ predicts μ in total variation if and only if μ is absolutely continuous with respect to ρ.*

Thus, for a class $\mathcal{C}$ of measures there is a predictor ρ that predicts every $\mu\in\mathcal{C}$ in total variation if and only if every $\mu\in\mathcal{C}$ has a density with respect to ρ. Although such sets of processes are rather large, they do not include even such basic examples as the set of all Bernoulli i.i.d. processes. That is, there is no ρ that would predict in total variation every Bernoulli i.i.d. process measure δ_p, $p\in[0,1]$, where p is the probability of 0 (see the Bernoulli i.i.d. example in Section 2.2.2 for a more detailed discussion). Therefore, perhaps for many (if not most) practical applications this measure of the quality of prediction is too strong, and one is interested in weaker measures of performance.

For two measures μ and ρ introduce the *expected cumulative Kullback-Leibler divergence (KL divergence)* as

$$\delta_n(\mu,\rho) := \mathbb{E}_\mu \sum_{t=1}^{n} \sum_{a\in\mathbf{X}} \mu(x_t=a|x_{1..t-1}) \log \frac{\mu(x_t=a|x_{1..t-1})}{\rho(x_t=a|x_{1..t-1})}, \tag{2.1}$$

In words, we take the expected (over data) average (over time) KL divergence between μ- and ρ-conditional (on the past data) probability distributions of the next outcome.

Definition 2.3. *Say that ρ predicts μ in expected average KL divergence if*

$$\frac{1}{n} d_n(\mu,\rho) \to 0.$$

This measure of performance is much weaker, in the sense that it requires good predictions only one step ahead, and not on every step but only on average; also, the convergence is not with probability 1, but in expectation. With prediction quality so measured, predictors exist for relatively large classes of measures; most notably, [78] provides a predictor which predicts every stationary process in expected average KL divergence. A simple but useful identity that we will need (in the context of sequence prediction introduced also by [78]) is the following

$$d_n(\mu,\rho) = -\sum_{x_{1..n}\in \mathbf{X}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})}, \tag{2.2}$$

where on the right-hand side we have simply the KL divergence between measures μ and ρ restricted to the first n observations.

Thus, the results of the following sections will be established with respect to two very different measures of prediction quality, one of which is very strong and the other rather weak. This suggests that the facts established reflect some fundamental properties of the problem of prediction, rather than those pertinent to particular measures of performance. On the other hand, it remains open to extend the results of this section to different measures of performance (e.g., to those introduced in Section 2.4).

The following sets of process measures will be used repeatedly in the examples.

Definition 2.4. *Consider the following classes of process measures: $\mathcal{P}$ is*

the set of all process measures, $\mathcal{D}$ is the set of all degenerate discrete process measures, $\mathcal{S}$ is the set of all stationary processes and $\mathcal{M}_k$ is the set of all stationary measures with memory not greater than k (k-order Markov processes, with $\mathcal{B}:=\mathcal{M}_0$ being the set of all i.i.d. processes):

$$\mathcal{D}:=\{\mu\in\mathcal{P}:\exists x\in\mathbf{X}^\infty \quad \mu(x)=1\}, \tag{2.3}$$

$$\mathcal{S}:=\{\mu\in\mathcal{P}:\forall n,k\geq 1\,\forall a_{1..n}\in\mathbf{X}^n\,\mu(x_{1..n}=a_{1..n})=\mu(x_{1+k..n+k}=a_{1..n})\}, \tag{2.4}$$

$$\mathcal{M}_k:=\{\mu\in\mathcal{S}:\forall n\geq k\,\forall a\in\mathbf{X}\,\forall a_{1..n}\in\mathbf{X}^n \\ \mu(x_{n+1}=a|x_{1..n}=a_{1..n})=\mu(x_{k+1}=a|x_{1..k}=a_{n-k+1..n})\}, \tag{2.5}$$

$$\mathcal{B}:=\mathcal{M}_0. \tag{2.6}$$

Abusing the notation, we will sometimes use elements of $\mathcal{D}$ and $\mathbf{X}^\infty$ interchangeably. The following (simple and well-known) statement will be used repeatedly in the examples.

Lemma 2.5. *For every $\rho\in\mathcal{P}$ there exists $\mu\in\mathcal{D}$ such that $d_n(\mu,\rho)\geq n\log|\mathbf{X}|$ for all $n\in\mathbb{N}$.*

Proof. Indeed, for each n we can select $\mu(x_n=a)=1$ for $a\in\mathbf{X}$ that minimizes $\rho(x_n=a|x_{1..n-1})$, so that $\rho(x_{1..n})\leq|\mathbf{X}|^{-n}$. □

2.2 If there exists a consistent predictor then there exists a consistent Bayesian predictor with a discrete prior [R3]

In this section we show that if there is a predictor that predicts every μ in some class $\mathcal{C}$, then there is a Bayesian mixture of countably many elements from $\mathcal{C}$ that predicts every $\mu\in\mathcal{C}$ too. This is established for the two notions

of prediction quality that were introduced: total variation and expected average KL divergence.

2.2.1 Introduction and related work

If the class $\mathcal{C}$ of measures is countable (that is, if $\mathcal{C}$ can be represented as $\mathcal{C} := \{\mu_k : k \in \mathbb{N}\}$), then there exists a predictor which performs well for any $\mu \in \mathcal{C}$. Such a predictor can be obtained as a Bayesian mixture $\rho_S := \sum_{k\in\mathbb{N}} w_k \mu_k$, where w_k are summable positive real weights, and it has very strong predictive properties; in particular, ρ_S predicts every $\mu \in \mathcal{C}$ in total variation distance, as follows from the result of [11]. Total variation distance measures the difference in (predicted and true) conditional probabilities of all future events, that is, not only the probabilities of the next observations, but also of observations that are arbitrary far off in the future (see formal definitions in Section 2.1). In the context of sequence prediction the measure ρ_S (introduced in [93]) was first studied by [86]. Since then, the idea of taking a convex combination of a finite or countable class of measures (or predictors) to obtain a predictor permeates most of the research on sequential prediction (see, for example, [18]) and more general learning problems (see [41] as well as Chapter 4 of this manuscript). In practice, it is clear that, on the one hand, countable models are not sufficient, since already the class $\{\mu_p : p \in [0,1]\}$ of Bernoulli i.i.d. processes, where p is the probability of 0, is not countable. On the other hand, prediction in total variation can be too strong to require: predicting probabilities of the next observation may be sufficient, maybe even not on every step but in the Cesaro sense. A key observation here is that a predictor $\rho_S = \sum w_k \mu_k$ may be a good predictor not only when the data is generated by one of the processes μ_k, $k \in \mathbb{N}$, but when it comes from a much larger class. Let us consider this point in more detail. Fix for simplicity $\mathbf{X} = \{0,1\}$. The Laplace predictor

$$\lambda(x_{n+1} = 0 | x_1,...,x_n) = \frac{\#\{i \le n : x_i = 0\} + 1}{n + |\mathbf{X}|} \tag{2.7}$$

predicts any Bernoulli i.i.d. process: although convergence in total variation distance of conditional probabilities does not hold, predicted probabilities of the next outcome converge to the correct ones. Moreover, generalizing the Laplace predictor, a predictor λ_k can be constructed for the class M_k of all k-order Markov measures, for any given k. As was found by [78], the combination $\rho_R := \sum w_k \lambda_k$ is a good predictor not only for the set $\cup_{k\in\mathbb{N}} M_k$ of all finite-memory processes, but also for any measure μ coming from a much larger class: that of all stationary measures on $\mathbf{X}^\infty$. Here prediction is possible only in the Cesaro sense (more precisely, ρ_R predicts every stationary process in expected time-average Kullback-Leibler divergence, see definitions below). The Laplace predictor itself can be obtained as a Bayes mixture over all Bernoulli i.i.d. measures with uniform prior on the parameter p (the probability of 0). However the same (asymptotic) predictive properties are possessed by a Bayes mixture with a countably supported prior which is dense in [0,1] (e.g., taking $\rho := \sum w_k \delta_k$ where $\delta_k, k \in \mathbb{N}$ ranges over all Bernoulli i.i.d. measures with rational probability of 0). For a given k, the set of k-order Markov processes is parametrized by finitely many [0,1]-valued parameters. Taking a dense subset of the values of these parameters, and a mixture of the corresponding measures, results in a predictor for the class of k-order Markov processes. Mixing over these (for all $k \in \mathbb{N}$) yields, as in [78], a predictor for the class of all stationary processes. Thus, for the mentioned classes of processes, a predictor can be obtained as a Bayes mixture of countably many measures in the class. An additional reason why this kind of analysis is interesting is because of the difficulties arising in trying to construct Bayesian predictors for classes of processes that can not be easily parametrized. Indeed, a natural way to obtain a predictor for a class $\mathcal{C}$ of stochastic processes is to take a Bayesian mixture of the class. To do this, one needs to define the structure of a probability space on $\mathcal{C}$. If the class $\mathcal{C}$ is well parametrized, as is the case with the set of all Bernoulli i.i.d. process, then one can integrate with respect to the parametrization. In general, when the problem lacks a natural parametrization, although one

can define the structure of the probability space on the set of (all) stochastic process measures in many different ways, the results one can obtain will then be with probability 1 with respect to the prior distribution (see, for example, [43]). Pointwise consistency cannot be assured (see, for example, [29]) in this case, meaning that some (well-defined) Bayesian predictors are not consistent on some (large) subset of $\mathcal{C}$. Results with prior probability 1 can be hard to interpret if one is not sure that the structure of the probability space defined on the set $\mathcal{C}$ is indeed a natural one for the problem at hand (whereas if one does have a natural parametrization, then usually results for every value of the parameter can be obtained, as in the case with Bernoulli i.i.d. processes mentioned above). The results of the present section show that when a predictor exists it can indeed be given as a Bayesian predictor, which predicts every (and not almost every) measure in the class, while its support is only a countable set.

It is worth noting that for the problem of sequence prediction the case of stationary ergodic data is relatively well-studied, and several methods of consistent prediction are available, both for discrete- and real-valued data; limitations of these methods are also relatively well-understood (besides the works cited above, see also [5, 79, 63, 64, 1]). This is why in this chapter we mostly concentrate on the general case of (non-stationary) sources of data.

After the theorems we present some examples of families of measures for which predictors exist.

2.2.2 Results

Theorem 2.6. *Let $\mathcal{C}$ be a set of probability measures on $(\mathbf{X}^\infty,\mathcal{F})$. If there is a measure ρ such that ρ predicts every $\mu\in\mathcal{C}$ in total variation, then there is a sequence $\mu_k\in\mathcal{C}$, $k\in\mathbb{N}$ such that the measure $\nu:=\sum_{k\in\mathbb{N}}w_k\mu_k$ predicts every $\mu\in\mathcal{C}$ in total variation, where w_k are any positive weights that sum to 1.*

This relatively simple fact can be proven in different ways, relying on the mentioned equivalence (Theorem 2.2) of the statements "ρ predicts μ in

total variation distance" and "μ is absolutely continuous with respect to ρ." The proof presented below is not the shortest possible, but it uses ideas and techniques that are then generalized to the case of prediction in expected average KL-divergence, which is more involved, since in all interesting cases all measures $\mu\in\mathcal{C}$ are singular with respect to any predictor that predicts all of them. Another proof of Theorem 2.6 can be obtained from Theorem 2.9 in the next section. Yet another way would be to derive it from algebraic properties of the relation of absolute continuity, given in [70].

Proof. We break the (relatively easy) proof of this theorem into three steps, which will make the proof of the next theorem more understandable.

Step 1: densities. For any $\mu\in\mathcal{C}$, since ρ predicts μ in total variation, by Theorem 2.2, μ has a density (Radon-Nikodym derivative) f_μ with respect to ρ. Thus, for the (measurable) set T_μ of all sequences $x_1,x_2,...\in\mathbf{X}^\infty$ on which $f_\mu(x_{1,2,...})>0$ (the limit $\lim_{n\to\infty}\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$ exists and is finite and positive) we have $\mu(T_\mu)=1$ and $\rho(T_\mu)>0$. Next we will construct a sequence of measures $\mu_k\in\mathcal{C}$, $k\in\mathbb{N}$ such that the union of the sets T_{μ_k} has probability 1 with respect to every $\mu\in\mathcal{C}$, and will show that this is a sequence of measures whose existence is asserted in the theorem statement.

Step 2: a countable cover and the resulting predictor. Let $\varepsilon_k:=2^{-k}$ and let $m_1:=\sup_{\mu\in\mathcal{C}}\rho(T_\mu)$. Clearly, $m_1>0$. Find any $\mu_1\in\mathcal{C}$ such that $\rho(T_{\mu_1})\geq m_1-\varepsilon_1$, and let $T_1=T_{\mu_1}$. For $k>1$ define $m_k:=\sup_{\mu\in\mathcal{C}}\rho(T_\mu\backslash T_{k-1})$. If $m_k=0$ then define $T_k:=T_{k-1}$, otherwise find any μ_k such that $\rho(T_{\mu_k}\backslash T_{k-1})\geq m_k-\varepsilon_k$, and let $T_k:=T_{k-1}\cup T_{\mu_k}$. Define the predictor ν as $\nu:=\sum_{k\in\mathbb{N}}w_k\mu_k$.

Step 3: ν predicts every $\mu\in\mathcal{C}$. Since the sets T_1, $T_2\backslash T_1$,...,$T_k\backslash T_{k-1}$,... are disjoint, we must have $\rho(T_k\backslash T_{k-1})\to 0$, so that $m_k\to 0$ (since $m_k\leq\rho(T_k\backslash T_{k-1})+\varepsilon_k\to 0$). Let

$$T:=\cup_{k\in\mathbb{N}}T_k.$$

Fix any $\mu\in\mathcal{C}$. Suppose that $\mu(T_\mu\backslash T)>0$. Since μ is absolutely continuous

with respect to ρ, we must have $\rho(T_\mu \setminus T) > 0$. Then for every $k > 1$ we have

$$m_k = \sup_{\mu' \in \mathcal{C}} \rho(T_{\mu'} \setminus T_{k-1}) \geq \rho(T_\mu \setminus T_{k-1}) \geq \rho(T_\mu \setminus T) > 0,$$

which contradicts $m_k \to 0$. Thus, we have shown that

$$\mu(T \cap T_\mu) = 1. \tag{2.8}$$

Let us show that every $\mu \in \mathcal{C}$ is absolutely continuous with respect to ν. Indeed, fix any $\mu \in \mathcal{C}$ and suppose $\mu(A) > 0$ for some $A \in \mathcal{F}$. Then from (2.8) we have $\mu(A \cap T) > 0$, and, by absolute continuity of μ with respect to ρ, also $\rho(A \cap T) > 0$. Since $T = \cup_{k \in \mathbb{N}} T_k$, we must have $\rho(A \cap T_k) > 0$ for some $k \in \mathbb{N}$. Since on the set T_k the measure μ_k has non-zero density f_{μ_k} with respect to ρ, we must have $\mu_k(A \cap T_k) > 0$. (Indeed, $\mu_k(A \cap T_k) = \int_{A \cap T_k} f_{\mu_k} d\rho > 0$.) Hence,

$$\nu(A \cap T_k) \geq w_k \mu_k(A \cap T_k) > 0,$$

so that $\nu(A) > 0$. Thus, μ is absolutely continuous with respect to ν, and so, by Theorem 2.2, ν predicts μ in total variation distance. □

Thus, examples of families $\mathcal{C}$ for which there is a ρ that predicts every $\mu \in \mathcal{C}$ in total variation, are limited to families of measures which have a density with respect to some measure ρ. On the one hand, from statistical point of view, such families are rather large: the assumption that the probabilistic law in question has a density with respect to some (nice) measure is a standard one in statistics. It should also be mentioned that such families can easily be uncountable. On the other hand, even such basic examples as the set of all Bernoulli i.i.d. measures does not allow for a predictor that predicts every measure in total variation. Indeed, all these processes are singular with respect to one another; in particular, each of the non-overlapping sets T_p of all sequences which have limiting fraction p of 0s has probability 1 with respect to one of the measures and 0 with respect to all others;

since there are uncountably many of these measures, there is no measure ρ with respect to which they all would have a density (since such a measure should have $\rho(T_p)>0$ for all p) . As it was mentioned, predicting in total variation distance means predicting with arbitrarily growing horizon [46], while prediction in expected average KL divergence is only concerned with the probabilities of the next observation, and only on time and data average. For the latter measure of prediction quality, consistent predictors exist not only for the class of all Bernoulli processes, but also for the class of all stationary processes [78]. The next theorem establishes the result similar to Theorem 2.29 for expected average KL divergence.

Theorem 2.7. *Let $\mathcal{C}$ be a set of probability measures on $(\mathbf{X}^\infty,\mathcal{F})$. If there is a measure ρ such that ρ predicts every $\mu\in\mathcal{C}$ in expected average KL divergence, then there exist a sequence $\mu_k\in\mathcal{C}$, $k\in\mathbb{N}$ and a sequence $w_k>0, k\in\mathbb{N}$, such that $\sum_{k,\in\mathbb{N}} w_k=1$, and the measure $\nu:=\sum_{k\in\mathbb{N}} w_k\mu_k$ predicts every $\mu\in\mathcal{C}$ in expected average KL divergence.*

A difference worth noting with respect to the formulation of Theorem 2.6 (apart from a different measure of divergence) is in that in the latter the weights w_k can be chosen arbitrarily, while in Theorem 2.7 this is not the case. In general, the statement "$\sum_{k\in\mathbb{N}} w_k\nu_k$ predicts μ in expected average KL divergence for some choice of w_k, $k\in\mathbb{N}$" does not imply "$\sum_{k\in\mathbb{N}} w'_k\nu_k$ predicts μ in expected average KL divergence for every summable sequence of positive $w'_k, k\in\mathbb{N}$," while the implication trivially holds true if the expected average KL divergence is replaced by the total variation. This is illustrated in the last example of this section.

The idea of the proof of Theorem 2.7 is as follows. For every μ and every n we consider the sets T^n_μ of those $x_{1..n}$ on which μ is greater than ρ. These sets have to have (from some n on) a high probability with respect to μ. Then since ρ predicts μ in expected average KL divergence, the ρ-probability of these sets cannot decrease exponentially fast (that is, it has to be quite large). (The sequences $\mu(x_{1..n})/\rho(x_{1..n})$, $n\in\mathbb{N}$ will play the role

of densities of the proof of Theorem 2.6, and the sets T_μ^n the role of sets T_μ on which the density is non-zero.) We then use, for each given n, the same scheme to cover the set $\mathbf{X}^n$ with countably many T_μ^n, as was used in the proof of Theorem 2.6 to construct a countable covering of the set $\mathbf{X}^\infty$, obtaining for each n a predictor ν_n. Then the predictor ν is obtained as $\sum_{n\in\mathbb{N}} w_n\nu_n$, where the weights decrease subexponentially. The latter fact ensures that, although the weights depend on n, they still play no role asymptotically. The technically most involved part of the proof is to show that the sets T_μ^n in asymptotic have sufficiently large weights in those countable covers that we construct for each n. This is used to demonstrate the implication "if a set has a high μ probability, then its ρ-probability does not decrease too fast, provided some regularity conditions."

The proof is deferred to Section 2.6.1.

Example: countable classes of measures. A very simple but rich example of a class $\mathcal{C}$ that satisfies the conditions of both the theorems above, is any countable family $\mathcal{C}=\{\mu_k : k\in\mathbb{N}\}$ of measures. In this case, any mixture predictor $\rho:=\sum_{k\in\mathbb{N}} w_k\mu_k$ predicts all $\mu\in\mathcal{C}$ both in total variation and in expected average KL divergence. A particular instance, that has gained much attention in the literature, is the family of all computable measures. Although countable, this family of processes is rather rich. The problem of predicting all computable measures was introduced in [86], where a mixture predictor was proposed.

Example: Bernoulli i.i.d. processes. Consider the class $\mathcal{B}=\{\mu_p : p\in[0,1]\}$ of all Bernoulli i.i.d. processes: $\mu_p(x_k=0)=p$ independently for all $k\in\mathbb{N}$. Clearly, this family is uncountable. Moreover, each set

$$T_p:=\{x\in\mathbf{X}^\infty:\text{ the limiting fraction of 0s in } x \text{ equals } p\},$$

has probability 1 with respect to μ_p and probability 0 with respect to any $\mu_{p'} : p'\neq p$. Since the sets T_p, $p\in[0,1]$ are non-overlapping, there is no measure ρ for which $\rho(T_p)>0$ for all $p\in[0,1]$. That is, there is no measure

ρ with respect to which all μ_p are absolutely continuous. Therefore, by Theorem 2.2, a predictor that predicts any $\mu\in\mathcal{B}$ in total variation does not exist, demonstrating that this notion of prediction is rather strong. However, we know (e.g., [54]) that the Laplace predictor (2.7) predicts every Bernoulli i.i.d. process in expected average KL divergence (and not only). Hence, Theorem 2.29 implies that there is a countable mixture predictor for this family too. Let us find such a predictor. Let $\mu_q : q\in Q$ be the family of all Bernoulli i.i.d. measures with rational probability of 0, and let $\rho := \sum_{q\in Q} w_q\mu_q$, where w_q are arbitrary positive weights that sum to 1. Let μ_p be any Bernoulli i.i.d. process. Let $h(p,q)$ denote the divergence $p\log(p/q)+(1-p)\log(1-p/1-q)$. For each ε we can find a $q\in Q$ such that $h(p,q)<\varepsilon$. Then

$$\frac{1}{n}d_n(\mu_p,\rho)=\frac{1}{n}\mathbb{E}_{\mu_p}\log\frac{\log\mu_p(x_{1..n})}{\log\rho(x_{1..n})}\le\frac{1}{n}\mathbb{E}_{\mu_p}\log\frac{\log\mu_p(x_{1..n})}{w_q\log\mu_q(x_{1..n})}$$
$$=-\frac{\log w_q}{n}+h(p,q)\le\varepsilon+o(1). \quad (2.9)$$

Since this holds for each ε, we conclude that $\frac{1}{n}d_n(\mu_p,\rho)\to 0$ and ρ predicts every $\mu\in\mathcal{B}$ in expected average KL divergence.

Example: stationary processes. In [78] a predictor ρ_R was constructed which predicts every stationary process $\rho\in\mathcal{S}$ in expected average KL divergence. (This predictor is obtained as a mixture of predictors for k-order Markov sources, for all $k\in\mathbb{N}$.) Therefore, Theorem 2.7 implies that there is also a countable mixture predictor for this family of processes. Such a predictor can be constructed as follows (the proof in this example is based on the proof in [80], Appendix 1). Observe that the family $\mathcal{M}_k$ of k-order stationary binary-valued Markov processes is parametrized by 2^k [0,1]-valued parameters: probability of observing 0 after observing $x_{1..k}$, for each $x_{1..k}\in\mathbf{X}^k$. For each $k\in\mathbb{N}$ let μ^k_q, $q\in Q^{2^k}$ be the (countable) family of all stationary k-order Markov processes with rational values of all the parameters. We will show that any predictor $\nu:=\sum_{k\in\mathbb{N}}\sum_{q\in Q^{2^k}} w_k w_q\mu^k_q$, where w_k, $k\in\mathbb{N}$ and $w_q, q\in Q^{2^k}$,

$k\in\mathbb{N}$ are any sequences of positive real weights that sum to 1, predicts every stationary $\mu\in\mathcal{S}$ in expected average KL divergence. For $\mu\in\mathcal{S}$ and $k\in\mathbb{N}$ define the k-order conditional Shannon entropy $h_k(\mu):=\mathbb{E}_\mu\log\mu(x_{k+1}|x_{1..k})$. We have $h_{k+1}(\mu)\geq h_k(\mu)$ for every $k\in\mathbb{N}$ and $\mu\in\mathcal{S}$, and the limit

$$h_\infty(\mu):=\lim_{k\to\infty}h_k(\mu) \tag{2.10}$$

is called the limit Shannon entropy; see, for example, [32]. Fix some $\mu\in\mathcal{S}$. It is easy to see that for every $\varepsilon>0$ and every $k\in\mathbb{N}$ we can find a k-order stationary Markov measure $\mu^k_{q_\varepsilon}$, $q_\varepsilon\in Q^{2^k}$ with rational values of the parameters, such that

$$\mathbb{E}_\mu\log\frac{\mu(x_{k+1}|x_{1..k})}{\mu^k_{q_\varepsilon}(x_{k+1}|x_{1..k})}<\varepsilon. \tag{2.11}$$

We have

$$\begin{aligned}\frac{1}{n}d_n(\mu,\nu)&\leq-\frac{\log w_k w_{q_\varepsilon}}{n}+\frac{1}{n}d_n(\mu,\mu^k_{q_\varepsilon})\\&=O(k/n)+\frac{1}{n}\mathbb{E}_\mu\log\mu(x_{1..n})-\frac{1}{n}\mathbb{E}_\mu\log\mu^k_{q_\varepsilon}(x_{1..n})\\&=o(1)+h_\infty(\mu)-\frac{1}{n}\mathbb{E}_\mu\sum_{k=1}^{n}\log\mu^k_{q_\varepsilon}(x_t|x_{1..t-1})\\&=o(1)+h_\infty(\mu)-\frac{1}{n}\mathbb{E}_\mu\sum_{t=1}^{k}\log\mu^k_{q_\varepsilon}(x_t|x_{1..t-1})-\frac{n-k}{n}\mathbb{E}_\mu\log\mu^k_{q_\varepsilon}(x_{k+1}|x_{1..k})\\&\leq o(1)+h_\infty(\mu)-\frac{n-k}{n}(h_k(\mu)-\varepsilon),\end{aligned} \tag{2.12}$$

where the first inequality is derived analogously to (2.9), the first equality follows from (2.2), the second equality follows from the Shannon-McMillan-Breiman theorem (e.g., [32]), that states that $\frac{1}{n}\log\mu(x_{1..n})\to h_\infty(\mu)$ in expectation (and a.s.) for every $\mu\in\mathcal{S}$, and (2.2); in the third equality we have used the fact that $\mu^k_{q_\varepsilon}$ is k-order Markov and μ is stationary, whereas the last inequality follows from (2.11). Finally, since the choice of k and ε was

arbitrary, from (2.12) and (2.10) we obtain $\lim_{n\to\infty}\frac{1}{n}d_n(\mu,\nu)=0$.

Example: weights may matter. Finally, we provide an example that illustrates the difference between the formulations of Theorems 2.6 and 2.7: in the latter the weights are not arbitrary. We will construct a sequence of measures $\nu_k, k\in\mathbb{N}$, a measure μ, and two sequences of positive weights w_k and w'_k with $\sum_{k\in\mathbb{N}}w_k=\sum_{k\in\mathbb{N}}w'_k=1$, for which $\nu:=\sum_{k\in\mathbb{N}}w_k\nu_k$ predicts μ in expected average KL divergence, but $\nu':=\sum_{k\in\mathbb{N}}w'_k\nu_k$ does not. Let ν_k be a deterministic measure that first outputs k 0s and then only 1s, $k\in\mathbb{N}$. Let $w_k=w/k^2$ with $w=6/\pi^2$ and $w'_k=2^{-k}$. Finally, let μ be a deterministic measure that outputs only 0s. We have $d_n(\mu,\nu)=-\log(\sum_{k\geq n}w_k)\leq-\log(wn^{-2})=o(n)$, but $d_n(\mu,\nu')=-\log(\sum_{k\geq n}w'_k)=-\log(2^{-n+1})=n-1\neq o(n)$, proving the claim.

2.3 Characterizing predictable classes [R3]

In this section we exhibit some sufficient conditions on the class $\mathcal{C}$, under which a predictor for all measures in $\mathcal{C}$ exists. It is important to note that none of these conditions relies on a parametrization of any kind. The conditions presented are of two types: conditions on asymptotic behaviour of measures in $\mathcal{C}$, and on their local (restricted to first n observations) behaviour. Conditions of the first type concern separability of $\mathcal{C}$ with respect to the total variation distance and the expected average KL divergence. We show that in the case of total variation separability is a necessary and sufficient condition for the existence of a predictor, whereas in the case of expected average KL divergence it is sufficient but is not necessary.

The conditions of the second kind concern the "capacity" of the sets $\mathcal{C}^n:=\{\mu^n:\mu\in\mathcal{C}\}$, $n\in\mathbb{N}$, where μ^n is the measure μ restricted to the first n observations. Intuitively, if $\mathcal{C}^n$ is small (in some sense), then prediction is possible. We measure the capacity of $\mathcal{C}^n$ in two ways. The first way is to find the maximum probability given to each sequence $x_1,...,x_n$ by some measure in the class, and then take a sum over $x_1,...,x_n$. Denoting the obtained quantity c_n, one can show that it grows polynomially in n for some important

classes of processes, such as i.i.d. or Markov processes. We show that, in general, if c_n grows subexponentially then a predictor exists that predicts any measure in $\mathcal{C}$ in expected average KL divergence. On the other hand, exponentially growing c_n are not sufficient for prediction. A more refined way to measure the capacity of $\mathcal{C}^n$ is using a concept of channel capacity from information theory, which was developed for a closely related problem of finding optimal codes for a class of sources. We extend corresponding results from information theory to show that sublinear growth of channel capacity is sufficient for the existence of a predictor, in the sense of expected average divergence. Moreover, the obtained bounds on the divergence are optimal up to an additive logarithmic term.

2.3.1 Separability

Knowing that a mixture of a countable subset gives a predictor if there is one, a notion that naturally comes to mind, when trying to characterize families of processes for which a predictor exists, is separability. Can we say that there is a predictor for a class $\mathcal{C}$ of measures if and only if $\mathcal{C}$ is separable? Of course, to talk about separability we need a suitable topology on the space of all measures, or at least on $\mathcal{C}$. If the formulated questions were to have a positive answer, we would need a different topology for each of the notions of predictive quality that we consider. Sometimes these measures of predictive quality indeed define a nice enough structure of a probability space, but sometimes they do not. The question whether there exists a topology on $\mathcal{C}$, separability with respect to which is equivalent to the existence of a predictor, is already more vague and less appealing. Nonetheless, in the case of total variation distance we obviously have a candidate topology: that of total variation distance, and indeed separability with respect to this topology is equivalent to the existence of a predictor, as the next theorem shows. This theorem also implies Theorem 2.6, thereby providing an alternative proof for the latter. In the case of expected average KL divergence the situation is

different. While one can introduce a topology based on it, separability with respect to this topology turns out to be a sufficient but not a necessary condition for the existence of a predictor, as is shown in Theorem 2.11.

Definition 2.8 (unconditional total variation distance)**.** *Introduce the (unconditional) total variation distance*

$$v(\mu,\rho) := \sup_{A\in\mathcal{F}} |\mu(A)-\rho(A)|.$$

Theorem 2.9. *Let $\mathcal{C}$ be a set of probability measures on $(\mathbf{X}^\infty,\mathcal{F})$. There is a measure ρ such that ρ predicts every $\mu\in\mathcal{C}$ in total variation if and only if $\mathcal{C}$ is separable with respect to the topology of total variation distance. In this case, any measure ν of the form $\nu=\sum_{k=1}^\infty w_k\mu_k$, where $\{\mu_k : k\in\mathbb{N}\}$ is any dense countable subset of $\mathcal{C}$ and w_k are any positive weights that sum to 1, predicts every $\mu\in\mathcal{C}$ in total variation.*

Proof. Sufficiency and the mixture predictor. Let $\mathcal{C}$ be separable in total variation distance, and let $\mathcal{D}=\{\nu_k : k\in\mathbb{N}\}$ be its dense countable subset. We have to show that $\nu := \sum_{k\in\mathbb{N}} w_k\nu_k$, where w_k are any positive real weights that sum to 1, predicts every $\mu\in\mathcal{C}$ in total variation. To do this, it is enough to show that $\mu(A)>0$ implies $\nu(A)>0$ for every $A\in\mathcal{F}$ and every $\mu\in\mathcal{C}$. Indeed, let A be such that $\mu(A)=\varepsilon>0$. Since $\mathcal{D}$ is dense in $\mathcal{C}$, there is a $k\in\mathbb{N}$ such that $v(\mu,\nu_k)<\varepsilon/2$. Hence $\nu_k(A)\geq\mu(A)-v(\mu,\nu_k)\geq\varepsilon/2$ and $\nu(A)\geq w_k\nu_k(A)\geq w_k\varepsilon/2>0$.

Necessity. For any $\mu\in\mathcal{C}$, since ρ predicts μ in total variation, μ has a density (Radon-Nikodym derivative) f_μ with respect to ρ. We can define L_1 distance with respect to ρ as $L_1^\rho(\mu,\nu)=\int_{\mathbf{X}^\infty}|f_\mu-f_\nu|d\rho$. The set of all measures that have a density with respect to ρ, is separable with respect to this distance (for example, a dense countable subset can be constructed based on measures whose densities are step-functions, that take only rational values, see, e.g., [53]); therefore, its subset $\mathcal{C}$ is also separable. Let $\mathcal{D}$ be any dense countable subset of $\mathcal{C}$. Thus, for every $\mu\in\mathcal{C}$ and every ε there is a

$\mu' \in \mathcal{D}$ such that $L_1^\rho(\mu,\mu') < \varepsilon$. For every measurable set A we have

$$|\mu(A) - \mu'(A)| = \left| \int_A f_\mu d\rho - \int_A f_{\mu'} d\rho \right| \le \int_A |f_\mu - f_{\mu'}| d\rho \le \int_{\mathbf{X}^\infty} |f_\mu - f_{\mu'}| d\rho < \varepsilon.$$

Therefore, $v(\mu,\mu') = \sup_{A\in\mathcal{F}} |\mu(A) - \mu'(A)| < \varepsilon$, and the set $\mathcal{C}$ is separable in total variation distance. □

Definition 2.10 (asymptotic KL "distance" D)**.** *Define asymptotic expected average KL divergence between measures μ and ρ as*

$$D(\mu,\rho) = \limsup_{n\to\infty} \frac{1}{n} d_n(\mu,\rho). \tag{2.13}$$

Theorem 2.11. *For any set $\mathcal{C}$ of probability measures on $(\mathbf{X}^\infty,\mathcal{F})$, separability with respect to the asymptotic expected average KL divergence D is a sufficient but not a necessary condition for the existence of a predictor:*

(i) If there exists a countable set $\mathcal{D} := \{\nu_k : k \in \mathbb{N}\} \subset \mathcal{C}$, such that for every $\mu \in \mathcal{C}$ and every $\varepsilon > 0$ there is a measure $\mu' \in \mathcal{D}$, such that $D(\mu,\mu') < \varepsilon$, then every measure ν of the form $\nu = \sum_{k=1}^\infty w_k \mu_k$, where w_k are any positive weights that sum to 1, predicts every $\mu \in \mathcal{C}$ in expected average KL divergence.

(ii) There is an uncountable set $\mathcal{C}$ of measures, and a measure ν, such that ν predicts every $\mu \in \mathcal{C}$ in expected average KL divergence, but $\mu_1 \ne \mu_2$ implies $D(\mu_1,\mu_2) = \infty$ for every $\mu_1,\mu_2 \in \mathcal{C}$; in particular, $\mathcal{C}$ is not separable with respect to D.

Proof. (i) Fix $\mu \in \mathcal{C}$. For every $\varepsilon > 0$ pick $k \in \mathbb{N}$ such that $D(\mu,\nu_k) < \varepsilon$. We have

$$d_n(\mu,\nu) = \mathbb{E}_\mu \log \frac{\mu(x_{1..n})}{\nu(x_{1..n})} \le \mathbb{E}_\mu \log \frac{\mu(x_{1..n})}{w_k \nu_k(x_{1..n})} = -\log w_k + d_n(\mu,\nu_k) \le n\varepsilon + o(n).$$

Since this holds for every ε, we conclude $\frac{1}{n} d_n(\mu,\nu) \to 0$.

(ii) Let $\mathcal{C}$ be the set of all deterministic sequences (measures concentrated on just one sequence) such that the number of 0s in the first n symbols is less than $\sqrt{n}$. Clearly, this set is uncountable. It is easy to check that $\mu_1 \neq \mu_2$ implies $D(\mu_1,\mu_2)=\infty$ for every $\mu_1,\mu_2 \in \mathcal{C}$, but the predictor ν, given by $\nu(x_n = 0) := 1/n$ independently for different n, predicts every $\mu \in \mathcal{C}$ in expected average KL divergence. □

Examples. Basically, the examples of the preceding section carry over here. Indeed, the example of countable families is trivially also an example of separable (with respect to either of the considered topologies) family. For Bernoulli i.i.d. and k-order Markov processes, the (countable) sets of processes that have rational values of the parameters, considered in the previous section, are dense both in the topology of the parametrization and with respect to the asymptotic average divergence D. It is also easy to check from the arguments presented in the corresponding example of Section 2.5.1, that the family of all k-order stationary Markov processes with rational values of the parameters, where we take all $k \in \mathbb{N}$, is dense with respect to D in the set $\mathcal{S}$ of all stationary processes, so that $\mathcal{S}$ is separable with respect to D. Thus, the sufficient but not necessary condition of separability is satisfied in this case. On the other hand, neither of these latter families is separable with respect to the topology of total variation distance.

2.3.2 Conditions based on local behaviour of measures

Next we provide some sufficient conditions for the existence of a predictor based on local characteristics of the class of measures, that is, measures truncated to the first n observations. First of all, it must be noted that necessary and sufficient conditions cannot be obtained this way. The basic example is that of a family $\mathcal{D}$ of all deterministic sequences that are 0 from some time on. This is a countable class of measures which is very easy to predict. Yet, the class of measures on $\mathbf{X}^n$, obtained by truncating all measures in $\mathcal{D}$ to the first n observations, coincides with what would be

obtained by truncating all deterministic measures to the first n observations, the latter class being obviously not predictable at all (see also examples below). Nevertheless, considering this kind of local behaviour of measures, one can obtain not only sufficient conditions for the existence of a predictor, but also rates of convergence of the prediction error. It also gives some ideas of how to construct predictors, for the cases when the sufficient conditions obtained are met.

For a class $\mathcal{C}$ of stochastic processes and a sequence $x_{1..n} \in \mathbf{X}^n$ introduce the coefficients

$$c_{x_{1..n}}(\mathcal{C}) := \sup_{\mu \in \mathcal{C}} \mu(x_{1..n}). \tag{2.14}$$

Define also the normalizer

$$c_n(\mathcal{C}) := \sum_{x_{1..n} \in \mathbf{X}^n} c_{x_{1..n}}(\mathcal{C}). \tag{2.15}$$

Definition 2.12 (NML estimate)**.** *The normalized maximum likelihood estimator λ is defined (e.g., [54]) as*

$$\lambda_{\mathcal{C}}(x_{1..n}) := \frac{1}{c_n(\mathcal{C})} c_{x_{1..n}}(\mathcal{C}), \tag{2.16}$$

for each $x_{1..n} \in \mathbf{X}^n$.

The family $\lambda_{\mathcal{C}}(x_{1..n})$ (indexed by n) in general does not immediately define a stochastic process over $\mathbf{X}^\infty$ ($\lambda_{\mathcal{C}}$ are not consistent for different n); thus, in particular, using average KL divergence for measuring prediction quality would not make sense, since

$$d_n(\mu(\cdot|x_{1..n-1}), \lambda_{\mathcal{C}}(\cdot|x_{1..n-1}))$$

can be negative, as the following example shows.

Example: negative d_n for NML estimates. Let the processes μ_i, $i \in \{1,...,4\}$ be defined on the steps $n = 1,2$ as follows. $\mu_1(00) = \mu_2(01) =$

$\mu_4(11)=1$, while $\mu_3(01)=\mu_3(00)=1/2$. We have $\lambda_{\mathcal{C}}(1)=\lambda_{\mathcal{C}}(0)=1/2$, while $\lambda_{\mathcal{C}}(00)=\lambda_{\mathcal{C}}(01)=\lambda_{\mathcal{C}}(11)=1/3$. If we define $\lambda_{\mathcal{C}}(x|y)=\lambda_{\mathcal{C}}(yx)/\lambda_{\mathcal{C}}(y)$, we obtain $\lambda_{\mathcal{C}}(1|0)=\lambda_{\mathcal{C}}(0|0)=2/3$. Then $d_2(\mu_3(\cdot|0),\lambda_{\mathcal{C}}(\cdot|0))=\log 3/4<0$.

Yet, by taking an appropriate mixture, it is still possible to construct a predictor (a stochastic process) based on λ, that predicts all the measures in the class.

Definition 2.13 (predictor ρ_c)**.** *Let $w:=6/\pi^2$ and let $w_k:=\frac{w}{k^2}$. Define a measure μ_k as follows. On the first k steps it is defined as $\lambda_{\mathcal{C}}$, and for $n>k$ it outputs only zeros with probability 1; so, $\mu_k(x_{1..k})=\lambda_{\mathcal{C}}(x_{1..k})$ and $\mu_k(x_n=0)=1$ for $n>k$. Define the measure ρ_c as*

$$\rho_c=\sum_{k=1}^{\infty} w_k\mu_k. \tag{2.17}$$

Thus, we have taken the normalized maximum likelihood estimates λ_n for each n and continued them arbitrarily (actually, by a deterministic sequence) to obtain a sequence of measures on $(\mathbf{X}^\infty,\mathcal{F})$ that can be summed.

Theorem 2.14. *For any set $\mathcal{C}$ of probability measures on $(\mathbf{X}^\infty,\mathcal{F})$, the predictor ρ_c defined above satisfies*

$$\frac{1}{n}d_n(\mu,\rho_c)\le\frac{\log c_n(\mathcal{C})}{n}+O\left(\frac{\log n}{n}\right); \tag{2.18}$$

in particular, if

$$\log c_n(\mathcal{C})=o(n), \tag{2.19}$$

then ρ_c predicts every $\mu\in\mathcal{C}$ in expected average KL divergence.

Proof. Indeed,

$$\begin{aligned}\frac{1}{n}d_n(\mu,\rho_c)=\frac{1}{n}\mathbb{E}\log\frac{\mu(x_{1..n})}{\rho_c(x_{1..n})}&\le\frac{1}{n}\mathbb{E}\log\frac{\mu(x_{1..n})}{w_n\mu_n(x_{1..n})}\\&\le\frac{1}{n}\log\frac{c_n(\mathcal{C})}{w_n}=\frac{1}{n}(\log c_n(\mathcal{C})+2\log n+\log w).\end{aligned} \tag{2.20}$$

□

Example: i.i.d., finite-memory. To illustrate the applicability of the theorem we first consider the class of i.i.d. processes $\mathcal{B}$ over the binary alphabet $\mathbf{X}=\{0,1\}$. It is easy to see that, for each $x_1,...,x_n$,

$$\sup_{\mu\in\mathcal{B}}\mu(x_{1..n})=(k/n)^k(1-k/n)^{n-k},$$

where $k=\#\{i\le n: x_i=0\}$ is the number of 0s in $x_1,...,x_n$. For the constants $c_n(\mathcal{C})$ we can derive

$$c_n(C)=\sum_{x_{1..n}\in\mathbf{X}^n}\sup_{\mu\in\mathcal{B}}\mu(x_{1..n})=\sum_{x_{1..n}\in\mathbf{X}^n}(k/n)^k(1-k/n)^{n-k}$$
$$=\sum_{k=0}^{n}\binom{n}{k}(k/n)^k(1-k/n)^{n-k}\le\sum_{k=0}^{n}\sum_{t=0}^{n}\binom{n}{k}(k/n)^t(1-k/n)^{n-t}=n+1,$$

so that $c_n(C)\le n+1$.

In general, for the class $\mathcal{M}_k$ of **processes with memory** k over a finite space $\mathbf{X}$ we can get polynomial coefficients $c_n(\mathcal{M}_k)$ (see, for example, [54] and also Section 2.4). Thus, with respect to finite-memory processes, the conditions of Theorem 2.14 leave ample space for the growth of $c_n(\mathcal{C})$, since (2.19) allows subexponential growth of $c_n(\mathcal{C})$. Moreover, these conditions are tight, as the following example shows.

Example: exponential coefficients are not sufficient. Observe that the condition (2.19) cannot be relaxed further, in the sense that exponential coefficients c_n are not sufficient for prediction. Indeed, for the class of all deterministic processes (that is, each process from the class produces some fixed sequence of observations with probability 1) we have $c_n=2^n$, while obviously for this class a predictor does not exist.

Example: stationary processes. For the set of all stationary processes we can obtain $c_n(C)\ge 2^n/n$ (as is easy to see by considering periodic n-order Markov processes, for each $n\in\mathbb{N}$), so that the conditions of Theorem 2.14

are not satisfied. This cannot be fixed, since uniform rates of convergence cannot be obtained for this family of processes, as was shown in [78].

Optimal rates of convergence. A natural question that arises with respect to the bound (2.18) is whether it can be matched by a lower bound. This question is closely related to the optimality of the normalized maximum likelihood estimates used in the construction of the predictor. In general, since NML estimates are not optimal, neither are the rates of convergence in (2.18). To obtain (close to) optimal rates one has to consider a different measure of capacity.

To do so, we make the following connection to a problem in information theory. Let $\mathcal{P}(\mathbf{X}^\infty)$ be the set of all stochastic processes (probability measures) on the space $(\mathbf{X}^\infty,\mathcal{F})$, and let $\mathcal{P}(\mathbf{X})$ be the set of probability distributions over a (finite) set $\mathbf{X}$. For a class $\mathcal{C}$ of measures we are interested in a predictor that has a small (or minimal) worst-case (with respect to the class $\mathcal{C}$) probability of error. Thus, we are interested in the quantity

$$\inf_{\rho\in\mathcal{P}(\mathbf{X}^\infty)}\sup_{\mu\in\mathcal{C}} D(\mu,\rho), \tag{2.21}$$

where the infimum is taken over all stochastic processes ρ, and D is the asymptotic expected average KL divergence (2.13). (In particular, we are interested in the conditions under which the quantity (2.21) equals zero.) This problem has been studied for the case when the probability measures are over a finite set $\mathbf{X}$, and D is replaced simply by the KL divergence d between the measures. Thus, the problem was to find the probability measure ρ (if it exists) on which the following minimax is attained

$$R(A) := \inf_{\rho\in\mathcal{P}(\mathbf{X})}\sup_{\mu\in A} d(\mu,\rho), \tag{2.22}$$

where $A\subset\mathcal{P}(\mathbf{X})$. This problem is closely related to the problem of finding the best code for the class of sources A, which was its original motivation. The normalized maximum likelihood distribution considered above does not

in general lead to the optimum solution for this problem. The optimum solution is obtained through the result that relates the minimax (2.22) to the so-called channel capacity.

Definition 2.15 (Channel capacity)**.** *For a set A of measures on a finite set $\mathbf{X}$ the* channel capacity *of A is defined as*

$$C(A) := \sup_{P \in \mathcal{P}_0(A)} \sum_{\mu \in S(P)} P(\mu) d(\mu, \rho_P), \tag{2.23}$$

where $\mathcal{P}_0(A)$ is the set of all probability distributions on A that have a finite support, $S(P)$ is the (finite) support of a distribution $P \in \mathcal{P}_0(A)$, and $\rho_P = \sum_{\mu \in S(P)} P(\mu)\mu$.

It is shown in [75, 33] that $C(A) = R(A)$, thus reducing the problem of finding a minimax to an optimization problem. For probability measures over infinite spaces this result ($R(A)=C(A)$) was generalized by [39], but the divergence between probability distributions is measured by KL divergence (and not asymptotic average KL divergence), which gives infinite $R(A)$ e.g. already for the class of i.i.d. processes.

However, truncating measures in a class $\mathcal{C}$ to the first n observations, we can use the results about channel capacity to analyse the predictive properties of the class. Moreover, the rates of convergence that can be obtained along these lines are close to optimal. In order to pass from measures minimizing the divergence for each individual n to a process that minimizes the divergence for all n we use the same idea as when constructing the process ρ_c.

Theorem 2.16. *Let $\mathcal{C}$ be a set of measures on $(\mathbf{X}^\infty, \mathcal{F})$, and let $\mathcal{C}^n$ be the class of measures from $\mathcal{C}$ restricted to $\mathbf{X}^n$. There exists a measure ρ_C such that*

$$\frac{1}{n} d_n(\mu, \rho_C) \leq \frac{C(\mathcal{C}^n)}{n} + O\left(\frac{\log n}{n}\right); \tag{2.24}$$

in particular, if $C(\mathcal{C}^n)/n \to 0$, then ρ_C predicts every $\mu \in \mathcal{C}$ in expected average KL divergence. Moreover, for any measure ρ_C and every $\varepsilon > 0$ there exists

$\mu\in\mathcal{C}$ *such that*

$$\frac{1}{n}d_n(\mu,\rho_C)\geq\frac{C(\mathcal{C}^n)}{n}-\varepsilon.$$

Proof. As shown in [33], for each n there exists a sequence ν_k^n, $k\in\mathbb{N}$ of measures on $\mathbf{X}^n$ such that

$$\limsup_{k\to\infty}\sup_{\mu\in\mathcal{C}^n}d_n(\mu,\nu_k^n)\to C(\mathcal{C}^n).$$

For each $n\in\mathbb{N}$ find an index k_n such that

$$|\sup_{\mu\in\mathcal{C}^n}d_n(\mu,\nu_{k_n}^n)-C(\mathcal{C}^n)|\leq 1.$$

Define the measure ρ_n as follows. On the first n symbols it coincides with $\nu_{k_n}^n$ and $\rho_n(x_m=0)=1$ for $m>n$. Finally, set $\rho_C=\sum_{n=1}^\infty w_n\rho_n$, where $w_k=\frac{w}{n^2}$,$w=6/\pi^2$. We have to show that $\lim_{n\to\infty}\frac{1}{n}d_n(\mu,\rho_C)=0$ for every $\mu\in\mathcal{C}$. Indeed, similarly to (2.20), we have

$$\begin{aligned}\frac{1}{n}d_n(\mu,\rho_C)&=\frac{1}{n}\mathbb{E}_\mu\log\frac{\mu(x_{1..n})}{\rho_C(x_{1..n})}\\&\leq\frac{\log w_k^{-1}}{n}+\frac{1}{n}\mathbb{E}_\mu\log\frac{\mu(x_{1..n})}{\rho_n(x_{1..n})}\leq\frac{\log w+2\log n}{n}+\frac{1}{n}d_n(\mu,\rho_n)\\&\leq o(1)+\frac{C(\mathcal{C}^n)}{n}.\quad(2.25)\end{aligned}$$

The second statement follows from the fact [75, 33] that $C(\mathcal{C}^n)=R(\mathcal{C}^n)$ (cf. (2.22)). □

Thus, if the channel capacity $C(\mathcal{C}^n)$ grows sublinearly, a predictor can be constructed for the class of processes $\mathcal{C}$. In this case the problem of constructing the predictor is reduced to finding the channel capacities for different n and finding the corresponding measures on which they are attained or approached.

Examples. For the class of all Bernoulli i.i.d. processes, the channel capacity $C(\mathcal{B}^n)$ is known to be $O(\log n)$ [54]. For the family of all stationary

processes it is $O(n)$, so that the conditions of Theorem 2.16 are satisfied for the former but not for the latter.

We also remark that the requirement of a sublinear channel capacity cannot be relaxed, in the sense that a linear channel capacity is not sufficient for prediction, since it is the maximal possible capacity for a set of measures on $\mathbf{X}^n$, achieved, for example, on the set of all measures, or on the set of all deterministic sequences.

2.4 Conditions under which one measure is a predictor for another [R8]

In this section we address the following question: what are the conditions under which a measure ρ is a good predictor for a measure μ? As it was mentioned, for prediction in total variation distance, this relationship is described by the relation of absolute continuity (see Theorem 2.2). Here we will attempt to establish similar conditions for other measures of predictive quality, including, but not limited to, expected average KL divergence.

We start with the following observation. For a Bayesian mixture ξ of a countable class of measures ν_i, $i \in \mathbb{N}$, we have $\xi(A) \geq w_i \nu_i(A)$ for any i and any measurable set A, where w_i is a constant. This condition is stronger than the assumption of absolute continuity and is sufficient for prediction in a very strong sense. Since we are willing to be satisfied with prediction in a weaker sense (e.g. convergence of conditional probabilities), let us make a weaker assumption: Say that *a measure ρ dominates a measure μ with coefficients $c_n > 0$* if

$$\rho(x_1,...,x_n) \geq c_n \mu(x_1,...,x_n) \tag{2.26}$$

for all $x_1,...,x_n$.

The first question we consider in this section is, under what conditions on c_n does (2.26) imply that ρ predicts μ? Observe that if $\rho(x_1,...,x_n) > 0$ for any $x_1,...,x_n$ then any measure μ is *locally* absolutely continuous with respect

to ρ (that is, the measure μ restricted to the first n trials $\mu|_{\mathbf{X}^n}$ is absolutely continuous w.r.t. $\rho|_{\mathbf{X}^n}$ for each n), and moreover, for any measure μ some constants c_n can be found that satisfy (2.26). For example, if ρ is Bernoulli i.i.d. measure with parameter $\frac{1}{2}$ and μ is any other measure, then (2.26) is (trivially) satisfied with $c_n = 2^{-n}$. Thus we know that if $c_n \equiv c$ then ρ predicts μ in a very strong sense, whereas exponentially decreasing c_n are not enough for prediction. Perhaps somewhat surprisingly, we will show that dominance with any subexponentially decreasing coefficients is sufficient for prediction in expected average KL divergence. Dominance with any polynomially decreasing coefficients, and also with coefficients decreasing (for example) as $c_n = \exp(-\sqrt{n}/\log n)$, is sufficient for (almost sure) prediction on average (i.e. in Cesaro sense). However, for prediction on every step we have a negative result: for any dominance coefficients that go to zero there exists a pair of measures ρ and μ which satisfy (2.26) but ρ does not predict μ in the sense of almost sure convergence of probabilities. Thus the situation is similar to that for predicting any stationary measure: prediction is possible in the average but not on every step.

Note also that for Laplace's measure ρ_L it can be shown that ρ_L dominates any i.i.d. measure μ with linearly decreasing coefficients $c_n = \frac{1}{n+1}$; a generalization of ρ_L for predicting all measures with memory k (for a given k) dominates them with polynomially decreasing coefficients. Thus dominance with decreasing coefficients generalizes (in a sense) predicting countable classes of measures (where we have dominance with a constant), absolute continuity (via local absolute continuity), and predicting i.i.d. and finite-memory measures.

Another way to look for generalizations is as follows. The Bayes mixture ξ, being a sum of countably many measures (predictors), possesses some of their predicting properties. In general, which predictive properties are preserved under summation? In particular, if we have two predictors ρ_1 and ρ_2 for two classes of measures, we are interested in the question whether $\frac{1}{2}(\rho_1 + \rho_2)$ is a predictor for the union of the two classes. An answer to this

question would improve our understanding of how far a class of measures for which a predicting measure exists can be extended without losing this property.

Thus, the second question we consider in this section is the following: suppose that a measure ρ predicts μ (in some weak sense), and let χ be some other probability measure (e.g. a predictor for a different class of measures). Does the measure $\rho' = \frac{1}{2}(\rho+\chi)$ still predict μ? That is, we ask to which prediction quality criteria does the idea of taking a Bayesian sum generalize. Absolute continuity is preserved under summation along with its (strong) prediction ability. It was mentioned in [80] that prediction in the (weak) sense of convergence of expected averages of conditional probabilities is preserved under summation. Here we find that several stronger notions of prediction are not preserved under summation.

Thus we address the following two questions. Is dominance with decreasing coefficients sufficient for prediction in some sense, under some conditions on the coefficients (Section 2.4.2)? And, if a measure ρ predicts a measure μ in some sense, does the measure $\frac{1}{2}(\rho+\chi)$ also predict μ in the same sense, where χ is an arbitrary measure (Section 2.4.3)? Considering different criteria of prediction (a.s. convergence of conditional probabilities, a.s. convergence of averages, etc.) in the above two questions we obtain not two but many different questions, for some of which we find positive answers and for some negative, yet some are left open.

The rest of this section is organized as follows. Section 2.4.1 introduces the measures of divergence of probability measures that we will consider. Section 2.4.2 addresses the question of whether dominance with decreasing coefficients is sufficient for prediction, while in Section 2.4.3 we consider the problem of summing a predictor with an arbitrary measure.

2.4.1 Measuring performance of prediction

In addition to the measures of performance of prediction used in the previous sections (expected average KL divergence and total variation), here we introduce several more.

For two measures μ and ρ define the following measures of divergence.

(δ) Kullback-Leibler (KL) divergence
$$\delta_n(\mu,\rho|x_{<n}) = \sum_{x\in\mathbf{X}} \mu(x_n = x|x_{<n}) \log \frac{\mu(x_n = x|x_{<n})}{\rho(x_n = x|x_{<n})},$$

($\bar{d}$) average KL divergence $\bar{d}_n(\mu,\rho|x_{1..n}) = \frac{1}{n} d_n(\mu,\rho) = \frac{1}{n}\sum_{t=1}^{n} \delta_t(\mu,\rho|x_{<t})$,

(a) absolute distance $a_n(\mu,\rho|x_{<n}) = \sum_{x\in\mathbf{X}} |\mu(x_n = x|x_{<n}) - \rho(x_n = x|x_{<n})|$,

($\bar{a}$) average absolute distance $\bar{a}_n(\mu,\rho|x_{1..n}) = \frac{1}{n}\sum_{t=1}^{n} a_t(\mu,\rho|x_{<t})$.

Definition 2.17. *We say that ρ predicts μ*

(d) *in (non-averaged) KL divergence if $\delta_n(\mu,\rho|x_{<n}) \to 0$ μ-a.s. as $t \to \infty$,*

($\bar{d}$) *in (time-average) average KL divergence if $\bar{d}_n(\mu,\rho|x_{1..n}) \to 0$ μ-a.s.,*

($\mathbb{E}\bar{d}$) *in expected average KL divergence if $\mathbb{E}_\mu \bar{d}_n(\mu,\rho|x_{1..n}) \to 0$,*

(a) *in absolute distance if $a_n(\mu,\rho|x_{<n}) \to 0$ μ-a.s.,*

($\bar{a}$) *in average absolute distance if $\bar{a}_n(\mu,\rho|x_{1..n}) \to 0$ μ-a.s.,*

($\mathbb{E}\bar{a}$) *in expected average absolute distance if $\mathbb{E}_\mu \bar{a}_n(\mu,\rho|x_{1..n}) \to 0$.*

The argument $x_{1..n}$ will be often left implicit in our notation. Recall (definition 2.1) measure ρ converges to a measure μ in *total variation* (tv)

if $\sup_{A\subset\sigma(\bigcup_{t=n}^{\infty}\mathbf{X}^t)}|\mu(A|x_{<n})-\rho(A|x_{<n})|\to 0$ μ-almost surely. The following implications hold (and are complete and strict):

$$\begin{array}{ccccccc} & & \delta & \Rightarrow & \bar{d} & & \mathbb{E}\bar{d} \\ & & \Downarrow & & \Downarrow & & \Downarrow \\ tv & \Rightarrow & a & \Rightarrow & \bar{a} & \Rightarrow & \mathbb{E}\bar{a} \end{array} \tag{2.27}$$

to be understood as e.g.: if $\bar{d}_n\to 0$ a.s. then $\bar{a}_n\to 0$ a.s, or, if $\mathbb{E}\bar{d}_n\to 0$ then $\mathbb{E}\bar{a}_n\to 0$. The horizontal implications $\Rightarrow$ follow immediately from the definitions, and the $\Downarrow$ follow from the following Lemma:

Lemma 2.18. *For all measures ρ and μ and sequences $x_{1:\infty}$ we have: $a_t^2\leq 2\delta_t$ and $\bar{a}_n^2\leq 2\bar{d}_n$ and $(\mathbb{E}\bar{a}_n)^2\leq 2\mathbb{E}\bar{d}_n$.*

Proof. Pinsker's inequality [41, Lem.3.11*a*] implies $a_t^2\leq 2\delta_t$. Using this and Jensen's inequality for the average $\frac{1}{n}\sum_{t=1}^n[...]$ we get

$$2\bar{d}_n \;=\; \frac{1}{n}\sum_{t=1}^n 2\delta_t \;\geq\; \frac{1}{n}\sum_{t=1}^n a_t^2 \;\geq\; \left(\frac{1}{n}\sum_{t=1}^n a_t\right)^2 \;=\; \bar{a}_n^2 \tag{2.28}$$

Using this and Jensen's inequality for the expectation $\mathbb{E}$ we get $2\mathbb{E}\bar{d}_n\geq\mathbb{E}\bar{a}_n^2\geq(\mathbb{E}\bar{a}_n)^2$. □

2.4.2 Dominance with decreasing coefficients

First we consider the question whether property (2.26) is sufficient for prediction.

Definition 2.19. *We say that a measure ρ dominates a measure μ with coefficients $c_n>0$ iff*

$$\rho(x_{1:n}) \;\geq\; c_n\mu(x_{1:n}). \tag{2.29}$$

for all $x_{1:n}$.

Suppose that ρ dominates μ with decreasing coefficients c_n. Does ρ predict μ in (expected, expected average) KL divergence (absolute distance)? First let us give an example.

Proposition 2.20. *Let ρ_L be the Laplace measure, given by $\rho_L(x_{n+1}=a|x_{1..n})=\frac{k+1}{n+|\mathbf{X}|}$ for any $a\in\mathbf{X}$ and any $x_{1..n}\in\mathbf{X}^n$, where k is the number of occurrences of a in $x_{1..n}$ (this is also well defined for $n=0$). Then*

$$\rho_L(x_{1..n}) \geq \frac{n!}{(n+|\mathbf{X}|-1)!}\,\mu(x_{1..n}) \tag{2.30}$$

for any measure μ which generates independently and identically distributed symbols. The equality is attained for some choices of μ.

Proof. We will only give the proof for $\mathbf{X}=\{0,1\}$, the general case is analogous. To calculate $\rho_L(x_{1..n})$ observe that it only depends on the number of 0s and 1s in $x_{1..n}$ and not on their order. Thus we compute $\rho_L(x_{1..n})=\frac{k!(n-k)!}{(n+1)!}$ where k is the number of 1s. For any measure μ such that $\mu(x_n=1)=p$ for some $p\in[0,1]$ independently for all n, and for Laplace measure ρ_L we have

$$\begin{aligned}\frac{\mu(x_{1..n})}{\rho_L(x_{1..n})} &= \frac{(n+1)!}{k!(n-k)!}p^k(1-p)^{n-k} = (n+1)\binom{n}{k}p^k(1-p)^{n-k}\\ &\leq (n+1)\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} = n+1,\end{aligned}$$

for any n-letter word $x_1,...,x_n$ where k is the number of 1s in it. The equality in the bound is attained when $p=1$, so that $k=n$, $\mu(x_{1..n})=1$, and $\rho_L(x_{1..n})=\frac{1}{n+1}$. □

Thus for Laplace's measure ρ_L and binary $\mathbf{X}$ we have $c_n=\mathcal{O}(\frac{1}{n})$. As mentioned in the introduction, in general, exponentially decreasing coefficients c_n are not sufficient for prediction, since (2.26) is satisfied with ρ being a Bernoulli i.i.d. measure and μ any other measure. On the other hand, the following proposition shows that in a weak sense of convergence in expected

average KL divergence (or absolute distance) the property (2.26) with subexponentially decreasing c_n is sufficient. We also remind that if c_n are bounded from below then prediction in the strong sense of total variation is possible.

Theorem 2.21. *Let μ and ρ be two measures on $\mathbf{X}^\infty$ and suppose that $\rho(x_{1..n}) \geq c_n \mu(x_{1..n})$ for any $x_{1..n}$, where c_n are positive constants satisfying $\frac{1}{n}\log c_n^{-1} \to 0$. Then ρ predicts μ in expected average KL divergence $\mathbb{E}_\mu \bar{d}_n(\mu,\rho) \to 0$ and in expected average absolute distance $\mathbb{E}_\mu \bar{a}_n(\mu,\rho) \to 0$.*

Proof. For convergence in average expected KL divergence, using (2.2) we derive

$$\mathbb{E}_\mu \bar{d}_n(\mu,\rho) = \frac{1}{n}\mathbb{E}\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})} \leq \frac{1}{n}\log c_n^{-1} \to 0.$$

The statement for expected average distance follows from this and Lemma 2.18. □

With a stronger condition on c_n prediction in average KL divergence can be established.

Theorem 2.22. *Let μ and ρ be two measures on $\mathbf{X}^\infty$ and suppose that $\rho(x_{1..n}) \geq c_n \mu(x_{1..n})$ for every $x_{1..n}$, where c_n are positive constants satisfying*

$$\sum_{n=1}^{\infty} \frac{(\log c_n^{-1})^2}{n^2} \;<\; \infty. \tag{2.31}$$

Then ρ predicts μ in average KL divergence $\bar{d}_n(\mu,\rho) \to 0$ μ-a.s. and in average absolute distance $\bar{a}_n(\mu,\rho) \to 0$ μ-a.s.

In particular, the condition (2.31) on the coefficients is satisfied for polynomially decreasing coefficients, or for $c_n = \exp(-\sqrt{n}/\log n)$.

The proof is deferred to Section 2.6.2.

However, no form of dominance with decreasing coefficients is sufficient for prediction in absolute distance or KL divergence, as the following negative result states.

Proposition 2.23. *For each sequence of positive numbers c_n that goes to 0 there exist measures μ and ρ and a number $\varepsilon>0$ such that $\rho(x_{1..n})\geq c_n\mu(x_{1..n})$ for all $x_{1..n}$, yet $a_n(\mu,\rho|x_{1..n})>\varepsilon$ and $\delta_n(\mu,\rho|x_{1..n})>\varepsilon$ infinitely often μ-a.s.*

Proof. Let μ be concentrated on the sequence 11111... (that is $\mu(x_n=1)=1$ for all n), and let $\rho(x_n=1)=1$ for all n except for a subsequence of steps $n=n_k$, $k\in\mathbb{N}$ on which $\rho(x_{n_k}=1)=1/2$ independently of each other. It is easy to see that choosing n_k sparse enough we can make $\rho(1_1...1_n)$ decrease to 0 arbitrary slowly; yet $|\mu(x_{n_k})-\rho(x_{n_k})|=1/2$ for all k. □

Thus for the first question — whether dominance with some coefficients decreasing to zero is sufficient for prediction, we have the following table of questions and answers, where, in fact, positive answers for a_n are implied by positive answers for δ_n and vice versa for the negative answers:

$\mathbb{E}\bar{d}_n$	$\bar{d}_n$	δ_n	$\mathbb{E}\bar{a}_n$	$\bar{a}_n$	a_n
+	+	−	+	+	−

However, if we take into account the conditions on the coefficients, we see some open problems left, and different answers for $\bar{d}_n$ and $\bar{a}_n$ may be obtained. Following is the table of conditions on dominance coefficients and answers to the questions whether these conditions are sufficient for prediction (coefficients bounded from below are included for the sake of completeness).

	$\mathbb{E}\bar{d}_n$	$\bar{d}_n$	δ_n	$\mathbb{E}\bar{a}_n$	$\bar{a}_n$	a_n
$\log c_n^{-1}=o(n)$	+	?	−	+	?	−
$\sum_{n=1}^{\infty}\frac{\log c_n^{-1}}{n^2}<\infty$	+	+	−	+	+	−
$c_n\geq c>0$	+	+	+	+	+	+

We know from Proposition 2.23 that the condition $c_n\geq c>0$ for convergence in δ_n can not be improved; thus the open problem left is to find whether $\log c_n^{-1}=o(n)$ is sufficient for prediction in $\bar{d}_n$ or at least in $\bar{a}_n$.

Another open problem is to find out whether any conditions on dominance coefficients are necessary for prediction; so far we only have some

sufficient conditions. On the one hand, the obtained results suggest that some form of dominance with decreasing coefficients may be necessary for prediction, at least in the sense of convergence of averages. On the other hand, the condition (2.26) is uniform over all sequences which probably is not necessary for prediction. As for prediction in the sense of almost sure convergence, perhaps more subtle behavior of the ratio $\frac{\mu(x_{1..n})}{\rho(x_{1..n})}$ should be analyzed, since dominance with decreasing coefficients is not sufficient for prediction in this sense.

2.4.3 Preservation of the predictive ability under summation with an arbitrary measure

Now we turn to the question whether, given a measure ρ that predicts a measure μ in some sense, the "contaminated" measure $(1-\varepsilon)\rho+\varepsilon\chi$ for some $0<\varepsilon<1$ also predicts μ in the same sense, where χ is an arbitrary probability measure. Since most considerations are independent of the choice of ε, in particular the results in this section, we set $\varepsilon=\frac{1}{2}$ for simplicity. We define

Definition 2.24. *By "ρ contaminated with χ" we mean $\rho':=\frac{1}{2}(\rho+\chi)$, where ρ and χ are probability measures.*

Positive results can be obtained for convergence in expected average KL divergence. The statement of the next proposition in a different form was mentioned in [80, 42]. Since the proof is simple we present it here for the sake of completeness; it is based on the same ideas as the proof of Theorem 2.21.

Proposition 2.25. *Let μ and ρ be two measures on $\mathbf{X}^\infty$ and suppose that ρ predicts μ in expected average KL divergence. Then so does the measure $\rho'=\frac{1}{2}(\rho+\chi)$ where χ is any other measure on $\mathbf{X}^\infty$.*

Proof.

$$0 \;\le\; \mathbb{E}\bar{d}_n(\mu,\rho') \;=\; \frac{1}{n}\mathbb{E}\sum_{t=1}^{n}\sum_{x_t\in\mathbf{X}}\mu(x_t|x_{<t})\log\frac{\mu(x_t|x_{<t})}{\rho'(x_t|x_{<t})} \;=\; \frac{1}{n}\mathbb{E}\log\frac{\mu(x_{1..n})}{\rho'(x_{1..n})}$$
$$= \frac{1}{n}\mathbb{E}\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})}\frac{\rho(x_{1..n})}{\rho'(x_{1..n})} \;=\; \mathbb{E}\bar{d}_n(\mu,\rho)+\frac{1}{n}\mathbb{E}\log\frac{\rho(x_{1..n})}{\rho'(x_{1..n})},$$

where the first term tends to 0 by assumption and the second term is bounded from above by $\frac{1}{n}\log 2\to 0$. Since the sum is bounded from below by 0 we obtain the statement of the proposition. □

Next we consider some negative results. An example of measures μ, ρ and χ such that ρ predicts μ in absolute distance (or KL divergence) but $\frac{1}{2}(\rho+\chi)$ does not, can be constructed similarly to the example from [45] (of a measure ρ which is a sum of distributions arbitrarily close to μ yet does not predict it). The idea is to take a measure χ that predicts μ much better than ρ on almost all steps, but on some steps gives grossly wrong probabilities.

Proposition 2.26. *There exist measures μ, ρ and χ such that ρ predicts μ in absolute distance (KL divergence) but $\frac{1}{2}(\rho+\chi)$ does not predict μ in absolute distance (KL divergence).*

Proof. Let μ be concentrated on the sequence 11111... (that is $\mu(x_n=1)=1$ for any n), and let $\rho(x_n=1)=\frac{n}{n+1}$ with probabilities independent on different trials. Clearly, ρ predicts μ in both absolute distance and KL divergence. Let $\chi(x_n=1)=1$ for all n except on the sequence $n=n_k=2^{2^k}=n_{k-1}^2$, $k\in\mathbb{N}$ on which $\chi(x_{n_k}=1)=n_{k-1}/n_k=2^{-2^{k-1}}$. This implies that $\chi(1_{1..n_k})=2/n_k$ and $\chi(1_{1..n_k-1})=\chi(1_{1..n_{k-1}})=2/n_{k-1}=2/\sqrt{n_k}$. It is now easy to see that $\frac{1}{2}(\rho+\chi)$ does not predict μ, neither in absolute distance nor in KL divergence. Indeed for $n=n_k$ for some k we have

$$\tfrac{1}{2}(\rho+\chi)(x_n=1|1_{<n}) \;=\; \frac{\rho(1_{1..n})+\chi(1_{1..n})}{\rho(1_{<n})+\chi(1_{<n})} \;\le\; \frac{1/(n+1)+2/n}{1/n \;+\; 2/\sqrt{n}} \;\to\; 0.$$

□

For the (expected) average absolute distance the negative result also holds:

Proposition 2.27. *There exist such measures μ, ρ and χ that ρ predicts μ in average absolute distance but $\frac{1}{2}(\rho+\chi)$ does not predict μ in (expected) average absolute distance.*

Proof. Let μ be Bernoulli $1/2$ distribution and let $\rho(x_n=1)=1/2$ for all n (independently of each other) except for some sequence n_k, $k\in\mathbb{N}$ on which $\rho(x_{n_k}=1)=0$. Choose n_k sparse enough for ρ to predict μ in the average absolute distance. Let χ be Bernoulli $1/3$. Observe that χ assigns non-zero probabilities to all finite sequences, whereas μ-a.s. from some n on $\rho(x_{1..n})=0$. Hence $\frac{1}{2}(\rho+\chi)(x_{1..n})=\frac{1}{2}\chi(x_{1..n})$ and so $\frac{1}{2}(\rho+\chi)$ does not predict μ. □

Thus for the question of whether predictive ability is preserved when an arbitrary measure is added to the predictive measure, we have the following table of answers.

$\mathbb{E}\bar{d}_n$	$\bar{d}_n$	δ_n	$\mathbb{E}\bar{a}_n$	$\bar{a}_n$	a_n
$+$	?	$-$	$-$	$-$	$-$

As it can be seen, there remains one open question: whether this property is preserved under almost sure convergence of the average KL divergence.

It can be inferred from the example in Proposition 2.26 that contaminating a predicting measure ρ with a measure χ spoils ρ if χ is better than ρ on almost every step. It thus can be conjectured that adding a measure can only spoil a predictor on sparse steps, not affecting the average.

2.5 Nonrealizable version of the sequence prediction problem [R1]

In this section we generalize our approach to sequence prediction even further. Instead of assuming that the measure μ generating the data belongs

to some set $\mathcal{C}$ of measures, we would like to assume that the measure μ is completely arbitrary. As we know, in this case we cannot hope to get the error converge to zero. However, we can try to make the error as small as the error of any predictor from a given set $\mathcal{C}$.

Recall that a predictor ρ is required to give conditional probabilities $\rho(x_{n+1}=a|x_1,...,x_n)$ for all possible histories $x_1,...,x_n$. Therefore, it defines itself a probability measure on the space Ω of one-way infinite sequences. In other words, a probability measure on Ω can be considered both as a data-generating mechanism and as a predictor.

Thus, given a set $\mathcal{C}$ of probability measures on Ω, one can ask two kinds of questions about $\mathcal{C}$. First, does there exist a predictor ρ, whose forecast probabilities converge (in a certain sense) to the μ-conditional probabilities, if an arbitrary $\mu\in\mathcal{C}$ is chosen to generate the data? Here we assume that the "true" measure that generates the data belongs to the set $\mathcal{C}$ of interest, and would like to construct a predictor that predicts all measures in $\mathcal{C}$. The second type of questions is as follows: does there exist a predictor that predicts at least as well as any predictor $\rho\in\mathcal{C}$, if the measure that generates the data comes possibly from outside of $\mathcal{C}$? Thus, here we consider elements of $\mathcal{C}$ as predictors, and we would like to combine their predictive properties, if this is possible. Note that in this setting the two questions above concern the same object: a set $\mathcal{C}$ of probability measures on Ω.

Each of these two questions, the realizable and the non-realizable one, have enjoyed much attention in the literature; the setting for the non-realizable case is usually slightly different, which is probably why it has not (to the best of the author's knowledge) been studied as another facet of the realizable case. The realizable case has been considered in detail in the previous sections (Section 2.2–2.4).

The non-realizable case is usually studied in a slightly different, non-probabilistic, setting. We refer to [18] for a comprehensive overview. It is usually assumed that the observed sequence of outcomes is an arbitrary (deterministic) sequence; it is required not to give conditional probabilities,

but just deterministic guesses (although these guesses can be selected using randomisation). Predictions result in a certain loss, which is required to be small as compared to the loss of a given set of reference predictors (experts) $\mathcal{C}$. The losses of the experts and the predictor are observed after each round. In this approach, it is mostly assumed that the set $\mathcal{C}$ is finite or countable. The main difference with the formulation considered in this section is that we require a predictor to give probabilities, and thus the loss is with respect to something never observed (probabilities, not outcomes). The loss itself is not completely observable in our setting. In this sense our non-realizable version of the problem is more difficult. Assuming that the data generating mechanism is probabilistic, even if it is completely unknown, makes sense in such problems as, for example, game playing, or market analysis. In these cases one may wish to assign smaller loss to those models or experts who give probabilities closer to the correct ones (which are never observed), even though different probability forecasts can often result in the same action. Aiming at predicting probabilities of outcomes also allows us to abstract from the actual use of the predictions (for example, making bets) and thus from considering losses in a general form; instead, we can concentrate on those forms of loss that are more convenient for the analysis. In this latter respect, the problems we consider are easier than those considered in prediction with expert advice. (However, in principle, nothing restricts us to considering the simple losses that we chose; they are just a convenient choice.) Noteworthy, the probabilistic approach also makes the machinery of probability theory applicable, hopefully making the problem easier. Another way to look at the difference between the non-realizable problems of this manuscript and prediction with expert advice is as follows: the latter is prequential (in the sense of [23]), whereas the former is not.

Let us further break the non-realizable case into two problems. The first one is as follows. Given a set $\mathcal{C}$ of predictors, we want to find a predictor whose prediction error converges to zero if there is at least one predictor in $\mathcal{C}$ whose prediction error converges to zero; we call this problem simply the

"non-realizable" case, or Problem 2 (leaving the name "Problem 1" to the realizable case). The second non-realizable problem is the "fully agnostic" problem: it is to make the prediction error asymptotically as small as that of the best (for the given process measure generating the data) predictor in $\mathcal{C}$ (we call this Problem 3). Thus, we now have three problems about a set of process measures $\mathcal{C}$ to address.

In this section we show that if the quality of prediction is measured in total variation, then all the three problems coincide: any solution to any one of them is a solution to the other two. For the case of expected average KL divergence, all the three problems are different: the realizable case is strictly easier than non-realizable (Problem 2), which is, in turn, strictly easier than the fully agnostic case (Problem 3). We then analyse which results concerning prediction in total variation can be transferred to which of the problems concerning prediction in average KL divergence. We will extend the result of Section 2.2 about the existence of Bayesian predictor from Problem 1 to the (non-realizable) case of Problem 2, for prediction in expected average KL divergence. We do not have an analogous result for Problem 3 (and, in fact, conjecture that the opposite statement holds true). However, for the fully agnostic case of Problem 3, we show that separability with respect to a certain topology given by KL divergence is a sufficient (though not a necessary) condition for the existence of a predictor. This is used to demonstrate that there is a solution to this problem for the set of all finite-memory process measures, complementing similar results obtained earlier in different settings. On the other hand, we show that there is no solution to this problem for the set of all stationary process measures, in contrast to a result of B. [78] that gives a solution to the realizable case of this problem (that is, a predictor whose expected average KL error goes to zero if any stationary process is chosen to generate the data). Finally, we also consider a modified version of Problem 3, in which the performance of predictors is only compared on individual sequences. For this problem, we obtain, using a result from [77], a characterisation of those sets $\mathcal{C}$ for which

a solution exists in terms of the Hausdorff dimension.

2.5.1 Sequence prediction problems

For the two notions of predictive quality introduced, we can now state formally the sequence prediction problems.

Problem 1(realizable case). Given a set of probability measures $\mathcal{C}$, find a measure ρ such that ρ predicts in total variation (expected average KL divergence) every $\mu\in\mathcal{C}$, if such a ρ exists.

This is the problem considered in Sections 2.2 (restated here to ease the comparison). Problem 1 is about finding a predictor for the case when the process generating the data is known to belong to a given class $\mathcal{C}$. That is, the set $\mathcal{C}$ here is a set of measures that generate the data. Next let us formulate the questions about $\mathcal{C}$ as a set of predictors.

Problem 2 (non-realizable case). Given a set of process measures (predictors) $\mathcal{C}$, find a process measure ρ such that ρ predicts in total variation (in expected average KL divergence) every measure $\nu\in\mathcal{P}$ such that there is $\mu\in\mathcal{C}$ which predicts (in the same sense) ν.

While Problem 2 is already quite general, it does not yet address what can be called the fully agnostic case: if nothing at all is known about the process ν generating the data, it means that there may be no $\mu\in\mathcal{C}$ such that μ predicts ν, and then, even if we have a solution ρ to the Problem 2, we still do not know what the performance of ρ is going to be on the data generated by ν, compared to the performance of the predictors from $\mathcal{C}$. To address this fully agnostic case we have to introduce the notion of loss.

Definition 2.28. *Introduce the almost sure total variation loss of ρ with respect to μ*

$$l_{tv}(\mu,\rho):=\inf\{\alpha\in[0,1]:\limsup_{n\to\infty}v(\mu,\rho,x_{1..n})\leq\alpha\ \mu\text{–}a.s.\},$$

and the asymptotic KL loss

$$l_{KL}(\nu,\rho) := \limsup_{n\to\infty} \frac{1}{n} d_n(\nu,\rho).$$

We can now formulate the fully agnostic version of the sequence prediction problem.

Problem 3. Given a set of process measures (predictors) $\mathcal{C}$, find a process measure ρ such that ρ predicts at least as well as any μ in $\mathcal{C}$, if any process measure $\nu \in \mathcal{P}$ is chosen to generate the data:

$$l(\nu,\rho) - l(\nu,\mu) \leq 0 \tag{2.32}$$

for every $\nu \in \mathcal{P}$ and every $\mu \in \mathcal{C}$, where $l(\cdot,\cdot)$ is either $l_{tv}(\cdot,\cdot)$ or $l_{KL}(\cdot,\cdot)$.

The three problems just formulated represent different conceptual approaches to the sequence prediction problem. Let us illustrate the difference by the following **informal example**. Suppose that the set $\mathcal{C}$ is that of all (ergodic, finite-state) Markov chains. Markov chains being a familiar object in probability and statistics, we can easily construct a predictor ρ that predicts every $\mu \in \mathcal{C}$ (for example, in expected average KL divergence, see [54]). That is, if we know that the process μ generating the data is Markovian, we know that our predictor is going to perform well. This is the realizable case of Problem 1. In reality, rarely can we be sure that the Markov assumption holds true for the data at hand. We may believe, however, that it is still a reasonable assumption, in the sense that there is a Markovian model which, for our purposes (for the purposes of prediction), is a good model of the data. Thus we may assume that there is a Markov model (a predictor) that predicts well the process that we observe, and we would like to combine the predictive qualities of all these Markov models. This is the "non-realizable" case of Problem 2. Note that this problem is more difficult than the first one; in particular, a process ν generating the data may be singular with respect to any Markov process, and still be predicted well (in the sense of

expected average KL divergence, for example) by some of them. Still, here we are making some assumptions about the process generating the data, and, if these assumptions are wrong, then we do not know anything about the performance of our predictor. Thus, we may ultimately wish to acknowledge that we do not know anything at all about the data; we still know a lot about Markov processes, and we would like to use this knowledge on our data. If there is anything at all Markovian in it (that is, anything that can be captured by a Markov model), then we would like our predictor to use it. In other words, we want to have a predictor that predicts any process measure whatsoever (at least) as well as any Markov predictor. This is the "fully agnostic" case of Problem 3.

Of course, Markov processes were just mentioned as an example, while in this section we are only concerned with the most general case of arbitrary (uncountable) sets $\mathcal{C}$ of process measures.

The following statement is rather obvious.

Proposition 2.29. *Any solution to Problem 3 is a solution to Problem 2, and any solution to Problem 2 is a solution to Problem 1.*

Despite the conceptual differences in formulations, it may be somewhat unclear whether the three problems are indeed different. It appears that this depends on the measure of predictive quality chosen: for the case of prediction in total variation distance all the three problems coincide, while for the case of prediction in expected average KL divergence they are different.

2.5.2 Characterizations of learnable classes for prediction in total variation

As it was mentioned, a measure μ is absolutely continuous with respect to a measure ρ if and only if ρ predicts μ in total variation distance. This reduces studying at least Problem 1 for total variation distance to studying the relation of absolute continuity. Introduce the notation $\rho \geq_{tv} \mu$ for this relation.

Let us briefly recall some facts we know about $\geq_{tv}$; details can be found, for example, in [70]. Let $[\mathcal{P}]_{tv}$ denote the set of equivalence classes of $\mathcal{P}$ with respect to $\geq_{tv}$, and for $\mu\in\mathcal{P}_{tv}$ denote $[\mu]$ the equivalence class that contains μ. Two elements $\sigma_1,\sigma_2\in[\mathcal{P}]_{tv}$ (or $\sigma_1,\sigma_2\in\mathcal{P}$) are called disjoint (or singular) if there is no $\nu\in[\mathcal{P}]_{tv}$ such that $\sigma_1\geq_{tv}\nu$ and $\sigma_2\geq_{tv}\nu$; in this case we write $\sigma_1\perp_{tv}\sigma_2$. We write $[\mu_1]+[\mu_2]$ for $[\frac{1}{2}(\mu_1+\mu_2)]$. Every pair $\sigma_1,\sigma_2\in[\mathcal{P}]_{tv}$ has a supremum $\sup(\sigma_1,\sigma_2)=\sigma_1+\sigma_2$. Introducing into $[\mathcal{P}]_{tv}$ an extra element 0 such that $\sigma\geq_{tv}0$ for all $\sigma\in[\mathcal{P}]_{tv}$, we can state that for every $\rho,\mu\in[\mathcal{P}]_{tv}$ there exists a unique pair of elements μ_s and μ_a such that $\mu=\mu_a+\mu_s$, $\rho\geq\mu_a$ and $\rho\perp_{tv}\mu_s$. (This is a form of Lebesgue decomposition.) Moreover, $\mu_a=\inf(\rho,\mu)$. Thus, every pair of elements has a supremum and an infimum. Moreover, every bounded set of disjoint elements of $[\mathcal{P}]_{tv}$ is at most countable.

Furthermore, we introduce the (unconditional) total variation distance between process measures.

Definition 2.30 (unconditional total variation distance)**.** *The (unconditional) total variation distance is defined as*

$$v(\mu,\rho):=\sup_{A\in\mathcal{B}}|\mu(A)-\rho(A)|.$$

Known characterizations of those sets $\mathcal{C}$ that are bounded with respect to $\geq_{tv}$ can now be related to our prediction problems 1-3 as follows.

Theorem 2.31. *Let $\mathcal{C}\subset\mathcal{P}$. The following statements about $\mathcal{C}$ are equivalent.*

(i) There exists a solution to Problem 1 in total variation.

(ii) There exists a solution to Problem 2 in total variation.

(iii) There exists a solution to Problem 3 in total variation.

(iv) $\mathcal{C}$ is upper-bounded with respect to $\geq_{tv}$.

(v) *There exists a sequence* $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$ *such that for some (equivalently, for every) sequence of weights* $w_k \in (0,1]$, $k \in \mathbb{N}$ *such that* $\sum_{k\in\mathbb{N}} w_k = 1$, *the measure* $\nu = \sum_{k\in\mathbb{N}} w_k \mu_k$ *satisfies* $\nu \geq_{tv} \mu$ *for every* $\mu \in \mathcal{C}$.

(vi) $\mathcal{C}$ *is separable with respect to the total variation distance.*

(vii) *Let* $\mathcal{C}^+ := \{\mu \in \mathcal{P} : \exists \rho \in \mathcal{C}\ \rho \geq_{tv} \mu\}$. *Every disjoint (with respect to* $\geq_{tv}$*) subset of* $\mathcal{C}^+$ *is at most countable.*

Moreover, every solution to any of the Problems 1-3 is a solution to the other two, as is any upper bound for $\mathcal{C}$. *The sequence* μ_k *in the statement (v) can be taken to be any dense (in the total variation distance) countable subset of* $\mathcal{C}$ *(cf. (vi)), or any maximal disjoint (with respect to* $\geq_{tv}$*) subset of* $\mathcal{C}^+$ *of statement (vii), in which every measure that is not in* $\mathcal{C}$ *is replaced by any measure from* $\mathcal{C}$ *that dominates it.*

Proof. The implications $(i) \Leftarrow (ii) \Leftarrow (iii)$ are obvious (cf. Proposition 2.29). The equivalence $(iv) \Leftrightarrow (i)$ is a reformulation of the result of Theorem 2.2. $(i) \Rightarrow (ii)$ follows from the equivalence $(i) \Leftrightarrow (iv)$ and the transitivity of $\geq_{tv}$; $(i) \Rightarrow (iii)$ follows from the transitivity of $\geq_{tv}$ and from Lemma 2.32 below: indeed, from Lemma 2.32 we have $l_{tv}(\nu,\mu) = 0$ if $\mu \geq_{tv} \nu$ and $l_{tv}(\nu,\mu) = 1$ otherwise. From this and the transitivity of $\geq_{tv}$ it follows that if $\rho \geq_{tv} \mu$ then also $l_{tv}(\nu,\rho) \leq l_{tv}(\nu,\mu)$ for all $\nu \in \mathcal{P}$. The equivalence of (v), (vi), and (i) was established in Theorems 2.6 and 2.9. The equivalence of (iv) and (vii) was proven in [70]. The concluding statements of the theorem are easy to demonstrate from the results cited above. □

The following lemma is an easy consequence of [11].

Lemma 2.32. *Let* μ,ρ *be two process measures. Then* $v(\mu,\rho,x_{1..n})$ *converges to either 0 or 1 with* μ*-probability 1.*

Proof. Assume that μ is not absolutely continuous with respect to ρ (the other case is covered by [11]). By Lebesgue decomposition theorem, the

measure μ admits a representation $\mu=\alpha\mu_a+(1-\alpha)\mu_s$ where $\alpha\in[0,1]$ and the measures μ_a and μ_s are such that μ_a is absolutely continuous with respect to ρ and μ_s is singular with respect to ρ. Let W be such a set that $\mu_a(W)=\rho(W)=1$ and $\mu_s(W)=0$. Note that we can take $\mu_a=\mu|_W$ and $\mu_s=\mu|_{\mathbf{X}^\infty\setminus W}$. From [11] we have $v(\mu_a,\rho,x_{1..n})\to 0$ μ_a-a.s., as well as $v(\mu_a,\mu,x_{1..n})\to 0$ μ_a-a.s. and $v(\mu_s,\mu,x_{1..n})\to 0$ μ_s-a.s. Moreover, $v(\mu_s,\rho,x_{1..n})\geq|\mu_s(W|x_{1..n})-\rho(W|x_{1..n})|=1$ so that $v(\mu_s,\rho,x_{1..n})\to 1$ μ_s-a.s. Furthermore,

$$v(\mu,\rho,x_{1..n})\leq v(\mu,\mu_a,x_{1..n})+v(\mu_a,\rho,x_{1..n})=I$$

and

$$v(\mu,\rho,x_{1..n})\geq -v(\mu,\mu_s,x_{1..n})+v(\mu_s,\rho,x_{1..n})=II.$$

We have $I\to 0$ μ_a-a.s. and hence $\mu|_W$-a.s., as well as $II\to 1$ μ_s-a.s. and hence $\mu|_{\mathbf{X}^\infty\setminus W}$-a.s. Thus,

$$\begin{aligned}&\mu(v(\mu,\rho,x_{1..n})\to 0 \text{ or } 1)\\&\leq\mu(W)\mu|_W(I\to 0)+\mu(\mathbf{X}^\infty\setminus W)\mu|_{\mathbf{X}^\infty\setminus W}(II\to 1)=\mu(W)+\mu(\mathbf{X}^\infty\setminus W)=1,\end{aligned}$$

which concludes the proof. □

Remark. Using Lemma 2.32 we can also define *expected* (rather than almost sure) total variation loss of ρ with respect to μ, as the μ-probability that $v(\mu,\rho)$ converges to 1:

$$l'_{tv}(\mu,\rho):=\mu\{x_1,x_2,\dots\in\mathbf{X}^\infty:v(\mu,\rho,x_{1..n})\to 1\}.$$

Then Problem 3 can be reformulated for this notion of loss. However, it is easy to see that for this reformulation Theorem 2.31 holds true as well.

Thus, we can see that, for the case of prediction in total variation, all the sequence prediction problems formulated reduce to studying the relation of absolute continuity for process measures and those families of measures that are absolutely continuous (have a density) with respect to some measure (a

predictor).

2.5.3 Characterizations of learnable classes for prediction in expected average KL divergence

First of all, we have to observe that for prediction in KL divergence Problems 1, 2, and 3 are different, as the following theorem shows. While the examples provided in the proof are artificial, there is a very important example illustrating the difference between Problem 1 and Problem 3 for expected average KL divergence: the set $\mathcal{S}$ of all stationary processes, given in Theorem 2.39 in the end of this section.

Theorem 2.33. *For the case of prediction in expected average KL divergence, Problems 1, 2 and 3 are different: there exists a set $\mathcal{C}_1 \subset \mathcal{P}$ for which there is a solution to Problem 1 but there is no solution to Problem 2, and there is a set $\mathcal{C}_2 \subset \mathcal{P}$ for which there is a solution to Problem 2 but there is no solution to Problem 3.*

Proof. We have to provide two examples. Fix the binary alphabet $\mathbf{X} = \{0,1\}$. For each deterministic sequence $t = t_1,t_2,\dots \in \mathbf{X}^\infty$ construct the process measure γ_t as follows: $\gamma_t(x_n = t_n | t_{1..n-1}) := 1 - \frac{1}{n+1}$ and for $x_{1..n-1} \neq t_{1..n-1}$ let $\gamma_t(x_n = 0 | x_{1..n-1}) = 1/2$, for all $n \in \mathbb{N}$. That is, γ_t is Bernoulli i.i.d. $1/2$ process measure strongly biased towards a specific deterministic sequence, t. Let also $\gamma(x_{1..n}) = 2^{-n}$ for all $x_{1..n} \in \mathbf{X}^n$, $n \in \mathbb{N}$ (the Bernoulli i.i.d. $1/2$). For the set $\mathcal{C}_1 := \{\gamma_t : t \in \mathbf{X}^\infty\}$ we have a solution to Problem 1: indeed, $d_n(\gamma_t,\gamma) \leq 1 = o(n)$. However, there is no solution to Problem 2. Indeed, for each $t \in \mathcal{D}$ we have $d_n(t,\gamma_t) = \log n = o(n)$ (that is, for every deterministic measure there is an element of $\mathcal{C}_1$ which predicts it), while by Lemma 2.5 for every $\rho \in \mathcal{P}$ there exists $t \in \mathcal{D}$ such that $d_n(t,\rho) \geq n$ for all $n \in \mathbb{N}$ (that is, there is no predictor which predicts every measure that is predicted by at least one element of $\mathcal{C}_1$).

The second example is similar. For each deterministic sequence $t = t_1,t_2,\dots \in \mathcal{D}$ construct the process measure γ_t as follows: $\gamma'_t(x_n = t_n | t_{1..n-1}) :=$

$2/3$ and for $x_{1..n-1}\neq t_{1..n-1}$ let $\gamma'_t(x_n=0|x_{1..n-1})=1/2$, for all $n\in\mathbb{N}$. It is easy to see that γ is a solution to Problem 2 for the set $\mathcal{C}_2:=\{\gamma'_t:t\in\mathbf{X}^\infty\}$. Indeed, if $\nu\in\mathcal{P}$ is such that $d_n(\nu,\gamma')=o(n)$ then we must have $\nu(t_{1..n})=o(1)$. From this and the fact that γ and γ' coincide (up to $O(1)$) on all other sequences we conclude $d_n(\nu,\gamma)=o(n)$. However, there is no solution to Problem 3 for $\mathcal{C}_2$. Indeed, for every $t\in\mathcal{D}$ we have $d_n(t,\gamma'_t)=n\log 3/2+o(n)$. Therefore, if ρ is a solution to Problem 3 then $\limsup\frac{1}{n}d_n(t,\rho)\leq\log 3/2<1$ which contradicts Lemma 2.5. □

Thus, prediction in expected average KL divergence turns out to be a more complicated matter than prediction in total variation. The next idea is to try and see which of the facts about prediction in total variation can be generalized to some of the problems concerning prediction in expected average KL divergence.

First, observe that, for the case of prediction in total variation, the equivalence of Problems 1 and 2 was derived from the transitivity of the relation $\geq_{tv}$ of absolute continuity. For the case of expected average KL divergence, the relation "ρ predicts μ in expected average KL divergence" is not transitive (and Problems 1 and 2 are not equivalent). However, for Problem 2 we are interested in the following relation: ρ "dominates" μ if ρ predicts every ν such that μ predicts ν. Denote this relation by $\geq_{KL}$:

Definition 2.34 ($\geq_{KL}$)**.** *We write $\rho\geq_{KL}\mu$ if for every $\nu\in\mathcal{P}$ the equality* $\limsup\frac{1}{n}d_n(\nu,\mu)=0$ *implies* $\limsup\frac{1}{n}d_n(\nu,\rho)=0$.

The relation $\geq_{KL}$ has some similarities with $\geq_{tv}$. First of all, $\geq_{KL}$ is also transitive (as can be easily seen from the definition). Moreover, similarly to $\geq_{tv}$, one can show that for any μ,ρ any strictly convex combination $\alpha\mu+(1-\alpha)\rho$ is a supremum of $\{\rho,\mu\}$ with respect to $\geq_{KL}$. Next we will obtain a characterization of predictability with respect to $\geq_{KL}$ similar to one of those obtained for $\geq_{tv}$.

The key observation is the following. If there is a solution to Problem 2 for a set $\mathcal{C}$ then a solution can be obtained as a Bayesian mixture over

a countable subset of $\mathcal{C}$. For total variation this is the statement (v) of Theorem 2.31.

Theorem 2.35. *Let $\mathcal{C}$ be a set of probability measures on Ω. If there is a measure ρ such that $\rho \geq_{KL} \mu$ for every $\mu \in \mathcal{C}$ (ρ is a solution to Problem 2) then there is a sequence $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$, such that $\sum_{k\in\mathbb{N}} w_k \mu_k \geq_{KL} \mu$ for every $\mu \in \mathcal{C}$, where w_k are some positive weights.*

The proof is deferred to Section 2.6.3.

For the case of Problem 3, we do not have results similar to Theorem 2.35 (or statement (v) of Theorem 2.31); in fact, we conjecture that the opposite is true: there exists a (measurable) set $\mathcal{C}$ of measures such that there is a solution to Problem 3 for $\mathcal{C}$, but there is no Bayesian solution to Problem 3, meaning that there is no probability distribution on $\mathcal{C}$ (discrete or not) such that the mixture over $\mathcal{C}$ with respect to this distribution is a solution to Problem 3 for $\mathcal{C}$.

However, we can take a different route and extend another part of Theorem 2.31 to obtain a characterization of sets $\mathcal{C}$ for which a solution to Problem 3 exists.

We have seen that, in the case of prediction in total variation, separability with respect to the topology of this distance is a necessary and sufficient condition for the existence of a solution to Problems 1-3. In the case of expected average KL divergence the situation is somewhat different, since, first of all, (asymptotic average) KL divergence is not a metric. While one can introduce a topology based on it, separability with respect to this topology turns out to be a sufficient but not a necessary condition for the existence of a predictor, as is shown in the next theorem.

Definition 2.36. *Define the distance $d_\infty(\mu_1,\mu_2)$ on process measures as follows*

$$d_\infty(\mu_1,\mu_2) = \limsup_{n\to\infty} \sup_{x_{1..n}\in \mathbf{X}^n} \frac{1}{n} \left| \log \frac{\mu_1(x_{1..n})}{\mu_2(x_{1..n})} \right|, \tag{2.33}$$

where we assume $\log 0/0 := 0$.

Clearly, d_∞ is symmetric and satisfies the triangle inequality, but it is not exact. Moreover, for every μ_1,μ_2 we have

$$\limsup_{n\to\infty}\frac{1}{n}d_n(\mu_1,\mu_2)\le d_\infty(\mu_1,\mu_2). \tag{2.34}$$

The distance $d_\infty(\mu_1,\mu_2)$ measures the difference in behaviour of μ_1 and μ_2 on all individual sequences. Thus, using this distance to analyse Problem 3 is most close to the traditional approach to the non-realizable case, which is formulated in terms of predicting individual deterministic sequences.

Theorem 2.37. *(i) Let $\mathcal{C}$ be a set of process measures. If $\mathcal{C}$ is separable with respect to d_∞ then there is a solution to Problem 3 for $\mathcal{C}$, for the case of prediction in expected average KL divergence.*

(ii) There exists a set of process measures $\mathcal{C}$ such that $\mathcal{C}$ is not separable with respect to d_∞, but there is a solution to Problem 3 for this set, for the case of prediction in expected average KL divergence.

Proof. For the first statement, let $\mathcal{C}$ be separable and let $(\mu_k)_{k\in\mathbb{N}}$ be a dense countable subset of $\mathcal{C}$. Define $\nu:=\sum_{k\in\mathbb{N}}w_k\mu_k$, where w_k are any positive summable weights. Fix any measure τ and any $\mu\in\mathcal{C}$. We will show that $\limsup_{n\to\infty}\frac{1}{n}d_n(\tau,\nu)\le\limsup_{n\to\infty}\frac{1}{n}d_n(\tau,\mu)$. For every ε, find such a $k\in\mathbb{N}$ that $d_\infty(\mu,\mu_k)\le\varepsilon$. We have

$$\begin{aligned}
d_n(\tau,\nu)\le d_n(\tau,w_k\mu_k)&=\mathbb{E}_\tau\log\frac{\tau(x_{1..n})}{\mu_k(x_{1..n})}-\log w_k\\
&=\mathbb{E}_\tau\log\frac{\tau(x_{1..n})}{\mu(x_{1..n})}+\mathbb{E}_\tau\log\frac{\mu(x_{1..n})}{\mu_k(x_{1..n})}-\log w_k\\
&\le d_n(\tau,\mu)+\sup_{x_{1..n}\in\mathbf{X}^n}\log\left|\frac{\mu(x_{1..n})}{\mu_k(x_{1..n})}\right|-\log w_k.
\end{aligned}$$

From this, dividing by n taking $\limsup_{n\to\infty}$ on both sides, we conclude

$$\limsup_{n\to\infty}\frac{1}{n}d_n(\tau,\nu)\le\limsup_{n\to\infty}\frac{1}{n}d_n(\tau,\mu)+\varepsilon.$$

Since this holds for every $\varepsilon > 0$ the first statement is proven.

The second statement is proven by the following example. Let $\mathcal{C}$ be the set of all deterministic sequences (measures concentrated on just one sequence) such that the number of 0s in the first n symbols is less than $\sqrt{n}$, for all $n \in \mathbb{N}$. Clearly, this set is uncountable. It is easy to check that $\mu_1 \neq \mu_2$ implies $d_\infty(\mu_1, \mu_2) = \infty$ for every $\mu_1, \mu_2 \in \mathcal{C}$, but the predictor ν, given by $\nu(x_n = 0) = 1/n$ independently for different n, predicts every $\mu \in \mathcal{C}$ in expected average KL divergence. Since all elements of $\mathcal{C}$ are deterministic, ν is also a solution to Problem 3 for $\mathcal{C}$. □

Although simple, Theorem 2.37 can be used to establish the existence of a solution to Problem 3 for an important class of process measures: that of all processes with finite memory, as the next theorem shows. Results similar to Theorem 2.38 are known in different settings, e.g., [92, 76, 17] and others.

Theorem 2.38. *There exists a solution to Problem 3 for prediction in expected average KL divergence for the set of all finite-memory process measures* $\mathcal{M} := \cup_{k \in \mathbb{N}} \mathcal{M}_k$.

Proof. We will show that the set $\mathcal{M}$ is separable with respect to d_∞. Then the statement will follow from Theorem 2.37. It is enough to show that each set $\mathcal{M}_k$ is separable with respect to d_∞.

For simplicity, assume that the alphabet is binary ($|\mathbf{X}| = 2$; the general case is analogous). Observe that the family $\mathcal{M}_k$ of k-order stationary binary-valued Markov processes is parametrized by $|\mathbf{X}|^k$ [0,1]-valued parameters: probability of observing 0 after observing $x_{1..k}$, for each $x_{1..k} \in \mathbf{X}^k$. Note that this parametrization is continuous (as a mapping from the parameter space with the Euclidean topology to $\mathcal{M}_k$ with the topology of d_∞). Indeed, for any $\mu_1, \mu_2 \in \mathcal{M}_k$ and every $x_{1..n} \in \mathbf{X}^n$ such that $\mu_i(x_{1..n}) \neq 0$, $i = 1,2$, it is easy to see that

$$\frac{1}{n}\left|\log\frac{\mu_1(x_{1..n})}{\mu_2(x_{1..n})}\right| \leq \sup_{x_{1..k+1}} \frac{1}{k+1}\left|\log\frac{\mu_1(x_{1..k+1})}{\mu_2(x_{1..k+1})}\right|, \tag{2.35}$$

so that the right-hand side of (2.35) also upper-bounds $d_\infty(\mu_1,\mu_2)$, implying continuity of the parametrization.

It follows that the set $\mu_q^k,\ q \in Q^{|\mathbf{X}|^k}$ of all stationary k-order Markov processes with rational values of all the parameters ($Q := \mathbb{Q} \cap [0,1]$) is dense in $\mathcal{M}_k$, proving the separability of the latter set. □

Another important example is the set of all stationary process measures $\mathcal{S}$. This example also illustrates the difference between the prediction problems that we consider. For this set a solution to Problem 1 was given in [78]. In contrast, here we show that there is no solution to Problem 3 for $\mathcal{S}$.

Theorem 2.39. *There is no solution to Problem 3 for the set of all stationary processes $\mathcal{S}$.*

Proof. This proof is based on the construction similar to the one used in [78] to demonstrate impossibility of consistent prediction of stationary processes without Cesaro averaging.

Let m be a Markov chain with states 0,1,2,... and state transitions defined as follows. From each sate $k \in \mathbb{N} \cup \{0\}$ the chain passes to the state $k+1$ with probability $2/3$ and to the state 0 with probability $1/3$. It is easy to see that this chain possesses a unique stationary distribution on the set of states (see, e.g., [84]); taken as the initial distribution it defines a stationary ergodic process with values in $\mathbb{N} \cup \{0\}$. Fix the ternary alphabet $\mathbf{X} = \{a,0,1\}$. For each sequence $t = t_1,t_2,\dots \in \{0,1\}^\infty$ define the process μ_t as follows. It is a deterministic function of the chain m. If the chain is in the state 0 then the process μ_t outputs a; if the chain m is in the state $k > 0$ then the process outputs t_k. That is, we have defined a hidden Markov process which in the state 0 of the underlying Markov chain always outputs a, while in other states it outputs either 0 or 1 according to the sequence t.

To show that there is no solution to Problem 3 for $\mathcal{S}$, we will show that there is no solution to Problem 3 for the smaller set $\mathcal{C} := \{\mu_t : t \in \{0,1\}^\infty\}$. Indeed, for any $t \in \{0,1\}^\infty$ we have $d_n(t,\mu_t) = n\log 3/2 + o(n)$. Then if ρ is a

solution to Problem 3 for $\mathcal{C}$ we should have $\limsup_{n\to\infty}\frac{1}{n}d_n(t,\rho)\le\log 3/2<1$ for every $t\in\mathcal{D}$, which contradicts Lemma 2.5. □

From the proof of Theorem 2.39 one can see that, in fact, the statement that is proven is stronger: there is no solution to Problem 3 for the set of all functions of stationary ergodic countable-state Markov chains. We conjecture that a solution to Problem 2 exists for the latter set, but not for the set of all stationary processes.

As we have seen in the statements above, the set of all deterministic measures $\mathcal{D}$ plays an important role in the analysis of the predictors in the sense of Problem 3. Therefore, an interesting question is to characterize those sets $\mathcal{C}$ of measures for which there is a predictor ρ that predicts *every individual sequence* at least as well as any measure from $\mathcal{C}$. Such a characterization can be obtained in terms of Hausdorff dimension, using a result of [77], that shows that Hausdorff dimension of a set characterizes the optimal prediction error that can be attained by any predictor.

For a set $A\subset\mathbf{X}^\infty$ denote $H(A)$ its Hausdorff dimension (see, for example, [10] for its definition).

Theorem 2.40. *Let $\mathcal{C}\subset\mathcal{P}$. The following statements are equivalent.*

(i) There is a measure $\rho\in\mathcal{P}$ that predicts every individual sequence at least as well as the best measure from $\mathcal{C}$: for every $\mu\in\mathcal{C}$ and every sequence $x_1,x_2,\dots\in\mathbf{X}^\infty$ we have

$$\liminf_{n\to\infty}-\frac{1}{n}\log\rho(x_{1..n})\le\liminf_{n\to\infty}-\frac{1}{n}\log\mu(x_{1..n}). \tag{2.36}$$

(ii) For every $\alpha\in[0,1]$ the Hausdorff dimension of the set of sequences on which the average prediction error of the best measure in $\mathcal{C}$ is not greater than α is bounded by $\alpha/\log|\mathbf{X}|$:

$$H(\{x_1,x_2,\dots\in\mathbf{X}^\infty:\inf_{\mu\in\mathcal{C}}\liminf_{n\to\infty}-\frac{1}{n}\log\mu(x_{1..n})\le\alpha\})\le\alpha/\log|\mathbf{X}|. \tag{2.37}$$

Proof. The implication $(i)\Rightarrow(ii)$ follows directly from [77] where it is shown that for every measure ρ one must have

$$H(\{x_1,x_2,\cdots\in\mathbf{X}^\infty:\liminf_{n\to\infty}-\frac{1}{n}\log\rho(x_{1..n})\leq\alpha\})\leq\alpha/\log|\mathbf{X}|.$$

To show the opposite implication, we again refer to [77]: for every set $A\subset\mathbf{X}^\infty$ there is a measure ρ_A such that

$$\liminf_{n\to\infty}-\frac{1}{n}\log\rho_A(x_{1..n})\leq H(A)\log|\mathbf{X}|. \tag{2.38}$$

For each $\alpha\in[0,1]$ define $A_\alpha:=\{x_1,x_2,\cdots\in\mathbf{X}^\infty:\inf_{\mu\in\mathcal{C}}\liminf_{n\to\infty}-\frac{1}{n}\log\mu(x_{1..n})\leq\alpha\}$). By assumption, $H(A_\alpha)\leq\alpha/\log|X|$, so that from (2.38) for all $x_1,x_2,\cdots\in A_\alpha$ we obtain

$$\liminf_{n\to\infty}-\frac{1}{n}\log\rho_A(x_{1..n})\leq\alpha. \tag{2.39}$$

Furthermore, define $\rho:=\sum_{q\in Q}w_q\rho_{A_q}$, where $Q=[0,1]\cap\mathbb{Q}$ is the set of rationals in $[0,1]$ and $(w_q)_{q\in Q}$ is any sequence of positive reals satisfying $\sum_{q\in Q}w_q=1$. For every $\alpha\in[0,1]$ let $q_k\in Q$, $k\in\mathbb{N}$ be such a sequence that $0\leq q_k-\alpha\leq 1/k$. Then, for every $n\in\mathbb{N}$ and every $x_1,x_2,\cdots\in A_{q_k}$ we have

$$-\frac{1}{n}\log\rho(x_{1..n})\leq-\frac{1}{n}\log\rho_q(x_{1..n})-\frac{\log w_{q_k}}{n}.$$

From this and (2.39) we get

$$\liminf_{n\to\infty}-\frac{1}{n}\log\rho(x_{1..n})\leq\liminf_{n\to\infty}\rho_{q_k}(x_{1..n})+1/k\leq q_k+1/k.$$

Since this holds for every $k\in\mathbb{N}$, it follows that for all $x_1,x_2,\cdots\in\cap_{k\in\mathbb{N}}A_{q_k}=A_\alpha$ we have

$$\liminf_{n\to\infty}-\frac{1}{n}\log\rho(x_{1..n})\leq\inf_{k\in\mathbb{N}}(q_k+1/k)=\alpha,$$

which completes the proof of the implication $(ii)\Rightarrow(i)$. □

2.6 Longer proofs

2.6.1 Proof of Theorem 2.7

The proof is broken into the same steps as the (simpler) proof of Theorem 2.6, to make the analogy explicit and the proof more understandable.

Proof. Define the weights $w_k := wk^{-2}$, where w is the normalizer $6/\pi^2$.
Step 1: densities. Define the sets

$$T_\mu^n := \left\{ x_{1..n} \in \mathbf{X}^n : \mu(x_{1..n}) \geq \frac{1}{n} \rho(x_{1..n}) \right\}. \tag{2.40}$$

Using Markov's inequality, we derive

$$\mu(\mathbf{X}^n \backslash T_\mu^n) = \mu\left(\frac{\rho(x_{1..n})}{\mu(x_{1..n})} > n \right) \leq \frac{1}{n} E_\mu \frac{\rho(x_{1..n})}{\mu(x_{1..n})} = \frac{1}{n}, \tag{2.41}$$

so that $\mu(T_\mu^n) \to 1$. (Note that if μ is singular with respect to ρ, as is typically the case, then $\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$ converges to 0 μ-a.e. and one can replace $\frac{1}{n}$ in (2.40) by 1, while still having $\mu(T_\mu^n) \to 1$.)

Step 2n: a countable cover, time n. Fix an $n \in \mathbb{N}$. Define $m_1^n := \max_{\mu \in \mathcal{C}} \rho(T_\mu^n)$ (since $\mathbf{X}^n$ are finite all suprema are reached). Find any μ_1^n such that $\rho_1^n(T_{\mu_1^n}^n) = m_1^n$ and let $T_1^n := T_{\mu_1^n}^n$. For $k > 1$, let $m_k^n := \max_{\mu \in \mathcal{C}} \rho(T_\mu^n \backslash T_{k-1}^n)$. If $m_k^n > 0$, let μ_k^n be any $\mu \in \mathcal{C}$ such that $\rho(T_{\mu_k^n}^n \backslash T_{k-1}^n) = m_k^n$, and let $T_k^n := T_{k-1}^n \cup T_{\mu_k^n}^n$; otherwise let $T_k^n := T_{k-1}^n$. Observe that (for each n) there is only a finite number of positive m_k^n, since the set $\mathbf{X}^n$ is finite; let K_n be the largest index k such that $m_k^n > 0$. Let

$$\nu_n := \sum_{k=1}^{K_n} w_k \mu_k^n. \tag{2.42}$$

As a result of this construction, for every $n \in \mathbb{N}$ every $k \leq K_n$ and every

$x_{1..n} \in T_k^n$ using (2.40) we obtain

$$\nu_n(x_{1..n}) \geq w_k \frac{1}{n} \rho(x_{1..n}). \tag{2.43}$$

Step 2: the resulting predictor. Finally, define

$$\nu := \frac{1}{2}\gamma + \frac{1}{2}\sum_{n\in\mathbb{N}} w_n \nu_n, \tag{2.44}$$

where γ is the i.i.d. measure with equal probabilities of all $x \in \mathbf{X}$ (that is, $\gamma(x_{1..n}) = |\mathbf{X}|^{-n}$ for every $n \in \mathbb{N}$ and every $x_{1..n} \in \mathbf{X}^n$). We will show that ν predicts every $\mu \in \mathcal{C}$, and then in the end of the proof (Step r) we will show how to replace γ by a combination of a countable set of elements of $\mathcal{C}$ (in fact, γ is just a regularizer which ensures that ν-probability of any word is never too close to 0).

Step 3: ν predicts every $\mu \in \mathcal{C}$. Fix any $\mu \in \mathcal{C}$. Introduce the parameters $\varepsilon_\mu^n \in (0,1)$, $n \in \mathbb{N}$, to be defined later, and let $j_\mu^n := 1/\varepsilon_\mu^n$. Observe that $\rho(T_k^n \backslash T_{k-1}^n) \geq \rho(T_{k+1}^n \backslash T_k^n)$, for any $k > 1$ and any $n \in \mathbb{N}$, by definition of these sets. Since the sets $T_k^n \backslash T_{k-1}^n$, $k \in \mathbb{N}$ are disjoint, we obtain $\rho(T_k^n \backslash T_{k-1}^n) \leq 1/k$. Hence, $\rho(T_\mu^n \backslash T_j^n) \leq \varepsilon_\mu^n$ for some $j \leq j_\mu^n$, since otherwise $m_j^n = \max_{\mu\in\mathcal{C}} \rho(T_\mu^n \backslash T_{j_\mu^n}^n) > \varepsilon_\mu^n$ so that $\rho(T_{j_\mu^n+1}^n \backslash T_{j_\mu^n}^n) > \varepsilon_\mu^n = 1/j_\mu^n$, which is a contradiction. Thus,

$$\rho(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \varepsilon_\mu^n. \tag{2.45}$$

We can upper-bound $\mu(T^n_\mu \backslash T^n_{j^n_\mu})$ as follows. First, observe that

$$\begin{aligned} d_n(\mu,\rho) = &- \sum_{x_{1..n} \in T^n_\mu \cap T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in T^n_\mu \backslash T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in \mathbf{X}^n \backslash T^n_\mu} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= I + II + III. \end{aligned} \tag{2.46}$$

Then, from (2.40) we get

$$I \geq -\log n. \tag{2.47}$$

Observe that for every $n \in \mathbb{N}$ and every set $A \subset \mathbf{X}^n$, using Jensen's inequality we can obtain

$$\begin{aligned} -\sum_{x_{1..n} \in A} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} &= -\mu(A) \sum_{x_{1..n} \in A} \frac{1}{\mu(A)} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &\geq -\mu(A) \log \frac{\rho(A)}{\mu(A)} \geq -\mu(A) \log \rho(A) - \frac{1}{2}. \end{aligned} \tag{2.48}$$

Thus, from (2.48) and (2.45) we get

$$II \geq -\mu(T^n_\mu \backslash T^n_{j^n_\mu}) \log \rho(T^n_\mu \backslash T^n_{j^n_\mu}) - 1/2 \geq -\mu(T^n_\mu \backslash T^n_{j^n_\mu}) \log \varepsilon^n_\mu - 1/2. \tag{2.49}$$

Furthermore,

$$\begin{aligned} III \geq \sum_{x_{1..n} \in \mathbf{X}^n \backslash T^n_\mu} \mu(x_{1..n}) \log \mu(x_{1..n}) &\geq \mu(\mathbf{X}^n \backslash T^n_\mu) \log \frac{\mu(\mathbf{X}^n \backslash T^n_\mu)}{|\mathbf{X}^n \backslash T^n_\mu|} \\ &\geq -\frac{1}{2} - \mu(\mathbf{X}^n \backslash T^n_\mu) n \log |\mathbf{X}| \geq -\frac{1}{2} - \log |\mathbf{X}|, \end{aligned} \tag{2.50}$$

where in the second inequality we have used the fact that entropy is maximized when all events are equiprobable, in the third one we used $|\mathbf{X}^n \backslash T_\mu^n| \leq |\mathbf{X}|^n$, while the last inequality follows from (2.41). Combining (2.46) with the bounds (2.47), (2.49) and (2.50) we obtain

$$d_n(\mu,\rho) \geq -\log n - \mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \varepsilon_\mu^n - 1 - \log|\mathbf{X}|,$$

so that

$$\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \frac{1}{-\log \varepsilon_\mu^n} \Big(d_n(\mu,\rho) + \log n + 1 + \log|\mathbf{X}| \Big). \tag{2.51}$$

Since $d_n(\mu,\rho) = o(n)$, we can define the parameters ε_μ^n in such a way that $-\log \varepsilon_\mu^n = o(n)$ while at the same time the bound (2.51) gives $\mu(T_\mu^n \backslash T_{j_\mu^n}^n) = o(1)$. Fix such a choice of ε_μ^n. Then, using $\mu(T_\mu^n) \to 1$, we can conclude

$$\mu(\mathbf{X}^n \backslash T_{j_\mu^n}^n) \leq \mu(\mathbf{X}^n \backslash T_\mu^n) + \mu(T_\mu^n \backslash T_{j_\mu^n}^n) = o(1). \tag{2.52}$$

We proceed with the proof of $d_n(\mu,\nu) = o(n)$. For any $x_{1..n} \in T_{j_\mu^n}^n$ we have

$$\nu(x_{1..n}) \geq \frac{1}{2} w_n \nu_n(x_{1..n}) \geq \frac{1}{2} w_n w_{j_\mu^n} \frac{1}{n} \rho(x_{1..n}) = \frac{w_n w}{2n} (\varepsilon_\mu^n)^2 \rho(x_{1..n}), \tag{2.53}$$

where the first inequality follows from (2.44), the second from (2.43), and in the equality we have used $w_{j_\mu^n} = w/(j_\mu^n)^2$ and $j_\mu^n = 1/\varepsilon_n^\mu$. Next we use the decomposition

$$d_n(\mu,\nu) = -\sum_{x_{1..n} \in T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} - \sum_{x_{1..n} \in \mathbf{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} = I + II. \tag{2.54}$$

From (2.53) we find

$$I \le -\log\Big(\frac{w_n w}{2n}(\varepsilon_\mu^n)^2\Big) - \sum_{x_{1..n}\in T^n_{j^n_\mu}} \mu(x_{1..n})\log\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$$

$$= (1+3\log n - 2\log\varepsilon_\mu^n - 2\log w) + \left(d_n(\mu,\rho) + \sum_{x_{1..n}\in \mathbf{X}^n\setminus T^n_{j^n_\mu}} \mu(x_{1..n})\log\frac{\rho(x_{1..n})}{\mu(x_{1..n})} \right)$$

$$\le o(n) - \sum_{x_{1..n}\in \mathbf{X}^n\setminus T^n_{j^n_\mu}} \mu(x_{1..n})\log\mu(x_{1..n})$$

$$\le o(n) + \mu(\mathbf{X}^n\setminus T^n_{j^n_\mu}) n\log|\mathbf{X}| = o(n), \quad (2.55)$$

where in the second inequality we have used $-\log\varepsilon_\mu^n = o(n)$ and $d_n(\mu,\rho) = o(n)$, in the last inequality we have again used the fact that the entropy is maximized when all events are equiprobable, while the last equality follows from (2.52). Moreover, from (2.44) we find

$$II \le \log 2 - \sum_{x_{1..n}\in \mathbf{X}^n\setminus T^n_{j^n_\mu}} \mu(x_{1..n})\log\frac{\gamma(x_{1..n})}{\mu(x_{1..n})} \le 1 + n\mu(\mathbf{X}^n\setminus T^n_{j^n_\mu})\log|\mathbf{X}| = o(n), \quad (2.56)$$

where in the last inequality we have used $\gamma(x_{1..n}) = |\mathbf{X}|^{-n}$ and $\mu(x_{1..n}) \le 1$, and the last equality follows from (2.52).

From (2.54), (2.55) and (2.56) we conclude $\frac{1}{n} d_n(\nu,\mu) \to 0$.

Step r: the regularizer γ. It remains to show that the i.i.d. regularizer γ in the definition of ν (2.44), can be replaced by a convex combination of a countably many elements from $\mathcal{C}$. Indeed, for each $n \in \mathbb{N}$, denote

$$A_n := \{x_{1..n} \in \mathbf{X}^n : \exists \mu \in \mathcal{C}\ \mu(x_{1..n}) \neq 0\},$$

and let for each $x_{1..n} \in \mathbf{X}^n$ the measure $\mu_{x_{1..n}}$ be any measure from $\mathcal{C}$ such

that $\mu_{x_{1..n}}(x_{1..n}) \geq \frac{1}{2}\sup_{\mu\in\mathcal{C}}\mu(x_{1..n})$. Define

$$\gamma'_n(x'_{1..n}) := \frac{1}{|A_n|}\sum_{x_{1..n}\in A_n}\mu_{x_{1..n}}(x'_{1..n}),$$

for each $x'_{1..n}\in A^n$, $n\in\mathbb{N}$, and let $\gamma' := \sum_{k\in\mathbb{N}} w_k\gamma'_k$. For every $\mu\in\mathcal{C}$ we have

$$\gamma'(x_{1..n}) \geq w_n|A_n|^{-1}\mu_{x_{1..n}}(x_{1..n}) \geq \frac{1}{2}w_n|\mathbf{X}|^{-n}\mu(x_{1..n})$$

for every $n\in\mathbb{N}$ and every $x_{1..n}\in A_n$, which clearly suffices to establish the bound $II = o(n)$ as in (2.56). □

2.6.2 Proof of Theorem 2.22

Proof. Again the second statement (about absolute distance) follows from the first one and Lemma 2.18, so that we only have to prove the statement about KL divergence.

Introduce the symbol $\mathbb{E}^n$ for μ-expectation over x_n conditional on $x_{<n}$. Consider random variables $l_n = \log\frac{\mu(x_n|x_{<n})}{\rho(x_n|x_{<n})}$ and $\bar{l}_n = \frac{1}{n}\sum_{t=1}^n l_t$. Observe that $\delta_n = \mathbb{E}^n l_n$, so that the random variables $m_n = l_n - \delta_n$ form a martingale difference sequence (that is, $\mathbb{E}^n m_n = 0$) with respect to the standard filtration defined by $x_1,...,x_n,...$. Let also $\bar{m}_n = \frac{1}{n}\sum_{t=1}^n m_t$. We will show that $\bar{m}_n \to 0$ μ-a.s. and $\bar{l}_n \to 0$ μ-a.s. which implies $\bar{d}_n \to 0$ μ-a.s.

Note that

$$\bar{l}_n \;=\; \frac{1}{n}\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})} \;\leq\; \frac{\log c_n^{-1}}{n} \;\to\; 0.$$

Thus to show that $\bar{l}_n$ goes to 0 we need to bound it from below. It is easy to see that $n\bar{l}_n$ is (μ-a.s.) bounded from below by a constant, since $\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$ is a positive μ-martingale whose expectation is 1, and so it converges to a finite limit μ-a.s. by Doob's submartingale convergence theorem, see e.g. [84, p.508].

Next we will show that $\bar{m}_n \to 0$ μ-a.s. We have

$$\begin{aligned} m_n &= \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} - \log \frac{\mu(x_{<n})}{\rho(x_{<n})} - \mathbb{E}^n \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} + \mathbb{E}^n \log \frac{\mu(x_{<n})}{\rho(x_{<n})} \\ &= \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} - \mathbb{E}^n \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})}. \end{aligned}$$

Let $f(n)$ be some function monotonically increasing to infinity such that

$$\sum_{n=1}^{\infty} \frac{(\log c_n^{-1} + f(n))^2}{n^2} \;<\; \infty \tag{2.57}$$

(e.g. choose $f(n) = \log n$ and exploit $(\log c_n^{-1} + f(n))^2 \leq 2(\log c_n^{-1})^2 + 2f(n)^2$ and (2.31).) For a sequence of random variables λ_n define

$$(\lambda_n)^{+(f)} \;=\; \begin{cases} \lambda_n & \text{if } \lambda_n \geq -f(n) \\ 0 & \text{otherwise} \end{cases}$$

and $\lambda_n^{-(f)} = \lambda_n - \lambda_n^{+(f)}$. Introduce also

$$m_n^+ \;=\; \left(\log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \right)^{+(f)} - \mathbb{E}^n \left(\log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \right)^{+(f)},$$

$m_n^- = m_n - m_n^+$ and the averages $\bar{m}_n^+$ and $\bar{m}_n^-$. Observe that m_n^+ is a martingale difference sequence. Hence to establish the convergence $\bar{m}_n^+ \to 0$ we can use the martingale strong law of large numbers [84, p.501], which states that, for a martingale difference sequence γ_n, if $\mathbb{E}(n\bar{\gamma}_n)^2 < \infty$ and $\sum_{n=1}^{\infty} \mathbb{E}\gamma_n^2/n^2 < \infty$ then $\bar{\gamma}_n \to 0$ a.s. Indeed, for m_n^+ the first condition is trivially satisfied (since the expectation in question is a finite sum of finite numbers), and the second follows from the fact that $|m_n^+| \leq \log c_n^{-1} + f(n)$ and (2.57).

Furthermore, we have

$$m_n^- \;=\; \left(\log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \right)^{-(f)} - \mathbb{E}^n \left(\log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \right)^{-(f)}.$$

As it was mentioned before, $\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})}$ converges μ-a.s. either to (positive) infinity or to a finite number. Hence $\left(\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})}\right)^{-(f)}$ is non-zero only a finite number of times, and so its average goes to zero. To see that

$$\mathbb{E}^n\left(\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})}\right)^{-(f)}\to 0$$

we write

$$\begin{aligned}\mathbb{E}^n\left(\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})}\right)^{-(f)} &= \sum_{x_n\in\mathbf{X}}\mu(x_n|x_{<n})\left(\log\frac{\mu(x_{<n})}{\rho(x_{<n})}+\log\frac{\mu(x_n|x_{<n})}{\rho(x_n|x_{<n})}\right)^{-(f)}\\ &\geq \sum_{x_n\in\mathbf{X}}\mu(x_n|x_{<n})\left(\log\frac{\mu(x_{<n})}{\rho(x_{<n})}+\log\mu(x_n|x_{<n})\right)^{-(f)}\end{aligned}$$

and note that the first term in brackets is bounded from below, and so for the sum in brackets to be less than $-f(n)$ (which is unbounded) the second term $\log\mu(x_n|x_{<n})$ has to go to $-\infty$, but then the expectation goes to zero since $\lim_{u\to 0}u\log u=0$.

Thus we conclude that $\bar{m}_n^-\to 0$ μ-a.s., which together with $\bar{m}_n^+\to 0$ μ-a.s. implies $\bar{m}_n\to 0$ μ-a.s., which, finally, together with $\bar{l}_n\to 0$ μ-a.s. implies $\bar{d}_n\to 0$ μ-a.s. □

2.6.3 Proof of Theorem 2.35

Proof. This proof follows the same step as the proof of Theorem 2.7 (presented in Section 2.6.1) but is a bit more involved.

Define the sets C_μ as the set of all measures $\tau\in\mathcal{P}$ such that μ predicts τ in expected average KL divergence. Let $\mathcal{C}^+:=\cup_{\mu\in\mathcal{C}}C_\mu$. For each $\tau\in\mathcal{C}^+$ let $p(\tau)$ be any (fixed) $\mu\in\mathcal{C}$ such that $\tau\in C_\mu$. In other words, $\mathcal{C}^+$ is the set of all measures that are predicted by some of the measures in $\mathcal{C}$, and for each measure τ in $\mathcal{C}^+$ we designate one "parent" measure $p(\tau)$ from $\mathcal{C}$ such that $p(\tau)$ predicts τ.

Define the weights $w_k := 1/k(k+1)$, for all $k \in \mathbb{N}$.

Step 1. For each $\mu \in \mathcal{C}^+$ let δ_n be any monotonically increasing function such that $\delta_n(\mu) = o(n)$ and $d_n(\mu, p(\mu)) = o(\delta_n(\mu))$. Define the sets

$$U_\mu^n := \left\{ x_{1..n} \in \mathbf{X}^n : \mu(x_{1..n}) \geq \frac{1}{n} \rho(x_{1..n}) \right\}, \tag{2.58}$$

$$V_\mu^n := \left\{ x_{1..n} \in \mathbf{X}^n : p(\mu)(x_{1..n}) \geq 2^{-\delta_n(\mu)} \mu(x_{1..n}) \right\}, \tag{2.59}$$

and

$$T_\mu^n := U_\mu^n \cap V_\mu^n. \tag{2.60}$$

We will upper-bound $\mu(T_\mu^n)$. First, using Markov's inequality, we derive

$$\mu(\mathbf{X}^n \setminus U_\mu^n) = \mu\left(\frac{\rho(x_{1..n})}{\mu(x_{1..n})} > n \right) \leq \frac{1}{n} E_\mu \frac{\rho(x_{1..n})}{\mu(x_{1..n})} = \frac{1}{n}. \tag{2.61}$$

Next, observe that for every $n \in \mathbb{N}$ and every set $A \subset \mathbf{X}^n$, using Jensen's inequality we can obtain

$$\begin{aligned} -\sum_{x_{1..n} \in A} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} &= -\mu(A) \sum_{x_{1..n} \in A} \frac{1}{\mu(A)} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &\geq -\mu(A) \log \frac{\rho(A)}{\mu(A)} \geq -\mu(A) \log \rho(A) - \frac{1}{2}. \end{aligned} \tag{2.62}$$

Moreover,

$$\begin{aligned} d_n(\mu, p(\mu)) = &- \sum_{x_{1..n} \in \mathbf{X}^n \setminus V_\mu^n} \mu(x_{1..n}) \log \frac{p(\mu)(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in V_\mu^n} \mu(x_{1..n}) \log \frac{p(\mu)(x_{1..n})}{\mu(x_{1..n})} \geq \delta_n(\mu_n) \mu(\mathbf{X}^n \setminus V_\mu^n) - 1/2, \end{aligned}$$

where in the inequality we have used (2.59) for the first summand and (2.62)

for the second. Thus,

$$\mu(\mathbf{X}^n \backslash V_\mu^n) \le \frac{d_n(\mu, p(\mu)) + 1/2}{\delta_n(\mu)} = o(1). \tag{2.63}$$

From (2.60), (2.61) and (2.63) we conclude

$$\mu(\mathbf{X}^n \backslash T_\mu^n) \le \mu(\mathbf{X}^n \backslash V_\mu^n) + \mu(\mathbf{X}^n \backslash U_\mu^n) = o(1). \tag{2.64}$$

Step 2n: a countable cover, time n. Fix an $n \in \mathbb{N}$. Define $m_1^n := \max_{\mu \in \mathcal{C}} \rho(T_\mu^n)$ (since $\mathbf{X}^n$ are finite all suprema are reached). Find any μ_1^n such that $\rho_1^n(T_{\mu_1^n}^n) = m_1^n$ and let $T_1^n := T_{\mu_1^n}^n$. For $k > 1$, let $m_k^n := \max_{\mu \in \mathcal{C}} \rho(T_\mu^n \backslash T_{k-1}^n)$. If $m_k^n > 0$, let μ_k^n be any $\mu \in \mathcal{C}$ such that $\rho(T_{\mu_k^n}^n \backslash T_{k-1}^n) = m_k^n$, and let $T_k^n := T_{k-1}^n \cup T_{\mu_k^n}^n$; otherwise let $T_k^n := T_{k-1}^n$. Observe that (for each n) there is only a finite number of positive m_k^n, since the set $\mathbf{X}^n$ is finite; let K_n be the largest index k such that $m_k^n > 0$. Let

$$\nu_n := \sum_{k=1}^{K_n} w_k p(\mu_k^n). \tag{2.65}$$

As a result of this construction, for every $n \in \mathbb{N}$ every $k \le K_n$ and every $x_{1..n} \in T_k^n$ using the definitions (2.60), (2.58) and (2.59) we obtain

$$\nu_n(x_{1..n}) \ge w_k \frac{1}{n} 2^{-\delta_n(\mu)} \rho(x_{1..n}). \tag{2.66}$$

Step 2: the resulting predictor. Finally, define

$$\nu := \frac{1}{2}\gamma + \frac{1}{2}\sum_{n \in \mathbb{N}} w_n \nu_n, \tag{2.67}$$

where γ is the i.i.d. measure with equal probabilities of all $x \in \mathbf{X}$ (that is, $\gamma(x_{1..n}) = |\mathbf{X}|^{-n}$ for every $n \in \mathbb{N}$ and every $x_{1..n} \in \mathbf{X}^n$). We will show that ν predicts every $\mu \in \mathcal{C}^+$, and then in the end of the proof (Step r) we will show how to replace γ by a combination of a countable set of elements of $\mathcal{C}$ (in

fact, γ is just a regularizer which ensures that ν-probability of any word is never too close to 0).

Step 3: ν predicts every $\mu \in \mathcal{C}^+$. Fix any $\mu \in \mathcal{C}^+$. Introduce the parameters $\varepsilon_\mu^n \in (0,1)$, $n \in \mathbb{N}$, to be defined later, and let $j_\mu^n := 1/\varepsilon_\mu^n$. Observe that $\rho(T_k^n \backslash T_{k-1}^n) \geq \rho(T_{k+1}^n \backslash T_k^n)$, for any $k > 1$ and any $n \in \mathbb{N}$, by definition of these sets. Since the sets $T_k^n \backslash T_{k-1}^n$, $k \in \mathbb{N}$ are disjoint, we obtain $\rho(T_k^n \backslash T_{k-1}^n) \leq 1/k$. Hence, $\rho(T_\mu^n \backslash T_j^n) \leq \varepsilon_\mu^n$ for some $j \leq j_\mu^n$, since otherwise $m_j^n = \max_{\mu \in \mathcal{C}} \rho(T_\mu^n \backslash T_{j_\mu^n}^n) > \varepsilon_\mu^n$ so that $\rho(T_{j_\mu^n+1}^n \backslash T_{j_\mu^n}^n) > \varepsilon_\mu^n = 1/j_\mu^n$, which is a contradiction. Thus,

$$\rho(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \varepsilon_\mu^n. \tag{2.68}$$

We can upper-bound $\mu(T_\mu^n \backslash T_{j_\mu^n}^n)$ as follows. First, observe that

$$\begin{aligned} d_n(\mu,\rho) = &- \sum_{x_{1..n} \in T_\mu^n \cap T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in T_\mu^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in \mathbf{X}^n \backslash T_\mu^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= I + II + III. \end{aligned} \tag{2.69}$$

Then, from (2.60) and (2.58) we get

$$I \geq -\log n. \tag{2.70}$$

From (2.62) and (2.68) we get

$$II \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \rho(T_\mu^n \backslash T_{j_\mu^n}^n) - 1/2 \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \varepsilon_\mu^n - 1/2. \tag{2.71}$$

Furthermore,

$$III \geq \sum_{x_{1..n}\in \mathbf{X}^n\setminus T_\mu^n} \mu(x_{1..n})\log\mu(x_{1..n})$$
$$\geq \mu(\mathbf{X}^n\setminus T_\mu^n)\log\frac{\mu(\mathbf{X}^n\setminus T_\mu^n)}{|\mathbf{X}^n\setminus T_\mu^n|} \geq -\frac{1}{2} - \mu(\mathbf{X}^n\setminus T_\mu^n)n\log|\mathbf{X}|, \quad (2.72)$$

where the first inequality is obvious, in the second inequality we have used the fact that entropy is maximized when all events are equiprobable and in the third one we used $|\mathbf{X}^n\setminus T_\mu^n| \leq |\mathbf{X}|^n$. Combining (2.69) with the bounds (2.70), (2.71) and (2.72) we obtain

$$d_n(\mu,\rho) \geq -\log n - \mu(T_\mu^n\setminus T_{j_\mu^n}^n)\log\varepsilon_\mu^n - 1 - \mu(\mathbf{X}^n\setminus T_\mu^n)n\log|\mathbf{X}|,$$

so that

$$\mu(T_\mu^n\setminus T_{j_\mu^n}^n) \leq \frac{1}{-\log\varepsilon_\mu^n}\Big(d_n(\mu,\rho)+\log n+1+\mu(\mathbf{X}^n\setminus T_\mu^n)n\log|\mathbf{X}|\Big). \quad (2.73)$$

From the fact that $d_n(\mu,\rho) = o(n)$ and (2.64) it follows that the term in brackets is $o(n)$, so that we can define the parameters ε_μ^n in such a way that $-\log\varepsilon_\mu^n = o(n)$ while at the same time the bound (2.73) gives $\mu(T_\mu^n\setminus T_{j_\mu^n}^n) = o(1)$. Fix such a choice of ε_μ^n. Then, using (2.64), we conclude

$$\mu(\mathbf{X}^n\setminus T_{j_\mu^n}^n) \leq \mu(\mathbf{X}^n\setminus T_\mu^n) + \mu(T_\mu^n\setminus T_{j_\mu^n}^n) = o(1). \quad (2.74)$$

We proceed with the proof of $d_n(\mu,\nu) = o(n)$. For any $x_{1..n} \in T_{j_\mu^n}^n$ we have

$$\nu(x_{1..n}) \geq \frac{1}{2}w_n\nu_n(x_{1..n}) \geq \frac{1}{2}w_n w_{j_\mu^n}\frac{1}{n}2^{-\delta_n(\mu)}\rho(x_{1..n}) \geq \frac{w_n}{4n}(\varepsilon_\mu^n)^2 2^{-\delta_n(\mu)}\rho(x_{1..n}), \quad (2.75)$$

where the first inequality follows from (2.67), the second from (2.66), and in the third we have used $w_{j_\mu^n} = 1/(j_\mu^n)(j_\mu^n+1)$ and $j_\mu^n = 1/\varepsilon_n^\mu$. Next we use the

decomposition

$$d_n(\mu,\nu)=-\sum_{x_{1..n}\in T^n_{j^n_\mu}}\mu(x_{1..n})\log\frac{\nu(x_{1..n})}{\mu(x_{1..n})}-\sum_{x_{1..n}\in\mathbf{X}^n\setminus T^n_{j^n_\mu}}\mu(x_{1..n})\log\frac{\nu(x_{1..n})}{\mu(x_{1..n})}=I+II. \tag{2.76}$$

From (2.75) we find

$$\begin{aligned} I &\le -\log\Big(\frac{w_n}{4n}(\varepsilon^n_\mu)^2 2^{-\delta_n(\mu)}\Big)-\sum_{x_{1..n}\in T^n_{j^n_\mu}}\mu(x_{1..n})\log\frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &=(o(n)-2\log\varepsilon^n_\mu+\delta_n(\mu))+\left(d_n(\mu,\rho)+\sum_{x_{1..n}\in\mathbf{X}^n\setminus T^n_{j^n_\mu}}\mu(x_{1..n})\log\frac{\rho(x_{1..n})}{\mu(x_{1..n})}\right) \\ &\le o(n)-\sum_{x_{1..n}\in\mathbf{X}^n\setminus T^n_{j^n_\mu}}\mu(x_{1..n})\log\mu(x_{1..n}) \\ &\le o(n)+\mu(\mathbf{X}^n\setminus T^n_{j^n_\mu})n\log|\mathbf{X}|=o(n), \end{aligned} \tag{2.77}$$

where in the second inequality we have used $-\log\varepsilon^n_\mu=o(n)$, $d_n(\mu,\rho)=o(n)$ and $\delta_n(\mu)=o(n)$, in the last inequality we have again used the fact that the entropy is maximized when all events are equiprobable, while the last equality follows from (2.74). Moreover, from (2.67) we find

$$II\le\log 2-\sum_{x_{1..n}\in\mathbf{X}^n\setminus T^n_{j^n_\mu}}\mu(x_{1..n})\log\frac{\gamma(x_{1..n})}{\mu(x_{1..n})}\le 1+n\mu(\mathbf{X}^n\setminus T^n_{j^n_\mu})\log|\mathbf{X}|=o(n), \tag{2.78}$$

where in the last inequality we have used $\gamma(x_{1..n})=|\mathbf{X}|^{-n}$ and $\mu(x_{1..n})\le 1$, and the last equality follows from (2.74).

From (2.76), (2.77) and (2.78) we conclude $\frac{1}{n}d_n(\nu,\mu)\to 0$.

Step r: the regularizer γ. It remains to show that the i.i.d. regularizer γ in the definition of ν (2.67), can be replaced by a convex combination of a countably many elements from $\mathcal{C}$. This can be done exactly as in the corresponding step (Step r) of the proof of Theorem 2.7 (Section 2.6.1). □

Chapter 3

Statistical analysis of stationary ergodic time series [R2, R4, R5, R11]

Numerous statistical inference problems can be formulated in the following way: given a sample $x_1,...,x_n$ (where the variables x_i are possibly multi-dimensional) we need to decide whether the (unknown) distribution that generated this sample belongs to a family of distributions H_0 versus it belongs to a family H_1. This formulation encompasses a broad range of problems, that can roughly be grouped into two categories: model verification and property testing. Model verification problems include goodness-of-fit testing (the case when H_0 consists of just one element $H_0 = \{\rho_0\}$), and testing membership to various parametric families, e.g., to the set of all Markov processes. Property testing problems include testing for homogeneity, component independence, and many others.

Most of the research on these and related problems, even in non-parametric settings, is traditionally concentrated on studying independent and identically distributed variables x_i. Unlike on sequence prediction, the research on hypothesis testing for general stationary ergodic processes is very scarce, and for many problems it has still remained unclear whether they can be solved in this setting. This is why in this chapter we concentrate on station-

ary ergodic time series, and do not venture beyond this model. Restricting our considerations to the set of stationary ergodic processes allows us to benefit from the structure imposed on it by the distributional distance: the space of all stationary process is separable with respect to it, and possesses numerous other useful properties.

The presented contribution is as follows. A new methodology for constructing statistical inference procedures is proposed, which is based on estimating the distributional distance. The developed method is used to construct consistent algorithms for such problems as time series classification, change point estimation and time series clustering, for real-valued data, under the only assumption that the sequences under study are generated by stationary ergodic distribution. Using this method, for discrete-valued data, a *complete characterization (necessary and sufficient conditions)* of those hypotheses $H_0 \subset \mathcal{E}$ for which there exist a consistent test against $\mathcal{E} \backslash H_0$ is proposed. Some generalizations of this results (e.g., to arbitrary families H_0,$H_1 \subset \mathcal{E}$) are also considered.

In addition, it is shown that there is no consistent test for homogeneity if the only assumption on the data is that it is stationary ergodic, and even if one makes a slightly stronger assumption that the distributions are B-processes. This result (for stationary ergodic data) has been claimed in [69]; however, what is really proven in that work is only that there is no consistent estimate of a certain process distance ($\bar{d}$-distance) for stationary ergodic processes; thus, the statement about homogeneity testing was only a conjecture. A proof of a stronger version (for B-processes) of this conjecture is presented in this chapter.

The rest of this chapter is organized as follows. Section 3.1 introduces additional definitions and results that we need, including the definition of distributional distance. Section 3.2 presents the proposed approach to statistical inference, which is based on empirical estimates of the distributional distance. This approach is used to obtain consistent goodness-of-fit tests, as well as change point estimates and a method for time series classifica-

tion. All these algorithms are very simple and serve as an illustration of the proposed approach; at the same time, these results are considerably more general than those available before. Section 3.3 presents the criterion for the existence of a consistent test for $H_0 \subset \mathcal{E}$ against $\mathcal{E} \backslash H_0$, and some generalizations. Section 3.4 further extends the results of Section 3.2 to obtain a consistent algorithm for time-series clustering. The computational complexity of the proposed algorithm is also analyzed. Finally, Section 3.5 shows that there is no asymptotically consistent test for homogeneity for B-processes (and hence for stationary ergodic processes).

3.1 Preliminaries

We are considering (stationary ergodic) processes with the alphabet $A = \mathbb{R}$. The generalization to $A = \mathbb{R}^d$ is straightforward; moreover, the results can be extended to the case when A is a complete separable metric space. For each $k \in \mathbb{N}$, let B^k be the set of all cylinders of the form $A_1 \times \cdots \times A_k$ where $A_i \subset A$ are intervals with rational endpoints. Let $\mathcal{B} = \cup_{k=1}^{\infty} B^k$; since this set is countable we can introduce an enumeration $\mathcal{B} = \{B_i : i \in \mathbb{N}\}$. The set $\{B_i \times A^\infty : i \in \mathbb{N}\}$ generates the Borel σ-algebra on $\mathbb{R}^\infty = A^\infty$. For a set $B \in \mathcal{B}$ let $|B|$ be the index k of the set B^k that B comes from: $|B| = k : B \in B^k$.

For a sequence $X \in A^n$ and a set $B \in \mathcal{B}$ denote $\nu(X,B)$ the frequency with which the sequence X falls in the set B

$$\nu(X,B) := \begin{cases} \frac{1}{n-|B|+1} \sum_{i=1}^{n-|B|+1} I_{\{(X_i,\ldots,X_{i+|B|-1}) \in B\}} & \text{if } n \geq |B|, \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $X = (X_1,\ldots,X_n)$. For example,

$$\nu\big((0.5,1.5,1.2,1.4,2.1),([1.0,2.0] \times [1.0,2.0])\big) = 1/2.$$

As before, we use the symbol $\mathcal{S}$ for the set of all stationary processes on A^∞. A stationary process ρ is called *(stationary) ergodic* if the frequency of

occurrence of each word B in a sequence $X_1,X_2,...$ generated by ρ tends to its a priori (or limiting) probability a.s.: $\rho(\lim_{n\to\infty}\nu(X_{1..n},B)=\rho(X_{1..|B|}=B))=1$. By virtue of the ergodic theorem (e.g. [10]), this definition can be shown to be equivalent to the standard definition of stationary ergodic processes (every shift-invariant set has measure 0 or 1; see e.g. [21]). Denote $\mathcal{E}$ the set of all stationary ergodic processes.

Definition 3.1 (distributional distance)**.** *The distributional distance is defined for a pair of processes ρ_1,ρ_2 as follows [36]:*

$$d(\rho_1,\rho_2)=\sum_{i=1}^{\infty}w_i|\rho_1(B_i)-\rho_2(B_i)|, \tag{3.2}$$

where w_i are summable positive real weights (e.g., $w_k=2^{-k}$).

It is easy to see that d is a metric. Equipped with this metric, the space of all stochastic processes is separable and complete; moreover, it is a compact. The set of stationary processes $\mathcal{S}$ is convex closed subset of the space of all stochastic processes (hence a compact too). The set of all finite-memory stationary distributions is dense in $\mathcal{S}$. (Taking only those that have rational transition probabilities we obtain a countable dense subset of $\mathcal{S}$.) The set $\mathcal{E}$ is not convex (a mixture of stationary ergodic distributions is always stationary but never ergodic) and is not closed (its closure is $\mathcal{S}$). We refer to [36] for more details and proofs of these facts.

When talking about closed and open subsets of $\mathcal{S}$ we assume the topology of d.

Definition 3.2 (empirical distributional distance)**.** *For $X,Y\in A^*$, define empirical distributional distance $\hat{d}(X,Y)$ as*

$$\hat{d}(X,Y):=\sum_{i=1}^{\infty}w_i|\nu(X,B_i)-\nu(Y,B_i)|. \tag{3.3}$$

Similarly, we can define the empirical distance when only one of the process

measures is unknown:

$$\hat{d}(X,\rho) := \sum_{i=1}^{\infty} w_i |\nu(X,B_i) - \rho(B_i)|, \tag{3.4}$$

where $\rho \in \mathcal{E}$ *and* $X \in A^*$.

The following lemma will play a key role in establishing the main results.

Lemma 3.3. *Let two samples* $X=(X_1,...,X_k)$ *and* $Y=(Y_1,...,Y_m)$ *be generated by stationary ergodic processes* ρ_X *and* ρ_Y *respectively. Then*

(i) $\lim_{k,m\to\infty} \hat{d}(X,Y) = d(\rho_X,\rho_Y)$ *a.s.*

(ii) $\lim_{k\to\infty} \hat{d}(X,\rho_Y) = d(\rho_X,\rho_Y)$ *a.s.*

Proof. For any $\varepsilon > 0$ we can find such an index J that $\sum_{i=J}^{\infty} w_i < \varepsilon/2$. Moreover, for each j we have $\nu((X_1,...,X_k),B_j) \to \rho_X(B_j)$ a.s., so that

$$|\nu((X_1,...,X_k),B_j) - \rho(B_j)| < \varepsilon/(4Jw_j)$$

from some step k on; define $K_j := k$. Let $K := \max_{j<J} K_j$ (K depends on the realization $X_1,X_2,...$). Define analogously M for the sequence $(Y_1,...,Y_m,...)$. Thus for $k > K$ and $m > M$ we have

$$\begin{aligned}
|\hat{d}(X,Y) - d(\rho_X,\rho_Y)| = & \\
& \left| \sum_{i=1}^{\infty} w_i \big(|\nu(X,B_i) - \nu(Y,B_i)| - |\rho_X(B_i) - \rho_Y(B_i)| \big) \right| \\
& \le \sum_{i=1}^{\infty} w_i \big(|\nu(X,B_i) - \rho_X(B_i)| + |\nu(Y,B_i) - \rho_Y(B_i)| \big) \\
& \le \sum_{i=1}^{J} w_i \big(|\nu(X,B_i) - \rho_X(B_i)| + |\nu(Y,B_i) - \rho_Y(B_i)| \big) + \varepsilon/2 \\
& \le \sum_{i=1}^{J} w_i (\varepsilon/(4Jw_i) + \varepsilon/(4Jw_i)) + \varepsilon/2 = \varepsilon,
\end{aligned}$$

which proves the first statement. The second statement can be proven analogously. □

Considering the Borel (with respect to the metric d) sigma-algebra $\mathcal{F}_{\mathcal{S}}$ on the set $\mathcal{S}$, we obtain a standard probability space $(\mathcal{S},\mathcal{F}_{\mathcal{S}})$. An important tool that will be used in the analysis is **ergodic decomposition** of stationary processes (see e.g. [36, 10]): any stationary process can be expressed as a mixture of stationary ergodic processes. More formally, for any $\rho\in\mathcal{S}$ there is a measure W_ρ on $(\mathcal{S},\mathcal{F}_{\mathcal{S}})$, such that

$$W_\rho(\mathcal{E})=1, \tag{3.5}$$

and $\rho(B)=\int dW_\rho(\mu)\mu(B)$, for any $B\in\mathcal{F}_{A^\infty}$.

3.2 Statistical analysis based on estimates of distributional distance [R4]

In this section we present our approach to the problem of statistical analysis of time series, when nothing is known about the underlying process generating the data, except that it is stationary ergodic. There is a vast literature on time series analysis under various parametric assumptions, and also under such non-parametric assumptions as that the processes is finite-memory or has certain mixing rates. While under these settings most of the problems of statistical analysis are clearly solvable and efficient algorithms exist, in the general setting of stationary ergodic processes it is far less clear what can be done in principle, which problems of statistical analysis admit a solution and which do not. In this chapter we propose a method of statistical analysis of time series, that allows us to demonstrate that some classical statistical problems indeed admit a solution under the only assumption that the data is stationary ergodic, whereas before solutions only for more restricted cases were known. The solutions are always constructive, that is, we present

asymptotically accurate algorithms for each of the considered problems. All the algorithms are based on empirical estimates of distributional distance, which is in the core of the suggested approach. We suggest that the proposed approach can be applied to other problems of statistical analysis of time series, with the view of establishing principled positive results, leaving the task of finding optimal algorithms for each particular problem as a topic for further research.

Here we concentrate on the following three conceptually simple problems: goodness-of-fit (or identity) testing, process classification, and the change point problem. A somewhat more technical problem of time-series clustering will be considered in Section 3.4.

Identity testing. The first problem is the following problem of hypothesis testing. A stationary ergodic process distribution ρ is known theoretically. Given a data sample, it is required to test whether it was generated by ρ, versus it was generated by any other stationary ergodic distribution that is different from ρ (goodness-of-fit, or identity testing). The case of i.i.d. or finite-memory processes was widely studied (see e.g. [21]); in particular, when ρ has a finite memory [81] proposes a test against any stationary ergodic alternative: a test that can be based on an arbitrary universal code. It was noted in [83] that an asymptotically accurate test for the case of stationary ergodic processes over finite alphabet exists (but no test was proposed). Here we propose a concrete and simple asymptotically accurate goodness-of-fit test, which demonstrates the proposed approach: to use empirical distributional distance for hypotheses testing. By asymptotically accurate test we mean the following. First, the Type I error of the test (or its size) is fixed and is given as a parameter to the test. That is, given any $\alpha > 0$ as an input, under H_0 (that is, if the data sample was indeed generated by ρ) the probability that the test says "H_1" is not greater than α. Second, under any hypothesis in H_1 (that is, if the distribution generating the data is different from ρ), the test will say "H_0" not more than a finite number of times, with probability 1. In other words, the Type I error of the test is

fixed and the Type II error can be made not more than a finite number of times, as the data sample increases, with probability 1 under any stationary ergodic alternative.

Process classification. In the next problem that we consider, we again have to decide whether a data sample was generated by a process satisfying a hypothesis H_0 or a hypothesis H_1. However, here H_0 and H_1 are not known theoretically, but are represented by two additional data samples. More precisely, the probelm is that of process classification, which can be formulated as follows. We are given three samples $X = (X_1,...,X_k)$, $Y = (Y_1,...,Y_m)$ and $Z = (Z_1,...,Z_n)$ generated by stationary ergodic processes with distributions ρ_X, ρ_Y and ρ_Z. It is known that $\rho_X \neq \rho_Y$ but either $\rho_Z = \rho_X$ or $\rho_Z = \rho_Y$. It is required to test which one is the case. That is, we have to decide whether the sample Z was generated by the same process as the sample X or by the same process as the sample Y. This problem for the case of dependent time series was considered for example in [37], where a solution is presented under the finite-memory assumption. It is closely related to many important problems in statistics and application areas, such as pattern recognition. Apparently, no asymptotically accurate procedure for process classification has been known so far for the general case of stationary ergodic processes. Here we propose a test that converges almost surely to the correct answer. In other words, the test makes only a finite number of errors with probability 1, with respect to any stationary ergodic processes generating the data. Unlike in the previous problem, here we do not explicitly distinguish between Type I and Type II error, since the hypotheses are by nature symmetric: H_0 is "$\rho_Z = \rho_X$" and H_1 is "$\rho_Z = \rho_Y$".

Change point estimation. Finally, we consider the change point problem. It is another classical problem, with vast literature on both parametric (see e.g. [7]) and non-parametric (see e.g. [15]) methods for solving it. In this section we address the case where the data is dependent, its form and the structure of dependence is unknown, and, importantly, marginal distributions before and after the change may be the same. We consider the

following (off-line) setting of the problem: a (real-valued) sample $Z_1,...,Z_n$ is given, where $Z_1,...,Z_k$ are generated according to some distribution ρ_X and $Z_{k+1},...,Z_n$ are generated according to some distribution ρ_Y which is different from ρ_X. It is known that the distributions ρ_X and ρ_Y are stationary ergodic, but nothing else is known about them.

Most literature on change point problem for dependent time series assumes that the marginal distributions before and after the change point are different, and often also make explicit restrictions on the dependence, such as requirements on mixing rates. Nonparametric methods used in these cases are typically based on Kolmogorov-Smirnov statistic, Cramer-von Mises statistic, or generalizations thereof [15, 16, 35]. The main difference with our results is that we do not assume that the single-dimensional marginals (or finite-dimensional marginals of any given fixed size) are different, and do not make any assumptions on the structure of dependence. The only assumption is that the (unknown) process distributions before and after the change point are stationary ergodic.

Methodology. All the tests that we construct are based on empirical estimates of the so-called distributional distance. For two processes ρ_1,ρ_2 a distributional distance is defined as $\sum_{k=1}^{\infty} w_k|\rho_1(B_k)-\rho_2(B_k)|$, where w_k are positive summable real weights, e.g. $w_k=2^{-k}$ and B_k range over a countable field that generates the sigma-algebra of the underlying probability space. For example, if we are talking about finite-alphabet processes with the binary alphabet $A=\{0,1\}$, B_k would range over the set $A^*=\cup_{k\in\mathbb{N}}A^k$; that is, over all tuples 0,00,01,10,000,001,... ; therefore, the distributional distance in this case is the weighted sum of differences of probabilities of all possible tuples. In this section we consider real-valued processes, $A=\mathbb{R}$, so B_k can be taken to range over all intervals with rational endpoints, all pairs of such intervals, triples, etc.

Although distributional distance is a natural concept that, for stochastic processes, has been studied for a while [36], its empirical estimates have not, to our knowledge, been used for statistical analysis of time series. We

argue that this distance is rather natural for this kind of problems, first of all, since it can be consistently estimated (unlike, for example, $\bar{d}$ distance, which cannot [69] be consistently estimated for the general case of stationary ergodic processes). Secondly, it is always bounded, unlike (empirical) KL divergence, which is often used for statistical inference for time series (e.g. [21, 81, 2, 20] and others). Other approaches to statistical analysis of stationary dependent time series include the use of (universal) codes [50, 81, 80]. Here we first show that distributional distance between stationary ergodic processes can be consistently estimated based on sampling, and then apply it to construct a consistent test for the three problems of statistical analysis described above.

Although empirical estimates of the distributional distance involve taking an infinite sum, in practice it is obvious that only a finite number of summands has to be calculated. This is due to the fact that empirical estimates have to be compared to each other or to theoretically known probabilities, and since the (bounded) summands have (exponentially) decreasing weights, the result of the comparison is known after only finitely many evaluations (see a more formal discussion on this in Section 3.4 on time-series clustering). Therefore, the algorithms presented can be applied in practice. On the other hand, the main value of the results is in the demonstration of what is possible in principle; finding practically efficient procedures for each of the considered problems is an interestring problem for further research.

3.2.1 Goodness-of-fit

For a given stationary ergodic process measure ρ and a sample $X=(X_1,...,X_n)$ we wish to test the hypothesis H_0 that the sample was generated by ρ versus H_1 that it was generated by a stationary ergodic distribution that is different from ρ. Thus, $H_0=\{\rho\}$ and $H_1=\mathcal{E}\setminus H_0$.

Define the set D_δ^n as the set of all samples of length n that are at least

δ-far from ρ in empirical distributional distance:

$$D_\delta^n := \{X \in A^n : \hat{d}(X,\rho) \geq \delta\}.$$

For each n and each given confidence level α define the critical region C_α^n of the test as $C_\alpha^n := D_\gamma^n$ where

$$\gamma := \inf\{\delta : \rho(D_\delta^n) \leq \alpha\}.$$

The test rejects H_0 at confidence level α if $(X_1,...,X_n) \in C_\alpha^n$ and accepts it otherwise. In words, for each sequence we measure the distance between the empirical probabilities (frequencies) and the measure ρ (that is, the theoretical ρ-probabilities); we then take a largest ball (with respect to this distance) around ρ that has ρ-probability not greater than $1-\alpha$. The test rejects all sequences outside this ball.

Definition 3.4 (Goodness-of-fit test)**.** *For each $n \in \mathbb{N}$ and $\alpha \in (0,1)$ the goodness-of-fit test $G_n^\alpha : A^n \to \{0,1\}$ is defined as*

$$G_n^\alpha(X_1,...,X_n) := \begin{cases} 1 & \text{if } (X_1,...,X_n) \in C_\alpha^n, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 3.5. *The test G_n^α has the following properties.*

(i) For every $\alpha \in (0,1)$ and every $n \in \mathbb{N}$ the Type I error of the test is not greater than α: $\rho(G_n^\alpha = 1) \leq \alpha$.

(ii) For every $\alpha \in (0,1)$ the Type II error goes to 0 almost surely: for every $\rho' \neq \rho$ we have $\lim_{n\to\infty} G_n^\alpha = 1$ with ρ' probability 1.

Note that using an appropriate randomization in the definition of C_α^n we can make the Type I error exactly α.

Proof. The first statement holds by construction. To prove the second statement, let the sample X be generated by $\rho' \in \mathcal{E}$, $\rho' \neq \rho$, and define $\delta = d(\rho,\rho')/2$.

By Lemma 3.3 we have $\rho(D_\delta^n) \to 0$, so that $\rho(D_\delta^n) < \alpha$ from some n on; denote it n_1. Thus, for $n > n_1$ we have $D_\delta^n \subset C_\alpha^n$. At the same time, by Lemma 3.3 we have $\hat{d}(X,\rho) > \delta$ from some n on, which we denote $n_2(X)$, with ρ'-probability 1. So, for $n > \max\{n_1, n_2(X)\}$ we have $X \in D_\delta^n \subset C_\alpha^n$, which proves the statement *(ii)*. □

3.2.2 Process classification

Let there be given three samples $X = (X_1,...,X_k)$, $Y = (Y_1,...,Y_m)$ and $Z = (Z_1,...,Z_n)$. Each sample is generated by a stationary ergodic process ρ_X, ρ_Y and ρ_Z respectively. Moreover, it is known that either $\rho_Z = \rho_X$ or $\rho_Z = \rho_Y$, but $\rho_X \neq \rho_Y$. We wish to construct a test that, based on the finite samples X,Y and Z will tell whether $\rho_Z = \rho_X$ or $\rho_Z = \rho_Y$.

The test chooses the sample X or Y according to whichever is closer to Z in $\hat{d}$. That is, we define the test $G(X,Y,Z)$ as follows. If $\hat{d}(X,Z) \leq \hat{d}(Y,Z)$ then the test says that the sample Z is generated by the same process as the sample X, otherwise it says that the sample Z is generated by the same process as the sample Y.

Definition 3.6 (Process classifier)**.** *Define the classifier* $L : A^* \times A^* \times A^* \to \{1,2\}$ *as follows*

$$L(X,Y,Z) := \begin{cases} 1 & \text{if } \hat{d}(X,Z) \leq \hat{d}(Y,Z) \\ 2 & \text{otherwise,} \end{cases}$$

for $X,Y,Z \in A^*$.

Theorem 3.7. *The test* $L(X,Y,Z)$ *makes only a finite number of errors when* $|X|,|Y|$ *and* $|Z|$ *go to infinity, with probability 1: if* $\rho_X = \rho_Z$ *then*

$$L(X,Y,Z) = 1$$

from some $|X|,|Y|,|Z|$ *on with probability 1; otherwise*

$$L(X,Y,Z)=2$$

from some $|X|,|Y|,|Z|$ *on with probability 1.*

Proof. From the fact that d is a metric and from Lemma 3.3 we conclude that $\hat{d}(X,Z)\to 0$ (with probability 1) if and only if $\rho_X=\rho_Z$. So, if $\rho_X=\rho_Z$ then by assumption $\rho_Y\neq\rho_Z$ and $\hat{d}(X,Z)\to 0$ a.s. while

$$\hat{d}(Y,Z)\to d(\rho_Y,\rho_Z)\neq 0.$$

Thus in this case $\hat{d}(Y,Z)>\hat{d}(X,Z)$ from some $|X|,|Y|,|Z|$ on with probability 1, from which moment we have $L(X,Y,Z)=1$. The opposite case is analogous. □

3.2.3 Change point problem

The sample $Z=(Z_1,...,Z_n)$ consists of two concatenated parts $X=(X_1,...,X_k)$ and $Y=(Y_1,...,Y_m)$, where $m=n-k$, so that $Z_i=X_i$ for $1\leq i\leq k$ and $Z_{k+j}=Y_j$ for $1\leq j\leq m$. The samples X and Y are generated independently by two different stationary ergodic processes with alphabet $A=\mathbb{R}$. The distributions of the processes are unknown. The value k is called the *change point*. It is assumed that k is linear in n; more precisely, $\alpha n<k<\beta n$ for some $0<\alpha\leq\beta<1$ from some n on.

It is required to estimate the change point k based on the sample Z.

For each t, $1\leq t\leq n$, denote U^t the sample $(Z_1,...,Z_t)$ consisting of the first t elements of the sample Z, and denote V^t the remainder $(Z_{t+1},...,Z_n)$.

Definition 3.8 (Change point estimator)**.** *Define the change point estimate* $\hat{k}:A^*\to\mathbb{N}$ *as follows:*

$$\hat{k}(X_1,...,X_n):=\operatorname{argmax}_{t\in[\alpha n,n-\beta n]}\hat{d}(U^t,V^t).$$

The following theorem establishes asymptotic consistency of this estimator.

Theorem 3.9. *For the estimate $\hat{k}$ of the change point k we have*

$$|\hat{k}-k| = o(n) \quad a.s.$$

where n is the size of the sample, and when $k, n-k \to \infty$ in such a way that $\alpha < \frac{k}{n} < \beta$ for some $\alpha, \beta \in (0,1)$ from some n on.

The proof is deferred to Section 3.6.1.

3.3 Characterizing families of stationary processes for which consistent tests exist [R2]

Given a sample $X_1,...,X_n$, where, for the sake of this section, X_i are from a finite alphabet A, we wish to decide whether it was generated by a distribution belonging to a family H_0, versus it was generated by a distribution belonging to a family H_1. The only assumption we are willing to make about the the distribution generating the sample is that it is stationary ergodic.

A test is a function that takes a sample and gives a binary (possibly incorrect) answer: the sample was generated by a distribution from H_0 or from H_1. An answer $i \in \{0,1\}$ is correct if the sample is generated by a distribution that belongs to H_i. Here we are concerned with characterizing those pairs of H_0 and H_1 for which consistent tests exist. There are several ways of formalizing what is a consistent test, from which we consider two. For these two notions of consistency we find some necessary and some sufficient conditions for the existence of a consistent test, expressed in topological terms. For one notion of consistency (asymmetric testing) considered, the necessary and sufficient conditions coincide when H_1 is the complement of H_0, thereby providing a complete characterization of the hypotheses for which consistent tests exist.

Examples. Before introducing the definitions of consistency, let us give some examples motivating the general problem in question. Most of these examples are classical problems studied in mathematical statistics and related fields, mostly for i.i.d. data, with much literature devoted to each of them. (An incomplete survey of related work for non-i.i.d. data is given further in this section.) The most basic case of the hypothesis testing problem is testing a simple hypothesis $H_0=\{\rho_0\}$ versus a simple hypothesis $H_1=\{\rho_1\}$, where ρ_0 and ρ_1 are two stationary ergodic process distributions (which are assumed completely known theoretically). A more complex but more realistic problem is when only one of the hypothesis is simple, $H_0=\{\rho_0\}$ but the alternative is general, for example H_1 is the set of all stationary ergodic processes that are different from ρ_0. This is the goodness-of-fit problem that we have considered in Section 3.2. One may also consider variants in which the alternative is the set of all stationary ergodic processes that differ from ρ_0 *by at least* ε in some distance. The described hypotheses are variants of the so-called goodness-of-fit, or identity testing problem. Another class of hypothesis testing problems is presented by the problem of *model verification*. Suppose we have some relatively simple (possibly parametric) set of assumptions, and we wish to test whether the process generating the given sample satisfies this assumptions. As an example, H_0 can be the set of all k-order Markov processes (fixed $k\in\mathbb{N}$) and H_1 is the set of all stationary ergodic processes that do not belong to H_0; one may also wish to consider more restrictive alternatives, for example H_1 is the set of all k'-order Markov processes where $k'>k$. Of course, instead of Markov processes one can consider other models, e.g. Hidden Markov processes. A similar problem is that of testing that the process has entropy less than some given ε versus its entropy exceeds ε, or versus its entropy is greater than $\varepsilon+\delta$ for some positive δ.

Yet another type of hypothesis testing problems concerns *property testing*. Suppose we are given two samples, generated independently of each other by stationary ergodic distributions, and we wish to test the hypoth-

esis that they are independent versus they are not independent. Or, that they are generated by the same process versus they are generated by different processes.

In all the considered cases, when the hypothesis testing problem turns out to be too difficult (i.e. there is no consistent test for the chosen notion of consistency) for the case of stationary ergodic processes, one may wish to restrict either H_0, H_1 or both H_0 and H_1 to some smaller class of processes. Thus, one may wish to test the hypothesis of independence when, for example, both processes are known to have finite memory, or to have certain mixing rates.

All the problems described above are special cases of the following general formulation: given two sets H_0 and H_1 which are contained in the set of all stationary ergodic process distributions, and given a sample generated by a process that comes from either H_0 or H_1, we would like have a test that tells us which one is the case: H_0 or H_1. The goal of this section is to characterize those pairs of H_0,H_1 for which a consistent test exists. Ideally, the characterization should be complete, that is, in the form of necessary and sufficient conditions, that can be verified for at least most of the problems outlined above. This goal is partially achieved: for two (rather natural) notions of consistency, we find some necessary and some sufficient conditions, that, for one of these notions of consistency, coincide in the case when H_1 is the complement of H_0. We show that these conditions are indeed relatively easy to verify for some of the considered hypotheses, such as identity testing, model verification and testing independence.

In this section we will use the following notions of consistency of tests (see Section 3.3.1 for formal definitions). The first one is the same that was introduced in Section 3.2: the Type I error (level) is fixed and the probability of Type II error is required to go to 0 as the sample size increases. To distinguish it from other notions of consistency that we will consider, we call it *asymmetric consistency*.

The second notion of consistency that we will consider is uniform consis-

tency. For two hypothesis H_0 and H_1, a test is called *uniformly consistent*, if for any $\varepsilon > 0$ there is a sample size n such that the probability of error on a sample of size larger than n is not greater than ε if any distribution from $H_0 \cup H_1$ is chosen to generate the sample. Thus, a uniformly consistent test provides performance guarantees for finite sample sizes.

Prior work. There is a vast body of literature on hypothesis testing for i.i.d. (real- or discrete-valued) data (see e.g. [58, 49]). In the context of discrete-valued i.i.d. data, the necessary and sufficient conditions for the two types of consistency introduced are rather simple. There is an asymmetrically consistent test if and only if the closure of H_0 does not intersect H_1, and there is a uniformly consistent test if and only if the closures of H_0 and H_1 are disjoint, where the topology is that of the parameter space (probabilities of each symbol), see e.g. [19]. Some extensions to Markov chains are also possible [9, 6].

There is, however, much less literature on hypothesis testing beyond i.i.d. or parametric models, while the question of determining whether a consistent test exists, for different notions of consistency and different hypotheses, is much less trivial. For a weaker notion of consistency, namely, requiring that the test should stabilize on the correct answer for a.e. realization of the process (under either H_0 or H_1), [50] constructs a consistent test for so-called constrained finite-state model classes (including finite-state Markov and hidden Markov processes), against the general alternative of stationary ergodic processes. For the same notion of consistency, [65] gives sufficient conditions on two hypotheses H_0 and H_1 that consist of stationary ergodic real-valued processes, under which a consistent test exists, extending the results of [25] for i.i.d. data. The latter condition is that H_0 and H_1 are contained in disjoint F_σ sets (countable unions of closed sets), with respect to the topology of weak convergence. In [62] some results are presented on testing the hypothesis that the process has a finite memory, and some related problems. Asymmetrically consistent tests for some specific hypotheses, but under the general alternative of stationary ergodic processes, have been

proposed in Section 3.2 above (see also references therein).

The results. Here we obtain some topological characterizations of the hypotheses for which consistent tests exist, for the case of stationary ergodic distributions. The obtained characterizations are rather similar to those mentioned above for the case of i.i.d. data, but are with respect to the topology of distributional distance. The fact that necessary and sufficient conditions are obtained for one of the notions of consistency, indicates that this topology is the right one to consider.

The tests that we construct are based on empirical estimates of distributional distance. In particular, the *uniform test* φ_{H_0,H_1} outputs 0 if the given sample is closer to the (closure of) H_0 than to the (closure of) H_1, and outputs 1 otherwise. The *asymmetric test* $\psi^{\alpha}_{H_0,H_1}$, for a given level α, takes the smallest ε-neighbourhood of the closure of H_0 that has probability not less than $1-\alpha$ with respect to any distribution in it, and outputs 0 if the sample falls into this neighbourhood, and 1 otherwise.

This is a generalization of the goodness-of-fit procedure introduced in Section 3.2.

For the case of testing H_0 against its complement to the set $\mathcal{E}$ of all stationary ergodic processes, we obtain the following necessary and sufficient condition (formalized in the next section).

Theorem. There exists an asymmetrically consistent test for H_0 against $H_1 := \mathcal{E} \setminus H_0$ if and only if H_1 has probability 0 with respect to ergodic decomposition of every distribution from the closure of H_0. In this case, the test $\psi^{\alpha}_{H_0,H_1}$ is asymmetrically consistent too.

For the general case, as well as for the case of uniform consistency, we obtain some necessary and some sufficient conditions, in the same terms. The main results are illustrated with derivations of several known and some new results for specific hypotheses. In particular, we show that any set of processes which is continuously parametrized by a compact set of parameters, and is closed under taking ergodic decompositions, can be tested with asymmetric consistency against its complement to the set of all stationary ergodic

processes. Such parametric families include k-order Markov processes and k-state Hidden Markov processes.

3.3.1 Definitions: consistency of tests

A **test** is a function $\varphi: A^* \to \{0,1\}$ that takes a sample and outputs a binary answer, where the answer i is interpreted as "the sample was generated by a distribution that belongs to H_i". The answer i is correct if the sample was indeed generated by a distribution from H_i, otherwise we say that the test made an **error**. A test φ makes the **Type I** error if it says 1 while H_0 is true, and it makes **Type II** error if it says 0 while H_1 is true.

Call a family of tests $\psi^\alpha, \alpha \in (0,1)$ **asymmetrically consistent** if: (i) The probability of Type I error is always bounded by α: $\rho\{X \in A^n : \psi^\alpha(X) = 1\} \leq \alpha$ for every $\rho \in H_0$, every $n \in \mathbb{N}$ and every $\alpha \in (0,1)$, and (ii) Type II error is made not more than a finite number of times with probability 1: $\rho(\lim_{n\to\infty} \psi^\alpha(X_{1..n}) = 1) = 1$ for every $\rho \in H_1$ and every $\alpha \in (0,1)$. (Abusing the notation, we will sometimes call families of tests $\psi^\alpha, \alpha \in (0,1)$ simply tests.)

A test φ is called **uniformly consistent** if for every α there is an $n_\alpha \in \mathbb{N}$ such that for every $n \geq n_\alpha$ the probability of error on a sample of size n is less than α: $\rho(X \in A^n : \varphi(X) = i) < \alpha$ for every $\rho \in H_{1-i}$ and every $i \in \{0,1\}$.

3.3.2 Topological characterizations

The tests presented below are based on *empirical estimates of the distributional distance* d:

$$\hat{d}(X_{1..n}, \rho) = \sum_{i=1}^{\infty} w_i |\nu(X_{1..n}, B_i) - \rho(B_i)|,$$

where $n \in \mathbb{N}$, $\rho \in \mathcal{S}$, $X_{1..n} \in A^n$. That is, $\hat{d}(X_{1..n}, \rho)$ measures the discrepancy between empirically estimated and theoretical probabilities. For a sample

$X_{1..n} \in A^n$ and a hypothesis $H \subset \mathcal{E}$ define

$$\hat{d}(X_{1..n},H) = \inf_{\rho \in H} \hat{d}(X_{1..n},\rho).$$

For $H \subset \mathcal{S}$, denote $\mathrm{cl}H$ the closure of H (with respect to the topology of d).

For $H_0,H_1 \subset \mathcal{S}$, the **uniform test** φ_{H_0,H_1} is constructed as follows. For each $n \in \mathbb{N}$ let

$$\varphi_{H_0,H_1}(X_{1..n}) := \begin{cases} 0 & \text{if } \hat{d}(X_{1..n},\mathrm{cl}H_0 \cap \mathcal{E}) < \hat{d}(X_{1..n},\mathrm{cl}H_1 \cap \mathcal{E}), \\ 1 & \text{otherwise.} \end{cases} \tag{3.6}$$

Since the set $\mathcal{S}$ is a complete separable metric space, it is easy to see that the function $\varphi_{H_0,H_1}(X_{1..n})$ is measurable provided $\mathrm{cl}H_0$ is measurable.

Theorem 3.10 (uniform testing)**.** *Let H_0,H_1 be measurable subsets of $\mathcal{E}$. If $W_\rho(H_i) = 1$ for every $\rho \in \mathrm{cl}H_i$ then the test φ_{H_0,H_1} is uniformly consistent. Conversely, if there exists a uniformly consistent test for H_0 against H_1 then $W_\rho(H_{1-i}) = 0$ for any $\rho \in clH_i$.*

The proofs are deferred to section 3.6.

Construct the **asymmetric test** $\psi^\alpha_{H_0,H_1}, \alpha \in (0,1)$ as follows. For each $n \in \mathbb{N}$, $\delta > 0$ and $H \subset \mathcal{E}$ define the neighbourhood $b^n_\delta(H)$ of n-tuples around H as

$$b^n_\delta(H) := \{X \in A^n : \hat{d}(X,H) \le \delta\}.$$

Moreover, let

$$\gamma_n(H,\theta) := \inf\{\delta : \inf_{\rho \in H} \rho(b^n_\delta(H)) \ge \theta\}$$

be the smallest radius of a neighbourhood around H that has probability not less than θ with respect to any process in H, and let $C^n(H,\theta) := b^n_{\gamma_n(H,\theta)}(H)$ be a neighbourhood of this radius. Define

$$\psi^\alpha_{H_0,H_1}(X_{1..n}) := \begin{cases} 0 & \text{if } X_{1..n} \in C^n(\mathrm{cl}H_0 \cap \mathcal{E}, 1-\alpha), \\ 1 & \text{otherwise.} \end{cases}$$

Again, it is easy to see that the function $\varphi_{H_0,H_1}(X_{1..n})$ is measurable, since the set $\mathcal{S}$ is separable. We will often omit the subscript H_0,H_1 from $\psi^\alpha_{H_0,H_1}$ when it can cause no confusion.

Theorem 3.11. *Let H_0,H_1 be measurable subsets of $\mathcal{E}$. If $W_\rho(H_0)=1$ for every $\rho\in \mathrm{cl}H_0$ then the test $\psi^\alpha_{H_0,H_1}$ is asymmetrically consistent. Conversely, if there is an asymmetrically consistent test for H_0 against H_1 then $W_\rho(H_1)=0$ for any $\rho\in \mathrm{cl}H_0$.*

For the case when H_1 is the complement of H_0 the necessary and sufficient conditions of Theorem 3.11 coincide and give the following criterion.

Corollary 3.12. *Let $H_0\subset\mathcal{E}$ be measurable and let $H_1=\mathcal{E}\setminus H_0$. The following statements are equivalent:*

(i) There exists an asymmetrically consistent test for H_0 against H_1.

(ii) The test $\psi^\alpha_{H_0,H_1}$ is asymmetrically consistent.

(iii) The set H_1 has probability 0 with respect to ergodic decomposition of every ρ in the closure of H_0: $W_\rho(H_1)=0$ for each $\rho\in\mathrm{cl}H_0$.

3.3.3 Examples

Theorems 3.11 and 3.10 can be used to check whether a consistent test exists for such problems as identity, independence, estimating the order of a (Hidden) Markov model, bounding entropy, bounding distance, uniformity, monotonicity, etc. Some of these examples are considered in this section.

Example 1: Simple hypotheses, Identity. First of all, it is obvious that sets that consisit of just one or finitely many stationary ergodic processes are closed and closed under ergodic decompositions; therefore, for any pair of disjoint sets of this type, there exists a uniformly consistent test. (In particular, there is a uniformly consistent test for $H_0=\{\rho_0\}$ against $H_1=\{\rho_1\}$.) A more interesting case is identity testing, or goodness-of-fit, introduced in

Section 3.2: the problem here consists in testing whether a distribution generating the sample obeys a certain given law, versus it does not. Let $\rho \in \mathcal{E}$, $H_0 = \{\rho\}$ and $H_1 = \mathcal{E} \backslash H_0$. Then there is an asymmetrically consistent test for H_0 against H_1. The conditions of Theorem 3.12 are easily verified for this case, so that we recover Theorem 3.5.

As far as uniform testing is concerned, it is, first of all, clear that for any ρ_0 there is no uniformly consistent test for identity. More generally, for any non-empty H_0 there is no uniformly consistent test for H_0 against $\mathcal{E} \backslash H_0$ provided the latter complement is also non-empty. Indeed, this follows from Theorem 3.10 since in these cases the closures of H_0 and H_1 are not disjoint. One might suggest at this point that a uniformly consistent test exists if we restrict H_1 to those processes that are sufficiently far from ρ_0. However, this is not true. We can prove an even stronger negative result.

Proposition 3.13. *Let $\rho,\nu \in \mathcal{E}$, $\rho \neq \nu$ and let $\varepsilon > 0$. There is no uniformly consistent test for $H_0 = \{\rho\}$ against $H_1 = \{\nu' \in \mathcal{E} : d(\nu',\nu) \leq \varepsilon\}$.*

The proof of the proposition is deferred to the Section 3.6.2. What it means is that, while distributional distance is well suited for characterizing those hypotheses for which consistent test exist, it is not suited for *formulating the actual hypotheses*. Apparently a stronger distance is needed for the latter.

Example 2: Markov and Hidden Markov processes: bounding the order. For any k, there is an asymmetrically consistent test of the hypothesis $\mathcal{M}_k =$ "the process is Markov of order not greater than k" against $\mathcal{E} \backslash \mathcal{M}_k$. For any k, there is an asymmetrically consistent test of $\mathcal{HM}_k =$"the process is given by a Hidden Markov process with not more than k states" against $H_1 = \mathcal{E} \backslash \mathcal{HM}_k$. Indeed, in both cases (k-order Markov, Hidden Markov with not more than k states), the hypothesis H_0 is a parametric family, with a compact set of parameters, and a continuous function mapping parameters to processes (that is, to the space $\mathcal{S}$). Weierstrass theorem then implies that the image of such a compact parameter set is closed (and compact). More-

over, in both cases H_0 is closed under taking ergodic decompositions. Thus, by Theorem 3.11, there exists an asymmetrically consistent test.

The problem of estimating the order of a (hidden) Markov process, based on a sample from it, was addressed in a number of works. In the contest of hypothesis testing, asymmetrically consistent tests for $\mathcal{M}_k$ against $\mathcal{M}^t$ with $t>k$ were given in [6], see also [9]. The existence of non-uniformly consistent tests (a notion weaker than that of asymmetric consistency) for $\mathcal{M}_k$ against $\mathcal{E}\setminus\mathcal{M}_k$, and of $\mathcal{HM}_k$ against $\mathcal{E}\setminus\mathcal{HM}_k$, was established in [50]. Asymmetrically consistent tests for $\mathcal{M}_k$ against $\mathcal{E}\setminus\mathcal{M}_k$ were obtained in [80], while for the case of asymmetric testing for $\mathcal{HM}_k$ against $\mathcal{E}\setminus\mathcal{HM}_k$ the positive result above is apparently new.

Example 3: Smooth parametric families. From the discussion in the previous example we can see that the following generalization is valid. Let $H_0\subset\mathcal{S}$ be a set of processes that is continuously parametrized by a compact set of parameters. If H_0 is closed under taking ergodic decompositions, then there is an asymmetrically consistent test for H_0 against $\mathcal{E}\setminus H_0$. In particular, this strengthens the mentioned result of [50], since a stronger notion of consistency is used, as well as a more general class of parametric families is considered.

Clearly, a similar statement can be derived for uniform testing: given two disjoint sets H_0 and H_1 each of which is continuously parametrized by a compact set of parameters and is closed under taking ergodic decompositions, there exists a uniformly consistent test of H_0 against H_1.

Example 4: Independence. Suppose that $A=A_1\times A_2$, so that a sample $X_{1..n}$ consists of two processes $X^1_{1..n}$ and $X^2_{1..n}$, which we call features. The hypothesis of independence is that the first feature is independent from the second: $\rho(X^1_{1..t}\in T_1,X^2_{1..t}\in T_2)=\rho(X^1_{1..t}\in T_1)\rho(X^2_{1..t}\in T_2)$ for any $(T_1,T_2)\in A^n$ and any $n\in\mathbb{N}$. Let $\mathcal{I}$ be the set of all stationary ergodic processes satisfying this property. It is easy to see that Theorem 3.11 implies, that there exists an asymmetrically consistent test for $\mathcal{I}\cap\mathcal{M}_k$ against $\mathcal{E}\setminus\mathcal{I}$, for any given $k\in\mathbb{N}$. Analogously, if we confine H_0 to Hidden Markov processes of a

given order, then asymmetric testing is possible. That is, there exists an an asymmetrically consistent test for $\mathcal{I} \cap \mathcal{HM}_k$ against $\mathcal{E} \setminus \mathcal{I}$, for any given $k \in \mathbb{N}$. As far as uniform testing is concerned, positive results can be obtained if we restrict both H_0 and H_1 to the corresponding subset of some set continuously parametrized by a compact set of parameters, such as the sets of (Hidden) Markov processes of given order.

The question of whether $\mathcal{I}$ can be tested against $\mathcal{E} \setminus \mathcal{I}$ is more difficult. It is clear that the closure of $\mathcal{I}$ only contains processes with independent features. It is not clear whether any of the limiting points of $\mathcal{I}$ has ergodic components whose features are not independent. If there are none, this would prove that there exists an asymmetrically consistent test for independence, for the class of stationary ergodic process.

3.4 Clustering time series [R11]

In this section we use the approach developed in the previous sections to construct an algorithm for clustering time-series data and show its consistency under the general assumption that the time series are stationary ergodic.

Given a finite set of objects the problem to "cluster" similar objects together, in the absence of any examples of "good" clusterings, is notoriously hard to formalize. Most of the work on clustering is concerned with particular parametric data generating models, or particular algorithms, a given similarity measure, and (very often) a given number of clusters. It is clear that, as in almost learning problems, in clustering finding the right similarity measure is an integral part of the problem. However, even if one assumes the similarity measure known, it is hard to define what a good clustering is [52, 90]. What is more, even if one assumes the similarity measure to be simply the Euclidean distance (on the plane), and the number of clusters k known, then clustering may still appear intractable for computational reasons. Indeed, in this case finding k centres (points which minimize the cumulative distance from each point in the sample to one of the centres)

seems to be a natural goal, but this problem is NP-hard [61].

In this section we consider the problem of clustering time-series data. That is, each data point is itself a sample generated by a certain discrete-time stochastic process. This version of the problem has numerous applications, such as clustering biological data, financial observations, or behavioural patterns, and as such it has gained a tremendous attention in the literature.

The main observation that we make here is that, in the case of clustering processes, one can benefit from the notion of ergodicity to define what appears to be a very natural notion of consistency. This notion of consistency is shown to be satisfied by simple algorithms that we present, which are polynomial in all arguments.

With these considerations in mind, define the clustering problem as follows. N samples are given: $\mathbf{x}_1 = (x_1^1,...,x_{n_1}^1),...,\mathbf{x}_N = (x_1^N,...,x_{n_N}^N)$. Each sample is drawn by one out of k different stationary ergodic distributions. The samples are *not* assumed to be drawn independently; rather, it is assumed that the joint distribution of the samples is stationary ergodic. The target clustering is as follows: those and only those samples are put into the same cluster that were generated by the same distribution. As is usual in the clustering literature, the number k of target clusters is assumed to be known. A clustering algorithm is called asymptotically consistent if the probability that it outputs the target clustering converges to 1, as the lengths $(n_1,...,n_N)$ of the samples tend to infinity (a variant of this definition is to require the algorithm to stabilize on the correct answer with probability 1). Note the particular regime of asymptotic: not with respect to the number of samples N, but with respect to the length of the samples $n_1,...,n_N$.

Similar formulations have appeared in the literature before. Perhaps the most close approach is mixture models [85, 91]: it is assumed that there are k different distributions that have a particular known form (such as Gaussian, Hidden Markov models, or graphical models) and each one out of N samples is generated independently according to one of these k distributions (with some fixed probability). Since the model of the data is specified quite well,

one can use likelihood-based distances (and then, for example, the k-means algorithm), or Bayesian inference, to cluster the data. Clearly, the main difference from our setting is in that we do not assume any known model of the data; not even between-sample independence is assumed.

The problem of clustering in our formulation is close to the following two classical problems of mathematical statistics. The first one is homogeneity testing, or the two-sample problem. Two samples $\mathbf{x}_1 = (x_1^1,...,x_{n_1}^1)$ and $\mathbf{x}_2 = (x_1^2,...,x_{n_2}^2)$ are given, and it is required to test whether they were generated by the same distribution, or by different distributions. This corresponds to clustering just two data points ($N = 2$) with the number k of clusters unknown: either $k = 1$ or $k = 2$. As we show in Section 3.5, this problem is impossible to solve for stationary ergodic (binary-valued) processes, which is why we assume known k in this section. The second problem is process classification, or the three-sample problem, that we have considered in detail in Section 3.2.2: Three samples $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3$ are given, it is known that two of them were generated by the same distribution, while the third one was generated by a different distribution. It is required to find out which two were generated by the same distribution. This corresponds to clustering three data points, with the number of clusters $k=2$. The clustering algorithm that we will present in this section is therefore a generalization of the simple procedure of Section 3.2.2.

In this section we will also consider in some detail the question of calculating empirical estimates of the distributional distance (on which all the algorithms in this chapter are based). Although its definition involves infinite summation, we show that it can be easily calculated.

3.4.1 Problem formulation

The clustering problem can be defined as follows. We are given N samples $\mathbf{x}_1,...,\mathbf{x}_N$, where each sample $\mathbf{x}_i$ is a string of length n_i of symbols from A: $\mathbf{x}_i = X_{1..n_i}^i$. Each sample is generated by one out of k different *unknown*

stationary ergodic distributions $\rho_1,...,\rho_k \in \mathcal{E}$. Thus, there is a partitioning $I=\{I_1,...,I_k\}$ of the set $\{1..N\}$ into k *disjoint* subsets $I_j, j=1..k$

$$\{1..N\}=\cup_{j=1}^{k} I_j,$$

such that $\mathbf{x}_j$, $1\leq j\leq N$ is generated by ρ_j if and only if $j\in I_j$. The partitioning I is called the *target clustering* and the sets $I_i, 1\leq i\leq k$, are called the *target clusters*. Given samples $\mathbf{x}_1,...,\mathbf{x}_N$ and a target clustering I, let $I(\mathbf{x})$ denote the cluster that contains $\mathbf{x}$.

A *clustering function* F takes a finite number of samples $\mathbf{x}_1,...,\mathbf{x}_N$ and an optional parameter k (the target number of clusters) and outputs a partition $F(\mathbf{x}_1,...,\mathbf{x}_N,(k))=\{T_1,...,T_k\}$ of the set $\{1..N\}$.

Definition 3.14 (asymptotic consistency)**.** *Let a finite number N of samples be given, and let the target clustering partition be I. Define $n=\min\{n_1,...,n_N\}$. A clustering function F is strongly asymptotically consistent if*

$$F(\mathbf{x}_1,...,\mathbf{x}_N,(k))=I$$

from some n on with probability 1. A clustering function is weakly asymptotically consistent if

$$P(F(\mathbf{x}_1,...,\mathbf{x}_N,(k))=I)\to 1.$$

Note that the consistency is asymptotic with respect to *the minimal length of the sample*, and not with respect to the *number of samples*.

Since in this section we will be also interested in analysing computation complexity of the proposed methods, we will use a slightly more detailed definition of the distributional distance.

Definition 3.15. *The distributional distance is defined for a pair of pro-*

cesses ρ_1,ρ_2 *as follows*

$$d(\rho_1,\rho_2)=\sum_{m,l=1}^{\infty} w_m w_l \sum_{B\in B^{m,l}} |\rho_1(B)-\rho_2(B)|,$$

where $w_j=2^{-j}$.

The clustering algorithm below is based on empirical estimates of d:

$$\hat{d}(X^1_{1..n_1},X^2_{1..n_2})=\sum_{m,l=1}^{\infty} w_m w_l \sum_{B\in B^{m,l}} |\nu(X^1_{1..n_1},B)-\nu(X^2_{1..n_2},B)|, \tag{3.7}$$

where $n_1,n_2\in\mathbb{N}$, $\rho\in\mathcal{S}$, $X^i_{1..n_i}\in A^{n_i}$.

It is easy to check that Lemma 3.3 holds for this modified definition of d as well.

3.4.2 Clustering algorithm

Algorithm 1 is a simple clustering algorithm, which, given the number k of clusters, will be shown to be consistent under most general assumptions. It works as follows. The point $\mathbf{x}_1$ is assigned to the first cluster. Next, find the point that is farthest away from $\mathbf{x}_1$ in the empirical distributional distance $\hat{d}$, and assign this point to the second cluster. For each $j=3..k$, find a point that maximizes the minimal distance to those points already assigned to clusters, and assign it to the cluster j. Thus we have one point in each of the k clusters. Next simply assign each of the remaining points to the cluster that contains the closest points from those k already assigned. One can notice that Algorithm 1 is just one iteration of the k-means algorithm, with so-called farthest-point initialization [47], and a specially designed distance.

Proposition 3.16 (calculating $\hat{d}(\mathbf{x}_1,\mathbf{x}_2)$)**.** *For two samples* $\mathbf{x}_1=X^1_{1..n_1}$ *and* $\mathbf{x}_2=X^2_{1..n_2}$ *the computational complexity (time and space) of calculating the*

Algorithm 1 The case of known number of clusters k

INPUT: The number of clusters k, samples $\mathbf{x}_1,...,x_N$.
Initialize: $j:=1$, $c_1:=1$, $T_1:=\{x_{c_1}\}$.
for $j:=2$ to k **do**
  $c_j:=\operatorname{argmax}\{i=1,...,N:\min_{t=1}^{j-1}\hat{d}(\mathbf{x}_i,\mathbf{x}_{c_t})\}$
  $T_j:=\{x_{c_j}\}$
end for
for $i=1$ to N **do**
  Put $\mathbf{x}_i$ into the set $T_{\operatorname{argmin}_{j=1}^{k}\hat{d}(\mathbf{x}_i,\mathbf{x}_{c_j})}$
end for
OUTPUT: the sets T_j, $j=1..k$.

empirical distributional distance $\hat{d}(\mathbf{x}_1,\mathbf{x}_2)$ *(3.7) is* $O(n^2\log s_{\min}^{-1})$, *where* $n=\max(n_1,n_2)$ *and*

$$s_{\min}=\min_{i=1..n_1,j=1..n_2,X_i^1\neq X_j^2}|X_i^1-X_j^2|.$$

Proof. First, observe that for fixed m and l, the sum

$$T^{m,l}:=\sum_{B\in B^{m,l}}|\nu(X^1_{1..n_1},B)-\nu(X^2_{1..n_2},B)| \tag{3.8}$$

has not more than n_1+n_2-2m+2 non-zero terms (assuming $m\le n_1,n_2$; the other case is obvious). Indeed, for each $i=0,1$, in the sample $\mathbf{x}_i$ there are n_i-m+1 tuples of size k: $X^i_{1..m},X^i_{2..m+1},...,X^i_{n_1-m+1..n_1}$. Therefore, the complexity of calculating $T^{m,l}$ is $O(n_1+n_2-2m+2)=O(n)$. Furthermore, observe that for each m, for all $l>\log s_{\min}^{-1}$ the term $T^{m,l}$ is constant. Therefore, it is enough to calculate $T^{m,1},...,T^{m,\log s_{\min}^{-1}}$, since for fixed m

$$\sum_{l=1}^{\infty}w_mw_lT^{m,l}=w_mw_{\log s_{\min}^{-1}}T^{m,\log s_{\min}^{-1}}+\sum_{l=1}^{\log s_{\min}^{-1}}w_mw_lT^{m,l}$$

(that is, we double the weight of the last non-zero term). Thus, the complexity of calculating $\sum_{l=1}^{\infty}w_mw_lT^{m,l}$ is $O(n\log s_{\min}^{-1})$. Finally, for all $m>n$ we have $T^{m,l}=0$. Since $\hat{d}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{m,l=1}^{\infty}w_m,w_lT^{m,l}$, the statement is

proven. □

Theorem 3.17. *Let $N \in \mathbb{N}$ and suppose that the samples $\mathbf{x}_1,...,\mathbf{x}_N$ are generated in such a way that the joint distribution is stationary ergodic. If the correct number of clusters k is known, then Algorithm 1 is strongly asymptotically consistent. Algorithm 1 makes $O(kN)$ calculations of $\hat{d}(\cdot,\cdot)$, so that its computational complexity is $O(kNn_{\max}^2 \log s_{\min}^{-1})$, where $n_{\max} = \max_{i=1}^k n_i$ and $s_{\min} = \min_{u,v=1..N, u\neq v, i=1..n_u, j=1..n_v, X_i^u \neq X_j^v} |X_i^u - X_j^v|$.*

Observe that the samples are not required to be generated independently. The only requirement on the distribution of samples is that the joint distribution is stationary ergodic. This is perhaps one of the mildest possible probabilistic assumptions.

Proof. By Lemma 3.3, $\hat{d}(\mathbf{x}_i,\mathbf{x}_j)$, $i,j \in \{1..N\}$ converges to 0 if and only if $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same cluster. Since there are only finitely many samples $\mathbf{x}_i$, there exists some $\delta > 0$ such that, from some n on, we will have $\hat{d}(\mathbf{x}_i,\mathbf{x}_j) < \delta$ if $\mathbf{x}_i,\mathbf{x}_j$ belong to the same target cluster ($I(\mathbf{x}_i) = I(\mathbf{x}_j)$), and $\hat{d}(\mathbf{x}_i,\mathbf{x}_j) > \delta$ otherwise ($I(\mathbf{x}_i) \neq I(\mathbf{x}_j)$). Therefore, from some n on, for every $j \leq k$ we will have $\max\{i = 1,...,N : \min_{t=1}^{j-1} \hat{d}(\mathbf{x}_i,\mathbf{x}_{c_t})\} > \delta$ and the sample $\mathbf{x}_{c_j}$, where $c_j = \operatorname{argmax}\{i = 1,...,N : \min_{t=1}^{j-1} \hat{d}(\mathbf{x}_i,\mathbf{x}_{c_t})\}$, will be selected from a target cluster that does not contain any $\mathbf{x}_{c_i}$, $i < j$. The consistency statement follows.

Next, let us find how many pairwise distance estimates $\hat{d}(\mathbf{x}_i,\mathbf{x}_j)$ the algorithm has to make. On the first iteration of the loop, it has to calculate $\hat{d}(\mathbf{x}_i,\mathbf{x}_{c_1})$ for all $i = 1..N$. On the second iteration, it needs again $\hat{d}(\mathbf{x}_i,\mathbf{x}_{c_1})$ for all $i = 1..N$, which are already calculated, and also $\hat{d}(\mathbf{x}_i,\mathbf{x}_{c_2})$ for all $i = 1..N$, and so on: on jth iteration of the loop we need to calculate $d(\mathbf{x}_i,\mathbf{x}_{c_j})$, $i = 1..N$, which gives at most kN pairwise distance calculations in total. The statement about computational complexity follows from this and Proposition 3.16: indeed, apart from the calculation of $\hat{d}$, the rest of the computations is of order $O(kN)$. □

3.5 Discrimination between B-processes is impossible [R5]

Two series of binary observations $x_1,x_1,...$ and $y_1,y_2,...$ are presented sequentially. A *discrimination procedure* (or homogeneity test) D is a family of mappings $D_n : X^n \times X^n \to \{0,1\}$, $n \in \mathbb{N}$, $X = \{0,1\}$, that maps a pair of samples $(x_1,...,x_n)$, $(y_1,...,y_n)$ into a binary ("yes" or "no") answer: the samples are generated by different distributions, or they are generated by the same distribution.

A discrimination procedure D is *asymptotically correct for a set* $\mathcal{C}$ *of process distributions* if for any two distributions $\rho_x,\rho_y \in \mathcal{C}$ independently generating the sequences $x_1,x_2,...$ and $y_1,y_2,...$ correspondingly the expected output converges to the correct answer: the following limit exists and the equality holds

$$\lim_{n\to\infty} \mathbb{E} D_n((x_1,...,x_n),(y_1,...,y_n)) = \begin{cases} 0 & \text{if } \rho_x = \rho_y, \\ 1 & \text{otherwise.} \end{cases}$$

This is perhaps the weakest notion of correctness one can consider. Clearly, asymptotically correct discriminating procedures exist for many classes of processes, for example for the class of all i.i.d. processes (e.g. [58]) and various parametric families.

We show that there is no asymptotically correct discrimination procedure for the class of all B-processes (see the definition below), meaning that for any discrimination the expected answer does not converge to the correct one for some processes. The class of B-processes is sufficiently wide to include, for example, k-order Markov processes and functions of them, but, on the other hand, it is a strict subset of the set of stationary ergodic processes. B-processes play important role in such fields as information theory and ergodic theory [82, 67].

Previously, in [69] and [68] it was shown that consistent estimates of

$\bar{d}$-distance (defined below) for B-processes exist, while it is impossible to estimate this distance outside this class. In [69] it is also claimed that consistent discrimination procedure does not exist for the set of all stationary ergodic processes; however, what is shown in that work is that consistent estimate of $\bar{d}$-distance do not exist for this set. The result of this section is stronger than this claim: a consistent discrimination procedure does not exist for a smaller set of processes, that of all B-processes.

Next we define the $\bar{d}$ distance and B-processes (mainly following [69] in our formulations) and give more precise formulations of some of the existing results mentioned above.

For two finite-valued stationary processes ρ_x and ρ_y the *$\bar{d}$-distance* $\bar{d}(\rho_x,\rho_y)$ is said to be less than ε if there exists a single stationary process ν_{xy} on pairs (x_n,y_n), $n\in\mathbb{N}$, such that x_n, $n\in\mathbb{N}$ are distributed according to ρ_x and y_n are distributed according to ρ_y while

$$\nu_{xy}(x_1\neq y_1)\leq\varepsilon. \tag{3.9}$$

The infimum of the ε's for which a coupling can be found such that (3.9) is satisfied is taken to be the $\bar{d}$-distance between ρ_x and ρ_y.

Definition 3.18. *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 [67]. As it was mentioned, [69] constructs an estimator $\bar{s}_n$ such that

$$\lim_{n\to\infty}\bar{s}_n((x_1,...,x_n),(y_1,...,y_n))=\bar{d}(\rho_1,\rho_2)\quad \rho_1\times\rho_2\text{–a.s.} \tag{3.10}$$

if both processes ρ_1 and ρ_2 generating the samples x_i and y_i respectively are B-processes. In the same work it is shown that there is no estimator $\bar{s}_n$ for which (3.10) holds for every pair ρ_1,ρ_2 of stationary ergodic processes. Some extensions of these results are given in [68].

It is interesting to compare these results to those that are obtained for the distributional distance. As we have shown (Lemma 3.3), this distance can be consistently estimated. Moreover, based on its estimate, we can also construct a consistent change point estimate, as was demonstrated in Section 3.2 above. On the other hand, the results of the present section implies that one cannot consistently tell whether there is a change in the sample or not.

Summarizing, we can say that the stronger the distance the harder it is to estimate: the distributional distance can be consistently estimated for stationary ergodic processes, the $\bar{d}$ distance can be consistently estimated for B-processes but not for stationary ergodic processes, while the strongest possible distance— the one that gives discrete topology, cannot be consistently estimated for B-processes, as is shown in this section.

The **main result** of this section is the following theorem.

Theorem 3.19. *There is no asymptotically correct discrimination procedure for the set of all B-processes.*

The proof, which we defer to Section3.6.3, is by contradiction. It is assumed that a consistent discrimination procedure exists, and a process is exhibited that will trick such a procedure to give divergent results. The construction on which the proof is based uses the ideas of the construction of B. Ryabko used in [78] to demonstrate that consistent prediction for stationary ergodic processes is impossible (see also the modification of this construction in [38]).

3.6 Longer proofs

3.6.1 Proof of Theorem 3.9

Proof. To prove the statement, we will show that for every γ, $0 < \gamma < 1$ with probability 1 the inequality $\hat{d}(U^t,V^t) < \hat{d}(X,Y)$ holds for each t such

that $\alpha k \le t < \gamma k$ possibly except for a finite number of times. Thus we will show that linear γ-underestimates occur only a finite number of times, and for overestimate it is analogous. Fix some γ, $0<\gamma<1$ and $\varepsilon>0$. Let J be big enough to have $\sum_{i=J}^{\infty} w_i < \varepsilon/2$ and also big enough to have an index $j<J$ for which $\rho_X(B_j) \ne \rho_Y(B_j)$. Take $M_\varepsilon \in \mathbb{N}$ large enough to have $|\nu(Y,B_i)-\rho_Y(B_i)| \le \varepsilon/2J$ for all $m > M_\varepsilon$ and for each i, $1 \le i \le J$, and also to have $|B_i|/m < \varepsilon/J$ for each i, $1 \le i \le J$. This is possible since empirical frequencies converge to the limiting probabilities a.s. (that is, M_ε depends on the realizations $Y_1,Y_2,\dots$) (cf. the proof of Lemma 3.3). Find a K_ε (that depends on X) such that for all $k > K_\varepsilon$ and for all i, $1 \le i \le J$ we have

$$|\nu(U^t,B_i)-\rho_X(B_i)| \le \varepsilon/2J \text{ for each } t \in [\alpha n,\dots,k] \tag{3.11}$$

(this is possible simply because $\alpha n \to \infty$). Furthermore, we can select K_ε large enough to have $|\nu((X_s,X_{s+1},\dots,X_k),B_i)-\rho_X(B_i)| \le \varepsilon/2J$ for each $s \le \gamma k$: this follows from (3.11) and the indentity $\nu((X_s,X_{s+1},\dots,X_k) = \frac{k}{k-s}\nu((X_1,\dots,X_k)-\frac{s-1}{k-s}\nu(X_1,\dots,X_{s-1})+o(1)$.

So, for each $s \in [\alpha n,\gamma k]$ we have

$$\begin{aligned}\left|\nu(V^s,B_j)-\frac{(1-\gamma)k\rho_X(B_j)+m\rho_Y(B_j)}{(1-\gamma)k+m}\right| & \\ \le \left|\frac{(1-\gamma)k\nu((X_s,\dots,X_k),B_j)+m\nu(Y,B_j)}{(1-\gamma)k+m}-\right. & \\ \left.\frac{(1-\gamma)k\rho_X(B_j)+m\rho_Y(B_j)}{(1-\gamma)k+m}\right| + \frac{|B_j|}{m+\gamma k} & \le 3\varepsilon/J,\end{aligned}$$

for $k > K_\varepsilon$ and $m > M_\varepsilon$ (from the definitions of K_ε and M_ε). Hence

$$\begin{aligned}
&|\nu(X,B_j)-\nu(Y,B_j)|-|\nu(U^s,B_j)-\nu(V^s,B_j)| \\
&\geq |\nu(X,B_j)-\nu(Y,B_j)| \\
&\qquad -\left|\nu(U^s,B_j)-\frac{(1-\gamma)k\rho_X(B_j)+m\rho_Y(B_j)}{(1-\gamma)k+m}\right|-3\varepsilon/J \\
&\geq |\rho_X(B_j)-\rho_Y(B_j)| \\
&\qquad -\left|\rho_X(B_j)-\frac{(1-\gamma)k\rho_X(B_j)+m\rho_Y(B_j)}{(1-\gamma)k+m}\right|-4\varepsilon/J \\
&\qquad\qquad =\delta_j-4\varepsilon/J,
\end{aligned}$$

for some δ_j that depends only on k/m and γ. Summing over all B_i, $i \in \mathbb{N}$, we get

$$\hat{d}(X,Y)-\hat{d}(U^s,V^s) \geq w_j\delta_j-5\varepsilon,$$

for all n such that $k > K_\varepsilon$ and $m > M_\varepsilon$, which is positive for small enough ε. □

3.6.2 Proofs for Section 3.3

The proofs will use the following lemmas.

Lemma 3.20 (smooth probabilities of deviation)**.** *Let $m > 2k > 1$, $\rho \in \mathcal{S}$, $H \subset \mathcal{S}$, and $\varepsilon > 0$. Then*

$$\rho(\hat{d}(X_{1..m},H) \geq \varepsilon) \leq \rho\left(\hat{d}(X_{1..k},H) \geq \varepsilon - \frac{2k}{m-k+1} - t_k\right), \tag{3.12}$$

where t_k is the sum of all the weights of tuples longer than k in the definition of d: $t_k := \sum_{i:|B_i|>n} w_i$, and

$$\rho(\hat{d}(X_{1..m},H) \leq \varepsilon) \leq \rho\left(\hat{d}(X_{1..k},H) \leq \frac{m}{m-k+1}\varepsilon + \frac{2k}{m-k+1}\right). \tag{3.13}$$

The meaning of this lemma is as follows. For any word $X_{1..m}$, if it is far away from (or close to) a given distribution μ (in the empirical distributional distance), then some of its shorter subwords $X_{i..i+k}$ is far from (close to) μ too. By stationarity, we may assume that $i=1$. Therefore, the probability of a δ-ball of samples of a given length is close to the probability of a δ-ball of samples of smaller size. In other words, for a stationary distribution μ, it cannot happen that a small sample is likely to be close to μ, but a larger sample is likely to be far.

Proof. Let B be a tuple such that $|B|<k$ and $X_{1..m}\in A^m$ be any sample of size $m>1$. The number of occurrences of B in X can be bounded by the number of occurrences of B in subwords of X of length k as follows:

$$\#(X_{1..m},B)\leq\frac{1}{k-|B|+1}\sum_{i=1}^{m-k+1}\#(X_{i..i+k-1},B)+2k$$
$$=\sum_{i=1}^{m-k+1}\nu(X_{i..i+k-1},B)+2k.$$

Indeed, summing over $i=1..m-k$ the number of occurrences of B in all $X_{i..i+k-1}$ we count each occurrence of B exactly $k-|B|+1$ times, except for those that occur in the first and last k symbols. Dividing by $m-|B|+1$, and using the definition (3.1), we obtain

$$\nu(X_{1..m},B)\leq\frac{1}{m-|B|+1}\left(\sum_{i=1}^{m-k+1}\nu(X_{i..i+k-1},B)|+2k\right). \tag{3.14}$$

Summing over all B, for any μ, we get

$$\hat{d}(X_{1..m},\mu)\leq\frac{1}{m-k+1}\sum_{i=1}^{m-k+1}\hat{d}(X_{i..i+n-1},\mu)+\frac{2k}{m-k+1}+t_k, \tag{3.15}$$

where in the right-hand side t_k corresponds to all the summands in the left-hand side for which $|B|>k$, where for the rest of the summands we used

$|B|\le k$. Since this holds for any μ, we conclude that

$$\hat{d}(X_{1..m},H)\le\frac{1}{m-k+1}\left(\sum_{i=1}^{m-k+1}\hat{d}(X_{i..i+k-1},H)\right)+\frac{2k}{m-k+1}+t_k.$$

Therefore, for any $X_{1..m}\in A^m$, if $\hat{d}(X_{1..m},H)>\varepsilon$ then there is an index $i\le m-k$ such that $\hat{d}(X_{i..i+k-1},H)>\varepsilon-\frac{2k}{m-k+1}-t_k$. Moreover, we have (by the definition of stationarity)

$$\rho(\hat{d}(X_{i..i+k-1},H)>\varepsilon')=\rho(\hat{d}(X_{1..k},H)>\varepsilon')$$

where $\varepsilon'=\varepsilon-\frac{2k}{m-k+1}-t_k$. So we have

$$\rho\Big(\hat{d}(X_{1..k},H)\ge\varepsilon'\Big)\ge\rho\Big(\hat{d}(X_{1..m},H)\ge\varepsilon\Big),$$

proving (3.12). The second statement can be proven similarly; indeed, analogously to (3.14) we have

$$\begin{aligned}\nu(X_{1..m},B)&\ge\frac{1}{m-|B|+1}\sum_{i=1}^{m-k+1}\nu(X_{i..i+k-1},B)-\frac{2k}{m-|B|+1}\\&\ge\frac{1}{m-k+1}\left(\frac{m-k+1}{m}\sum_{i=1}^{m-k+1}\nu(X_{i..i+k-1},B)\right)-\frac{2k}{m},\end{aligned}$$

where we have used $|B|\ge 1$. Summing over different B, we obtain (similar to (3.15)),

$$\hat{d}(X_{1..m},\mu)\ge\frac{1}{m-k+1}\sum_{i=1}^{m-k+1}\frac{m-k+1}{m}\hat{d}_k(X_{i..i+n-1},\mu)-\frac{2k}{m}$$

(since the frequencies are non-negative, there is no t_n term here), which, using stationarity of ρ, implies (3.13). □

Lemma 3.21. *Let $\rho_k\in\mathcal{S}$, $k\in\mathbb{N}$ be a sequence of processes that converges to*

a process ρ_*. Then, for any $T\in A^*$ and $\varepsilon>0$ if $\rho_k(T)>\varepsilon$ for infinitely many indices k, then $\rho_*(T)\geq\varepsilon$

Proof. The statement follows from the fact that $\rho(T)$ is continuous as a function of ρ. □

Proof of Theorem 3.11. To establish the first statement of Theorem 3.11, we have to show that the family of tests ψ^α is consistent. By construction, for any $\rho\in \mathrm{cl}H_0\cap\mathcal{E}$ we have $\rho(\psi^\alpha(X_{1..n})=1)\leq\alpha$.

To prove the consistency of ψ, it remains to show that

$$\xi(\lim_{n\to\infty}\psi^\alpha(X_{1..n})=1)$$

for any $\xi\in H_1$ and $\alpha>0$. To do this, fix any $\xi\in H_1$ and let

$$\Delta:=d(\xi,\mathrm{cl}H_0):=\inf_{\rho\in\mathrm{cl}H_0\cap\mathcal{E}}d(\xi,\rho).$$

Since $\xi\notin \mathrm{cl}H_0$, we have $\Delta>0$. Suppose that there exists an $\alpha>0$, such that, for infinitely many n, some samples from the $\Delta/2$-neighbourhood of n-samples around ξ are sorted as H_0 by ψ, that is, $C^n(\mathrm{cl}H_0\cap\mathcal{E},1-\alpha)\cap b^n_{\Delta/2}(\xi)\neq\varnothing$. Then for these n we have $\gamma_n(\mathrm{cl}H_0\cap\mathcal{E},1-\alpha)\geq\Delta/2$.

This means that there exists an increasing sequence $n_k,k\in\mathbb{N}$, and a sequence $\rho_k\in\mathrm{cl}H_0$, $k\in\mathbb{N}$, such that

$$\rho_k(b^{n_k}_{\Delta/2}(\mathrm{cl}H_0\cap\mathcal{E}))<1-\alpha.$$

Since the set $\mathrm{cl}H_0$ is compact (as a closed subset of a compact set $\mathcal{S}$), we may assume (passing to a subsequence, if necessary) that ρ_k converges to a certain $\rho_*\in\mathrm{cl}H_0$. Using Lemma 3.20, (3.13), for every m large enough to satisfy $\frac{n_m}{n_m-n_k+1}\delta/4+\frac{n_k}{n_m-n_k+1}<\delta/2$ we have

$$\rho_m(b^{n_k}_{\Delta/4}(\mathrm{cl}H_0\cap\mathcal{E}))<1-\alpha.$$

Since this holds for infinitely many m, using Lemma 3.21 (with $T=b^{n_k}_{\Delta/4}(\mathrm{cl}H_0\cap\mathcal{E})$) we conclude that

$$\rho_*(b^{n_k}_{\Delta/4}(\mathrm{cl}H_0\cap\mathcal{E}))\le 1-\alpha.$$

Since the latter inequality holds for infinitely many indices k we also have

$$\rho_*(\limsup_{n\to\infty}\hat{d}(X_{1..n},\mathrm{cl}H_0\cap\mathcal{E})>\Delta/4)>0.$$

However, we must have $\rho_*(\lim_{n\to\infty}\hat{d}(X_{1..n},\mathrm{cl}H_0\cap\mathcal{E})=0)=1$ for every $\rho_*\in\mathrm{cl}H_0$: indeed, for $\rho_*\in\mathrm{cl}H_0\cap\mathcal{E}$ it follows from Lemma 3.3, and for $\rho_*\in\mathrm{cl}H_0\setminus\mathcal{E}$ from Lemma 3.3, ergodic decomposition and the conditions of the theorem ($W_\rho(H_0)=1$ for $\rho\in\mathrm{cl}H_0$).

This contradiction shows that for every α there are not more than finitely many n for which $C^n(\mathrm{cl}H_0\cap\mathcal{E},1-\alpha)\cap b^n_{\Delta/2}(\xi)\ne\varnothing$. To finish the proof of the first statement, it remains to note that, as follows from Lemma 3.3,

$$\xi\{X_1,X_2,\ldots.:X_{1..n}\in b^n_{\Delta/2}(\xi)\text{ from some }n\text{ on}\}\ge\xi\Big(\lim_{n\to\infty}\hat{d}(X_{1..n},\xi)=0\Big)=1.$$

To establish the second statement of Theorem 3.11 we assume that there exists a consistent test φ for H_0 against H_1, and we will show that $W_\rho(H_1)=0$ for every $\rho\in\mathrm{cl}H_0$. Take $\rho\in\mathrm{cl}H_0$ and suppose that $W_\rho(H_1)=\delta>0$. We have

$$\limsup_{n\to\infty}\int_{H_1}dW_\rho(\mu)\mu(\psi^{\delta/2}_n=0)\le\int_{H_1}dW_\rho(\mu)\limsup_{n\to\infty}\mu(\psi^{\delta/2}_n=0)=0,$$

where the inequality follows from Fatou's lemma (the functions under integral are all bounded by 1), and the equality from the consistency of ψ. Thus, from some n on we will have $\int_{H_1}dW_\rho\mu(\psi^{\delta/2}_n=0)<1/4$ so that $\rho(\psi^{\delta/2}_n=0)<1-3\delta/4$. For any set $T\in A^n$ the function $\mu(T)$ is continuous as a function of T. In particular, it holds for the set $T:=\{X_{1..n}:\psi^{\delta/2}_n(X_{1..n})=0\}$. Therefore, since $\rho\in\mathrm{cl}H_0$, for any n large enough we can find a $\rho'\in H_0$ such

that $\rho'(\psi_n^{\delta/2}=0)<1-3\delta/4$, which contradicts the consistency of ψ. Thus, $W_\rho(H_1)=0$, and Theorem 3.11 is proven. □

Proof of Theorem 3.10. To prove the first statement of the theorem, we will show that the test φ_{H_0,H_1} is a uniformly consistent test for $\mathrm{cl}H_0\cap\mathcal{E}$ against $\mathrm{cl}H_1\cap\mathcal{E}$ (and hence for H_0 against H_1), under the conditions of the theorem. Suppose that, on the contrary, for some $\alpha>0$ for every $n'\in\mathbb{N}$ there is a process $\rho\in\mathrm{cl}H_0$ such that $\rho(\varphi(X_{1..n})=1)>\alpha$ for some $n>n'$. Define

$$\Delta:=d(\mathrm{cl}H_0,\mathrm{cl}H_1):=\inf_{\rho_0\in\mathrm{cl}H_0\cap\mathcal{E},\rho_1\in\mathrm{cl}H_1\cap\mathcal{E}}d(\rho_0,\rho_1),$$

which is positive since $\mathrm{cl}H_0$ and $\mathrm{cl}H_1$ are closed and disjoint. We have

$$\begin{aligned}\alpha<\rho(\varphi(X_{1..n})=1)&\\ \le\rho(\hat d(X_{1..n},H_0)\ge\Delta/2\ &or\ \hat d(X_{1..n},H_1)<\Delta/2)\\ \le\rho(\hat d(X_{1..n},H_0)\ge\Delta/2)&+\rho(\hat d(X_{1..n},H_1)<\Delta/2).\quad(3.16)\end{aligned}$$

This implies that either $\rho(\hat d(X_{1..n},\mathrm{cl}H_0)\ge\Delta/2)>\alpha/2$ or $\rho(\hat d(X_{1..n},\mathrm{cl}H_1)<\Delta/2)>\alpha/2$, so that, by assumption, at least one of these inequalities holds for infinitely many $n\in\mathbb{N}$ for some sequence $\rho_n\in H_0$. Suppose that it is the first one, that is, there is an increasing sequence n_i, $i\in\mathbb{N}$ and a sequence $\rho_i\in\mathrm{cl}H_0$, $i\in\mathbb{N}$ such that

$$\rho_i(\hat d(X_{1..n_i},\mathrm{cl}H_0)\ge\Delta/2)>\alpha/2\text{ for all }i\in\mathbb{N}.\qquad(3.17)$$

The set $\mathcal{S}$ is compact, hence so is its closed subset $\mathrm{cl}H_0$. Therefore, the sequence ρ_i, $i\in\mathbb{N}$ must contain a subsequence that converges to a certain process $\rho_*\in\mathrm{cl}H_0$. Passing to a subsequence if necessary, we may assume that this convergent subsequence is the sequence ρ_i, $i\in\mathbb{N}$ itself.

Using Lemma 3.20, (3.12) (with $\rho=\rho_{n_m}$, $m=n_m$, $k=n_k$, and $H=\mathrm{cl}H_0$), and taking k large enough to have $t_{n_k}<\Delta/4$, for every m large enough to

have $\frac{2n_k}{n_m-n_k+1}<\Delta/4$, we obtain

$$\rho_{n_m}\Big(\hat{d}(X_{1..n_k},\mathrm{cl}H_0)\geq\Delta/4\Big)\geq\rho_{n_m}\Big(\hat{d}(X_{1..n_m},\mathrm{cl}H_0)\geq\Delta/2\Big)>\alpha/2. \qquad (3.18)$$

That is, we have shown that for any large enough index n_k the inequality $\rho_{n_m}(\hat{d}(X_{1..n_k},\mathrm{cl}H_0)\geq\Delta/4)>\alpha/2$ holds for infinitely many indices n_m. From this and Lemma 3.21 with $T=T_k:=\{X:\hat{d}(X_{1..n_k},\mathrm{cl}H_0)\geq\Delta/4\}$ we conclude that $\rho_*(T_k)>\alpha/2$. The latter holds for infinitely many k; that is, $\rho_*(\hat{d}(X_{1..n_k},\mathrm{cl}H_0)\geq\Delta/4)>\alpha/2$ infinitely often. Therefore,

$$\rho_*(\limsup_{n\to\infty} d(X_{1..n},\mathrm{cl}H_0)\geq\Delta/4)>0.$$

However, we must have

$$\rho_*(\lim_{n\to\infty} d(X_{1..n},\mathrm{cl}H_0)=0)=1$$

for every $\rho_*\in\mathrm{cl}H_0$: indeed, for $\rho_*\in\mathrm{cl}H_0\cap\mathcal{E}$ it follows from Lemma 3.3, and for $\rho_*\in\mathrm{cl}H_0\backslash\mathcal{E}$ from Lemma 3.3, ergodic decomposition and the conditions of the theorem.

Thus, we have arrived at a contradiction that shows that $\rho_n(\hat{d}(X_{1..n},\mathrm{cl}H_0)>\Delta/2)>\alpha/2$ cannot hold for infinitely many $n\in\mathbb{N}$ for any sequence of $\rho_n\in\mathrm{cl}H_0$. Analogously, we can show that $\rho_n(\hat{d}(X_{1..n},\mathrm{cl}H_1)<\Delta/2)>\alpha/2$ cannot hold for infinitely many $n\in\mathbb{N}$ for any sequence of $\rho_n\in\mathrm{cl}H_0$. Indeed, using Lemma 3.20, equation (3.13), we can show that $\rho_{n_m}(\hat{d}(X_{1..n_m},\mathrm{cl}H_1)\leq\Delta/2)>\alpha/2$ for a large enough n_m implies $\rho_{n_m}(\hat{d}(X_{1..n_k},\mathrm{cl}H_1)\leq 3\Delta/4)>\alpha/2$ for a smaller n_k. Therefore, if we assume that $\rho_n(\hat{d}(X_{1..n},\mathrm{cl}H_1)<\Delta/2)>\alpha/2$ for infinitely many $n\in\mathbb{N}$ for some sequence of $\rho_n\in\mathrm{cl}H_0$, then we will also find a ρ_* for which $\rho_*(\hat{d}(X_{1..n},\mathrm{cl}H_1)\leq 3\Delta/4)>\alpha/2$ for infinitely many n, which, using Lemma 3.3 and ergodic decomposition, can be shown to contradict the fact that $\rho_*(\lim_{n\to\infty}d(X_{1..n},\mathrm{cl}H_1)\geq\Delta)=1$.

Thus, returning to (3.16), we have shown that from some n on there is no $\rho\in\mathrm{cl}H_0$ for which $\rho(\varphi=1)>\alpha$ holds true. The statement for $\rho\in\mathrm{cl}H_1$ can

be proven analogously, thereby finishing the proof of the first statement.

To prove the second statement of the theorem, we assume that there exists a uniformly consistent test φ for H_0 against H_1, and we will show that $W_\rho(H_{1-i})=0$ for every $\rho\in \mathrm{cl}H_i$. Indeed, let $\rho\in \mathrm{cl}H_0$, that is, suppose that there is a sequence $\xi_i\in H_0, i\in\mathbb{N}$ such that $\xi_i\to\rho$. Assume $W_\rho(H_1)=\delta>0$ and take $\alpha:=\delta/2$. Since the test φ is uniformly consistent, there is an $N\in\mathbb{N}$ such that for every $n>N$ we have

$$\rho(\varphi(X_{1..n}=0))\le\int_{H_1}\varphi(X_{1..n}=0)dW_\rho+\int_{\mathcal{E}\setminus H_1}\varphi(X_{1..n}=0)dW_\rho$$
$$\le\delta\alpha+1-\delta\le 1-\delta/2.$$

Recall that, for $T\in A^*$, $\mu(T)$ is a continuous function in μ. In particular, this holds for the set $T=\{X\in A^n:\varphi(X)=0\}$, for any given $n\in\mathbb{N}$. Therefore, for every $n>N$ and for every i large enough, $\rho_i(\varphi(X_{1..n})=0)<1-\delta/2$ implies also $\xi_i(\varphi(X_{1..n})=0)<1-\delta/2$ which contradicts $\xi_i\in H_0$. This contradiction shows $W_\rho(H_1)=0$ for every $\rho\in\mathrm{cl}H_0$. The case $\rho\in\mathrm{cl}H_1$ is analogous. □

Proof of Proposition 3.13. Consider the process $(x_1,y_1),(x_2,y_2),...$ on pairs $(x_i,y_i)\in A^2$, such that the distribution of $x_1,x_2,...$ is ν, the distribution of $y_1,y_2,...$ is ρ and the two components x_i and y_i are independent; in other words, the distribution of (x_i,y_i) is $\nu\times\rho$. Consider also a two-state stationary ergodic Markov chain μ, with two states 1 and 2, whose transition probabilities are $\begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$, where $0<p<q<1$. The limiting (and initial) probability of the state 1 is $p/(p+q)$ and that of the state 2 is $q/(p+q)$. Finally, the process $z_1,z_2,...$ is constructed as follows: $z_i=x_i$ if μ is in the state a and $z_i=y_i$ otherwise (here it is assumed that the chain μ generates a sequence of outcomes independently of (x_i,y_i). Clearly, for every p,q satisfying $0<p<q<1$ the process $z_1,z_2,...$ is stationary ergodic; denote ζ its distribution. Let $p_n:=1/(n+1)$, $n\in\mathbb{N}$. Since $d(\rho,\nu)>\varepsilon$, we can find a $\delta>0$ such that $d(\rho,\zeta_n)>\varepsilon$ where ζ_n is the distribution ζ with parameters p_n and

q_n, where q_n satisfies $q_n/(p_n+q_n)=\delta$. Thus, $\zeta_n \in H_1$ for all $n \in \mathbb{N}$. However, $\lim_{n\to\infty}\zeta_n=\zeta_\infty$ where ζ_∞ is the stationary distribution with $W_{\zeta_\infty}(\rho)=\delta$ and $W_{\zeta_\infty}(\nu)=1-\delta$. Therefore, $\zeta_\infty \in \mathrm{cl}H_1$ and $W_{\zeta_\infty}(H_0)>0$, so that by Theorem 3.10 there is no uniformly consistent test for H_0 against H_1. □

3.6.3 Proof of Theorem 3.19

We will assume that asymptotically correct discrimination procedure D for the class of all B-processes exists, and will construct a B-process ρ such that if both sequences x_i and y_i, $i \in \mathbb{N}$ are generated by ρ then $\mathbb{E}D_n$ diverges; this contradiction will prove the theorem.

The scheme of the proof is as follows. On Step 1 we construct a sequence of processes ρ_{2k}, ρ_{d2k+1}, and ρ_{u2k+1}, where $k=0,1,\dots$. On Step 2 we construct a process ρ, which is shown to be the limit of the sequence ρ_{2k}, $k \in \mathbb{N}$, in $\bar{d}$-distance. On Step 3 we show that two independent runs of the process ρ have a property that (with high probability) they first behave like two runs of a single process ρ_0, then like two runs of two different processes ρ_{u1} and ρ_{d1}, then like two runs of a single process ρ_2, and so on, thereby showing that the test D diverges and obtaining the desired contradiction.

Assume that there exists an asymptotically correct discriminating procedure D. Fix some $\varepsilon \in (0,1/2)$ and $\delta \in [1/2,1)$, to be defined on Step 3.

Step 1. We will construct the sequence of process ρ_{2k}, ρ_{u2k+1}, and ρ_{d2k+1}, where $k=0,1,\dots$.

Step 1.0. Construct the process ρ_0 as follows. A Markov chain m_0 is defined on the set $\mathbb{N}$ of states. From each state $i \in \mathbb{N}$ the chain passes to the state 0 with probability δ and to the state $i+1$ with probability $1-\delta$. With transition probabilities so defined, the chain possesses a unique stationary distribution M_0 on the set $\mathbb{N}$, which can be calculated explicitly using e.g. [84, Theorem VIII.4.1], and is as follows: $M_0(0)=\delta$, $M_0(k)=\delta(1-\delta)^k$, for all $k \in \mathbb{N}$. Take this distribution as the initial distribution over the states.

The function f_0 maps the states to the output alphabet $\{0,1\}$ as follows:

$f_0(i)=1$ for every $i\in\mathbb{N}$. Let s_t be the state of the chain at time t. The process ρ_0 is defined as $\rho_0=f_0(s_t)$ for $t\in\mathbb{N}$. As a result of this definition, the process ρ_0 simply outputs 1 with probability 1 on every time step (however, by using different functions f we will have less trivial processes in the sequel). Clearly, the constructed process is stationary ergodic and a B-process. So, we have defined the chain m_0 (and the process ρ_0) up to a parameter δ.

Step 1.1. We begin with the process ρ_0 and the chain m_0 of the previous step. Since the test D is asymptotically correct we will have

$$\mathbb{E}_{\rho_0\times\rho_0}D_{t_0}((x_1,...,x_{t_0}),(y_1,...,y_{t_0}))<\varepsilon,$$

from some t_0 on, where both samples x_i and y_i are generated by ρ_0 (that is, both samples consist of 1s only). Let k_0 be such an index that the chain m_0 starting from the state 0 with probability 1 does not reach the state k_0-1 by time t_0 (we can take $k_0=t_0+2$).

Construct two processes ρ_{u1} and ρ_{d1} as follows. They are also based on the Markov chain m_0, but the functions f are different. The function $f_{u1}:\mathbb{N}\to\{0,1\}$ is defined as follows: $f_{u1}(i)=f_0(i)=1$ for $i\le k_0$ and $f_{u1}(i)=0$ for $i>k_0$. The function f_{d1} is identically 1 ($f_{d1}(i)=1$, $i\in\mathbb{N}$). The processes ρ_{u1} and ρ_{d1} are defined as $\rho_{u1}=f_{u1}(s_t)$ and $\rho_{d1}=f_{d1}(s_t)$ for $t\in\mathbb{N}$. Thus the process ρ_{d1} will again produce only 1s, but the process ρ_{u1} will occasionally produce 0s.

Step 1.2. Being run on two samples generated by the processes ρ_{u1} and ρ_{d1} which both start from the state 0, the test D_n on the first t_0 steps produces many 0s, since on these first k_0 states all the functions f, f_{u1} and f_{d1} coincide. However, since the processes are different and the test is asymptotically correct (by assumption), the test starts producing 1s, until by a certain time step t_1 almost all answers are 1s. Next we will construct the process ρ_2 by "gluing" together ρ_{u1} and ρ_{d1} and continuing them in such a way that, being run on two samples produced by ρ_2 the test first produces 0s (as if the samples were drawn from ρ_0), then, with probability close to

1/2 it will produce many 1s (as if the samples were from ρ_{u1} and ρ_{d1}) and then again 0s.

The process ρ_2 is the pivotal point of the construction, so we give it in some detail. On step 1.2a we present the construction of the process, and on step 1.2b we show that this process is a B-process by demonstrating that it is equivalent to a (deterministic) function of a Markov chain.

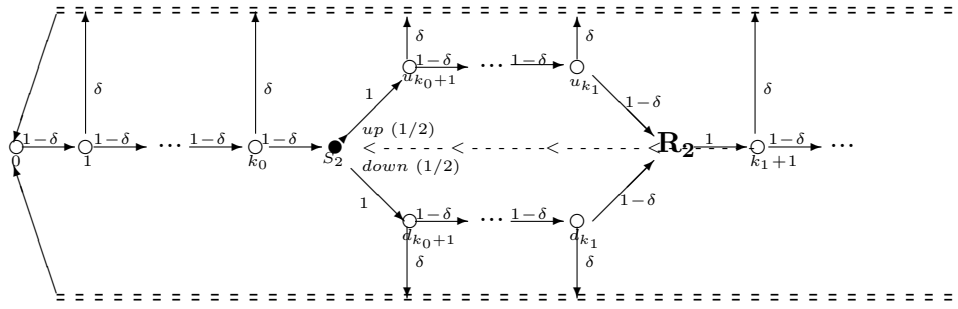
Step 1.2a. Let $t_1 > t_0$ be such a time index that

$$\mathbb{E}_{\rho_{u1}\times\rho_{d1}} D_k((x_1,...,x_{t_1}),(y_1,...,y_{t_1})) > 1-\varepsilon,$$

where the samples x_i and y_i are generated by ρ_{u1} and ρ_{d1} correspondingly (the samples are generated independently; that is, the process are based on two independent copies of the Markov chain m_0). Let $k_1 > k_0$ be such an index that the chain m starting from the state 0 with probability 1 does not reach the state k_1-1 by time t_1.

Construct the process ρ_2 as follows (see fig. 3.1). It is based on a chain

Figure 3.1: The processes m_2 and ρ_2. The states are depicted as circles, the arrows symbolize transition probabilities: from every state the process returns to 0 with probability δ or goes to the next state with probability $1-\delta$. From the switch S_2 the process passes to the state indicated by the switch (with probability 1); here it is the state u_{k_0+1}. When the process passes through the reset $\mathbf{R_2}$ the switch S_2 is set to either *up* or *down* with equal probabilities. (Here S_2 is in the position *up*.) The function f_2 is 1 on all states except $u_{k0+1},...,u_{k1}$ where it is 0; f_2 applied to the states output by m_2 defines ρ_2.



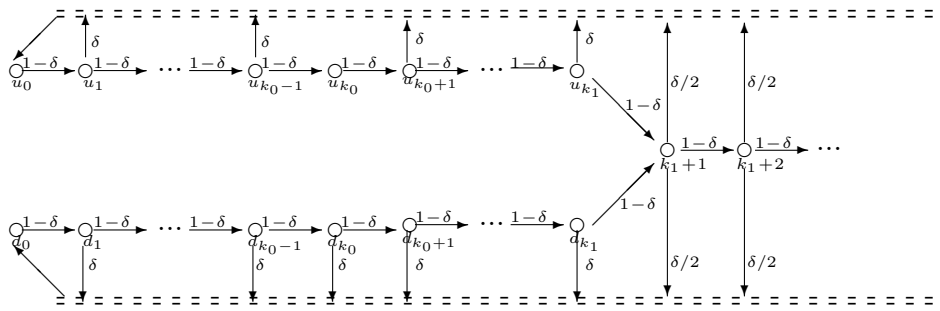
m_2 on which Markov assumption is violated. The transition probabilities on states 0,...,k_0 are the same as for the Markov chain m (from each state return to 0 with probability δ or go to the next state with probability $1-\delta$).

There are two "special" states: the "switch" S_2 and the "reset" R_2. From the state k_0 the chain passes with probability $1-\delta$ to the "switch" state S_2. The switch S_2 can itself have two values: *up* and *down*. If S_2 has the value *up* then from S_2 the chain passes to the state u_{k_0+1} with probability 1, while if $S_2 = down$ the chain goes to d_{k_0+1}, with probability 1. If the chain reaches the state R_2 then the value of S_2 is set to *up* with probability $1/2$ and with probability $1/2$ it is set to *down*. In other words, the first transition from S_2 is random (either to u_{k_0+1} or to d_{k_0+1} with equal probabilities) and then this decision is remembered until the "reset" state R_2 is visited, whereupon the switch again assumes the values *up* and *down* with equal probabilities.

The rest of the transitions are as follows. From each state u_i, $k_0 \le i \le k_1$ the chain passes to the state 0 with probability δ and to the next state u_{i+1} with probability $1-\delta$. From the state u_{k_1} the process goes with probability δ to 0 and with probability $1-\delta$ to the "reset" state R_2. The same with states d_i: for $k_0 < i \le k_1$ the process returns to 0 with probability δ or goes to the next state d_{i+1} with probability $1-\delta$, where the next state for d_{k_1} is the "reset" state R_2. From R_2 the process goes with probability 1 to the state k_1+1 where from the chain continues ad infinitum: to the state 0 with probability δ or to the next state k_1+2 etc. with probability $1-\delta$.

The initial distribution on the states is defined as follows. The probabilities of the states $0..k_0,k_1+1,k_1+2,...$ are the same as in the Markov chain m_0, that is, $\delta(1-\delta)^j$, for $j = 0..k_0,k_1+1,k_1+2,...$. For the states u_j and d_j, $k_0 < j \le k_1$ define their initial probabilities to be $1/2$ of the probability of the corresponding state in the chain m_0, that is $m_2(u_j) = m_2(d_j) = m_0(j)/2 = \delta(1-\delta)^j/2$. Furthermore, if the chain starts in a state u_j, $k_0 < j \le k_1$, then the value of the switch S_2 is *up*, and if it starts in the state d_j then the value of the switch S_2 is *down*, whereas if the chain starts in any other state then the probability distribution on the values of the switch S_2 is $1/2$ for either

Figure 3.2: The process m'_2. The function f_2 is 1 everywhere except the states u_{k_0+1},...,u_{k_1}, where it is 0.



up or *down*.

The function f_2 is defined as follows: $f_2(i)=1$ for $0\le i\le k_0$ and $i>k_1$ (before the switch and after the reset); $f_2(u_i)=0$ for all i, $k_0<i\le k_1$ and $f_2(d_i)=1$ for all i, $k_0<i\le k_1$. The function f_2 is undefined on S_2 and R_2, therefore there is no output on these states (we also assume that passing through S_2 and R_2 does not increment time). As before, the process ρ_2 is defined as $\rho_2=f_2(s_t)$ where s_t is the state of m_2 at time t, omitting the states S_2 and R_2. The resulting process s illustrated on fig. 3.1.

Step 1.2b. To show that the process ρ_2 is stationary ergodic and a B-process, we will show that it is equivalent to a function of a stationary ergodic Markov chain, whereas all such process are known to be B (e.g. [82]). The construction is as follows (see fig. 3.2). This chain has states k_1+1,... and also u_0,...,u_{k_0},u_{k_0+1},...,u_{k_1} and d_0,...,d_{k_0},d_{k_0+1},...,d_{k_1}. From the states u_i, $i=0$,...,k_1 the chain passes with probability $1-\delta$ to the next state u_{i+1}, where the next state for u_{k_1} is $k+1$ and with probability δ returns to the state u_0 (and not to the state 0). Transitions for the state d_0,...,d_{k_1-1} are defined analogously. Thus the states u_{k_i} correspond to the state *up* of the switch S_2 and the states d_{k_i} — to the state *down* of the switch. Transitions for the states $k+1$,$k+2$,... are defined as follows: with probability $\delta/2$ to the state u_0, with probability $\delta/2$ to the state d_0, and with probability $1-\delta$ to the next state. Thus, transitions to 0 from the states with indices

greater than k_1 corresponds to the reset R_2. Clearly, the chain m'_2 as defined possesses a unique stationary distribution M_2 over the set of states and $M_2(i) > 0$ for every state i. Moreover, this distribution is the same as the initial distribution on the states of the chain m_0, except for the states u_i and d_i, for which we have $m'_2(u_i) = m'_2(d_i) = m_0(i)/2 = \delta(1-\delta)^i/2$, for $0 \leq i \leq k_0$. We take this distribution as its initial distribution on the states of m'_2. The resulting process m'_2 is stationary ergodic, and a B-process, since it is a function of a Markov chain [82]. It is easy to see that if we define the function f_2 on the states of m'_2 as 1 on all states except $u_{k_0+1},...,u_{k_1}$, then the resulting process is exactly the process ρ_2. Therefore, ρ_2 is stationary ergodic and a B-process.

Step 1.k. As before, we can continue the construction of the processes ρ_{u3} and ρ_{d3}, that start with a segment of ρ_2. Let $t_2 > t_1$ be a time index such that

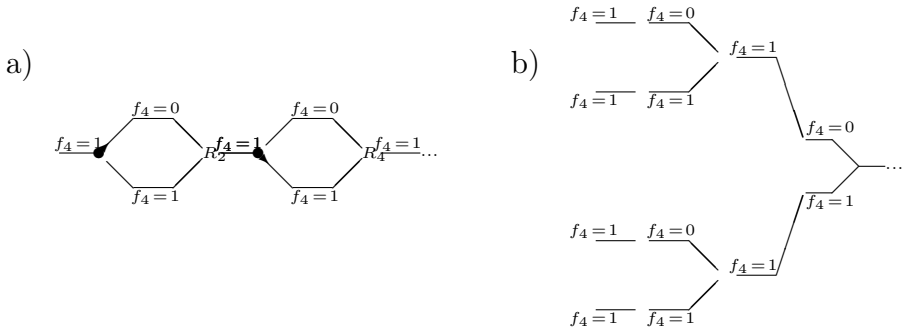
$$\mathbb{E}_{\rho_2 \times \rho_2} D_{t_2} < \varepsilon,$$

where both samples are generated by ρ_2. Let $k_2 > k_1$ be such an index that when starting from the state 0 the process m_2 with probability 1 does not reach $k_2 - 1$ by time t_2 (equivalently: the process m'_2 does not reach $k_2 - 1$ when starting from either u_0 or d_0). The processes ρ_{u3} and ρ_{d3} are based on the same process m_2 as ρ_2. The functions f_{u3} and f_{d3} coincide with f_2 on all states up to the state k_2 (including the states u_i and d_i, $k_0 < i \leq k_1$). After k_2 the function f_{u3} outputs 0s while f_{d3} outputs 1s: $f_{u3}(i) = 0$, $f_{d3}(i) = 1$ for $i > k_2$.

Furthermore, we find a time $t_3 > t_2$ by which we have $\mathbb{E}_{\rho_{u3} \times \rho_{d3}} D_{t_3} > 1 - \varepsilon$, where the samples are generated by ρ_{u3} and ρ_{d3}, which is possible since D is consistent. Next, find an index $k_3 > k_2$ such that the process m_2 does not reach $k_3 - 1$ with probability 1 if the processes ρ_{u3} and ρ_{d3} are used to produce two independent sequences and both start from the state 0. We then construct the process ρ_4 based on a (non-Markovian) process m_4 by "gluing" together ρ_{u3} and ρ_{d3} after the step k_3 with a switch S_4 and a reset

R_4 exactly as was done when constructing the process ρ_2. The process m_4 is illustrated on fig. 3.3a). The process m_4 can be shown to be equivalent to a Markov chain m'_4, which is constructed analogously to the chain m'_2 (see fig. 3.3b). Thus, the process ρ_4 is can be shown to be a B-process.

Figure 3.3: a) The processes m_4. b) The Markov chain m'_4



Proceeding this way we can construct the processes ρ_{2j}, ρ_{u2j+1} and ρ_{d2j+1}, $j\in\mathbb{N}$ choosing the time steps $t_j > t_{j-1}$ so that the expected output of the test approaches 0 by the time t_j being run on two samples produced by ρ_j for even j, and approaches 1 by the time t_j being run on samples produced by ρ_{uj} and ρ_{dj} for odd j:

$$\mathbb{E}_{\rho_{2j}\times\rho_{2j}} D_{t_{2j}} < \varepsilon \tag{3.19}$$

and

$$\mathbb{E}_{\rho_{u2j+1}\times\rho_{d2j+1}} D_{t_{2j+1}} > (1-\varepsilon). \tag{3.20}$$

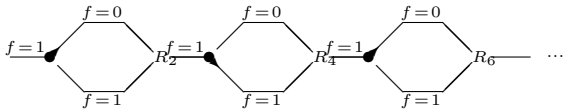
For each j the number $k_j > k_{j-1}$ is selected in a such a way that the state k_j-1 is not reached (with probability 1) by the time t_j when starting from the state 0. Each of the processes ρ_{2j}, ρ_{u2j+1} and ρ_{dj2+1}, $j\in\mathbb{N}$ can be shown to be stationary ergodic and a B-process by demonstrating equivalence to a Markov chain, analogously to the Step 1.2. The initial state distribution of each of the processes ρ_t,$t\in\mathbb{N}$ is $M_t(k)=\delta(1-\delta)^k$ and $M_t(u_k)=M_t(d_k)=\delta(1-\delta)^k/2$ for those $k\in\mathbb{N}$ for which the corresponding states are defined.

Step 2. Having defined k_j, $j\in\mathbb{N}$ we can define the process ρ. The construction is given on Step 2a, while on Step 2b we show that ρ is stationary

ergodic and a B-process, by showing that it is the limit of the sequence ρ_{2j}, $j \in \mathbb{N}$.

Step 2a. The process ρ can be constructed as follows (see fig. 3.4). The

Figure 3.4: The processes m_ρ and ρ. The states are on horizontal lines. The function f being applied to the states of m_ρ defines the process ρ. Its value is 0 on the states on the upper lines (states $u_{k_{2j}+1},\dots,u_{k_{2j+1}}$, where $k \in \mathbb{N}$) and 1 on the rest of the states.



construction is based on the (non-Markovian) process m_ρ that has states $0,\dots,k_0$, $k_{2j+1}{+}1,\dots,k_{2(j+1)}$, $u_{k_{2j}+1},\dots,u_{k_{2j+1}}$ and $d_{k_{2j}+1},\dots,d_{k_{2j+1}}$ for $j \in \mathbb{N}$, along with switch states S_{2j} and reset states R_{2j}. Each switch S_{2j} diverts the process to the state $u_{k_{2j}+1}$ if the switch has value *up* and to $d_{k_{2j}+1}$ if it has the value *down*. The reset R_{2j} sets S_{2j} to *up* with probability $1/2$ and to *down* also with probability $1/2$. From each state that is neither a reset nor a switch, the process goes to the next state with probability $1-\delta$ and returns to the state 0 with probability δ (cf. Step 1k).

The initial distribution M_ρ on the states of m_ρ is defined as follows. For every state i such that $0 \le i \le k_0$ and $k_{2j+1} < i \le k_{2_j+2}$, $j = 0,1,\dots$, define the initial probability of the state i as $M_\rho(i) = \delta(1-\delta)^i$ (the same as in the chain m_0), and for the sets u_j and d_j (for those j for which these sets are defined) let $M_\rho(u_j) = M_\rho(d_j) := \delta(1-\delta)^i/2$ (that is, $1/2$ of the probability of the corresponding state of m_0).

The function f is defined as 1 everywhere except for the states u_j (for all $j \in \mathbb{N}$ for which u_j is defined) on which f takes the value 0. The process ρ is defined at time t as $f(s_t)$, where s_t is the state of m_ρ at time t.

Step 2b. To show that ρ is a B-process, let us first show that it is stationary. Recall the definition 3.2 of the distributional distance between (arbitrary) process distributions. The set of all stochastic processes, equipped with this distance, is complete, and the set of all stationary processes is its

closed subset [36]. Thus, to show that the process ρ is stationary it suffices to show that $\lim_{j\to\infty} d(\rho_{2j},\rho)=0$, since the processes ρ_{2j}, $j\in\mathbb{N}$, are stationary. To do this, it is enough to demonstrate that

$$\lim_{j\to\infty} |\rho((x_1,...,x_{|B|})=B)-\rho_{2j}((x_1,...,x_{|B|})=B)|=0 \tag{3.21}$$

for each $B\in X^*$. Since the processes m_ρ and m_{2j} coincide on all states up to k_{2j+1}, we have

$$|\rho(x_n=a)-\rho_{2j}(x_n=a)|=|\rho(x_1=a)-\rho_{2j}(x_1=a)| \\ \leq \sum_{k>k_{2j+1}} M_\rho(k)+\sum_{k>k_{2j+1}} M_{2j}(k)$$

for every $n\in\mathbb{N}$ and $a\in X$. Moreover, for any tuple $B\in X^*$ we obtain

$$|\rho((x_1,...,x_{|B|})=B)-\rho_{2j}((x_1,...,x_{|B|})=B)| \\ \leq |B|\left(\sum_{k>k_{2j+1}} M_\rho(k)+\sum_{k>k_{2j+1}} M_{2j}(k)\right)\to 0$$

where the convergence follows from $k_{2j}\to\infty$. We conclude that (3.21) holds true, so that $d(\rho,\rho_{2j})\to 0$ and ρ is stationary.

To show that ρ is a B-process, we will demonstrate that it is the limit of the sequence ρ_{2k}, $k\in\mathbb{N}$ in the $\bar{d}$ distance (which was only defined for stationary processes). Since the set of all B-process is a closed subset of all stationary processes, it will follow that ρ itself is a B-process. (Observe that this way we get ergodicity of ρ "for free", since the set of all ergodic processes is closed in $\bar{d}$ distance, and all the processes ρ_{2j} are ergodic.) In order to show that $\bar{d}(\rho,\rho_{2k})\to 0$ we have to find for each j a processes ν_{2j} on pairs $(x_1,y_1),(x_2,y_2),...$, such that x_i are distributed according to ρ and y_i are distributed according to ρ_{2j}, and such that $\lim_{j\to\infty}\nu_{2j}(x_1\neq y_1)=0$. Construct such a coupling as follows. Consider the chains m_ρ and m_{2j}, which start in

the same state (with initial distribution being M_ρ) and always take state transitions together, where if the process m_ρ is in the state u_t or d_t, $t \geq k_{2j+1}$ (that is, one of the states which the chain m_{2j} does not have) then the chain m_{2j} is in the state t. The first coordinate of the process ν_{2j} is obtained by applying the function f to the process m_ρ and the second by applying f_{2j} to the chain m_{2j}. Clearly, the distribution of the first coordinate is ρ and the distribution of the second is ρ_{2j}. Since the chains start in the same state and always take state transitions together, and since the chains m_ρ and m_{2j} coincide up to the state k_{2j+1} we have $\nu_{2j}(x_1 \neq y_1) \leq \sum_{k>k_{2j+1}} M_\rho(k) \to 0$. Thus, $\bar{d}(\rho,\rho_{2j}) \to 0$, so that ρ is a B-process.

Step 3. Finally, it remains to show that the expected output of the test D diverges if the test is run on two independent samples produced by ρ.

Recall that for all the chains m_{2j}, m_{u2j+1} and m_{d2j+1} as well as for the chain m_ρ, the initial probability of the state 0 is δ. By construction, if the process m_ρ starts at the state 0 then up to the time step k_{2j} it behaves exactly as ρ_{2j} that has started at the state 0. In symbols, we have

$$E_{\rho\times\rho}(D_{t_{2j}}|s_0^x=0,s_0^y=0)=E_{\rho_{2j}\times\rho_{2j}}(D_{t_{2j}}|s_0^x=0,s_0^y=0) \tag{3.22}$$

for $j\in\mathbb{N}$, where s_0^x and s_0^y denote the initial states of the processes generating the samples x and y correspondingly.

We will use the following simple decomposition

$$\mathbb{E}(D_{t_j})=\delta^2\mathbb{E}(D_{t_j}|s_0^x=0,s_0^y=0)+(1-\delta^2)\mathbb{E}(D_{t_j}|s_0^x\neq 0 \text{ or } s_0^y\neq 0), \tag{3.23}$$

From this, (3.22) and (3.19) we have

$$\begin{aligned}\mathbb{E}_{\rho\times\rho}(D_{t_{2j}}) &\leq \delta^2\mathbb{E}_{\rho\times\rho}(D_{t_{2j}}|s_0^x=0,s_0^y=0)+(1-\delta^2)\\ &=\delta^2\mathbb{E}_{\rho_{2j}\times\rho_{2j}}(D_{t_{2j}}|s_0^x=0,s_0^y=0)+(1-\delta^2)\\ &\leq \mathbb{E}_{\rho_{2j}\times\rho_{2j}}+(1-\delta^2)<\varepsilon+(1-\delta^2).\end{aligned} \tag{3.24}$$

For odd indices, if the process ρ starts at the state 0 then (from the definition of t_{2j+1}) by the time t_{2j+1} it does not reach the reset R_{2j}; therefore, in this case the value of the switch S_{2j} does not change up to the time t_{2j+1}. Since the definition of m_ρ is symmetric with respect to the values *up* and *down* of each switch, the probability that two samples $x_1,...,x_{t_{2j+1}}$ and $y_1,...,y_{t_{2j+1}}$ generated independently by (two runs of) the process ρ produced different values of the switch S_{2j} when passing through it for the first time is $1/2$. In other words, with probability $1/2$ two samples generated by ρ starting at the state 0 will look by the time t_{2j+1} as two samples generated by ρ_{u2j+1} and ρ_{d2j+1} that has started at state 0. Thus

$$E_{\rho\times\rho}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0)\geq\frac{1}{2}E_{\rho_{u2j+1}\times\rho_{d2j+1}}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0) \qquad (3.25)$$

for $j\in\mathbb{N}$. Using this, (3.23), and (3.20) we obtain

$$\begin{aligned}\mathbb{E}_{\rho\times\rho}(D_{t_{2j+1}})&\geq\delta^2\mathbb{E}_{\rho\times\rho}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0)\\&\geq\frac{1}{2}\delta^2\mathbb{E}_{\rho_{2j+1}\times\rho_{2j+1}}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0)\\&\geq\frac{1}{2}\big(\mathbb{E}_{\rho_{2j+1}\times\rho_{2j+1}}(D_{t_{2j+1}})-(1-\delta^2)\big)>\frac{1}{2}(\delta^2-\varepsilon).\end{aligned} \qquad (3.26)$$

Taking δ large and ε small (e.g. $\delta=0.9$ and $\varepsilon=0.1$), we can make the bound (3.24) close to 0 and the bound (3.26) close to $1/2$, and the expected output of the test will cross these values infinitely often. Therefore, we have shown that the expected output of the test D diverges on two independent runs of the process ρ, contradicting the consistency of D. This contradiction concludes the proof.

Chapter 4

Finding an optimal strategy in a reactive environment [R7]

Many real-world "learning" problems (like learning to drive a car or playing a game) can be modelled as an agent π that interacts with an environment μ and is (occasionally) rewarded for its behaviour. We are interested in agents which perform well in the sense of having high long-term reward, also called the value $V(\mu,\pi)$ of agent π in environment μ. If μ is known, it is a pure (non-learning) computational problem to determine the optimal agent $\pi^\mu := \text{argmax}_\pi V(\mu,\pi)$. It is far less clear what an "optimal" agent means, if μ is unknown. A reasonable objective is to have a single policy π with high value simultaneously in many environments. We will formalize and call this criterion *self-optimizing* later.

This problem generalizes dramatically the problem of sequence prediction, as well as those of classification and (to a large extent) hypothesis testing. For example, the problem of sequence prediction is a special case in which actions have no impact on the environment. While being very general, the problem of finding an optimal strategy in a reactive environment is also very difficult; the results of this chapter present some first steps in attacking this problem in full generality.

Some related work. Reinforcement learning, sequential decision theory, adaptive control theory, and active expert advice, are theories dealing with this problem. They overlap but have different core focus: Reinforcement learning algorithms [87] are developed to learn μ or directly its value. Temporal difference learning is computationally very efficient, but has slow asymptotic guarantees (only) in (effectively) small observable MDPs. Others have faster guarantee in finite state MDPs [14]. There are algorithms [31] which are optimal for any finite connected POMDP, and this is apparently the largest class of environments considered. In sequential decision theory, a Bayes-optimal agent π^* that maximizes $V(\xi,\pi)$ is considered, where ξ is a mixture of environments $\nu \in \mathcal{C}$ and $\mathcal{C}$ is a class of environments that contains the true environment $\mu \in \mathcal{C}$ [41]. Policy π^* is self-optimizing in an arbitrary class $\mathcal{C}$, provided $\mathcal{C}$ allows for self-optimizingness [40]. Adaptive control theory [57] considers very simple (from an AI perspective) or special systems (e.g. linear with quadratic loss function), which sometimes allow computationally and data efficient solutions. Action with expert advice [24, 71, 72, 18] constructs an agent (called master) that performs nearly as well as the best agent (best expert in hindsight) from some class of experts, in *any* environment ν.

The difficulty in active learning problems can be identified (at least, for countable classes) with *traps* in the environments. Initially the agent does not know μ, so has asymptotically to be forgiven in taking initial "wrong" actions. A well-studied such class are ergodic MDPs which guarantee that, from any action history, every state can be (re)visited [40].

The results. The aim of this chapter is to characterize as general as possible classes $\mathcal{C}$ in which self-optimizing behaviour is possible (more general than POMDPs). To do this, we need to characterize classes of environments that forgive. For instance, exact state recovery is unnecessarily strong; it is sufficient being able to recover high rewards, from whatever states. Further, in many real world problems there is no information available about the "states" of the environment (e.g. in POMDPs) or the environment may

exhibit long history dependencies.

We propose to consider only environments in which, after any arbitrary finite sequence of actions, the best value is still achievable. The performance criterion here is asymptotic average reward. Thus we consider such environments for which there exists a policy whose asymptotic average reward exists and upper-bounds asymptotic average reward of any other policy. Moreover, the same property should hold after any finite sequence of actions has been taken (no traps). We call such environments *recoverable*. If we only want to get ε-close to the optimal value infinitely often with decreasing ε (that is, to have the same upper limit for the average value), then this property is already sufficient.

Yet recoverability in itself is not sufficient for identifying behaviour which results in optimal limiting average value. We require further that, from any sequence of k actions, it is possible to return to the optimal level of reward in $o(k)$ steps; that is, it is not just possible to recover after any sequence of (wrong) actions, but it is possible to recover fast. Environments which possess this property are called *value-stable*. (These conditions will be formulated in a probabilistic form.)

We show that for any countable class of value-stable environments there exists a policy which achieves best possible value in any of the environments from the class (i.e. is *self-optimizing* for this class).

Furthermore, we present some examples of environments which possess value-stability and/or recoverability. In particular, any ergodic MDP can be easily shown to be value-stable. A mixing-type condition which implies value-stability is also demonstrated. In addition, we provide a construction allowing to build examples of value-stable and/or recoverable environments which are not isomorphic to a finite POMDP, thus demonstrating that the class of value-stable environments is quite general.

Finally, we consider environments which are not recoverable but still are value-stable. In other words, we consider the question of what does it mean to be optimal in an environment which does not "forgive" wrong actions.

Even in such cases some policies are better than others, and we identify some conditions which are sufficient for learning a policy that is optimal from some point on.

It is important in our argument that the class of environments for which we seek a self-optimizing policy is countable, although the class of all value-stable environments is uncountable. To find a set of conditions necessary and sufficient for learning which do not rely on countability of the class is yet an open problem. However, from a computational perspective countable classes are sufficiently large (e.g. the class of all computable probability measures is countable). In view of the results of the previous chapters, in particular, of the results on sequence prediction, countable classes of environments are a natural first step to solve the general problem of characterizing learnability.

4.1 Problem formulation

The agent framework is general enough to allow modelling nearly any kind of (intelligent) system. In cycle k, an agent performs *action* $y_k \in \mathbf{Y}$ (output) which results in *observation* $o_k \in \mathcal{O}$ and *reward* $r_k \in \mathbb{R}$, followed by cycle $k+1$ and so on. We assume that the action space $\mathbf{Y}$, the observation space $\mathcal{O}$, and the reward space $\mathbb{R} \subset \mathbb{R}$ are finite, w.l.g. $\mathbb{R} = \{0,...,r_{max}\}$. We abbreviate $z_k := y_k r_k o_k \in \mathbf{Z} := \mathbf{Y} \times \mathbb{R} \times \mathcal{O}$ and $x_k = r_k o_k \in \mathbf{X} := \mathbb{R} \times \mathcal{O}$. An agent is identified with a (probabilistic) *policy* π. Given *history* $z_{<k}$, the probability that agent π acts y_k in cycle k is (by definition) $\pi(y_k|z_{<k})$. Thereafter, *environment* μ provides (probabilistic) reward r_k and observation o_k, i.e. the probability that the agent perceives x_k is (by definition) $\mu(x_k|z_{<k}y_k)$. Note that policy and environment are allowed to depend on the complete history. We do not make any MDP or POMDP assumption here, and we do not talk about states of the environment, only about observations. Each (policy,environment) pair (π,μ) generates an I/O sequence $z_1^{\pi\mu} z_2^{\pi\mu} ...$. More

formally, history $z_{1..k}^{\pi\mu}$ is a random variable with probability

$$P\big(z_{1..k}^{\pi\mu} = z_{1..k}\big) \;=\; \pi(y_1)\cdot\mu(x_1|y_1)\cdot\ldots\cdot\pi(y_k|z_{<k})\cdot\mu(x_k|z_{<k}y_k).$$

Since value maximizing policies can always be chosen deterministic, there is no real need to consider probabilistic policies, and henceforth we consider deterministic policies p. We assume that $\mu\in\mathcal{C}$ is the true, but unknown, environment, and $\nu\in\mathcal{C}$ a generic environment.

For an environment ν and a policy p define random variables (upper and lower average value)

$$\overline{V}(\nu,p) \;:=\; \limsup_m\big\{\tfrac{1}{m}r_{1..m}^{p\nu}\big\} \quad\text{and}\quad \underline{V}(\nu,p) \;:=\; \liminf_m\big\{\tfrac{1}{m}r_{1..m}^{p\nu}\big\}$$

where $r_{1..m}:=r_1+\ldots+r_m$. If there exists a constant $\overline{V}$ or a constant $\underline{V}$ such that

$$\overline{V}(\nu,p) \;=\; \overline{V} \text{ a.s., or } \underline{V}(\nu,p) \;=\; \underline{V} \text{ a.s.}$$

then we say that the upper limiting average or (respectively) lower average value exists, and denote it by $\overline{V}(\nu,p):=\overline{V}$ (or $\underline{V}(\nu,p):=\underline{V}$). If both upper and lower average limiting values exist and are equal then we simply say that average limiting value exist and denote it by $V(\nu,p):=\overline{V}(\nu,p)=\underline{V}(\nu,p)$

An environment ν is *explorable* if there exists a policy p_ν such that $V(\nu,p_\nu)$ exists and $\overline{V}(\nu,p)\leq V(\nu,p_\nu)$ with probability 1 for every policy p. In this case define $V_\nu^*:=V(\nu,p_\nu)$. An environment ν is *upper explorable* if there exists a policy p_ν such that $\overline{V}(\nu,p_\nu)$ exists and $\overline{V}(\nu,p)\leq\overline{V}(\nu,p_\nu)$ with probability 1 for every policy p. In this case define $\overline{V}_\nu^*:=\overline{V}(\nu,p_\nu)$.

A policy p is *self-optimizing* for a set of explorable environments $\mathcal{C}$ if $V(\nu,p)=V_\nu^*$ for every $\nu\in\mathcal{C}$. A policy p is *upper self-optimizing* for a set of explorable environments $\mathcal{C}$ if $\overline{V}(\nu,p)=\overline{V}_\nu^*$ for every $\nu\in\mathcal{C}$.

In the case when we we wish to obtain the optimal average value for any environment in the class we will speak about self-optimizing policies, whereas if we are only interested in obtaining the upper limit of the average value

then we will speak about upper self-optimizing policies. It turns out that the latter case is much more simple. The next two definitions present conditions on the environments which will be shown to be sufficient to achieve the two respective goals.

Definition 4.1 (recoverable)**.** *We call an upper explorable environment ν recoverable if for any history $z_{<k}$ such that $\nu(x_{<k}|y_{<k})>0$ there exists a policy p such that*

$$P(\overline{V}(\nu,p)=\overline{V}^*_\nu|z_{<k})=1.$$

Conditioning on the history $z_{<k}$ means that we take ν-conditional probabilities (conditional on $x_{<k}$) and first $k-1$ actions of the policy p are replaced by $y_{<k}$.

Recoverability means that after taking any finite sequence of (possibly sub-optimal) actions it is still possible to obtain the same upper limiting average value as an optimal policy would obtain. The next definition is somewhat more complex.

Definition 4.2 (value-stable environments)**.** *An explorable environment ν is* value-stable *if there exist a sequence of numbers $r^\nu_i\in[0,r_{max}]$ and two functions $d_\nu(k,\varepsilon)$ and $\varphi_\nu(n,\varepsilon)$ such that $\frac{1}{n}r^\nu_{1..n}\to V^*_\nu$, $d_\nu(k,\varepsilon)=o(k)$, $\sum_{n=1}^\infty\varphi_\nu(n,\varepsilon)<\infty$ for every fixed ε, and for every k and every history $z_{<k}$ there exists a policy $p=p^{z_{<k}}_\nu$ such that*

$$P\left(r^\nu_{k..k+n}-r^{p\nu}_{k..k+n}>d_\nu(k,\varepsilon)+n\varepsilon\mid z_{<k}\right)\leq\varphi_\nu(n,\varepsilon). \tag{4.1}$$

First of all, this condition means that the strong law of large numbers for rewards holds uniformly over histories $z_{<k}$; the numbers r^ν_i here can be thought of as expected rewards of an optimal policy. Furthermore, the environment is "forgiving" in the following sense: from any (bad) sequence of k actions it is possible (knowing the environment) to recover up to $o(k)$ reward loss; to recover means to reach the level of reward obtained by the optimal policy which from the beginning was taking only optimal actions.

That is, suppose that a person A has made k possibly suboptimal actions and after that "realized" what the true environment was and how to act optimally in it. Suppose that a person B was from the beginning taking only optimal actions. We want to compare the performance of A and B on first n steps after the step k. An environment is value stable if A can catch up with B except for $o(k)$ gain. The numbers r_i^ν can be thought of as expected rewards of B; A can catch up with B up to the reward loss $d_\nu(k,\varepsilon)$ with probability $\varphi_\nu(n,\varepsilon)$, where the latter does not depend on past actions and observations (the law of large numbers holds uniformly).

Examples of value-stable environments will be considered in Section 4.4.

4.2 Self-optimizing policies for a set of value-stable environments

In this section we present the main self-optimizingness result along with an informal explanation of its proof, and a result on upper self-optimizingness, which turns out to have much more simple conditions.

Theorem 4.3 (value-stable⇒self-optimizing)**.** *For any countable set $\mathcal{C}$ of value-stable environments, there exists a policy which is self-optimizing for $\mathcal{C}$.*

A formal proof is given in Section 4.6; here we give some intuitive justification. Suppose that all environments in $\mathcal{C}$ are deterministic. We will construct a self-optimizing policy p as follows: Let ν^t be the first environment in $\mathcal{C}$. The algorithm assumes that the true environment is ν^t and tries to get ε-close to its optimal value for some (small) ε. This is called an exploitation part. If it succeeds, it does some exploration as follows. It picks the first environment ν^e which has higher average asymptotic value than ν^t ($V^*_{\nu^e} > V^*_{\nu^t}$) and tries to get ε-close to this value acting optimally under ν^e. If it can not get close to the ν^e-optimal value then ν^e is not the true environment, and the next environment can be picked for exploration (here we call

"exploration" successive attempts to exploit an environment which differs from the current hypothesis about the true environment and has a higher average reward). If it can, then it switches to exploitation of ν^t, exploits it until it is ε'-close to $V^*_{\nu^t}$, $\varepsilon' < \varepsilon$ and switches to ν^e again this time trying to get ε'-close to V_{ν^e}; and so on. This can happen only a finite number of times if the true environment is ν^t, since $V^*_{\nu^t} < V^*_{\nu^e}$. Thus after exploration either ν^t or ν^e is found to be inconsistent with the current history. If it is ν^e then just the next environment ν^e such that $V^*_{\nu^e} > V^*_{\nu^t}$ is picked for exploration. If it is ν^t then the first consistent environment is picked for exploitation (and denoted ν^t). This in turn can happen only a finite number of times before the true environment ν is picked as ν^t. After this, the algorithm still continues its exploration attempts, but can always keep within $\varepsilon_k \to 0$ of the optimal value. This is ensured by $d(k) = o(k)$.

The probabilistic case is somewhat more complicated since we can not say whether an environment is "consistent" with the current history. Instead we test each environment for consistency as follows. Let ξ be a mixture of all environments in $\mathcal{C}$. Observe that together with some fixed policy each environment μ can be considered as a measure on $\mathbf{Z}^\infty$. Moreover, it can be shown that (for any fixed policy) the ratio $\frac{\nu(z_{<n})}{\xi(z_{<n})}$ is bounded away from zero if ν is the true environment μ and tends to zero if ν is singular with μ (in fact, here singularity is a probabilistic analogue of inconsistency). The exploration part of the algorithm ensures that at least one of the environments ν^t and ν^e is singular with ν on the current history, and a succession of tests $\frac{\nu(z_{<n})}{\xi(z_{<n})} \geq \alpha_s$ with $\alpha_s \to 0$ is used to exclude such environments from consideration.

Upper self-optimizingness. Next we consider the task in which our goal is more moderate. Rather than trying to find a policy which will obtain the same average limiting value as an optimal one for any environment in a certain class, we will try to obtain only the optimum upper limiting average. That is, we will try to find a policy which infinitely often gets as close as desirable to the maximum possible average value. It turns out that in this case a much simpler condition is sufficient: recoverability instead of value-

stability.

Theorem 4.4 (recoverable⇒upper self-optimizing)**.** *For any countable class $\mathcal{C}$ of recoverable environments, there exists a policy which is upper self-optimizing for $\mathcal{C}$.*

A formal proof can be found in Section 4.6; its idea is as follows. The upper self-optimizing policy p to be constructed will loop through all environments in $\mathcal{C}$ in such a way that each environment is tried infinitely often, and for each environment the agent will try to get ε-close (with decreasing ε) to the upper-limiting average value, until it either manages to do so, or a special stopping condition holds: $\frac{\nu(z_{<n})}{\xi(z_{<n})} < \alpha_s$, where α_s is decreasing accordingly. This condition necessarily breaks if the upper limiting average value cannot be achieved.

4.3 Non-recoverable environments

Before proceeding with examples of value-stable environments, we briefly discuss what can be achieved if an environment does not forgive initial wrong actions, that is, is not recoverable. It turns out that value-stability can be defined for non-recoverable environments as well, and optimal — in a worst-case sense — policies can be identified.

For an environment ν, a policy p and a history $z_{<k}$ such that $\nu(x_{<k}|y_{<k}) > 0$, if there exists a constant $\overline{V}$ or a constant $\underline{V}$ such that

$$P(\overline{V}(\nu,p) = \overline{V}|z_{<k}) = 1, \text{ or } \; P(\underline{V}(\nu,p) = \underline{V}|z_{<k}) = 1,$$

then we say that the upper conditional (on $z_{<k}$) limiting average or (respectively) lower conditional average value exists, and denote it by $\overline{V}(\nu,p,z_{<k}) := \overline{V}$ (or $\underline{V}(\nu,p,z_{<k}) := \underline{V}$). If both upper and lower conditional average limiting values exist and are equal then we say that that average conditional value exist and denote it by $V(\nu,p,(z_{<k})) := \overline{V}(\nu,p,z_{<k}) = \underline{V}(\nu,p,z_{<k})$

Call an environment ν *strongly (upper) explorable* if for any history $z_{<k}$ such that $\nu(x_{<k}|y_{<k})>0$ there exists a policy $p_\nu^{z_{<k}}$ such that $V(\nu,p_\nu^{z_{<k}})$ ($\overline{V}(\nu,p_\nu^{z_{<k}})$) exists and $\overline{V}(\nu,p,z_{<k})\leq V(\nu,p_\nu^{z_{<k}},z_{<k})$ (respectively $\overline{V}(\nu,p,z_{<k})\leq \overline{V}(\nu,p_\nu^{z_{<k}},z_{<k})$) with probability 1 for every policy p. In this case define $V_\nu^*(z_{<k}):=V(\nu,p_\nu^{z_{<k}})$ (respectively $\overline{V}_\nu^*(z_{<k}):=\overline{V}(\nu,p_\nu^{z_{<k}})$).

For a strongly explorable environment ν define the worst-case optimal value

$$W_\nu^*:=\inf_{k,z_{<k}:\nu(x_{<k}>0)} V_\nu^*(z_{<k}),$$

and for a strongly upper explorable ν define the worst-case upper optimal value

$$\overline{W}_\nu^*:=\inf_{k,z_{<k}:\nu(x_{<k}>0)} \overline{V}_\nu^*(z_{<k}).$$

In words, the worst-case optimal value is the asymptotic average reward which is attainable with certainty after any finite sequence of actions has been taken.

Note that a recoverable explorable environment is also strongly explorable.

A policy p will be called *worst-case self-optimizing* or *worst-case upper self-optimizing* for a class of environments $\mathcal{C}$ if $\liminf \frac{1}{m} r_{1m}^{p\nu} \geq W_\nu^*$, or (respectively) $\limsup \frac{1}{m} r_{1m}^{p\nu} \geq \overline{W}_\nu^*$ with probability 1 for every $\nu\in\mathcal{C}$, where $r_{1..m}:=r_1+...+r_m$.

Definition 4.5 (worst-case value-stable environments)**.** *A strongly explorable environment ν is* worst-case value-stable *if there exists a sequence of numbers $r_i^\nu\in[0,r_{max}]$ and two functions $d_\nu(k,\varepsilon)$ and $\varphi_\nu(n,\varepsilon)$ such that $\frac{1}{n}r_{1..n}^\nu\to W_\nu^*$, $d_\nu(k,\varepsilon)=o(k)$, $\sum_{n=1}^\infty \varphi_\nu(n,\varepsilon)<\infty$ for every fixed ε, and for every k and every history $z_{<k}$ there exists a policy $p=p_\nu^{z_{<k}}$ such that*

$$P\left(r_{k..k+n}^\nu - r_{k..k+n}^{p\nu} > d_\nu(k,\varepsilon)+n\varepsilon \mid z_{<k}\right)\leq \varphi_\nu(n,\varepsilon). \tag{4.2}$$

Note that a recoverable environment is value-stable if and only if it is worst-case value-stable.

Worst-case value stability helps to distinguish between irreversible ac-

tions (or "traps") and actions which result only in a temporary loss in performance; moreover, worst-case value-stability means that a temporary loss in performance can only be short (sublinear).

Finally, we can establish the following result (cf. Theorems 4.3 and 4.4).

Theorem 4.6 (worst-case self-optimizing)**.** *(i) For any countable set of worst-case value-stable environments $\mathcal{C}$ there exist a policy p which is worst-case self-optimizing for $\mathcal{C}$.*

(ii) For any countable set of strongly upper explorable environments $\mathcal{C}$ there exist a policy p which is worst-case upper self-optimizing for $\mathcal{C}$.

The proof of this theorem is analogous to the proofs of Theorems 4.3 and 4.4; the differences are explained in Section 4.6.

4.4 Examples

In this section we illustrate the applicability of the results of the previous section by classes of value-stable environments. These are also examples of recoverable environments, since recoverability is strictly weaker than value-stability. In the end of the section we also give some simple examples of recoverable but not value-stable environments.

We first note that passive environments are value-stable. An environment is called *passive* if the observations and rewards do not depend on the actions of the agent. Sequence prediction task provides a well-studied (and perhaps the only reasonable) class of passive environments: in this task the agent is required to give the probability distribution of the next observation given the previous observations. The true distribution of observations depends only on the previous observations (and does not depend on actions and rewards). Since we have confined ourselves to considering finite action spaces, the agent is required to give ranges of probabilities for the next observation, where the ranges are fixed beforehand. The reward 1 is given if all the ranges are

correct and the reward 0 is given otherwise. It is easy to check that any such environment is value-stable with $r_i^\nu \equiv 1$, $d(k,\varepsilon) \equiv 1$, $\varphi(n,\varepsilon) \equiv 0$, since, knowing the distribution, one can always start giving the correct probability ranges (this defines the policy p_ν).

Obviously, there are active value stable environments too. The next proposition provides some conditions on mixing rates which are sufficient for value-stability; we do not intend to provide sharp conditions on mixing rates but rather to illustrate the relation of value-stability with mixing conditions.

We say that a stochastic process h_k, $k \in \mathbb{N}$ satisfies strong α-mixing conditions with coefficients $\alpha(k)$ if (see e.g. [12])

$$\sup_{n\in\mathbb{N}} \sup_{B\in\sigma(h_1,\dots,h_n),C\in\sigma(h_{n+k},\dots)} |P(B\cap C) - P(B)P(C)| \le \alpha(k),$$

where $\sigma()$ stands for the sigma-algebra generated by the random variables in brackets. Loosely speaking, mixing coefficients α reflect the speed with which the process "forgets" about its past.

Proposition 4.7 (mixing and value-stability)**.** *Suppose that an explorable environment ν is such that there exist a sequence of numbers r_i^ν and a function $d(k)$ such that $\frac{1}{n} r^\nu_{1..n} \to V^*_\nu$, $d(k) = o(k)$, and for each $z_{<k}$ there exists a policy p such that the sequence $r_i^{p\nu}$ satisfies strong α-mixing conditions with coefficients $\alpha(k) = \frac{1}{k^{1+\varepsilon}}$ for some $\varepsilon > 0$ and*

$$r^\nu_{k..k+n} - \mathbb{E}\left(r^{p\nu}_{k..k+n} \mid z_{<k}\right) \le d(k)$$

for any n. Then ν is value-stable.

Proof. Using the union bound we obtain

$$\begin{aligned} &P\left(r^\nu_{k..k+n} - r^{p\nu}_{k..k+n} > d(k) + n\varepsilon\right) \\ &\le I\left(r^\nu_{k..k+n} - \mathbb{E} r^{p\nu}_{k..k+n} > d(k)\right) + P\left(\left|r^{p\nu}_{k..k+n} - \mathbb{E} r^{p\nu}_{k..k+n}\right| > n\varepsilon\right). \end{aligned}$$

The first term equals 0 by assumption and the second term for each ε can

be shown to be summable using [12, Thm.1.3]: for a sequence of uniformly bounded zero-mean random variables r_i satisfying strong α-mixing conditions the following bound holds true for any integer $q \in [1,n/2]$

$$P\left(|r_{1..n}| > n\varepsilon\right) \leq ce^{-\varepsilon^2 q/c} + cq\alpha\left(\frac{n}{2q}\right)$$

for some constant c; in our case we just set $q = n^{\frac{\varepsilon}{2+\varepsilon}}$. □

(PO)MDPs. Applicability of Theorem 4.3 and Proposition 4.7 can be illustrated on (PO)MDPs. We note that self-optimizing policies for (uncountable) classes of finite ergodic MDPs and POMDPs are known [14, 31]; the aim of the present section is to show that value-stability is a weaker requirement than the requirements of these models, and also to illustrate applicability of our results. We call μ a (stationary) *Markov decision process* (MDP) if the probability of perceiving $x_k \in \mathbf{X}$, given history $z_{<k}y_k$ only depends on $y_k \in \mathbf{Y}$ and x_{k-1}. In this case $x_k \in \mathbf{X}$ is called a *state*, $\mathbf{X}$ the *state space*. An MDP μ is called *ergodic* if there exists a policy under which every state is visited infinitely often with probability 1. An MDP with a stationary policy forms a Markov chain.

An environment is called a (finite) *partially observable MDP* (POMDP) if there is a sequence of random variables s_k taking values in a finite space $\mathcal{S}$ called the state space, such that x_k depends only on s_k and y_k, and s_{k+1} is independent of $s_{<k}$ given s_k. Abusing notation the sequence $s_{1..k}$ is called the underlying Markov chain. A POMDP is called *ergodic* if there exists a policy such that the underlying Markov chain visits each state infinitely often with probability 1.

In particular, any ergodic POMDP ν satisfies strong α-mixing conditions with coefficients decaying exponentially fast in case there is a set $H \subset \mathbb{R}$ such that $\nu(r_i \in H) = 1$ and $\nu(r_i = r | s_i = s, y_i = y) \neq 0$ for each $y \in \mathbf{Y}, s \in \mathcal{S}, r \in H, i \in \mathbb{N}$. Thus for any such POMDP ν we can use Proposition 4.7 with $d(k,\varepsilon)$ a constant function to show that ν is value-stable:

Corollary 4.8 (POMDP⇒value-stable)**.** *Suppose that a POMDP ν is ergodic and there exists a set $H\subset\mathbb{R}$ such that $\nu(r_i\in H)=1$ and $\nu(r_i=r|s_i=s,y_i=y)\neq 0$ for each $y\in\mathbf{Y},h\in\mathcal{S},r\in H$, where $\mathcal{S}$ is the finite state space of the underlying Markov chain. Then ν is value-stable.*

However, it is illustrative to obtain this result for MDPs directly, and in a slightly stronger form.

Proposition 4.9 (MDP⇒value-stable)**.** *Any finite-state ergodic MDP ν is a value-stable environment.*

Proof. Let $d(k,\varepsilon)=0$. Denote by μ the true environment, let $z_{<k}$ be the current history and let the current state (the observation x_k) of the environment be $a\in\mathbf{X}$, where $\mathbf{X}$ is the set of all possible states. Observe that for an MDP there is an optimal policy which depends only on the current state. Moreover, such a policy is optimal for any history. Let p_μ be such a policy. Let r_i^μ be the expected reward of p_μ on step i. Let $l(a,b)=\min\{n:x_{k+n}=b|x_k=a\}$. By ergodicity of μ there exists a policy p for which $\mathbb{E}l(b,a)$ is finite (and does not depend on k). A policy p needs to get from the state b to one of the states visited by an optimal policy, and then acts according to p_μ. Let $f(n):=\frac{nr_{\max}}{\log n}$. We have

$$
\begin{aligned}
&P\left(\left|r^{\mu}_{k..k+n}-r^{p\mu}_{k..k+n}\right|>n\varepsilon\right)\leq\sup_{a\in\mathbf{X}}P\left(\left|\mathbb{E}\left(r^{p_\mu\mu}_{k..k+n}|x_k=a\right)-r^{p\mu}_{k..k+n}\right|>n\varepsilon\right)\right)\\
&\leq\sup_{a,b\in\mathbf{X}}P(l(a,b)>f(n)/r_{\max})\\
&+\sup_{a\in\mathbf{X}}P\left(\left|\mathbb{E}\left(r^{p_\mu\mu}_{k..k+n}|x_k=a\right)-r^{p_\mu\mu}_{k..k+n}\right|>n\varepsilon-2f(n)\middle|x_k=a\right).
\end{aligned}
$$

In the last term we have the deviation of the reward attained by the optimal policy from its expectation. Clearly, both terms are bounded exponentially in n. □

In the examples above the function $d(k,\varepsilon)$ is a constant and $\varphi(n,\varepsilon)$ decays exponentially fast. This suggests that the class of value-stable environments

stretches beyond finite (PO)MDPs. We illustrate this guess by the construction that follows.

A general scheme for constructing **value-stable environment or recoverable environments**: infinitely armed bandit. Next we present a construction of environments which cannot be modelled as finite POMDPs but are value-stable and/or recoverable. Consider the following environment ν. There is a countable family $\mathcal{C}' = \{\zeta_i : i \in \mathbb{N}\}$ of *arms*, that is, sources generating i.i.d. rewards 0 and 1 (and, say, empty observations) with some probability δ_i of the reward being 1. The action space $\mathbf{Y}$ consists of three actions $\mathbf{Y} = \{g,u,d\}$. To get the next reward from the current arm ζ_i an agent can use the action g. Let i denote the index of the current arm. At the beginning $i = 0$, the current arm is ζ_0 and then the agent can move between arms as follows: it can move $U(i)$ arms "up" using the action u (i.e. $i := i + U(i)$) or it can move $D(i)$ arms "down" using the action d (i.e. $i := i - D(i)$ or 0 if the result is negative). The reward for actions u and d is 0. In all the examples below $U(i) \equiv 1$, that is, the action u takes the agent one arm up.

Clearly, ν is a POMDP with countably infinite number of states in the underlying Markov chain, which (in general) is not isomorphic to a finite POMDP.

Proposition 4.10. *If $D(i) = i$ for all $i \in \mathbb{N}$ then the environment ν just constructed is value-stable. If $D(i) \equiv 1$ then ν is recoverable but not necessarily value-stable; that is, there are choices of the probabilities δ_i such that ν is not value-stable.*

Proof. First we show that in either case ($D(i) = i$ or $D(i) = 1$) ν is explorable. Let $\delta = \sup_{i \in \mathbb{N}} \delta_i$. Clearly, $\overline{V}(\nu,p') \le \delta$ with probability 1 for any policy p' . A policy p which, knowing all the probabilities δ_i, achieves $\overline{V}(\nu,p) = \underline{V}(\nu,p) = \delta =: V^*_\nu$ a.s., can be easily constructed. Indeed, find a sequence ζ'_j, $j \in \mathbb{N}$, where for each j there is $i =: i_j$ such that $\zeta'_j = \zeta_i$, satisfying $\lim_{j\to\infty} \delta_{i_j} = \delta$. The policy p should carefully exploit one by one the arms ζ_j, staying with each

arm long enough to ensure that the average reward is close to the expected reward with ε_j probability, where ε_j quickly tends to 0, and so that switching between arms has a negligible impact on the average reward. Thus ν can be shown to be explorable. Moreover, a policy p just sketched can be made independent on (observation and) rewards.

Next we show if $D(i)=i$, that is, the action d always takes the agent down to the first arm, then the environment in value-stable. Indeed, one can modify the policy p (possibly allowing it to exploit each arm longer) so that on each time step t (from some t on) we have $j(t)\leq\sqrt{t}$, where $j(t)$ is the number of the current arm on step t. Thus, after any actions-perceptions history $z_{<k}$ one needs about $\sqrt{k}$ actions (one action u and enough actions d) to catch up with p. So, (4.1) can be shown to hold with $d(k,\varepsilon)=\sqrt{k}$, r_i the expected reward of p on step i (since p is independent of rewards, $r_i^{p\nu}$ are independent), and the rates $\varphi(n,\varepsilon)$ exponential in n.

To construct a non-value-stable environment with $D(i)\equiv 1$, simply set $\delta_0=1$ and $\delta_j=0$ for $j>0$; then after taking n actions u one can only return to optimal rewards with n actions (d), that is $d(k)=o(n)$ cannot be obtained. Still it is easy to check that recoverability is preserved, whatever the choice of δ_i. □

In the above construction we can also allow the action d to bring the agent $d(i)<i$ steps down, where i is the number of the current environment ζ, according to some (possibly randomized) function $d(i)$, thus changing the function $d_\nu(k,\varepsilon)$ and possibly making it non-constant in ε and as close as desirable to linear.

4.5 Necessity of value-stability

Now we turn to the question of how tight the conditions of value-stability are. The following proposition shows that the requirement $d(k,\varepsilon)=o(k)$ in (4.1) can not be relaxed.

Proposition 4.11 (necessity of $d(k,\varepsilon)=o(k)$)**.** *There exists a countable family of deterministic explorable environments $\mathcal{C}$ such that*

- *for any $\nu\in\mathcal{C}$ for any sequence of actions $y_{<k}$ there exists a policy p such that*
$$r^{\nu}_{n..k+n}=r^{p\nu}_{k..k+n} \text{ for all } n\geq k,$$
where r^{ν}_i are the rewards attained by an optimal policy p_ν (which from the beginning was acting optimally), but

- *for any policy p there exists an environment $\nu\in\mathcal{C}$ such that $\underline{V}(\nu,p)<V^*_\nu$.*

Clearly, each environment from such a class $\mathcal{C}$ satisfies the value stability conditions with $\varphi(n,\varepsilon)\equiv 0$ except $d(k,\varepsilon)=k\neq o(k)$.

Proof. There are two possible actions $y_i\in\{a,b\}$, three possible rewards $r_i\in\{0,1,2\}$ and no observations.

Construct the environment ν_0 as follows: if $y_i=a$ then $r_i=1$ and if $y_i=b$ then $r_i=0$ for any $i\in\mathbb{N}$.

For each i let n_i denote the number of actions a taken up to step i: $n_i:=\#\{j\leq i: y_j=a\}$. For each $s>0$ construct the environment ν_s as follows: $r_i(a)=1$ for any i, $r_i(b)=2$ if the longest consecutive sequence of action b taken has length greater than n_i and $n_i\geq s$; otherwise $r_i(b)=0$.

Suppose that there exists a policy p such that $\underline{V}(\nu_i,p)=V^*_{\nu_i}$ for each $i>0$ and let the true environment be ν_0. By assumption, for each s there exists such n that

$$\#\{i\leq n: y_i=b, r_i=0\}\geq s>\#\{i\leq n: y_i=a, r_i=1\}$$

which implies $\underline{V}(\nu_0,p)\leq 1/2<1=V^*_{\nu_0}$. □

It is also easy to show that the *uniformity of convergence in (4.1)* can not be dropped. That is, if in the definition of value-stability we allow the function $\varphi(n,\varepsilon)$ to depend additionally on the past history $z_{<k}$ then

Theorem 4.3 does not hold. This can be shown with the same example as constructed in the proof of Proposition 4.11, letting $d(k,\varepsilon)\equiv 0$ but instead allowing $\varphi(n,\varepsilon,z_{<k})$ to take values 0 and 1 according to the number of actions a taken, achieving the same behaviour as in the example provided in the last proof.

Moreover, we show that the requirement that the class $\mathcal{C}$ to be learnt is countable can not be easily withdrawn. Indeed, consider the class of all deterministic passive environments in the sequence prediction setting. In this task an agent gets the reward 1 if $y_i = o_{i+1}$ and the reward 0 otherwise, where the sequence of observation o_i is deterministic. Different sequences correspond to different environments. As it was mentioned before, any such environment ν is value-stable with $d_\nu(k,\varepsilon)\equiv 1$, $\varphi_\nu(n,\varepsilon)\equiv 0$ and $r_i^\nu \equiv 1$. Obviously, the class of all deterministic passive environments is not countable. Since for every policy p there is an environment on which p errs exactly on each step, the class of all deterministic passive environments can not be learned. Therefore, the following statement is valid:

Proposition 4.12. *There exist (uncountable) classes of value-stable environments for which there are no self-optimizing policies.*

However, strictly speaking, even for countable classes value-stability is not necessary for self-optimizingness. This can be demonstrated on the class $\nu_i : i > 0$ from the proof of Proposition 4.11. (Whereas if we add ν_0 to the class a self-optimizing policy no longer exists.) So we have the following:

Proposition 4.13. *There are countable classes of not value-stable environments for which self-optimizing policies exist.*

4.6 Longer proofs

In each of the proofs, a self-optimizing (or upper self-optimizing) policy p will be constructed. When the policy p has been defined up to a step k, an

environment μ, endowed with this policy, can be considered as a measure on $\mathbf{Z}^k$. We assume this meaning when we use environments as measures on $\mathbf{Z}^k$ (e.g. $\mu(z_{<i})$).

Proof of Theorem 4.3. A self-optimizing policy p will be constructed as follows. On each step we will have two polices: p^t which exploits and p^e which explores; for each i the policy p either takes an action according to p^t ($p(z_{<i})=p^t(z_{<i})$) or according to p^e ($p(z_{<i})=p^e(z_{<i})$), as will be specified below.

In the algorithm below, i denotes the number of the current step in the sequence of actions-observations. Let $n=1$, $s=1$, and $j^t=j^e=0$. Let also $\alpha_s=2^{-s}$ for $s\in\mathbb{N}$. For each environment ν, find such a sequence of real numbers ε_n^ν that $\varepsilon_n^\nu\to 0$ and $\sum_{n=1}^\infty \varphi_\nu(n,\varepsilon_n^\nu)\le\infty$.

Let $ß:\mathbb{N}\to\mathcal{C}$ be such a numbering that each $\nu\in\mathcal{C}$ has infinitely many indices. For all $i>1$ define a measure ξ as follows

$$\xi(z_{<i}) = \sum_{\nu\in\mathcal{C}} w_\nu \nu(z_{<i}), \tag{4.3}$$

where $w_\nu\in\mathbb{R}$ are (any) such numbers that $\sum_\nu w_\nu=1$ and $w_\nu>0$ for all $\nu\in\mathcal{C}$.
Define T. On each step i let

$$T \;\equiv\; T_i \;:=\; \left\{\nu\in\mathcal{C} : \frac{\nu(z_{<i})}{\xi(z_{<i})}\ge\alpha_s\right\}$$

Define ν^t. Set ν^t to be the first environment in T with index greater than $ß(j^t)$. In case this is impossible (that is, if T is empty), increment s, (re)define T and try again. Increment j^t.
Define ν^e. Set ν^e to be the first environment with index greater than $ß(j^e)$ such that $V^*_{\nu^e}>V^*_{\nu^t}$ and $\nu^e(z_{<k})>0$, if such an environment exists. Otherwise proceed one step (according to p^t) and try again. Increment j^e.
Consistency. On each step i (re)define T. If $\nu^t\notin T$, define ν^t, increment s and iterate the infinite loop. (Thus s is incremented only if ν^t is not in T or

if T is empty.)

Start the **infinite loop**. Increment n.

Let $\delta := (V^*_{\nu^e} - V^*_{\nu^t})/2$. Let $\varepsilon := \varepsilon^{\nu^t}_n$. If $\varepsilon < \delta$ set $\delta = \varepsilon$. Let $h = j^e$.

Prepare for exploration.

Increment h. The index h is incremented with each next attempt of exploring ν^e. Each attempt will be at least h steps in length.

Let $p^t = p^{y_{<i}}_{\nu^t}$ and set $p = p^t$.

Let i_h be the current step. Find k_1 such that

$$\frac{i_h}{k_1} V^*_{\nu^t} \leq \varepsilon/8 \tag{4.4}$$

Find $k_2 > 2i_h$ such that for all $m > k_2$

$$\left| \frac{1}{m - i_h} r^{\nu^t}_{i_h+1..m} - V^*_{\nu^t} \right| \leq \varepsilon/8. \tag{4.5}$$

Find k_3 such that

$$h r_{max}/k_3 < \varepsilon/8. \tag{4.6}$$

Find k_4 such that for all $m > k_4$

$$\frac{1}{m} d_{\nu^e}(m, \varepsilon/4) \leq \varepsilon/8, \quad \frac{1}{m} d_{\nu^t}(m, \varepsilon/8) \leq \varepsilon/8 \quad \text{and} \quad \frac{1}{m} d_{\nu^t}(i_h, \varepsilon/8) \leq \varepsilon/8. \tag{4.7}$$

Moreover, it is always possible to find such $k > \max\{k_1, k_2, k_3, k_4\}$ that

$$\frac{1}{2k} r^{\nu^e}_{k..3k} \geq \frac{1}{2k} r^{\nu^t}_{k..3k} + \delta. \tag{4.8}$$

Iterate up to the step k.

Exploration. Set $p^e = p^{y_{<n}}_{\nu^e}$. Iterate h steps according to $p = p^e$. Iterate further until either of the following conditions breaks

(i) $\left| r^{\nu^e}_{k..i} - r^{p\nu}_{k..i} \right| < (i-k)\varepsilon/4 + d_{\nu^e}(k, \varepsilon/4)$,

(ii) $i < 3k$.

(iii) $\nu^e \in T$.

Observe that either (i) or (ii) is necessarily broken.

If on some step ν^t is excluded from T then the infinite loop is iterated. If after exploration ν^e is not in T then redefine ν^e and **iterate the infinite loop**. If both ν^t and ν^e are still in T then **return** to "Prepare for exploration" (otherwise the loop is iterated with either ν^t or ν^e changed).
End of the infinite loop and the algorithm.

Let us show that with probability 1 the "Exploration" part is iterated only a finite number of times in a row with the same ν^t and ν^e.

Suppose the contrary, that is, suppose that (with some non-zero probability) the "Exploration" part is iterated infinitely often while $\nu^t, \nu^e \in T$. Observe that (4.1) implies that the ν^e-probability that (i) breaks is not greater than $\varphi_{\nu_e}(i-k, \varepsilon/4)$; hence by Borel-Cantelli lemma the event that (i) breaks infinitely often has probability 0 under ν^e.

Suppose that (i) holds almost every time. Then (ii) should be broken except for a finite number of times. We can use (4.4), (4.5), (4.7) and (4.8) to show that with probability at least $1-\varphi_{\nu^t}(k-i_h, \varepsilon/4)$ under ν^t we have $\frac{1}{3k} r^{p\nu^t}_{1..3k} \geq V^*_{\nu^t} + \varepsilon/2$. Again using Borel-Cantelli lemma and $k > 2i_h$ we obtain that the event that (ii) breaks infinitely often has probability 0 under ν^t.

Thus (at least) one of the environments ν^t and ν^e is singular with respect to the true environment ν given the described policy and current history. Denote this environment by ν'. It is known (see e.g. [21, Thm.26]) that if measures μ and ν are mutually singular then $\frac{\mu(x_1,\dots,x_n)}{\nu(x_1,\dots,x_n)} \to \infty$ μ-a.s. Thus

$$\frac{\nu'(z_{<i})}{\nu(z_{<i})} \to 0 \ \nu\text{-a.s.} \tag{4.9}$$

Observe that (by definition of ξ) $\frac{\nu(z_{<i})}{\xi(z_{<i})}$ is bounded. Hence using (4.9) we can see that

$$\frac{\nu'(z_{<i})}{\xi(z_{<i})} \to 0 \ \nu\text{-a.s.}$$

Since s and α_s are not changed during the exploration phase this implies

that on some step ν' will be excluded from T according to the "consistency" condition, which contradicts the assumption. Thus the "Exploration" part is iterated only a finite number of times in a row with the same ν^t and ν^e.

Observe that s is incremented only a finite number of times since $\frac{\nu'(z_{<i})}{\xi(z_{<i})}$ is bounded away from 0 where ν' is either the true environment ν or any environment from $\mathcal{C}$ which is equivalent to ν on the current history. The latter follows from the fact that $\frac{\xi(z_{<i})}{\nu(z_{<i})}$ is a submartingale with bounded expectation, and hence, by the submartingale convergence theorem (see e.g. [30]) converges with ν-probability 1.

Let us show that from some step on ν (or an environment equivalent to it) is always in T and selected as ν^t. Consider the environment ν^t on some step i. If $V^*_{\nu^t} > V^*_\nu$ then ν^t will be excluded from T since on any optimal for ν^t sequence of actions (policy) measures ν and ν^t are singular. If $V^*_{\nu^t} < V^*_\nu$ than ν^e will be equal to ν at some point, and, after this happens sufficient number of times, ν^t will be excluded from T by the "exploration" part of the algorithm, s will be decremented and ν will be included into T. Finally, if $V^*_{\nu^t} = V^*_\nu$ then either the optimal value V^*_ν is (asymptotically) attained by the policy p_t of the algorithm, or (if p_{ν^t} is suboptimal for ν) $\frac{1}{i} r^{p\nu^t}_{1..i} < V^*_{\nu^t} - \varepsilon$ infinitely often for some ε, which has probability 0 under ν^t and consequently ν^t is excluded from T.

Thus, the exploration part ensures that all environments not equivalent to ν with indices smaller than $ß(\nu)$ are removed from T and so from some step on ν^t is equal to (an environment equivalent to) the true environment ν.

We have shown in the "Exploration" part that $n \to \infty$, and so $\varepsilon_n^{\nu^t} \to 0$. Finally, using the same argument as before (Borel-Cantelli lemma, (i) and the definition of k) we can show that in the "exploration" and "prepare for exploration" parts of the algorithm the average value is within $\varepsilon_n^{\nu^t}$ of $V^*_{\nu^t}$ provided the true environment is (equivalent to) ν^t. □

Proof of Theorem 4.4. Let $ß: \mathbb{N} \to \mathcal{C}$ be such a numbering that each $\nu \in \mathcal{C}$ has infinitely many indices. Define the measure ξ as in (4.3). The policy p

acts according to the following algorithm.

Set $\varepsilon_s = \alpha_s = 2^{-s}$ for $s \in \mathbb{N}$, set $j = 1$, $s = n = 1$. The integer i will denote the current step in time.

Do the following *ad infinitum*. Set ν to be the first environment in $\mathcal{C}$ with index greater than $ß(j)$. Find the policy p_ν which achieves the upper limiting average value with probability one (such policy exists by definition of recoverability). Act according to p_ν until either

$$\left| \frac{1}{i} r^{p\nu}_{1..i} - \overline{V}^*(p, p_\nu) \right| < \varepsilon_n \tag{4.10}$$

or

$$\frac{\nu(z_{<i})}{\xi(z_{<i})} < \alpha_s. \tag{4.11}$$

Increment n, s, i.

It can be easily seen that one of the conditions necessarily breaks. Indeed, either in the true environment the optimal upper limiting average value for the current environment ν can be achieved by the optimal policy p_ν, in which case (4.10) breaks; or it cannot be achieved, which means that ν and ξ are singular, which implies that (4.11) will be broken (see e.g. [21, Thm.26]; cf. the same argument in the proof of Theorem 4.3). Since ν equals the true environment infinitely often and $\varepsilon_n \to 0$ we get the statement of the theorem. □

Proof of Theorem 4.6 is analogous to the proofs of Theorems 4.3 and 4.4, except for the following. Instead of the optimal average value V^*_ν and upper optimal average value $\overline{V}^*_\nu$ the values $V^*_\nu(z_{<k})$ and $\overline{V}^*_\nu(z_{<k})$ should be used, and they should be updated after each step k. □

Chapter 5

Classification [R9, R10]

The problem of classification (or pattern recognition) consists in assigning a (discrete-valued) label Y for an object X, on the basis of a sample of object-label pairs $(X_1,Y_1),...,(X_n,Y_n)$.

Classification is perhaps the one learning problem for which the question of learnability is well-understood; at least, this is the case under the assumption that the examples (X_i,Y_i) are independent and identically distributed. While this assumption is very strong, it may be considered reasonable for a variety of applications (with some notable exceptions which we will consider below). In this setting, the question of statistical learnability is effectively solved by the Vapnik-Chervonenkis theory; those classes of functions f can be learned (from finite samples) that have finite VC dimension [89]. In this case, one can use empirical risk minimization as a learning rule.

The contribution to this area presented here is two-fold. First, it is shown that the i.i.d. requirement on the distribution of samples is redundant, in the sense that most of the learning algorithms developed to work under this setting can be provably used in a more general (and, as it is argued below, more natural for many applications) setting. These results (extracted from my Ph.D. thesis) are presented in Section 5.1.

Second, it is demonstrated that the characterization of learnability presented by the VC theory is not quite complete, in the sense the number

of samples required to learn a classification function increases from linear in the VC dimension to arbitrary fast-growing functions, if one limits the consideration to computable methods only. This result is presented in Section 5.2.

5.1 Relaxing the i.i.d. assumption in classification [R10]

As it was mentioned, the majority of methods developed for solving the problem of classification rely on the assumptions that the examples are independent and identically distributed. This section is devoted to relaxing this assumption.

Consider the following example, that helps to **motivate** this problem. Suppose we are trying to recognise a hand-written text. Obviously, letters in the text are dependent (for example, we strongly expect to meet "u" after "q"). This seemingly implies that classification methods can not be applied to this task, which is, however, one of their classical applications.

We show that the following two assumptions on the distribution of examples are sufficient for classification. First, that the dependence between objects is only that between their labels and the type of object-label dependence does not change in time. Second, the rate of occurrence of each label should keep above some positive threshold.

These intuitive ideas lead us to the following model (to which we refer as "the conditional model"). The labels $y \in \mathbf{Y}$ are drawn according to some unknown (but fixed) distribution over the set of all infinite sequences of labels. There can be any type of dependence between labels; moreover, we can assume that we are dealing with any (fixed) combinatorial sequence of labels. However, in this sequence the rate of occurrence of each label should keep above some positive threshold. For each label y the corresponding object $x \in \mathbf{X}$ is generated according to some (unknown but fixed) probability

distribution $P(x|y)$. All the rest is as in the i.i.d. model.

The main difference from the i.i.d. model is in that in the conditional model the distribution of labels is arbitrary (apart from the frequency threshold requirement).

In this section we provide a tool for obtaining estimations of probability of error of a predictor in the conditional model from an estimation of the probability of error in the i.i.d. model. The general theorems about extending results concerning performance of a predictor to the conditional model are illustrated on two classes of predictors. First, we extend weak consistency results concerning partitioning and nearest neighbour estimates from the i.i.d. model to the conditional model. Second, we use some results of Vapnik-Chervonenkis theory to estimate performance in the conditional model (on finite amount of data) of predictors minimising empirical risk, and also obtain some strong consistency results.

These results on specific predictions methods are obtained as applications of the following observation. The only assumption on a predictor under which a predictor works in the new model as well as in the i.i.d. model is what we call *tolerance to data* (a stability-kind condition): in any large dataset there is no small subset which strongly changes the probability of error. This property should also hold with respect to permutations. This assumption on a predictor should be valid in the i.i.d. model. Thus, the results achieved in the i.i.d. model can be extended to the conditional model; this concerns distribution–free results as well as distribution–specific, results on the performance on finite samples as well as asymptotic results.

Related work. Various approaches to relaxing the i.i.d. assumption in learning tasks have been proposed in the literature. Thus, in [56, 55] the the nearest neighbour and kernel estimators are studied in the setting of regression estimation with continuous regression function, under the assumption that labels are conditionally independent given their objects, while objects form any individual sequence (which is the opposite to what we do). There are also several approaches in which different types of assumptions on the

joint distribution of objects and labels are made; then the authors construct a predictor or a class of predictors, to work well under the assumptions made. Thus, in [34, 3] a generalisation of PAC approach to Markov chains with finite or countable state space is presented. There is also a track of research on prediction under the assumption that the distribution generating examples is stationary and ergodic. The basic difference from our learning task, apart from different probabilistic assumption, is in that we are only concerned with object-label dependence, while in predicting ergodic sequences it is label-label (time series) dependence that is of primary interest. On this task see [78, 4, 64, 66] and references therein.

5.1.1 Definitions and general results

Consider a sequence of examples $(x_1,y_1),(x_2,y_2),...$; where each example $z_i := (x_i,y_i)$ consists of an object $x_i \in \mathbf{X}$ and a label $y_i := \eta(x_i) \in \mathbf{Y}$. Here $\mathbf{X}$ is a measurable space, $\mathbf{Y} := \{0,1\}$, and $\eta : \mathbf{X} \to \mathbf{Y}$ is some deterministic function. For simplicity, we made the assumption that the space $\mathbf{Y}$ is binary, but all results easily extend to the case of any finite space $\mathbf{Y}$. The notation $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is used for the measurable space of examples. Objects are drawn according to some probability distribution $\mathbf{P}$ on $\mathbf{X}^\infty$ (and labels are defined by η). Thus we consider only the case of deterministically defined labels (that is, the noise-free model).

The notation $\mathbf{P}$ is used for distributions on $\mathbf{X}^\infty$ while the symbol P is reserved for distributions on $\mathbf{X}$. In the latter case P^∞ denotes the i.i.d. distribution on $\mathbf{X}^\infty$ generated by P. Correspondingly we will use symbols $\mathbb{E}$, E and $\mathbb{E}^\infty$ for expectations over spaces $\mathbf{X}^\infty$ and $\mathbf{X}$.

The traditional assumption about the distribution $\mathbf{P}$ generating objects is that examples are independently and identically distributed (i.i.d.) according to some distribution P on $\mathbf{X}$ (i.e. $\mathbf{P} = P^\infty$).

Here we replace this assumption with the following two conditions.

First, for any $n \in \mathbb{N}$ and for any measurable set $A \subset \mathbf{X}$

$$\mathbf{P}(X_n \in A \,|\, Y_n, X_1, Y_1, ..., X_{n-1}, Y_{n-1}) = \mathbf{P}(X_n \in A \,|\, Y_n) \tag{5.1}$$

(i.e. some versions of conditional probabilities coincide). This condition looks very much like Markov condition which requires that each object depends on the past only through its immediate predecessor. The condition (5.1) says that each object depends on the past only through its label.

Second, for any $y \in \mathbf{Y}$, for any $n_1, n_2 \in \mathbb{N}$ and for any measurable set $A \subset \mathbf{X}$

$$\mathbf{P}(X_{n_1} \in A \,|\, Y_{n_1} = y) = \mathbf{P}(X_{n_2} \in A \,|\, Y_{n_2} = y) \tag{5.2}$$

(i.e. the process is uniform in time; (5.1) allows dependence in n).

Note that the first condition means that objects are conditionally independent given labels (on conditional independence see [22]).

Definition 5.1. *Under the conditions (5.1) and (5.2) we say that* objects are conditionally independent and identically distributed *(conditionally i.i.d).*

For each $y \in \mathbf{Y}$ denote the distribution $\mathbf{P}(X_n \mid Y_n = y)$ by P_y (it does not depend on n by (5.2)). Clearly, the distributions P_0 and P_1 define some distributions P on $\mathbf{X}$ up to a parameter $p \in [0,1]$. That is, $P_p(A) = pP_1(A) + (1-p)P_0(A)$ for any measurable set $A \subset \mathbf{X}$ and for each $p \in [0,1]$. Thus with each distribution $\mathbf{P}$ satisfying the assumptions (5.1) and (5.2) we will associate a family of distributions P_p, $p \in [0,1]$.

The assumptions of the conditional model can be also interpreted as follows. Assume that we have some individual sequence $(y_n)_{n \in \mathbb{N}}$ of labels and two probability distributions P_0 and P_1 on $\mathbf{X}$, such that there exists sets X_0 and X_1 in $\mathbf{X}$ such that $P_1(X_1) = P_0(X_0) = 1$ and $P_0(X_1) = P_1(X_0) = 0$ (i.e. X_0 and X_1 define some function η). Each example $x_n \in \mathbf{X}$ is drawn according to the distribution P_{y_n}; examples are drawn independently of each other.

A *predictor* is a measurable function $\Gamma_n := \Gamma(x_1,y_1,...,x_n,y_n,x_{n+1})$ taking values in $\mathbf{Y}$ (more formally, a family of functions indexed by n).

The probability of error of a predictor Γ on each step n is defined as

$$\mathrm{err}_n(\Gamma,\mathbf{P},z_1,...,z_n) := \mathbf{P}\big\{(x,y)\in\mathbf{Z} : y\neq\Gamma_n(z_1,...,z_n,x)\big\}$$

(z_i, $1\leq i\leq n$ are fixed and the probability is taken over z_{n+1}). We will sometimes omit some of the arguments of err_n when it can cause no confusion; in particular, we will often use a short notation $\mathbf{P}(\mathrm{err}_n(\Gamma,Z_1,...,Z_n)>\varepsilon)$ and an even shorter one $\mathbf{P}(\mathrm{err}_n(\Gamma)>\varepsilon)$ in place of

$$\mathbf{P}\big\{z_1,...,z_n : \mathrm{err}_n(\Gamma,\mathbf{P},z_1,...,z_n)>\varepsilon\big\}.$$

For a pair of distributions P_0 and P_1 and any $\delta\in(0,1/2)$ define

$$\bigtriangledown_\delta(P_0,P_1,n,\varepsilon) := \sup_{p\in[\delta,1-\delta]} P_p^\infty(\mathrm{err}_n(\Gamma)>\varepsilon) \tag{5.3}$$

For a predictor Γ and a distribution P on $\mathbf{X}$ define

$$\Delta(P,n,z_1,...,z_n) := \max_{j\leq\varkappa_n;\ \pi:\{1,...,n\}\to\{1,...,n\}} \big|\mathrm{err}_n(\Gamma,P^\infty,z_1,...,z_n) - \mathrm{err}_{n-j}(\Gamma,P^\infty,z_{\pi(1)},...,z_{\pi(n-j)})\big|.$$

Define the *tolerance to data* of Γ as

$$\Delta(P,n,\varepsilon) := P^n\big(\Delta(P,n,Z_1,...,Z_n)>\varepsilon\big) \tag{5.4}$$

for any $n\in\mathbb{N}$, any $\varepsilon>0$ and $\varkappa_n := \sqrt{n\log n}$. Furthermore, for a pair of distributions P_0 and P_1 and any $\delta\in(0,1/2)$ define

$$\Delta_\delta(P_0,P_1,n,\varepsilon) := \sup_{p\in[\delta,1-\delta]} \Delta(P_p,n,\varepsilon).$$

Tolerance to data means, in effect, that in any typical large portion of

data there is no small portion that changes strongly the probability of error. This property should also hold with respect to permutations.

We will also use another version of tolerance to data, in which instead of removing some examples we replace them with an arbitrary sample $z'_j,...,z'_n$ consistent with η:

$$\bar{\Delta}(P,z_1,...,z_n) := \sup_{j<\varkappa_n;\pi:\{1,...,n\}\to\{1,...,n\};z'_{n-j},...,z'_n} |\mathrm{err}_n(\Gamma,P^\infty,z_1,...,z_n) - \mathrm{err}_n(\Gamma,P^\infty,\zeta_1,...,\zeta_n)|,$$

where $\zeta_{\pi(i)} := z_{\pi(i)}$ if $i<n-j$ and $\zeta_{\pi(i)} := z'_i$ otherwise; the maximum is taken over all z'_i, $n-j<i\le n$ consistent with η. Define

$$\bar{\Delta}(P,n,\varepsilon) := P^n\big(\bar{\Delta}(P,n,Z_1,...,Z_n) > \varepsilon\big)$$

and

$$\bar{\Delta}_\delta(P_0,P_1,n,\varepsilon) := \sup_{p\in[\delta,1-\delta]} \bar{\Delta}(P_p,n,\varepsilon).$$

The same notational convention will be applied to Δ and $\bar{\Delta}$ as to err_n.

Various notions similar to tolerance to data have been studied in literature. Perhaps first they appeared in connection with deleted or condensed estimates (see e.g. [74]), and were later called stability (see [13, 48] for present studies of different kinds of stability, and for extensive overviews). Naturally, such notions arise when there is a need to study the behaviour of a predictor when some of the training examples are removed. These notions are much similar to what we call tolerance to data, only we are interested in the maximal deviation of probability of error while usually it is the average or minimal deviations that are estimated.

A predictor developed to work in the off-line setting should be, loosely speaking, tolerant to small changes in a training sample. The next theorem shows under which conditions this property of a predictor can be utilised.

Theorem 5.2. *Suppose that a distribution* $\mathbf{P}$ *generating examples is such*

that the objects are conditionally i.i.d, i.e. $\mathbf{P}$ *satisfies (5.1) and (5.2). Fix some* $\delta\in(0,1/2]$*, let* $p(n):=\frac{1}{n}\#\{i\le n: Y_i=1\}$ *and* $C_n:=\mathbf{P}(\delta\le p(n)\le 1-\delta)$ *for each* $n\in\mathbb{N}$*. Let also* $\alpha_n:=\frac{1}{1-1/\sqrt{n}}$*. For any predictor* Γ *and any* $\varepsilon>0$ *we have*

$$\mathbf{P}(\mathrm{err}_n(\Gamma)>\varepsilon)\le C_n^{-1}\alpha_n\big(\bigtriangledown_\delta(P_0,P_1,n+\varkappa_n,\delta\varepsilon/2) + \Delta_\delta(P_0,P_1,n+\varkappa_n,\delta\varepsilon/2)\big)+(1-C_n), \quad (5.5)$$

and

$$\mathbf{P}(\mathrm{err}_n(\Gamma)>\varepsilon)\le C_n^{-1}\alpha_n\big(\bigtriangledown_\delta(P_0,P_1,n,\delta\varepsilon/2) + \bar{\Delta}_\delta(P_0,P_1,n,\delta\varepsilon/2)\big)+(1-C_n). \quad (5.6)$$

The proofs for this section can be found in Section 5.3.1.

The theorem says that if we know with some confidence C_n that the rate of occurrence of each label is not less than some (small) δ, and have some bounds on the error rate and tolerance to data of a predictor in the i.i.d. model, then we can obtain bounds on its error rate in the conditional model.

Thus we have a tool for estimating the performance of a predictor on each finite step n. In Section 5.1.3 we will show how this result can be applied to predictors minimising empirical risk. However, if we are only interested in asymptotic results the formulations can be somewhat simplified.

Consider the following asymptotic condition on the frequencies of labels. Define $p(n):=\frac{1}{n}\#\{i\le n: Y_i=1\}$. We say that the *rates of occurrence of labels are bounded from below* if there exist such δ, $0<\delta<1/2$ that

$$\lim_{n\to\infty}\mathbf{P}(p(n)\in[\delta,1-\delta])=1. \quad (5.7)$$

As the condition (5.7) means $C_n\to 1$ we can derive from Theorem 5.2 the following corollary.

Corollary 5.3. *Suppose that a distribution* $\mathbf{P}$ *satisfies (5.1), (5.2), and*

(5.7) for some $\delta \in (0,1/2]$. *Let* Γ *be such a predictor that*

$$\lim_{n\to\infty} \bigtriangledown_\delta(P_0,P_1,n,\varepsilon) = 0 \tag{5.8}$$

and either

$$\lim_{n\to\infty} \Delta_\delta(P_0,P_1,n,\varepsilon) = 0 \tag{5.9}$$

or

$$\lim_{n\to\infty} \bar{\Delta}_\delta(P_0,P_1,n,\varepsilon) = 0 \tag{5.10}$$

for any $\varepsilon > 0$. *Then*

$$\mathbb{E}(\mathrm{err}_n(\Gamma,\mathbf{P},Z_1,...,Z_n)) \to 0.$$

In Section 5.1.2 we show how this statement can be applied to prove weak consistence of some classical nonparametric predictors in the conditional model.

5.1.2 Application to classical nonparametric predictors

In this section we will consider two types of classical nonparametric predictors: partitioning and nearest neighbour classifiers.

The nearest neighbour predictor assigns to a new object x_{n+1} the label of its nearest neighbour among $x_1,...,x_n$:

$$\Gamma_n(x_1,y_1,...,x_n,y_n,x_{n+1}) := y_j,$$

where $j := \mathrm{argmin}_{i=1,...,n} \|x - x_i\|$.

For i.i.d. distributions this predictor is also consistent, i.e.

$$E^\infty(\mathrm{err}_n(\Gamma,P^\infty)) \to 0,$$

for any distribution P on $\mathbf{X}$ (see [27]).

We generalise this result as follows.

Theorem 5.4. *Let* Γ *be the nearest neighbour classifier. Let* $\mathbf{P}$ *be some distribution on* $\mathbf{X}^\infty$ *satisfying (5.1), (5.2) and (5.7). Then*

$$\mathbb{E}(\mathrm{err}_n(\Gamma,\mathbf{P}))\to 0.$$

The proofs for this section can be found in Section 5.3.2.

A partitioning predictor on each step n partitions the object space $\mathbf{X}=\mathbb{R}^d$, $d\in\mathbb{N}$ into disjoint cells $A_1^n,A_2^n,\ldots$ and classifies in each cell according to the majority vote:

$$\Gamma(z_1,\ldots,z_n,x):=\begin{cases} 0 & \text{if } \sum_{i=1}^n I_{y_i=1}I_{x_i\in A(x)}\le\sum_{i=1}^n I_{y_i=0}I_{x_i\in A(x)} \\ 1 & \text{otherwise,}\end{cases}$$

where $A(x)$ denotes the cell containing x. Define

$$\mathrm{diam}(A):=\sup_{x,y\in A}\|x-y\|$$

and

$$N(x):=\sum_{i=1}^n I_{x_i\in A(x)}.$$

It is a well known result (see, e.g. [26]) that a partitioning predictor is weakly consistent, provided certain regulatory conditions on the size of cells. More precisely, let Γ be a partitioning predictor such that $\mathrm{diam}(A(X))\to 0$ in probability and $N(X)\to\infty$ in probability. Then for any distribution P on $\mathbf{X}$

$$E^\infty(\mathrm{err}_n(\Gamma,P^\infty))\to 0.$$

We generalise this result to the case of conditionally i.i.d. examples as follows.

Theorem 5.5. *Let* Γ *be a partitioning predictor such that* $\mathrm{diam}(A(X))\to 0$ *in probability and* $N(X)\to\infty$ *in probability, for any distribution generating*

i.i.d. examples. Then

$$\mathbb{E}(\mathrm{err}_n(\Gamma,\mathbf{P}))\to 0$$

for any distribution $\mathbf{P}$ *on* $\mathbf{X}^\infty$ *satisfying (5.1), (5.2) and (5.7).*

Observe that we only generalise results concerning weak consistency of (one) nearest neighbour and non-data-dependent partitioning rules. More general results exist (see e.g. [28]), in particular for data-dependent rules. However, we do not aim to generalise state-of-the-art results in nonparametric classification, but rather to illustrate that weak consistency results can be extended to the conditional model.

5.1.3 Application to empirical risk minimisation.

In this section we show how to estimate the performance of a predictor minimising empirical risk (over certain class of functions) using Theorem 5.2. To do this we estimate the tolerance to data of such predictors, using some results from Vapnik-Chervonenkis theory. For the overviews of Vapnik-Chervonenkis theory see [89, 26].

Let $\mathbf{X}=\mathbb{R}^d$ for some $d\in\mathbb{N}$ and let $\mathcal{C}$ be a class of measurable functions of the form $\varphi:\mathbf{X}\to\mathbf{Y}=\{0,1\}$, called *decision functions*. For a probability distribution P on $\mathbf{X}$ define $\mathrm{err}(\varphi,P):=P(\varphi(X_i)\neq Y_i)$. If the examples are generated i.i.d. according to some distribution P, the aim is to find a function φ from $\mathcal{C}$ for which $\mathrm{err}(\varphi,P)$ is minimal:

$$\varphi_P=\mathrm{argmin}_{\varphi\in\mathcal{C}}\mathrm{err}(\varphi,P).$$

In the theory of empirical risk minimisation this function is approximated by the function

$$\varphi_n^*:=\mathrm{argmin}_{\varphi\in\mathcal{C}}\overline{\mathrm{err}}_n(\varphi)$$

where $\overline{\mathrm{err}}_n(\varphi):=\sum_{i=1}^n I_{\varphi(X_i)\neq Y_i}$ is the empirical error functional, based on a sample (X_i,Y_i), $i=1,...,n$. Thus, $\Gamma_n(z_1,...,z_n,x_{n+1}):=\varphi_n^*(x_{n+1})$ is a predictor

minimising empirical risk over the class of functions $\mathcal{C}$.

One of the basic results of Vapnik-Chervonenkis theory is the estimation of the difference of probabilities of error between the best possible function in the class (φ_P) and the function which minimises empirical error:

$$P\big(\mathrm{err}_n(\Gamma,P^\infty)-\mathrm{err}(\varphi_P,P)>\varepsilon\big)\leq 8\mathcal{S}(\mathcal{C},n)e^{-n\varepsilon^2/128},$$

where the symbol $\mathcal{S}(\mathcal{C},n)$ is used for the n-th shatter coefficient of the class $\mathcal{C}$:

$$\mathcal{S}(\mathcal{C},n):=\max_{A:=\{x_1,\ldots,x_n\}\subset\mathbf{X}}\#\{C\cap A:C\in\mathcal{C}\}.$$

Thus,

$$P(\mathrm{err}_n(\Gamma)>\varepsilon)\leq I_{\mathrm{err}(\varphi_P,P)>\varepsilon/2}+8\mathcal{S}(\mathcal{C},n)e^{-n\varepsilon^2/512}.$$

A particularly interesting case is when the optimal rule belongs to $\mathcal{C}$, i.e. when $\eta\in\mathcal{C}$. This situation was investigated in e.g. [88]. Obviously, in this case $\varphi_P\in\mathcal{C}$ and $\mathrm{err}(\varphi_P,P)=0$ for any P. Moreover, a better bound exists (see [89, 26])

$$P(\mathrm{err}_n(\Gamma,P)>\varepsilon)\leq 2\mathcal{S}(\mathcal{C},n)e^{-n\varepsilon/2}.$$

Theorem 5.6. *Let $\mathcal{C}$ be a class of decision functions and let Γ be a predictor which for each $n\in\mathbb{N}$ minimises $\overline{\mathrm{err}}_n$ over $\mathcal{C}$ on the observed examples $(z_1,\ldots,z_n)$. Fix some $\delta\in(0,1/2]$, let $p(n):=\frac{1}{n}\#\{i\leq n:Y_i=0\}$ and $C_n:=\mathbf{P}(\delta\leq p(n)\leq 1-\delta)$ for each $n\in\mathbb{N}$. Assume $n>4/\varepsilon^2$ and let $\alpha_n:=\frac{1}{1-1/\sqrt{n}}$. We have*

$$\Delta(P_0,P_1,n,\varepsilon)\leq 16\mathcal{S}(\mathcal{C},n)e^{-n\varepsilon^2/512}. \tag{5.11}$$

(which does not depend on the distributions P_0 and P_1) and

$$\begin{aligned}\mathbf{P}(\mathrm{err}_n(\Gamma,\mathbf{P})>\varepsilon)&\leq I_{2\mathrm{err}(\varphi_{P_{1/2}},P_{1/2})>\varepsilon/2}\\&\quad+16\alpha_nC_n^{-1}\mathcal{S}(\mathcal{C},n)e^{-n\delta^2\varepsilon^2/2048}+(1-C_n).\end{aligned} \tag{5.12}$$

If in addition $\eta \in \mathcal{C}$ *then*

$$\Delta(n,\varepsilon) \leq 4\mathcal{S}(\mathcal{C},2n)2^{-n\varepsilon)/8} \tag{5.13}$$

and

$$\mathbf{P}(\mathrm{err}_n(\Gamma,\mathbf{P}) > \varepsilon) \leq 4\alpha_n C_n^{-1}\mathcal{S}(\mathcal{C},n)e^{-n\delta\varepsilon/16} + (1-C_n). \tag{5.14}$$

Thus, if we have bounds on the VC dimension of some class of classifiers, we can obtain bounds on the performance of predictors minimising empirical error for the conditional model.

Next we show how strong consistency results can be achieved in the conditional model. For general strong universal consistency results (with examples) see [60, 89].

Denote the VC dimension of $\mathcal{C}$ by $V(\mathcal{C})$:

$$V(\mathcal{C}) := \max\{n \in \mathbb{N} : \mathcal{S}(\mathcal{C},n) = 2^n\}.$$

Using Theorem 5.6 and Borel-Cantelli lemma, we obtain the following corollary.

Corollary 5.7. *Let* $\mathcal{C}^k$, $k \in \mathbb{N}$ *be a sequence of classes of decision functions with finite VC dimension such that* $\lim_{k\to 0}\inf_{\varphi\in\mathcal{C}^k}\mathrm{err}(\varphi,P) = 0$ *for any distribution* P *on* $\mathbf{X}$. *If* $k_n \to \infty$ *and* $\frac{V(\mathcal{C}^{k_n})\log n}{n} \to 0$ *as* $n \to \infty$ *then*

$$\mathrm{err}(\Gamma,\mathbf{P}) \to 0 \mathbf{P}-a.s.$$

where Γ *is a predictor which in each trial* n *minimises empirical risk over* $\mathcal{C}^{k_n}$ *and* $\mathbf{P}$ *is any distribution satisfying (5.1), (5.2) and* $\sum_{n=1}^{\infty}(1-C_n) < \infty$.

In particular, if we use bound on the VC dimension on classes of neural networks provided in [8] then we obtain the following corollary.

Corollary 5.8. *Let Γ be a classifier that minimises the empirical error over the class $\mathcal{C}^{(k)}$, where $\mathcal{C}^{(k)}$ is the class of neural net classifiers with k nodes in the hidden layer and the threshold sigmoid, and $k \to \infty$ so that $k\log n/n \to 0$ as $n \to \infty$. Let $\mathbf{P}$ be any distribution on $\mathbf{X}^\infty$ satisfying (5.1) and (5.2) such that $\sum_{n=1}^{\infty}(1-C_n) < \infty$. Then*

$$\lim_{n\to\infty} \mathrm{err}_n(\Gamma) = 0 \ \mathbf{P}-a.s.$$

5.2 Computational limitations on the statistical characterizations of learnability in classification [R9]

In this section we investigate the question of whether finite-step performance guarantees can be obtained if we consider the class of computable (on some Turing machine) classification methods. To make the problem more realistic, we assume that the target classification function η (that maps objects to labels) is also computable. Two definitions of target functions are considered: they are either of the form $\{0,1\}^\infty \to \{0,1\}$ or $\{0,1\}^t \to \mathbf{Y}$ for some t (which can be different for different target functions).

We show that there are classes $\mathcal{C}_k$ of functions for which the number of examples needed to approximate the classification problem to a certain accuracy grows faster in the VC dimension of the class than any computable function (rather than being linear as in the statistical setting, [89]). In particular this holds if $\mathcal{C}_k$ is the class of all computable functions of length not greater than k, in which case k is a (trivial) upper bound of the VC dimension.

Importantly, the same negative result holds even if we allow the data to be generated "actively," e.g., by some algorithm, rather than just by some fixed probability distribution.

To obtain this negative result we consider the task of data compression:

an impossibility result for the task of data compression allows us to estimate the sample complexity for classification. We also analyse how tight the negative result is, and show that for some simple computable rule (based on the nearest neighbour estimate) the sample complexity is finite in k, under different definitions of computational classification task.

Related work. In comparison to the vast literature on classification, relatively little attention had been paid to the "computable" version of the task. There is a track of research in which different concepts of computable learnability of functions on countable domains are studied, see [44]. Another approach is to consider classification methods as functions computable in polynomial time, or under other resource constraints. This approach leads to many interesting results, but it usually considers more specific settings of a learning problem, such as learning DNFs, finite automata, etc.

It may be interesting to observe the connection of the results for classification with another learning problem, sequence prediction. In one of its simplest forms this task is as follows: it is required to predict the next outcome of a deterministic sequence of symbols, where the sequence is assumed to be computable (is generated by some program). There is a predictor which can solve any such problem and the number of errors it makes is at most linear in the length of the program generating the sequence (see, e.g. [41], Section 3.2.3). Such a predictor is not computable. Trivially, there is no computable predictor for all computable sequences, since for any computable predictor a computable sequence can be constructed on which it errs at every trial, simply by reversing the predictions. Thus we have linear number of errors for non-computable predictor versus infinitely many errors for any computable one; whereas in classification, as we show, it is linear for a non-computable predictor versus growing faster than any computable function for any computable predictor.

5.2.1 Notation and definitions

By computable functions we mean functions computable on a Turing machine with an input tape, output tape, and some working tapes, the number of which is supposed to be fixed throughout the section.

All computable functions can be encoded (in a canonical way) and thus the set of computable functions can be effectively enumerated. Fix some canonical enumeration and define the *length* of a computable function η as $l(\eta) := |n|$ where n is the minimal number of η in such enumeration. For an introduction to the computability theory see, for example, [73].

From the set of all computable functions we are interested in labelling functions, that is, in functions which represent classification problems. In classification a labelling function is usually a function from the interval [0,1] or $[0,1]^d$ (sometimes more general spaces are considered) to a finite space $Y := \{0,1\}$. As we are interested in computable functions, we should consider instead total computable functions of the form $\{0,1\}^\infty \to \mathbf{Y}$. However, since we require that labelling functions are total (defined on all inputs) and computable, it can be easily shown (e.g. with König's lemma [51]) that any such function never scans its input tape further than a certain position independent of the input. Thus apparently the smallest meaningful class of computable labelling functions that we can consider is the class of functions of the form $\{0,1\}^t \to \mathbf{Y}$ for some t. So, we call a partial recursive function (or program) η a *labelling function* if there exists such $t =: t(\eta) \in \mathbb{N}$ that η accepts all strings from $X_t := \{0,1\}^t$ and only such strings. (It is not essential for this definition that η is not a total function. An equivalent for our purposes definition would be as follows: a labelling function is any total function which outputs the string 00 on all inputs except on the strings of some length $t =: t(\eta)$, on each of which it outputs either 0 or 1.)

It can be argued that this definition of a labelling function is too restrictive to approximate well the notion of a real function. However, as we are after negative results (for the class of all labelling functions), it is not a dis-

advantage. Other possible definitions are discussed in Section 5.2.3, where we are concerned with tightness of our negative results. In particular, all the results hold true if a target function is any total computable function of the form $\{0,1\}^\infty \to \mathbf{Y}$.

Define the task of computational classification as follows. An (unknown) labelling function η is fixed. The objects $x_1,...,x_n \in X$ are drawn according to some distribution P on $X_{t(\eta)}$. The labels y_i are defined according to η, that is $y_i := \eta(x_i)$.

A *predictor* is a family of functions $\varphi_n(x_1,y_1,...,x_n,y_n,x)$ (indexed by n) taking values in Y, such that for any n and any $t \in \mathbb{N}$, if $x_i \in X_t$ for each i, $1 \le i \le n$, then the marginal $\varphi(x)$ is a total function on X_t. We will often identify φ_n with its marginal $\varphi_n(x)$ when the values of other variables are clear. Thus, given a *sample* $x_1,y_1,...,x_n,y_n$ of labelled objects of the same size t a predictor produces a labelling function on X_t which is supposed to approximate η.

A *computable predictor* is a total computable function from $X_t \times Y \times \cdots \times X_t \times Y \times X_t$ to $\{0,1\}$, where the arguments are assumed to be encoded into a single input in a certain fixed (simple canonical) way.

We are interested in what sample size is required to approximate a labelling function η.

For a (computable) predictor φ, a labelling function η and $0 < \varepsilon \in \mathbb{R}$ define

$$\delta_n(\varphi,\eta,\varepsilon) := \sup_{P_t} P_t\Big\{ x_1,...,x_n \in X_t : \\ P_t\big\{ x \in X_t : \varphi_n(x_1,y_1,...,x_n,y_n,x) \ne \eta(x) \big\} > \varepsilon \Big\},$$

where $t = t(\eta)$ and P_t ranges over all distributions on X_t (i.i.d. on X_t^n). As usual in PAC theory we have two probabilities here: the P_t-probability over a training sample of size n that the P_t-probability of error of a predictor φ exceeds ε; then the supremum is taken over all possible distributions P_t.

For any $\delta > 0$ define the *sample complexity* of η with respect to φ as

$$N(\varphi,\eta,\delta,\varepsilon) := \min\{n \in \mathbb{N} : \delta_n(\varphi,\eta,\varepsilon) \leq \delta\}.$$

The number $N(\varphi,\eta,\delta,\varepsilon)$ is the minimal sample size required for a predictor φ to achieve ε-accuracy with probability $1-\delta$ when the (unknown) labelling function is η, under all probability distributions.

With the use of statistical learning theory [89] we can easily derive the following statement

Proposition 5.9. *There exists a predictor φ such that*

$$N(\varphi,\eta,\delta,\varepsilon) \leq \frac{const}{\varepsilon} l(\eta) \log \frac{1}{\delta}$$

for any labelling function η and any $\varepsilon,\delta > 0$.

Observe that the bound is linear in the length of η.

In the next section we investigate the question of whether any such bounds exist if we restrict our attention to computable predictors.

Proof. The predictor φ is defined as follows. For each sample $x_1,y_1,\ldots,x_n,y_n$ it finds a shortest program $\bar{\eta}$ such that $\bar{\eta}(x_i) = y_i$ for all $i \leq n$. Clearly, $l(\bar{\eta}) \leq l(\eta)$. Observe that the VC-dimension of the class of all computable functions of length not greater than $l(\eta)$ is bounded from above by $l(\eta)$, as there are not more than $2^{l(\eta)}$ such functions. Moreover, φ minimizes empirical risk over this class of functions. It remains to use the bound (see e.g. [26], Corollary 12.4) $\sup_{\eta \in \mathcal{C}} N(\varphi,\eta,\delta,\varepsilon) \leq \max\left(V(\mathcal{C})\frac{8}{\varepsilon}\log\frac{13}{\delta}, \frac{4}{\varepsilon}\log\frac{2}{\delta}\right)$, where $V(\mathcal{C})$ is the VC-dimension of the class $\mathcal{C}$. □

5.2.2 Sample complexity explosion for computable learning rules

The main result of this section is that for any computable predictor φ there is no computable upper bound in terms of $l(\eta)$ on the sample complexity of

the function η with respect to φ:

Theorem 5.10. *For every computable predictor φ and every partial computable function $\beta:\mathbb{N}\to\mathbb{N}$ that has infinite domain and goes to infinity, there are infinitely many functions η, such that for some $n>\beta(l(\eta))$*

$$P\{x\in X_{t(\eta)}:\varphi(x_1,y_1,...,x_n,y_n,x)\neq\eta(x)\}>0.05,$$

for any $x_1,...,x_n\in X_{t(\eta)}$, where $y_i=\eta(x_i)$ and P is the uniform distribution on $X_{t(\eta)}$.

For example, we can take $\beta(n)=2^n$, or 2^{2^n}.

Corollary 5.11. *For any computable predictor φ, any total computable function $\beta:\mathbb{N}\to\mathbb{N}$ and any $\delta<1$*

$$\sup_{\eta:l(\eta)\leq k} N(\varphi,\eta,\delta,0.05)>\beta(k)$$

from some k on.

Observe that there is no δ in the formulation of Theorem 5.10. Moreover, it is not important how the objects $(x_1,...,x_n)$ are generated — it can be any individual sample. In fact, we can assume that the sample is chosen in any manner, for example by some algorithm. This means that no computable upper bound on sample complexity exists even for *active learning algorithms*.

It appears that the task of classification is closely related to another learning task — data compression. Moreover, to prove Theorem 4.3 we need a similar negative result for this task. Thus before proceeding with the proof of the theorem, we introduce the task of data compression and derive a negative result for it. We call a total computable function $\psi:\mathbf{X}\to\mathbf{X}$ a *data compressor* if it is an injection (i.e. $x_1\neq x_2$ implies $\psi(x_1)\neq\psi(x_2)$). We say that a data compressor *compresses* the string x if $|\psi(x)|<|x|$. Clearly, for any natural n any data compressor compresses not more than half of the strings of size up to n.

Next we introduce Kolmogorov complexity; for fine details see [59]. The complexity of a string $x \in \mathbf{X}$ with respect to a Turing machine ζ is defined as

$$C_\zeta(x) = \min_p \{l(p) : \zeta(p) = x\},$$

where p ranges over all binary strings (interpreted as partial computable computable functions; minimum over empty set is defined as ∞). There exists such a machine ζ that $C_\zeta(x) \leq C_{\zeta'}(x) + c_{\zeta'}$ for any x and any machine ζ' (the constant $c_{\zeta'}$ depends on ζ' but not on x). Fix any such ζ and define the *Kolmogorov complexity* of a string $x \in \mathbf{X}$ as

$$C(x) := C_\zeta(x).$$

Clearly, $C(x) \leq |x| + b$ for any x and for some b depending only on ζ. A string is called c-incompressible if $C(x) \geq |x| - c$. Obviously, any data compressor can not compresses many c-incompressible strings, for any c. However, highly compressible strings (that is, strings with Kolmogorov complexity low relatively to their length) might be expected to be compressed well by some sensible data compressor. The following lemma shows that this cannot be always the case, no matter what we mean by "relatively low".

The lemma is proven using the fact that there are no non-trivial computable lower bounds on Kolmogorov complexity; the lemma itself can be considered as a different formulation of this statement. The proof of the lemma is followed by the proof of Theorem 5.10.

Lemma 5.12. *For every data compressor ψ and every partial computable function $\gamma : \mathbb{N} \to \mathbb{N}$ which has an infinite domain and goes to infinity there exist infinitely many strings x such that $C(x) \leq \gamma(|x|)$ and $|\psi(x)| \geq |x|$.*

For example, we can take $\gamma(n) = \log\log n$.

Proof. Suppose the contrary, i.e. that there exist a data compressor ψ and some function $\gamma : \mathbb{N} \to \mathbb{N}$ monotonically increasing to infinity such that if

$C(x) \leq \gamma(|x|)$ then $\psi(x) < |x|$ except for finitely many x. Let T be the set of all strings which are not compressed by ψ

$$T := \{x : |\psi(x)| \geq |x|\}.$$

Define the function τ on the set T as follows: $\tau(x)$ is the number of the element x in T

$$\tau(x) := \#\{x' \in T : x' \leq x\}$$

for each $x \in T$. Obviously, the set T is infinite. Moreover, $\tau(x) \leq x$ for any $x \in T$ (recall that we identify $\mathbf{X}$ and $\mathbb{N}$ via length-lexicographical ordering). Observe that τ is a total computable function on T and onto $\mathbb{N}$. Thus $\tau^{-1} : \mathbb{N} \to \mathbf{X}$ is a total computable function on $\mathbb{N}$. Hence, for any $x \in T$ for which $\gamma(|x|)$ is defined we have, except for finitely many x:

$$C(\tau(x)) \geq C(\tau^{-1}(\tau(x)) - c = C(x) - c > \gamma(|x|) - c, \tag{5.15}$$

for a constant c depending only on τ, where the first inequality follows from computability of τ^{-1} and the last from the definition of T. Since τ is computable we also have $C(\tau(x)) \leq C(x) + c'$ for some constant c'.

It is a well-known result (see e.g. [59]) that for any unbounded partial computable function δ with infinite domain there are infinitely many $x \in \mathbf{X}$ such that $C(x) \leq \delta(|x|)$. In particular, allowing $\delta(|x|) = \gamma(|x|) - c' - 2c$, we conclude that there are infinitely many $x \in T$ such that

$$C(\tau(x)) \leq C(x) + c' \leq \gamma(|\tau(x)|) - 2c \leq \gamma(|x|) - 2c,$$

which contradicts (5.15). □

5.2.3 Different settings and tightness of the negative results

In this section we discuss how tight the conditions of the statements are and to what extend they depend on the definitions.

Let us consider the question of whether there exists some (not necessarily computable) total sample-complexity function

$$\mathcal{N}_\varphi(k,\delta,\varepsilon) := \sup_{\eta: l(\eta)\le k} N(\varphi,\eta,\delta,\varepsilon),$$

at least for some predictor φ.

Proposition 5.13. *There exists a predictor φ such that $\mathcal{N}_\varphi(k,\delta,\varepsilon)<\infty$ for any $\varepsilon,\delta>0$ and any $k\in\mathbb{N}$.*

Indeed it is easy to see that the "pointwise" predictor

$$\varphi(x_1,y_1,\ldots,x_n,y_n,x) = \begin{cases} y_i & \text{if } x=x_i, 1\le i\le n \\ 0 & x\notin\{x_1,\ldots,x_n\} \end{cases} \tag{5.16}$$

satisfies the conditions of the proposition.

It can be argued that probably this statement is due to our definition of a labelling function. Next we will discuss some other variants of this definition.

First, observe that if we define a labelling function as any total computable function on $\{0,1\}^*$ then some labelling functions will not approximate any function on $[0,1]$; for example the function η_+ which counts bitwise sum of its input: $\eta_+(x) := \sum_{i=1}^{|x|} x_i \mod 2$. That is why we require a labelling function to be defined only on X_t for some t.

Another way to define a labelling function (which perhaps makes labelling functions most close to real functions) is as a function which accepts any *infinite* binary string. Let us call an *i-labelling function* any total recursive function $\eta: \mathbf{Y}^\infty \to \mathbf{Y}$. That is, η is computable on a Turing machine

with an input tape on which one way infinite input is written, an output tape and possibly some working tapes. The program η is required to halt on any input. As it was mentioned earlier, in this case the situation essentially does not change, since (as it is easy to show) for any i-labelling function η there exist $n_\eta \in \mathbb{N}$ such that η does not scan its input tape beyond position n_η. In particular, $\eta(x) = \eta(x')$ as soon as $x_i = x'_i$ for any $i \leq n_\eta$. Moreover, it is easy to check that Theorem 5.10 holds for i-labelling functions as well. Finally, it can be easily verified that Proposition 5.13 holds true if we consider i-labelling functions instead of labelling functions, constructing the required predictor based on the nearest neighbour predictor. Indeed, it suffices to replace the "pointwise" predictor in the proof of Proposition 5.13 by the predictor φ, which assigns to the object x the label of that object among $x_1,...,x_n$ with whom x has longest mutual prefix (where the prefixes are compared up to some growing horizon).

5.3 Longer proofs

5.3.1 Proofs for Section 5.1.1

Before proceeding with the proof of Theorem 5.2 we give some definitions and supplementary facts.

Define the conditional probabilities of error of Γ as follows

$$\mathrm{err}_n^0(\Gamma,\mathbf{P},z_0,...,z_n) := \mathbf{P}(Y_{n+1} \neq \Gamma(z_1,...,z_n,X_{n+1}) | Y_{n+1} = 0),$$

$$\mathrm{err}_n^1(\Gamma,\mathbf{P},z_0,...,z_n) := \mathbf{P}(Y_{n+1} \neq \Gamma(z_1,...,z_n,X_{n+1}) | Y_{n+1} = 1),$$

(with the same notational convention as used with the definition of $\mathrm{err}_n(\Gamma)$). In words, for each $y \in \mathbf{Y} = \{0,1\}$ we define err_n^y as the probability of all $x \in \mathbf{X}$, such that Γ makes an error on n'th trial, given that $Y_{n+1} = y$ and fixed $z_1,...,z_n$.

For any $\mathbf{y} := (y_1,y_2,...) \in \mathbf{Y}^\infty$, define $\mathbf{y}_n := (y_1,...,y_n)$ and $p_n(\mathbf{y}) := \frac{1}{n}\#\{i \leq$

$n : y_i = 0\}$, for $n > 1$.

Clearly (from the assumption (5.1)) the random variables $X_1,...,X_n$ are mutually conditionally independent given $Y_1,...,Y_n$, and by (5.2) they are distributed according to P_{Y_i}, $1 \le i \le n$. Hence, the following statement is valid.

Lemma 5.14. *Fix some $n > 1$ and some $\mathbf{y} \in \mathbf{Y}^\infty$ such that $\mathbf{P}((Y_1,....Y_{n+1}) = \mathbf{y}_{n+1}) \neq 0$. Then*

$$\mathbf{P}\left(\mathrm{err}_n^{y_{n+1}}(\Gamma) > \varepsilon \,\middle|\, (Y_1,...,Y_n) = \mathbf{y}_n\right) = P_p^n\left(\mathrm{err}_n^{y_{n+1}}(\Gamma) > \varepsilon \,\middle|\, (Y_1,...,Y_n) = \mathbf{y}_n\right)$$

for any $p \in (0,1)$.

Proof of Theorem 5.2. Fix some $n > 1$, some $y \in \mathbf{Y}$ and such $\mathbf{y}^1 \in \mathbf{Y}^\infty$ that $\delta \le p_n(\mathbf{y}^1) \le (1-\delta)$ and $\mathbf{P}((Y_1,...,Y_n) = \mathbf{y}_n^1) \neq 0$. Let $p := p_n(\mathbf{y}^1)$. We will find bounds on $\mathbf{P}\left(\mathrm{err}_n(\Gamma) > \varepsilon \,|\, (Y_1,...,Y_n) = \mathbf{y}_n^1\right)$, first in terms of Δ and then in terms of $\bar{\Delta}$.

Lemma 5.14 allows us to pass to the i.i.d. case:

$$\mathbf{P}\left(\mathrm{err}_n^y(\Gamma, X_1, y_1^1, ..., X_n, y_n^1, X_{n+1}) > \varepsilon\right) = P_p^n\left(\mathrm{err}_n^y(\Gamma, X_1, y_1^1, ..., X_n, y_n^1, X_{n+1}) > \varepsilon\right)$$

for any y such that $\mathbf{P}(Y_1 = y_1^1, ..., Y_n = y_n^1, Y_{n+1} = y) \neq 0$ (recall that we use upper-case letters for random variables and lower-case for fixed variables, so that the probabilities in the above formula are labels-conditional).

Clearly, for $\delta \le p \le 1-\delta$ we have $\mathrm{err}_n(\Gamma, P_p) \le \max_{y \in \mathbf{Y}}(\mathrm{err}_n^y(\Gamma, P_p))$, and if $\mathrm{err}_n(\Gamma, P_p) < \varepsilon$ then $\mathrm{err}_n^y(\Gamma, P_p) < \varepsilon/\delta$ for each $y \in \mathbf{Y}$.

Let m be such number that $m - \varkappa_m = n$. For any $\mathbf{y}^2 \in \mathbf{Y}^\infty$ such that $|mp_m(\mathbf{y}^2) - mp| \le \varkappa_m/2$ there exist such mapping $\pi : \{1,...,n\} \to \{1,...,m\}$ that $y^2_{\pi(i)} = y_i^1$ for any $i \le n$. Define random variables $X'_1 ... X'_m$ as follows: $X'_{\pi(i)} := X_i$ for $i \le n$, while the rest $\varkappa_m$ of X'_i are some random variables

independent from $X_1,...,X_n$ and from each other, and distributed according to P_p (a "ghost sample"). We have

$$\begin{aligned}
&P_p^n\big(\mathrm{err}_n^y(X_1,y_1^1,...,X_n,y_n^1)>\varepsilon\big)\\
&\quad=P_p^m\Big(\mathrm{err}_n^y(X_1,y_1^1,...,X_n,y_n^1)-\mathrm{err}_n^y(X_1',y_1^2,...,X_m',y_m^2)\\
&\quad+\mathrm{err}_n^y(X_1',y_1^2,...,X_m',y_m^2)>\varepsilon\Big)\\
&\le P_p^m\Big(\big|\mathrm{err}_n^y(X_1',y_1^2,...,X_n',y_n^2)-\mathrm{err}_n^y(X_1,y_1^1,...,X_n,y_n^1)\big|>\varepsilon/2\Big)\\
&\quad+P_p^n\Big(\mathrm{err}_n^y(X_1',y_1^2,...,X_n',y_n^2)>\varepsilon/2\Big).
\end{aligned}$$

Observe that $\mathbf{y}^2$ was chosen arbitrary (among sequences for which $|mp_m(\mathbf{y}^2)-mp|\le\varkappa_m/2$) and $(X_1,y_1^1,...,X_ny_n^1)$ can be obtained from $(X_1',y_1^2,...,X_m'y_m^2)$ by removing at most $\varkappa_m$ elements and applying some permutation. Thus the first term is bounded by

$$\begin{aligned}
&P_p^m\big(\max_{j\le\varkappa_m;\ \pi:\{1,...,m\}\to\{1,...,m\}}|\mathrm{err}_m^y(\Gamma,Z_1,...,Z_m)-\\
&\qquad\mathrm{err}_{m-j}^y(\Gamma,Z_{\pi(1)},...,Z_{\pi(m-j)})|>\varepsilon/2\ \big|\ |mp(m)-mp|\le\varkappa_m/2\big)\\
&\qquad\qquad\le\frac{\Delta(P_p,m,\delta\varepsilon/2)}{P_p^n(|mp(m)-mp|\le\varkappa_m)}\le\frac{1}{1-1/\sqrt{m}}\Delta(P_p,m,\delta\varepsilon/2),
\end{aligned}$$

and the second term is bounded by $\frac{1}{1-1/\sqrt{m}}P_p^m(\mathrm{err}_m(\Gamma)>\delta\varepsilon/2)$. Hence

$$\begin{aligned}
&P_p^n\big(\mathrm{err}_n^y(X_1,y_1^1,...,X_n,y_n^1)>\varepsilon\big)\\
&\qquad\qquad\le\alpha_n\big(\Delta(P_p,m,\delta\varepsilon/2)+P_p^m(\mathrm{err}_m(\Gamma)>\delta\varepsilon/2)\big).\quad(5.17)
\end{aligned}$$

Next we establish a similar bound in terms of $\bar\Delta$. For any $\mathbf{y}_n^2\in\mathbf{Y}^n$ such that $|np_n(\mathbf{y}^2)-np|\le\varkappa_n/2$ there exist such permutations π_1,π_2 of the set $\{1,...,n\}$ that $y^1_{\pi_1(i)}=y^2_{\pi_2(i)}$ for any $i\le n-\delta\varkappa_n$. Denote $n-\delta\varkappa_n$ by n' and define random variables $X_1'...X_n'$ as follows: $X'_{\pi_2(i)}:=X_{\pi_1(i)}$ for $i\le n'$, while for $n'<i\le n$ X_i' are some "ghost" random variables independent from

$X_1,...,X_n$ and from each other, and distributed according to P_p. We have

$$P_p^n\big(\mathrm{err}_n^y(X_1,y_1^1,...,X_n,y_n^1)>\varepsilon\big)$$
$$\le P_p^{n+\varkappa_n}\Big(\big|\mathrm{err}_n^y(X_1',y_1^2,...,X_n',y_n^2)-\mathrm{err}_n^y(X_1,y_1^1,...,X_n,y_n^1)\big|>\varepsilon/2\Big)$$
$$+P_p^n\Big(\mathrm{err}_n^y(X_1',y_1^2,...,X_n',y_n^2)>\varepsilon/2\Big),$$

Again, as $\mathbf{y}^2$ was chosen arbitrary (among sequences for which $|np_n(\mathbf{y}^2)-np|\le\varkappa_n/2$) and $(X_1,y_1^1,...,X_ny_n^1)$ differs from $(X_1',y_1^2,...,X_n'y_n^2)$ in at most $\varkappa_n$ elements, up to some permutation. Thus the first term is bounded by

$$P_p^n\Big(\sup_{j<\varkappa_n;\pi:\{1,...,n\}\to\{1,...,n\};z'_{n-j},...,z'_n}|\mathrm{err}_n^y(Z_1,...,Z_n)$$
$$-\mathrm{err}_n^y(\zeta_1,...,\zeta_n)|>\varepsilon/2\;\Big|\;|np(n)-np|\le\varkappa_n/2\Big)$$
$$\le\alpha_n\bar\Delta(P_p,n,\delta\varepsilon/2),$$

and the second term is bounded by $\alpha_nP_p^n(\mathrm{err}_n(\Gamma)>\delta\varepsilon/2)$. Hence

$$P_p^n\big(\mathrm{err}_n^y(X_1,y_1^1,...,X_n,y_n^1)>\varepsilon\big)$$
$$\le\alpha_n\big(\bar\Delta(P_p,n,\delta\varepsilon/2)+P_p^n(\mathrm{err}_n(\Gamma)>\delta\varepsilon/2)\big).\quad(5.18)$$

Finally, as $\mathbf{y}^1$ was chosen arbitrary among sequences $\mathbf{y}\in\mathbf{Y}^\infty$ such that $n\delta\le p_n(\mathbf{y}^1)\le n(1-\delta)$ from (5.17) and (5.18) we obtain (5.5) and (5.6). $\square$

5.3.2 Proofs for Section 5.1.2

The first part of the proof is common for theorems 5.4 and 5.5. Let us fix some distribution $\mathbf{P}$ satisfying conditions of the theorems. It is enough to show that

$$\sup_{p\in[\delta,1-\delta]}E^\infty(\mathrm{err}_n(\Gamma,P_p,Z_1,...,Z_n))\to0$$

and

$$\sup_{p\in[\delta,1-\delta]} E^\infty(\bar\Delta(P_p,n,Z_1,...,Z_n))\to 0$$

for nearest neighbour and partitioning predictor, and apply Corollary 5.3.

Observe that both predictors are symmetric, i.e. do not depend on the order of $Z_1,...,Z_n$. Thus, for any $z_1,...,z_n$

$$\bar\Delta(P_p,n,z_1,...,z_n)=\sup_{j\le\varkappa_n;\ \pi:\{1,...,n\}\to\{1,...,n\},z'_{n-j},...,z'_n} |\mathrm{err}_n(\Gamma,P_p,z_1,...,z_n)-\mathrm{err}_n(\Gamma,P_p,z_{\pi(1)},...,z_{\pi(n-j)},z'_{n-j},...,z'_n)|,$$

where the maximum is taken over all z'_i consistent with η, $n-j\le i\le n$. Define also the class-conditional versions of $\bar\Delta$:

$$\bar\Delta^y(P_p,n,z_1,...,z_n):=\sup_{j\le\varkappa_n;\ \pi:\{1,...,n\}\to\{1,...,n\},z'_{n-j},...,z'_n} |\mathrm{err}^y_n(\Gamma,P_p,z_1,...,z_n)-\mathrm{err}^y_n(\Gamma,P_p,z_{\pi(1)},...,z_{\pi(n-j)},z'_{n-j},...,z'_n)|.$$

Note that (omitting $z_1,...,z_n$ from the notation) $\mathrm{err}_n(\Gamma,P_p)\le\mathrm{err}^0_n(\Gamma,P_p)+\mathrm{err}^1_n(\Gamma,P_p)$ and $\bar\Delta(P_p,n)\le\bar\Delta^0(P_p,n)+\bar\Delta^1(P_p,n)$. Thus, it is enough to show that

$$\sup_{p\in[\delta,1-\delta]} E^\infty(\mathrm{err}^1_n(\Gamma,P_p))\to 0 \tag{5.19}$$

and

$$\sup_{p\in[\delta,1-\delta]} E^\infty(\bar\Delta^1(P_p,n))\to 0. \tag{5.20}$$

Observe that for each of the predictors in question the probability of error given that the true label is 1 will not decrease if an arbitrary (possibly large) portion of training examples labelled with ones is replaced with an arbitrary (but consistent with η) portion of the same size of examples labelled with zeros. Thus, for any n and any $p\in[\delta,1-\delta]$ we can decrease the number of ones in our sample (by replacing the corresponding examples with examples

from the other class) down to (say) $\delta/2$, not decreasing the probability of error on examples labelled with 1. So,

$$E^\infty(\mathrm{err}_n^1(\Gamma,P_p)) \le E^\infty(\mathrm{err}_n^1(\Gamma,P_{\delta/2}|p_n=\delta/2))+P_p(p_n\le\delta/2), \tag{5.21}$$

where as usual $p_n := \frac{1}{n}\#\{i\le n : y_i = 1\}$. Obviously, the last term (quickly) tends to zero. Moreover, it is easy to see that

$$\begin{aligned} E^\infty&(\mathrm{err}_n^1(\Gamma,P_{\delta/2})|p_n=n(\delta/2)) \\ &\le E^\infty\big(\mathrm{err}_n^1(\Gamma,P_{\delta/2})\big||n(\delta/2)-p_n|\le\varkappa_n/2\big)+E^\infty(\bar\Delta^1(P_{\delta/2},n)) \\ &\le \frac{1}{1-1/\sqrt{n}}E^\infty(\mathrm{err}_n^1(\Gamma,P_{\delta/2}))+E^\infty(\bar\Delta^1(P_{\delta/2},n)). \end{aligned} \tag{5.22}$$

The first term tends to zero, as it is known from the results for i.i.d. processes; thus, to establish (5.19) we have to show that

$$E(\bar\Delta^1(P_p,n,Z_1,...,Z_n))\to 0 \tag{5.23}$$

for any $p\in(0,1)$.

We will also show that (5.23) is sufficient to prove (5.20). Indeed,

$$\begin{aligned} \bar\Delta^1(P_p,n,z_1,...,z_n) \le{}& \mathrm{err}_n^1(\Gamma,P_p,z_1,...,z_n)+ \\ &\sup_{j\le\varkappa_n;\ \pi:\{1,...,n\}\to\{1,...,n\},z'_{n-j},...,z'_n} \mathrm{err}_n^1(\Gamma,P_p,z_{\pi(1)},...,z_{\pi(n-j)},z'_{n-j},...,z'_n) \end{aligned}$$

Denote the last summand by D. Again, we observe that D will not decrease if an arbitrary (possibly large) portion of training examples labelled with ones is replaced with an arbitrary (but consistent with η) portion of the same size of examples labelled with zeros. Introduce $\widetilde\Delta^1(P_p,n,z_1,...,z_n)$ as $\bar\Delta^1(P_p,n,z_1,...,z_n)$ with $\varkappa_n$ in the definition replaced by $\frac{2}{\delta}\varkappa_n$. Using the same

argument as in (5.21) and (5.22) we have

$$E^\infty(D) \le \frac{1}{1-1/\sqrt{n}}\big(E^\infty(\widetilde{\Delta}^1(P_{\delta/2},n)) + E^\infty(\mathrm{err}_n(\Gamma,P_{\delta/2})\big) + P_p(p_n \le \delta/2).$$

Thus, (5.20) holds true if (5.23) and

$$E^\infty(\widetilde{\Delta}^1(P_p,n,Z_1,...,Z_n)) \to 0. \tag{5.24}$$

Finally, we will prove (5.23); it will be seen that the proof of (5.24) is analogous (i.e. replacing $\varkappa_n$ by $\frac{2}{\delta}\varkappa_n$ does not affect the proof). Note that

$$E^\infty(\bar{\Delta}(P_p,n,Z_1,...,Z_n)) \le P_p\Big(\sup_{j\le\varkappa_n;\ \pi:\{1,...,n\}\to\{1,...,n\},z'_{n-j},...,z'_n} \big|\mathrm{err}_n(\Gamma,P_p,Z_1,...,Z_n) \ne \mathrm{err}_n(\Gamma,P_p,Z_{\pi(1)},...,Z_{\pi(n-j)},z'_{n-j},...,z')\big|\Big),$$

where the maximum is taken over all z'_i consistent with η, $n-j \le i \le n$. The last expression should be shown to tend to zero. This we will prove for each of the predictors separately.

Nearest Neighbour predictor. Fix some distribution P_p, $0<p<1$ and some $\varepsilon>0$. Fix also some $n\in\mathbb{N}$ and define (leaving $x_1,...,x_n$ implicit)

$$B_n(x) := P_p^{n+1}\{t\in\mathbf{X} : t \text{ and } x \text{ have the same nearest neighbour among } x_1,...,x_n\}$$

and $B_n := E(B_n(X))$ Note that $E^\infty(B_n)=1/n$, where the expectation is taken over $X_1,...,X_n$. Define $\mathcal{B} := \{(x_1,...,x_n)\in\mathbf{X}^n : B_n \le 1/n\varepsilon\}$ and $\mathcal{A}(x_1,...,x_n) :=$

$\{x: B_n(x) \le 1/n\varepsilon^2\}$. Applying Markov's inequality twice, we obtain

$$\begin{aligned} E^\infty(\bar{\Delta}(P_p,n)) &\le E^\infty(\bar{\Delta}(P_p,n)|(X_1,...,X_n)\in\mathcal{B})+\varepsilon \\ &\le E^\infty\Big(\sup_{j\le\varkappa_n;\ \pi:\{1,...,n\}\to\{1,...,n\},z'_{n-j},...,z'_n} \\ P_p\big\{x: \mathrm{err}_n(\Gamma,P_p,Z_1,...,Z_n) &\ne \mathrm{err}_n(\Gamma,P_p,Z_{\pi(1)},...,Z_{\pi(n-j)},z'_{n-j},...,z'_n) \\ &\big| x\in\mathcal{A}(X_1,...,X_n)\big\}\big|(X_1,...,X_n)\in\mathcal{B}\Big)+2\varepsilon. \end{aligned} \tag{5.25}$$

Removing one point x_i from a sample $x_1,...,x_n$ we can only change the value of Γ in the area

$$\{x\in\mathbf{X}: x_i \text{ is the nearest neighbour of x}\} = B_n(x_i),$$

while adding one point x_0 to the sample we can change the value of Γ in the area

$$D_n(x_0) := \{x\in\mathbf{X}: x_0 \text{ is the nearest neighbour of x}\}.$$

It can be shown that the number of examples (among $x_1,...,x_n$) for which a point x_0 is the nearest neighbour is not greater than a constant γ which depends only the space $\mathbf{X}$ (see [26], Corollary 11.1). Thus,

$$D_n(x_0) \subset \cup_{i=j_1,...,j_\gamma} B_n(x_i)$$

for some $j_1,...,j_\gamma$, and so

$$\begin{aligned} E^\infty(\bar{\Delta}(P_p,n)) &\le 2\varepsilon+2(\gamma+1)\varkappa_n E^\infty(\max_{x\in\mathcal{A}(X_1,...,X_n)} B_n(x)|(X_1,...,X_n)\in\mathcal{B}) \\ &\le 2\varkappa_n\frac{\gamma+1}{n\varepsilon^2}+2\varepsilon, \end{aligned}$$

which, increasing n, can be made less than 3ε. □

Partitioning predictor. For any measurable sets $\mathcal{B}\subset\mathbf{X}^n$ and $\mathcal{A}\subset\mathbf{X}$ define

$$D(\mathcal{B},\mathcal{A}):=E^{\infty}\Big(\sup_{j\le\varkappa_n;\ \pi:\{1,...,n\}\to\{1,...,n\},z'_{n-j},...,z'_n}$$
$$P_p\big\{x:\mathrm{err}_n(\Gamma,P_p,Z_1,...,Z_n)\ne\mathrm{err}_n(\Gamma,P_p,Z_{\pi(1)},...,Z_{\pi(n-j)},z'_{n-j},...,z'_n)$$
$$\big|x\in\mathcal{A}\big\}\big|(X_1,...,X_n)\in\mathcal{B}\Big)+2\varepsilon.$$

and $D:=D(\mathbf{X}^n,\mathbf{X})$.

Fix some distribution P_p, $0<p<1$ and some $\varepsilon>0$. Introduce

$$\hat{\eta}(x,X_1,...,X_n):=\frac{1}{N(x)}\sum_{i=1}^{n}I_{Y_i=1}I_{X_i\in A(x)}$$

($X_1,...X_n$ will usually be omitted). From the consistency results for i.i.d. model (see, e.g. [26], Theorem 6.1) we know that $E^{n+1}|\hat{\eta}_n(X)-\eta(X)|\to 0$ (the upper index in E^{n+1} indicating the number of examples it is taken over).

Thus, $E|\hat{\eta}_n(X)-\eta(X)|\le\varepsilon^4$ from some n on. Fix any such n and let $\mathcal{B}:=\{(x_1,...,x_n):E|\hat{\eta}_n(X)-\eta(X)|\le\varepsilon^2\}$. By Markov inequality we obtain $P_p(\mathcal{B})\ge 1-\varepsilon^2$. For any $(x_1,...,x_n)\in\mathcal{B}$ let $\mathcal{A}(x_1,...,x_n)$ be the union of all cells A_i^n for which $E(|\hat{\eta}_n(X)-\eta(X)||X\in A_i^n)\le\varepsilon$. Clearly, with $x_1,...,x_n$ fixed, $P_p(X\in\mathcal{A}(x_1,...,x_n))\ge 1-\varepsilon$. Moreover, $D\le D(\mathcal{B},\mathcal{A})+\varepsilon+\varepsilon^2$.

Fix $\mathcal{A}:=(x_1,...,x_n)$ for some $(x_1,...,x_n)\in\mathcal{B}$. Since $\eta(x)$ is always either 0 or 1, to change a decision in any cell $A\subset\mathcal{A}$ we need to add or remove at least $(1-\varepsilon)N(A)$ examples, where $N(A):=N(x)$ for any $x\in A$. Let $N(n):=E(N(X))$ and $A(n):=E(P_p(A(X))$. Clearly, $\frac{N(n)}{nA(n)}=1$ for any n, as $E\frac{N(X)}{n}=A(n)$.

As before, using Markov inequality and shrinking $\mathcal{A}$ if necessary we can have $P_p(\frac{\varepsilon^2 nA(X)}{N(n)}\le\varepsilon|X\in\mathcal{A})=1$, $P_p(\frac{\varepsilon^2 nA(n)}{N(X)}\le\varepsilon|X\in\mathcal{A})=1$, and $D\le D(\mathcal{B},\mathcal{A})+3\varepsilon+\varepsilon^2$. Thus, for all cells $A\subset\mathcal{A}$ we have $N(A)\ge\varepsilon nA(n)$, so that the probability of error can be changed in at most $2\frac{\varkappa_n}{(1-\varepsilon)\varepsilon nA(n)}$ cells; but the probability of each cell is not greater than $\frac{N(n)}{\varepsilon n}$. Hence $E^{\infty}(\bar{\Delta}(P_p,n))\le 2\frac{\varkappa_n}{n(1-\varepsilon)\varepsilon^2}+3\varepsilon+\varepsilon^2$. □

5.3.3 Proofs for Section 5.1.3

Proof of Theorem 5.6. Fix some probability distribution P_p and some $n\in\mathbb{N}$. Let $\varphi^\times$ be any decision rule $\varphi\in\mathcal{C}$ picked by $\Gamma_{n-\varkappa_n}$ on which (along with the corresponding permutation) the maximum

$$\max_{j\le\varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}}|\mathrm{err}_n(\Gamma,z_1,\dots,z_n)-\mathrm{err}_{n-j}(\Gamma,z_{\pi(1)},\dots,z_{\pi(n-j)})|$$

is reached. We need to estimate $P^n(|\mathrm{err}(\varphi^*)-\mathrm{err}(\varphi^\times)|>\varepsilon)$.

Clearly, $|\overline{\mathrm{err}}_n(\varphi^\times)-\overline{\mathrm{err}}_n(\varphi^*)|\le\varkappa_n$, as $\varkappa_n$ is the maximal number of errors which can be made on the difference of the two samples.

Moreover,

$$\begin{aligned}&P^n\big(|\mathrm{err}(\varphi_n^*)-\mathrm{err}(\varphi^\times)|>\varepsilon\big)\\&\quad\le P^n\big(|\mathrm{err}(\varphi_n^*)-\frac{1}{n}\overline{\mathrm{err}}_n(\varphi^*)|>\varepsilon/2\big)\\&\quad+P^n\big(|\frac{1}{n}\overline{\mathrm{err}}_n(\varphi^\times)-\mathrm{err}(\varphi^\times)|>\varepsilon/2-\varkappa_n/n\big)\end{aligned}$$

Observe that

$$P^n(\sup_{\varphi\in\mathcal{C}}|\frac{1}{n}\overline{\mathrm{err}}_n(\varphi)-\mathrm{err}(\varphi)|>\varepsilon)\le 8\mathcal{S}(\mathcal{C},n)e^{-n\varepsilon^2/32},\tag{5.26}$$

see [26], Theorem 12.6. Thus,

$$\Delta(P_p,n,\varepsilon)\le 16\mathcal{S}(\mathcal{C},n)e^{-n(\varepsilon/2-\varkappa_n/n)^2/32}\le 16\mathcal{S}(\mathcal{C},n)e^{-n\varepsilon^2/512}$$

for $n>4/\varepsilon^2$. So,

$$\begin{aligned}\mathbf{P}(\mathrm{err}_n(\Gamma,\mathbf{P})>\varepsilon)&\le I_{\sup_{p\in[\delta,1-\delta]}\mathrm{err}(\varphi_{P_p},P_p)>\varepsilon/2}\\&\quad+16\alpha C_n^{-1}\mathcal{S}(\mathcal{C},n)e^{-n\delta^2\varepsilon^2/2048}+(1-C_n).\end{aligned}$$

It remains to notice that

$$\mathrm{err}(\varphi_{P_p},P_p)=\inf_{\varphi\in\mathcal{C}}(p\mathrm{err}^1(\varphi,P_p)+(1-p)\mathrm{err}^0(\varphi,P_p))$$
$$\leq\inf_{\varphi\in\mathcal{C}}(\mathrm{err}^1(\varphi,P_{1/2})+\mathrm{err}^0(\varphi,P_{1/2}))=2\mathrm{err}(\varphi_{P_{1/2}},P_{1/2})$$

for any $p\in[0,1]$.

So far we have proven (5.11) and (5.12); (5.13) and (5.14) can be proven analogously, only for the case $\eta\in\mathcal{C}$ we have

$$P^n(\sup_{\varphi\in\mathcal{C}}|\frac{1}{n}\overline{\mathrm{err}}_n(\varphi)-\mathrm{err}(\varphi)|>\varepsilon)\leq\mathcal{S}(\mathcal{C},n)e^{-n\varepsilon}$$

instead of (5.26), and $\mathrm{err}(\varphi_{P_p},P_p)=0$. □

5.3.4 Proof of Theorem 5.10

Suppose the contrary, that is that there exists such a computable predictor φ and a partial computable function $\beta:\mathbb{N}\to\mathbb{N}$ such that for any except finitely many labelling functions η for which $\beta(l(\eta))$ is defined and all $n>\beta(l(\eta))$ we have

$$P\{x:\varphi(x_1,y_1,...,x_n,y_n,x)\neq\eta(x)\}\leq 0.05,$$

for some $x_i\in X_{t(\eta)}$, $y_i=\eta(x_i)$, $i\in\mathbb{N}$, where P is the uniform distribution on $X_{t(\eta)}$.

Define $\varepsilon:=0.05$. We will construct a data compressor ψ which contradicts Lemma 5.12. For each $y\in\mathbf{X}$ define $m:=|y|$, $t:=\lceil\log m\rceil$. Generate (lexicographically) first m strings of length t and denote them by x_i, $1\leq i\leq m$. Define the labelling function η_y as follows: $\eta_y(x)=y^i$, if x starts with x_i, where $1\leq i\leq m$. Clearly, $C(\eta_y)\geq C(y)-c$, where c is some universal constant capturing the above description. Let the distribution P be uniform on X_t.

Set $n:=\sqrt{m}$. Next we run the predictor φ on all possible tuples $\mathbf{x}=(x_1,...,x_n)\in {X_t}^n$ and each time count the errors that φ makes on all elements

of X_t:

$$E(\mathbf{x}) := \{x \in X_t : \varphi(x_1, y^1, ..., x_n, y^n, x) \neq \eta_y(x)\}.$$

Thus $E(\mathbf{x})$ is the set of all objects on which φ errs after being trained on $\mathbf{x}$. If $|E(\mathbf{x})| > \varepsilon m$ for all $\mathbf{x} \in X_t$ then $\psi(y) := 0y$.

Otherwise proceed as follows. Fix some tuple $\mathbf{x} = (x'_1, ..., x'_n)$ such that $|E(\mathbf{x})| \leq \varepsilon m$, and let $H := \{x'_1, ..., x'_n\}$ be the unordered tuple $\mathbf{x}$. Define

$$\varkappa^i := \begin{cases} e & \text{if } \ x_i \in E(\mathbf{x}) \setminus H \\ c_0 & \text{if } \ x_i \in H, y^i = 0 \\ c_1 & \text{if } \ x_i \in H, y^i = 1 \\ * & \text{otherwise} \end{cases}$$

for $1 \leq i \leq m$. Thus, each $\varkappa^i$ is a member of a five-letter alphabet (a four-element set) $\{e, c_0, c_1, *\}$. Denote the string $\varkappa^1 ... \varkappa^m$ by K.

So K contains the information about the (unordered) training set and the elements on which φ errs after being trained on this training set. Hence the string K, the predictor φ and the order of $(x'_1, ..., x'_n)$ (which is not contained in K) are sufficient to restore the string y. Furthermore, the n-tuple $(x'_1, ..., x'_n)$ can be obtained from H (the un-ordered tuple) by the appropriate permutation; let r be the number of this permutation in some fixed ordering of all $n!$ such permutations. Using Stirling's formula, we have $|r| \leq 2n \log n = \sqrt{m} \log m$; moreover, to encode r with some self-delimiting code we need not more than $2\sqrt{m} \log m$ symbols (for $m > 3$). Denote such an encoding of r by ρ.

Next, as there are at least $(1 - \varepsilon - \frac{1}{\sqrt{m}})m$ symbols $*$ in the m-element string K (at most εm symbols e_0 and e_1, and $n = \sqrt{m}$ symbols c_0 and c_1), it can be encoded by some simple binary code σ in such a way that

$$|\sigma(K)| \leq \frac{1}{2} m + 8(\varepsilon m + n). \tag{5.27}$$

Indeed, construct σ as follows. First replace all occurrences of the string

$**$ with 0. Encode the rest of the symbols with any fixed 3-bit encoding such that the code of each letter starts with 1. Clearly, $\sigma(K)$ is uniquely decodable. Moreover, it is easy to check that (5.27) is satisfied, as there are not less than $\frac{1}{2}(m-2(\varepsilon m+n))$ occurrences of the string $**$. We also need to write m in a self-delimiting way (denote it by s); clearly, $|s| \leq 2\log m$.

We can define a monotone increasing function β' with an infinite domain on which it coincides with β. Indeed, this can be done by executing in a quasi-parallel fashion β on all inputs and defining $\beta'(k) = \beta(k)$ if $\beta(k)$ was found and $\beta'(l) < \beta'(k)$ for all l on which β' is already defined. Next we can define a function $\beta^{-1}(n)$ with infinite domain such that β^{-1} goes monotonically to infinity and such that $\beta^{-1}(\beta'(n)) = n$. This can be done by running in a quasi-parallel fashion β on all inputs m and stopping when $\beta(m) = n$ with m as an output.

Finally, $\psi(y) = 1s\rho\sigma(K)$ and $|\psi(y)| \leq |y|$, for $m > 2^{10}$. Thus, ψ compresses any (except finitely many) y such that $n > \beta'(C(\eta_y))$; i.e. such that $\sqrt{m} > \beta'(C(\eta_y)) \geq \beta'(C(y) - c)$. This contradicts Lemma 5.12 with $\gamma(k) := \beta^{-1}(\sqrt{k}) + c$. □

Bibliography

[1] Terrence M. Adams and Andrew B. Nobel. On density estimation from ergodic processes. *The Annals of Probability*, 26(2):pp. 794–804, 1998.

[2] R Ahlswede and I Csiszar. Hypothesis testing with communication constraints. *IEEE Transactions on Information Theory*, 32(4):533–542, 1986.

[3] David Aldous and Umesh Vazirani. A markovian extension of valiant's learning model. *Inf. Comput.*, 117:181–186, March 1995.

[4] P. Algoet. Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *Information Theory, IEEE Transactions on*, 45(4):1165 –1185, may 1999.

[5] P.H. Algoet. Universal schemes for prediction, gambling and portfolio selection. *The Annals of Probability*, 20(2):901–941, 1992.

[6] T. Anderson and L. Goodman. Statistical inference about Markov chains. *Ann. Math. Stat.*, 28(1):89–110, 1957.

[7] M. Basseville and I.V. Nikiforov. *Detection of abrupt changes: theory and application*. Prentice Hall information and system sciences series. Prentice Hall, 1993.

[8] Eric B. Baum and David Haussler. What size net gives valid generalization? *Neural Comput.*, 1:151–160, March 1989.

[9] P. Billingsley. Statistical inference about Markov chains. *Ann. Math. Stat.*, 32(1):12–40, 1961.

[10] P. Billingsley. *Ergodic theory and information*. Wiley, New York, 1965.

[11] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

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

[13] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, March 2002.

[14] R. I. Brafman and M. Tennenholtz. A general polynomial time algorithm for near-optimal reinforcement learning. In *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, pages 734–739, 1999.

[15] B.E. Brodsky and B.S. Darkhovsky. *Nonparametric methods in change-point problems*. Mathematics and its applications. Kluwer Academic Publishers, 1993.

[16] E. Carlstein and S. Lele. Nonparametric change-point estimation for data from an ergodic sequence. *Teor. Veroyatnost. i Primenen.*, 38:910–917, 1993.

[17] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, 27:1865–1895, 1999.

[18] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[19] I. Csiszar. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

[20] I. Csiszar. Information theoretic methods in probability and statistics. In *Information Theory. 1997. Proceedings., 1997 IEEE International Symposium on*, page 2, jun-4 jul 1997.

[21] I. Csiszar and P.C. Shields. Notes on information theory and statistics. In *Foundations and Trends in Communications and Information Theory*, 2004.

[22] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):pp. 1–31, 1979.

[23] A. P. Dawid. Prequential data analysis. *Lecture Notes-Monograph Series*, 17:113–126, 1992.

[24] D. Pucci de Farias and N. Megiddo. How to combine expert (and novice) advice when actions impact the environment? In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[25] A. Dembo and Y. Peres. A topological criterion for hypothesis testing. *Ann. Math. Stat.*, 22:106–117, 1994.

[26] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Applications of mathematics. Springer, 1996.

[27] Luc Devroye. On the asymptotic probability of error in nonparametric discrimination. *The Annals of Statistics*, 9(6):pp. 1320–1327, 1981.

[28] Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gabor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22(3):pp. 1371–1385, 1994.

[29] P. Diaconis and D. Freedman. On the consistency of bayes estimates. *Annals of Statistics*, 14(1):1–26, 1986.

[30] J. L. Doob. *Stochastic Processes*. John Wiley & Sons, New York, 1953.

[31] E. Even-Dar, S. M. Kakade, and Y. Mansour. Reinforcement learning in POMDPs without resets. In *IJCAI*, pages 690–695, 2005.

[32] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, NY, USA, 1968.

[33] R. G. Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, M.I.T., 1976 (revised 1979).

[34] David Gamarnik. Extension of the pac framework to finite and countable markov chains. In *Proceedings of the twelfth annual conference on Computational learning theory*, COLT '99, pages 308–317, New York, NY, USA, 1999. ACM.

[35] L Giraitis, R Leipus, and D Surgailis. The change-point problem for dependent observations. *JStat Plan and Infer*, pages 1–15, 1995.

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

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

[38] L. Gyorfi, G. Morvai, and S. Yakowitz. Limits to consistent on-line forecasting for ergodic time series. *IEEE Transactions on Information Theory*, 44(2):886–892, 1998.

[39] D. Haussler. A general minimax result for relative entropy. *IEEE Trans. on Information Theory*, 43(4):1276–1280, 1997.

[40] M. Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proc. 15th Annual Conference on*

Computational Learning Theory (COLT 2002), Lecture Notes in Artificial Intelligence, pages 364–379, Sydney, Australia, July 2002. Springer.

[41] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[42] M. Hutter. On the foundations of universal sequence prediction. In *Proc. 3rd Annual Conference on Theory and Applications of Models of Computation (TAMC'06)*, volume 3959 of *LNCS*, pages 408–420. Springer, 2006.

[43] M. Jackson, E. Kalai, and R. Smorodinsky. Bayesian representation of stochastic processes under learning: de finetti revisited. *Econometrica*, 67(4):875–794, 1999.

[44] S. Jain. *Systems that learn: an introduction to learning theory*. Learning, development, and conceptual change. MIT Press, 1999.

[45] E. Kalai and E. Lehrer. Bayesian forecasting. Discussion paper 998, 1992.

[46] E. Kalai and E. Lehrer. Weak and strong merging of opinions. *Journal of Mathematical Economics*, 23:73–86, 1994.

[47] I. Katsavounidis, C.-C. Jay Kuo, and Zhen Zhang. A new initialization technique for generalized lloyd iteration. *IEEE Signal Processing Letters*, 1:144–146, 1994.

[48] Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Comput.*, 11:1427–1453, August 1999.

[49] M.G. Kendall and A. Stuart. *The advanced theory of statistics; Vol.2: Inference and relationship*. London, 1961.

[50] J.C. Kieffer. Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Transactions on Information Theory*, 39(3):893–902, 1993.

[51] S.C. Kleene. *Mathematical logic*. Wiley, 1967.

[52] J. Kleinberg. An impossibility theorem for clustering. In *15th Conf. Neiral Information Processing Systems (NIPS'02)*, pages 446–453, Montreal, Canada, 2002. MIT Press.

[53] A. N. Kolmogorov and S. V. Fomin. *Elements of the Theory of Functions and Functional Analysis*. Dover, 1975.

[54] R. Krichevsky. *Universal Compression and Retrival*. Kluwer Academic Publishers, 1993.

[55] S.R. Kulkarni and S.E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *Information Theory, IEEE Transactions on*, 41(4):1028 –1039, jul 1995.

[56] S.R. Kulkarni, S.E. Posner, and S. Sandilya. Data-dependent kn-nn and kernel estimators consistent for arbitrary processes. *Information Theory, IEEE Transactions on*, 48(10):2785 – 2788, oct 2002.

[57] P. R. Kumar and P. P. Varaiya. *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice Hall, Englewood Cliffs, NJ, 1986.

[58] E. Lehmann. *Testing Statistical Hypotheses, 2nd edition*. Wiley, New York, 1986.

[59] M. Li and P.M.B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, 2008.

[60] G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *Information Theory, IEEE Transactions on*, 41(3):677 –687, may 1995.

[61] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is np-hard. In *WALCOM '09: Proceedings of the 3rd International Workshop on Algorithms and Computation*, pages 274–285, Berlin, Heidelberg, 2009. Springer-Verlag.

[62] G. Morvai and B. Weiss. On classifying processes. *Bernoulli*, 11(3):523–532, 2005.

[63] G. Morvai, S. Yakowitz, and L. Gyorfi. Nonparametric inference for ergodic, stationary time series. *Ann. Statist*, 24(1):370–379, 1996.

[64] G. Morvai, S.J. Yakowitz, and P. Algoet. Weakly convergent nonparametric forecasting of stationary time series. *Information Theory, IEEE Transactions on*, 43(2):483 –498, March 1997.

[65] A. Nobel. Hypothesis testing for families of ergodic processes. *Bernoulli*, 12(2):251–269, 2006.

[66] Andrew B. Nobel. Limits to classification and regression estimation from ergodic processes. *The Annals of Statistics*, 27(1):pp. 262–273, 1999.

[67] D. Ornstein. *Ergodic Theory, Randomness, and Dynamical Systems*. Yale Mathematical Monographs. Yale University Press, 1974.

[68] D. S. Ornstein and P. C. Shields. The $\bar{d}$-recognition of processes. *Advances in Mathematics*, 104(2):182 – 224, 1994.

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

[70] A.I. Plesner and V.A. Rokhlin. Spectral theory of linear operators, II. *Uspekhi Matematicheskikh Nauk*, 1:71–191, 1946.

[71] J. Poland and M. Hutter. Defensive universal learning with experts. In *Proc. 16th International Conf. on Algorithmic Learning Theory (ALT'05)*, volume 3734 of *LNAI*, pages 356–370, Singapore, 2005. Springer, Berlin.

[72] J. Poland and M. Hutter. Universal learning of repeated matrix games. In *Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'06)*, Ghent, 2006.

[73] H. Rogers. *Theory of recursive functions and effective computability*. McGraw-Hill series in higher mathematics. McGraw-Hill, 1967.

[74] W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):pp. 506–514, 1978.

[75] B. Ryabko. Coding of a source with unknown but ordered probabilities. *Problems of Information Transmission*, 15(2):134–138, 1979.

[76] B. Ryabko. Twice-universal coding. *Problems of Information Transmission*, 3:173–177, 1984.

[77] B. Ryabko. Noiseless coding of combinatorial sources, Hausdorff dimension, and Kolmogorov complexity. *Problems of Information Transmission*, 22:16–26, 1986.

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

[79] 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.

[80] B. Ryabko and J. Astola. Universal codes as a basis for time series testing. *Statistical Methodology*, 3:375–397, 2006.

[81] B. Ryabko, J. Astola, and A. Gammerman. Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *Theoretical Computer Science*, 359:440–448, 2006.

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

[83] P. Shields. The interactions between ergodic theory and information theory. *IEEE Trans. on Information Theory*, 44(6):2079–2093, 1998.

[84] A. N. Shiryaev. *Probability*. Springer, 1996.

[85] P. Smyth. Clustering sequences with hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 648–654. MIT Press, 1997.

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

[87] R. Sutton and A. Barto. *Reinforcement learning: An introduction*. Cambridge, MA, MIT Press, 1998.

[88] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27:1134–1142, November 1984.

[89] V.N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998.

[90] R. Zadeh and S. Ben-David. A uniqueness theorem for clustering. In A. Ng J. Bilmes, editor, *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*, Montreal, Canada, 2009.

[91] Shi Zhong and Joydeep Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4:1001–1037, 2003.

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

[93] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83–124, 1970.