Learning Theory for Conditional Risk Minimization

Alexander Zimin
IST Austria
azimin@ist.ac.at

Christoph H. Lampert
IST Austria
chl@ist.ac.at

Abstract

In this work we study the *learnability of stochastic processes* with respect to the conditional risk, i.e. the existence of a learning algorithm that improves its next-step performance with the amount of observed data. We introduce a notion of pairwise discrepancy between conditional distributions at different times steps and show how certain properties of these discrepancies can be used to construct a successful learning algorithm. Our main results are two theorems that establish criteria for learnability for many classes of stochastic processes, including all special cases studied previously in the literature.

1 Introduction

A large part of machine learning relies on the assumption that data is independent and identically distributed. However, there are also many situations in which this assumption is violated, either because data points are statistically dependent, or because the underlying data distribution changes over time. In this paper, we study a scenario, in which the observed data is a realization of a *stochastic process*, i.e. a time series.

Stochastic processes can be conveniently understood from a generative point of view: each data point is sampled from a *conditional data distribution*, where the conditioning is on the sequence of observations so far. Several common situations are naturally represented in this way, including the case of i.i.d. data (the conditioning has no effect), Markov sequences (only the last element of the sequence influences the conditioning), deterministic or stochastic dynamical systems (the conditional distribution is a deterministic function of the history, potentially plus an independent noise term). In this work, however, we add a discriminative aspect: given a loss function and a set of hypotheses, the task is to find the best hypothesis to apply at the next time step. Formally, we study the problem of *conditional risk minimization*, where the risk at any time step is defined with respect to the conditional data distribution, i.e. as the expectation of the loss of a hypothesis on the next data point, conditioned on the data observed so far.

For a given class of processes, the task is to find an algorithm that outputs one hypothesis for each step with the property that the difference of their risks to the optimal ones converges to zero for every process in the class. We call a class of stochastic processes *learnable*, if it admits such an algorithm, and we call the algorithm itself a *limit learner* for this class.

Learnability has been established for a number of specific classes, see our overview in Section 2. However, these is no dedicated study of what makes stochastic processes learnable in general.

In this work we provide a general view on the problem by identifying a key characteristic of a stochastic process that allows the construction of a limit learner. Our main insight is the importance of the *pairwise discrepancies*, a notion of distance between the conditional data distributions of the process at different time steps. These are incomputable quantities, but they can be controlled in various ways depending on the properties of a particular process.

We distinguish between two situations: convergent and non-convergent discrepancies. For the former we show that if the pairwise discrepancies exhibit a specific type of convergence, then a standard empirical risk minimization algorithm can be shown to be a limit learner. This result covers the existing results on the learnability of particular classes. For the non-convergent situation we prove a general theorem that says that if one has tight control of the individual discrepancies for every process in a class, then a modification of an empirical risk minimization algorithm is a limit learner.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

2 Related work

While initially the statistical learning theory heavily relied on the i.i.d. assumption [Vapnik and Chervonenkis, 1974], extensions to time series have soon become a viable and important research direction. Early work on the learning theory for the stochastic processes considered the *marginal risk*, where the expectation of the loss is taken with respect to the marginal distribution of the process [Yu, 1994, Meir, 2000]. While the marginal risk is representative for the long term behaviour of the process, it was argued [Pestov, 2010, Shalizi and Kontorovitch, 2013] that the conditional risk is the more relevant quantity in many situations, namely when one is interested in the short term behaviour.

Conditional risk minimization is related to forecasting future values of time series, which is a well-studied topic with long history. Traditional approaches include forecasting by fitting different parametric models to the data, such as ARMA or ARIMA, or using spectral methods, see, e.g., [Box et al., 2015]. Alternative approaches include nonparametric prediction of time series [Modha and Masry, 1996, 1998, Alquier and Wintenberger, 2012], and prediction by statistical learning [Alquier et al., 2013, McDonald et al., 2012]. A related line of work comes from the field of dynamical systems, where one tries to estimate the underlying transformation that governs the transitions, see [Nobel, 2001, Farmer and Sidorowich, 1987, Casdagli, 1989, Steinwart and Anghel, 2009].

A number of papers established the learnability of particular classes: i.i.d., exchangeable, some special cases of stochastic processes, see [Steinwart, 2005, Pestov, 2010, Berti and Rigo, 2016, Mohri and Rostamizadeh, 2013], however, all these results look at the processes for which the conditional risk can be estimated by an uniformly weighted average over the previous observations. We cover these classes in our Theorem 1 by providing a general condition, which is fulfilled in these special cases. However, there are a lot of classes of our interest, such as dynamical systems, ergodic processes or distributional drift scenarios, which are not covered by the condition of Theorem 1 and require a different treatment. [Kuznetsov and Mohri, 2014] also looked at the possibility of estimating the conditional risk by an empirical average, but the convergence of their bound requires a different definition of a risk when one looks not at the next step distribution, but at some number steps into the future with this number increasing with the amount of observed data.

The above works follow a classical path of establishing learnability using the empirical risk minimizer, i.e. the minimizer of the training error, an estimate of a risk. In our work we use the same principle, minimization of an estimator. However, since the training error usually is a poor estimate of a conditional risk, the central problem in our setting is a choice of an estimator. The problem of the estimation of a conditional risk could be solved by estimating the next-step conditional distribution, which, however, is a hard problem in general. For example, [Györfi et al., 1998] show that it is not possible for all stationary and ergodic processes. On the opposite side, [Morvai, 2003] provides a scheme that can estimate the conditional probabilities at some random points in time, which, unfortunately, is not good enough for our setting since we need an estimator at every time step.

Conditional risk minimization was considered by [Kuznetsov and Mohri, 2015] and later extended in [Kuznetsov and Mohri, 2016]. They do not try to construct a limit learner, but rather consider the behaviour of the empirical risk minimization algorithm at each fixed time step by taking a non-adaptive estimator. Unfortunately, their methods can not be used to show learnability, because the generalization bounds have a constant term in the upper bound, which prevents it from converging. [Zimin and Lampert, 2015] considered a conditional risk minimization problem with a different notion of risk, when at each time step one conditions not on the whole sample, but on some fixed number of observations. For this problem setting, they proved the learnability of stationary mixing processes under a number of assumptions using NadarayaâĂŞWatson kernel estimator. [Wintenberger, 2014] looked at the problem of bounding the cumulative conditional risk in the online learning scenario. The main difference is that we study the harder problem of minimizing the risk at each step and not in the cumulative sense. We discuss the relation to their work in more details in the supplementary material.

3 Learnability of Stochastic Processes

3.1 Conditional risk minimization problem

We start by describing our notations. We are given a sequence of observations $\{\mathbf{z}_t\}_{t=1}^n$ from a stochastic process taking values in some space $\mathcal{Z}$. We will write $\mathbf{z}_{i:j}$ as a shorthand for $(\mathbf{z}_i, \dots, \mathbf{z}_j)$ for $i \le j$. We consider a hypotheses class $\mathcal{H}$, which is usually a subset of $\{h : \mathcal{Z} \to \mathcal{D}\}$ with $\mathcal{D}$ being a decision space. We also fix a loss function $\ell : \mathcal{D} \times \mathcal{Z} \to [0, 1]$, that allows us to evaluate the loss of a given hypothesis h on a given point z as $\ell(h(z), z) = \ell(h, z)$, where the latter version is used to shorten the notation. Whenever needed, we will use $\mathcal{L}(\mathcal{H})$ to denote the induced space of functions $\{\ell(h, \cdot), \forall h \in \mathcal{H}\}$. Throughout the paper we assume that $\mathcal{L}(\mathcal{H})$ has a finite sequential fat-shattering

dimension, a notion of complexity of a function class (the definition can be found in the supplementary material). For any time step t, we denote by $\mathbb{E}_t[f(\mathbf{z}_{t+1})]$ the expectation of a function f with respect to the distribution of the next step, conditioned on the data so far, i.e. $\mathbb{E}[f(\mathbf{z}_{t+1})|\,\mathbf{z}_{1:t}]$

Example. *A typical example of a setup is a classification problem, where $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ with $\mathcal{X}$ being an input space, $\mathcal{H} \subseteq \{h : \mathcal{X} \to [0,1]\}$ is a space of classifiers that output the probability of an input belonging to one of the classes, i.e. $\mathcal{D} = [0,1]$, and $\ell(h,(x,y)) = (h(x)-y)^2$ is the squared loss. However, our general formulation allows us to model not only standard machine learning tasks, but also, for example, time series prediction problems. To predict the next value of a discrete valued process taking values in a space S, we can define the hypotheses set as a set of constant functions: $\mathcal{H} = \{h_s(z) = s, \forall s \in S\}$. Then choosing a hypothesis is equivalent to choosing a value in S. In this situation it makes sense to use 0-1 loss $\ell(h_s, z) = \mathbb{I}[h_s(z) \neq z] = \mathbb{I}[s \neq z]$.*

Our goal at each step is to find a hypothesis with the minimal conditional risk, i.e. the minimizer of the expected loss on the next point conditioned on the observed data so far. Formally, the risk at step n is $R_n(h) = \mathbb{E}_n[\ell(h, \mathbf{z}_{n+1})]$ and we want to perform the minimization

$$\min_{h\in\mathcal{H}} R_n(h). \tag{1}$$

Example. *For the above-mentioned example of predicting the next step of discrete valued process, let us assume that the process is a Markov chain with a state space S and fix a transition function $\pi : S \to \Delta_S$, where Δ_S is a space of distributions over S. Then finding the most probable value on the next step can be stated as a conditional risk minimization problem with the defined $\mathcal{H}$ and ℓ:*

$$\min_{h\in\mathcal{H}} \mathbb{E}[\ell(h, \mathbf{z}_{n+1})|\,\mathbf{z}_{1:n}] = \min_{s\in S} \mathbb{P}[s \neq \mathbf{z}_{n+1}|\,\mathbf{z}_n], \tag{2}$$

which is equivalent to $\max_{s\in S} \pi(s|\mathbf{z}_n)$.

Since, in practice, the distribution of the process is unknown, we are looking for a method that can perform (1) based only on the observed data, i.e. that produces a sequence of hypotheses h_n, where each h_n can be computed from the data observed up to step n and h_n approximates the minimum of (1). In view of our task, the minimization of the conditional risk, we need control over the quality of the approximations. Ideally, the quality should improve with the method observing more data. We formalize this goal in the following definition.

Definition 1 (**Learnability**)**.** For a fixed loss function ℓ and a hypotheses class $\mathcal{H}$, we call a class of processes $\mathcal{C}$ *conditionally learnable in the limit* if there exists an algorithm that, for every process P in $\mathcal{C}$, produces a sequence of hypotheses, h_n, each based on $\mathbf{z}_{1:n}$, satisfying

$$R_n(h_n) - \inf_{h\in\mathcal{H}} R_n(h) \to 0 \tag{3}$$

in probability over the samples drawn from P. We call an algorithm that satisfies this condition a limit learner for the class $\mathcal{C}$.

Throughout the paper we will call conditionally learnable in the limit classes just learnable. It is also possible to consider almost sure convergence in the definition of learnability with a minor modifications of our statements.

Some classes of processes are known to be learnable, for example, i.i.d. sequences by support vector machines, [Steinwart, 2005], or exchangeable sequences by empirical risk minimization, [Pestov, 2010, Berti and Rigo, 2016]. On the opposite side, the class of all stationary and ergodic binary processes is not learnable in the particular prediction setting, as we show in the supplementary material based on the results of [Györfi et al., 1998].

3.2 Empirical risk minimization

In this paper we focus on the empirical risk minimization (ERM) principle, which governs us to construct an estimator, $\hat{R}_n$, of the risk based on the data and use the minimizer of the estimator, $h_n = \operatorname{argmin}_{h\in\mathcal{H}} \hat{R}_n(h)$, as an output hypothesis. The main question is how to construct this estimator. In the case of an i.i.d. process, it is common to use $\frac{1}{n}\sum_{t=1}^n \ell(h, \mathbf{z}_t)$. For general processes, however, this quantity is not a good choice as it does not converge to the conditional risk, except for some special cases, which are covered by Theorem 1. For other situations we consider linear estimators of the form $\hat{R}_n(h) = \sum_{t=1}^n w_t \ell(h, \mathbf{z}_t)$ (we omit the dependence of w_t's on n, but emphasize that at each step the weights can be different, because we estimate different quantities). We study the question of finding "good" weights w based on the observed sample that make empirical risk minimization a limit learner. This makes the weights a function of the observed data, so they must be treated as random variables.

Example. *For the Markov chain example the estimator has the form $\hat{R}_n(h_s) = \sum_{t=1}^n w_t \mathbb{I}[s \neq \mathbf{z}_t]$. Clearly, there is no fixed choice of weights that would approximate the conditional risk well for every realization of the process. The empirical average with uniform weights, $w_t \equiv \frac{1}{n}$, for example, converges to the risk with respect to the stationary distribution of the chain, not the conditional one. Instead, we should choose w_t*

to be large, if the (conditional) distribution of $\mathbf{z}_t$ *is similar to the distribution of* $\mathbf{z}_{n+1}$*, i.e.* $\pi(\cdot|\mathbf{z}_{t-1}) \approx \pi(\cdot|\mathbf{z}_n)$*. Otherwise,* w_t *should be small. The same intuition holds for general processes, as we will show in Section 4.2.*

To study the properties of the ERM, we use the fact that

$$R_n(h_n) - \inf_{h\in\mathcal{H}} R_n(h) \leq 2 \sup_{h\in\mathcal{H}} \left|\hat{R}_n(h) - R_n(h)\right| \quad (4)$$

and, henceforward, focus on the right hand side, i.e. uniform deviations of the estimator. Our starting point is the following decomposition, proposed in [Kuznetsov and Mohri, 2015]:

$$\begin{aligned} &\sup_{h\in\mathcal{H}} \left|\hat{R}_n(h) - R_n(h)\right| \\ &\leq \sup_{h\in\mathcal{H}} \left|\sum_{t=1}^{n} w_t \left(\ell(h,\mathbf{z}_t) - R_{t-1}(h)\right)\right| \quad (5) \\ &\quad + \sup_{h\in\mathcal{H}} \left|\sum_{t=1}^{n} w_t R_{t-1}(h) - R_n(h)\right|. \end{aligned}$$

For fixed data-independent weights, these two summands are well understood. The first term represents the stochastic part of the problem and can be shown to converge under very general conditions using the machinery of [Rakhlin et al., 2014]. An important fact is that the rate of convergence is determined by $\sqrt{\sum_{t=1}^n w_t^2}$. For example, uniform weights $w_t = \frac{1}{n}$ yield an optimal, $\mathcal{O}(\frac{1}{\sqrt{n}})$, rate. On the contrary, if only one sample is present, i.e. $w_t = 1$ for some t and $w_t = 0$ for all others, then there is no convergence as $\sqrt{\sum_{t=1}^n w_t^2}$ does not decrease as a function of n.

The second term measures how well $R_n(h)$ is approximated by the mixture of the previous expectations. As a distance measure between two distributions, it is also known by the name of an integral probability metric and is studied, for example, in [Zolotarev, 1983, Müller, 1997]. For risk minimization problems it is a very suitable measure of distance, because it is adapted to the underlying hypothesis set, making it a popular choice in the domain adaptation literature [Kifer et al., 2004, Ben-David et al., 2007, Mansour et al., 2009, Ben-David et al., 2010, Mohri and Medina, 2012].

The decomposition in (5) highlights a trade-off between two desirable properties of the weights: they should offer good statistical power (have small $\sqrt{\sum_{t=1}^n w_t^2}$), while achieving a high approximation quality (small discrepancy).

To our best knowledge, no studies has gone beyond the decomposition in (5) and even the behaviour of the terms in (5) is understood only for fixed weights.

4 Our contribution

Our major insight is that while the second term in (5) is hard to control directly, i.e. it is hard to find good weights based only on the data, controlling an upper bound is a more plausible task. To this end, we introduce the key notion of *pairwise discrepancies*, a measure of distance between conditional distributions at different time steps.

Definition 2 (Pairwise discrepancy). For a sample $\mathbf{z}_1, \mathbf{z}_2, \ldots$ from a fixed stochastic process, the pairwise discrepancy between time points i and j is

$$d_{i,j} = \sup_{h\in\mathcal{H}} |R_i(h) - R_j(h)|. \quad (6)$$

For weights that satisfy $w_t \geq 0$ and $\sum_{t=1}^n w_t = 1$, we can further upper bound the second term in (5).

$$\sup_{h\in\mathcal{H}} \left|\sum_{t=1}^{n} w_t R_{t-1}(h) - R_n(h)\right| \leq \sum_{t=1}^{n} w_t d_{t-1,n}, \quad (7)$$

which suggests that tight control of the pairwise discrepancies is desirable.

From this point we distinguish between two situations: when the discrepancies exhibit a special form of convergence and when they do not. In the former case, we show that using uniform weights is sufficient. For the latter, we construct a special weighting scheme that is based on the notion of an M-bound, which we introduce later, allowing to control the discrepancies.

4.1 Convergent case

The intuition behind this situation is that if a sequence is convergent then the average of elements in the sequence also converges to the same limit. However, in our situation we do not have a single sequence, but rather a double array of discrepancies, $d_{t,n}$. We use the following definition of convergence, which is a modification of standard convergence in probability to our special case of double array.

Definition 3. A double array of random variables $d_{t,n}$ with $n \in \mathbb{N}$ and $0 \leq t < n$ is called *convergent* if

$$\begin{aligned} \forall\varepsilon > 0, \forall\delta > 0, \exists n_0, t_0 : 0 \leq t_0 < n_0, \forall n \geq n_0, \quad (8) \\ \forall t_0 \leq t < n : \mathbb{P}\left[d_{t,n} > \varepsilon\right] \leq \delta. \end{aligned}$$

With this definition in hand, we can state the following theorem.

Theorem 1. *If every process in the class* $\mathcal{C}$ *has convergent discrepancies, then the ERM algorithm with uniform weights, i.e.* $w_t = \frac{1}{n}$ *at step* n*, is a limit learner.*

We will present a few examples of classes with convergent discrepancies in Section 5.1.

4.2 Non-convergent case

As we have observed in the example of a Markov chain, in the non-convergent situation there is no single choice of weights that can be applied irrespectively of the process and the sample. Rather we should adjust the weights to the data at hand.

Example. *Let us use the same Markov example to build intuition of how that can be achieved. First, we observe that the pairwise discrepancies can be written using the transition function,* $d_{i,j} = \max_{s\in S} |\pi(s|\mathbf{z}_i) - \pi(s|\mathbf{z}_j)|$. *Then it is immediate that* $d_{t-1,n} \le \mathbb{I}[\mathbf{z}_{t-1} \neq \mathbf{z}_n]$, *hence, setting* $w_t = \frac{\mathbb{I}[\mathbf{z}_{t-1}=\mathbf{z}_n]}{\sum_{j=1}^n \mathbb{I}[\mathbf{z}_{j-1}=\mathbf{z}_n]}$ *seems like a good choice: it uses only samples from the distribution we are trying to predict for, and it distributes the mass evenly among those. Consequently, the discrepancy term in* (5) *is zero and* $\sum_{t=1}^n w_t^2$ *is minimal. For each fixed value* s *of* $\mathbf{z}_n$, *the above upper bound* $\mathbb{I}[\mathbf{z}_{t-1} \neq s]$ *depends only on the past values of the process and* $\mathbf{z}_n$ *is actually the only thing that ties the weights to future values.*

We summarize the important properties of the example in the following definition, which allows more flexibility and does not require the space to be discrete or the process to be Markov.

Definition 4 (**M-bound**). An *M-bound* is a double array of random variables $M_{t,j}$, where $t, j \in \mathbb{N}$, such that

1. $M_{t,j}$ is a function of $\mathbf{z}_1, \dots, \mathbf{z}_t$ for any $j \in \mathbb{N}$,
2. there exists a sequence of random variable J_n taking values in $\mathbb{N}$ and a deterministic sequence Δ_n such that
$$d_{t-1,n} \le \Delta_n + M_{t-1,J_n}$$
for $1 \le t \le n$.

Note that for any process there is a trivial, but not that useful, M-bound with $\Delta_n = 1$ and $M_{t,j} = 0$ for all t, j.

Example. *In our running example, let* $s_1, s_2, \dots, s_{|S|}$ *be any enumeration of* S. *Then we can define* $M_{t,j} = \mathbb{I}[\mathbf{z}_t \neq s_j]$ *for* $1 \le j \le |S|$, $\Delta_n = 0$ *and set* $J_n = k$ *for* k *that satisfies* $s_k = \mathbf{z}_n$.

For a given M-bound, we can further upper bound (7):

$$\sup_{h\in\mathcal{H}} \left| \sum_{t=1}^n w_t R_{t-1}(h) - R_n(h) \right| \le \Delta_n + \sum_{t=1}^n w_t M_{t-1,J_n}. \tag{9}$$

This expression would be minimized by setting $w_t = 1$ for the index t with minimal value of M_{t-1,J_n}, while keeping the other weights at 0. However, as discussed above, such a choice is disastrous for the stochastic part of (5). Therefore, we suggest to use a version of soft-min with some smoothing function $g_n : \mathbb{R} \to [0, 1]$ that also could be different at each step.

$$w_t(J_n) = \frac{g_n(M_{t-1,J_n})}{\sum_{j=1}^n g_n(M_{j-1,J_n})} \tag{10}$$

for $1 \le t \le n$. A popular smoothing function is $g_n(x) = e^{-\gamma_n x}$ for some $\gamma_n > 0$. In the example of a Markov chain, the simpler $g_n(x) = \mathbb{I}[x = 0] = \lim_{\gamma\to 0} e^{-\gamma x}$ was sufficient.

Due to the stochastic nature of the process, it may not be possible to have a good bound for each possible realization. This can be seen even for simple Markov chains. Imagine a situation in a Markov chain when at step n we observe the state $\mathbf{z}_n$ for the first time. Then we have no information in the sample about the distribution of the next step. Nevertheless, if such realizations are rare, the process can still be learnable. We formalize this idea in an exceptional set of realizations, which we are going to ignore and require that they appear with small probability.

Definition 5 (**Exceptional set**). For a fixed n, for any $k \ge 1$ and $1 \le m \le n$, set

$$E_{k,m} = \Big\{ J_n \le k \ \wedge \ \sum_{t=1}^n g_n(M_{t-1,J_n}) \ge m \Big\}. \tag{11}$$

We define $E_{k,m}^c$, the complement of $E_{k,m}$, as an *exceptional set* of the realizations.

Note that this set is also different at each step, but we omit the index n to avoid cluttering of the notations. The condition in (11) mainly requires to have a lower bound on the denominator of $w_t(J_n)$'s, thereby avoiding the situation observed in a Markov chain example. We will discuss the behaviour of $\mathbb{P}[E_{k,m}]$ for discrete state Markov chains (and other processes) in Section 6.

Our main result is the following theorem, which provides guarantees on the performance of empirical risk minimization with our proposed choice of weights.

Theorem 2. *For any fixed* n, *for any* $k, m \ge 1$, $\alpha \in [0, 1]$ *and* $\beta \in [0, \alpha/4]$ *the following inequality holds*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \left| \hat{R}_n(h) - R_n(h) \right| - \Delta_n - \Lambda_n \ge \alpha \right] \tag{12}$$
$$\le \frac{2k\mathcal{N}_\infty(\mathcal{L}(\mathcal{H}), \beta, n)}{(\alpha - 4\beta)^2} e^{-\frac{1}{2}m(\alpha-4\beta)^2} + \mathbb{P}\left[E_{k,m}^c\right],$$

where $\Lambda_n = \sum_{t=1}^n w_t(J_n) M_{t-1,J_n}$ *and* $\mathcal{N}_\infty(\mathcal{L}(\mathcal{H}), \beta, n)$ *is a maximal* β*-cover of* $\mathcal{L}(\mathcal{H})$ *with respect to the* ℓ_∞ *norm (the definition is given in the appendix).*

It may be instructive to look at the different form of Theorem 2. For any $k, m \geq 1$, any $\delta > \mathbb{P}[E^c_{k,m}]$ and $\beta > 0$, with probability $1 - \delta$ the following inequality holds:

$$R_n(h_n) - \inf_{h \in \mathcal{H}} R_n(h) \leq \Delta_n + \Lambda_n + 4\beta \tag{13}$$

$$+ \sqrt{\frac{2 \log \frac{4m}{\delta} + 2 \log 2k \mathcal{N}_\infty(\mathcal{L}(\mathcal{H}), \beta, n)}{m}}. \tag{14}$$

From this form we can read off conditions under which the ERM algorithm becomes a limit learner. As this means that the right hand side converges to 0, in particular we need that $\Delta_n + \Lambda_n$ (which is simply the M-bound (10) for the chosen weights) vanishes in the limit.

Corollary 1. *Assume that for every the process P in the class $\mathcal{C}$ there exists a sequence of M-bounds and smoothing functions satisfying $\Delta_n \to 0$ and $\Lambda_n \to 0$. In addition, if there exist k_n, m_n satisfying $\frac{m_n}{\log n} \to \infty$ and $\mathbb{P}[E^c_{k_n,m_n}] \to 0$, then $\mathcal{C}$ is conditionally learnable in the limit by the ERM algorithm based on the given M-bounds and smoothing functions.*

Note that the algorithm does not require knowledge of the parameters k and m, merely the existence of good values.

The above results show that the quality of the M-bounds is of crucial importance for establishing the learnability. In Section 5.2 we highlight constructions of M-bounds based on prior knowledge (or assumptions) on the processes. Moreover, each process of a class should not produce unfavorable sequences very often. We will look into this property of processes in more details in Section 6.

5 Controlling pairwise discrepancies

In this section we look at different ways in which it is possible to control the pairwise discrepancies.

5.1 Convergent case

A trivial example of the convergent situation is an i.i.d. process, because all the discrepancies are zero. More generally, we consider a class of *uniformly convergent martingales*. This class consists of processes that form martingales for every function $f \in \mathcal{L}(\mathcal{H})$ applied to its values, that is $\mathbb{E}_s\left[\mathbb{E}_t\left[f(\mathbf{z}_{t+1})\right]\right] = \mathbb{E}_s\left[f(\mathbf{z}_{s+1})\right]$ for $s < t$. By standard results in the theory of martingales, e.g. [Williams, 1991], for every f there is a limit random variable $\mathbf{x}_f$, such that $\mathbb{E}_t f \to \mathbf{x}_f$ in probability and $\mathbb{E}_t f = \mathbb{E}_t \mathbf{x}_f$. We call a stochastic process a uniformly convergent martingale if $\sup_{f \in \mathcal{L}(\mathcal{H})} |\mathbb{E}_t f - \mathbf{x}_f| \to 0$ in probability. For such classes we have the following results.

Lemma 1. *A uniformly convergent martingale has convergent discrepancies.*

Corollary 2. *If a class $\mathcal{C}$ consists of a uniformly convergent martingales, then it is learnable by the ERM algorithm with uniform weights.*

As shown in [Berti et al., 2002], a prominent example of uniformly convergent martingales is a class of exchangeable sequences that is widely used in the statistical literature. Exchangeability means that the sampled data has the same distribution irrespectively of the order of variables. In addition to i.i.d., this assumption also covers an important case of complete dependence, when one observes copies of the same random variable. From this perspective, Corollary 2 can be seen as a generalization of the results proven by [Berti et al., 2002] for exchangeable sequences and, even further, by [Berti and Rigo, 2016] for conditionally identically distributed sequences.

Another example of a class with convergent discrepancies is a class of processes used in [Mohri and Rostamizadeh, 2013]. Their assumption (equation 6) states for any hypothesis h that depends only on $\mathbf{z}_{1:t}$ and any $n \geq s > t$ we have

$$\mathbb{E}\left[\ell(h, \mathbf{z}_{n+1}) \middle| \mathbf{z}_{1:s}\right] = \mathbb{E}\left[\ell(h, \mathbf{z}_{n+1}) \middle| \mathbf{z}_{1:t}\right]. \tag{15}$$

In particular, that means that for any fixed h (not depending on the sample)

$$\mathbb{E}\left[\ell(h, \mathbf{z}_{n+1}) \middle| \mathbf{z}_{1:n}\right] = \mathbb{E}\left[\ell(h, \mathbf{z}_{n+1})\right]. \tag{16}$$

If the marginal distributions at each step are the same, as assumed in [Mohri and Rostamizadeh, 2013], then (16) yields that all $d_{t,n}$ are zero, so they are convergent.

5.2 Non-convergent case

For the non-convergent case we give two examples of how M-bounds can be constructed.

We already discussed an example of an M-bound for predicting the next state of a discrete state Markov processes in Section 4.2. The construction can be extended to more general situations, when the discrepancy between two time steps can be related to the similarity of their histories, namely for all $h \in \mathcal{H}$,

$$\left|\mathbb{E}\left[\ell(h, \mathbf{z}_{i+1}) \middle| \mathbf{z}_{1:i}\right] - \mathbb{E}\left[\ell(h, \mathbf{z}_{j+1}) \middle| \mathbf{z}_{1:j}\right]\right| \leq \lambda(\mathbf{z}_{1:i}, \mathbf{z}_{1:j}), \tag{17}$$

where $\lambda : \mathcal{Z}^i \times \mathcal{Z}^j \to \mathbb{R}_+$ is a form of distance measure. This property can be derived, e.g., from the continuity of the conditional distribution with respect to the history, which is a common assumption in the literature on nonparametric estimation, e.g. [Györfi et al., 1989, Hansen, 2008, Linton and Sancetta, 2009].

For example, if λ takes only the q most recent values into account and is a metric, we can rewrite inequality (17) as

$$\sup_{h\in\mathcal{H}} |R_i(h) - R_j(h)| \le \lambda(\mathbf{z}_{i-q+1:i}, \mathbf{z}_{j-q+1:j}). \quad (18)$$

Now, for fixed $\Delta_n > 0$ let $\mathcal{M}$ be a Δ_n-cover of $\mathcal{Z}^q$ with respect to λ, i.e. a countable subset of $\mathcal{Z}^q$, such that for every element $\bar{z} \in \mathcal{Z}^q$ there is an element of the cover $\bar{m} \in \mathcal{M}$ with $\lambda(\bar{z}, \bar{m}) \le \Delta_n$. For any $\bar{z}$, let $c(\bar{z})$ denote the closest element of $\mathcal{M}$. Then we have

$$\sup_{h\in\mathcal{H}} |R_t(h) - R_n(h)| \le \Delta_n + \lambda(\mathbf{z}_{t-q+1:t}, c(\mathbf{z}_{n-q+1:n})). \quad (19)$$

Now, let $\bar{m}_1, \bar{m}_2, \dots$ be an enumeration of $\mathcal{M}$, and define $M_{t,j} = \lambda(\mathbf{z}_{t-q+1:t}, \bar{m}_j)$. Then we obtain an M-bound by setting $J_n = k$ for k that satisfies $m_k = c(\mathbf{z}_{n-q+1:n})$. In a similar way, we can obtain M-bounds for related statistical settings, as the assumption of continuity is a fundamental ingredient for many theoretical results in nonparametric statistics.

Another example of an M-bound can be given in the scenario of *rarely changing distributions*. In this case we observe independent samples, however, the distribution from which these are sampled may occasionally change [Tartakovsky et al., 2014]. Formally, for a fixed sequence of distribution $D_1, \dots, D_k$ and change points, $1 = c_1 < \dots < c_{k+1} = n + 1$, the samples $\mathbf{z}_{c_i:c_{i+1}-1}$ are drawn independently from the distribution D_i, for $i = 1, \dots, k$. A simple strategy for this task is to perform change point detection, for example [Tartakovsky et al., 2014, Khaleghi and Ryabko, 2016, Kifer et al., 2004], and then distribute the weight uniformly over the samples since the last change, i.e. $w_t = 0$ for $t = 1, \dots, c_k - 1$ and $w_t = \frac{1}{n-c_k+1}$ for $t = c_k, \dots, n$. A more elaborate scheme in this situation would be to estimate the discrepancies between segments and use the estimates in the place of real discrepancies. A similar approach was studied in the active learning scenario by [Pentina and Lampert, 2016].

6 Controlling the exceptional set

For classes of processes with non-convergent discrepancies, learnability requires not just the existance of an M-bound, but also control of the exceptional set (Corollary 1). In this section, we connect this property to some well-known properties of stochastic processes.

To isolate the properties of the process from the assumptions required to get an M-bound, we will analyze the ERM algorithm with a universal (though unfortunately incomputable) M-bound: we assume that we have access to the individual discrepancies $d_{i,j}$ and define the bound in the following way. Fix some $b_n > 0$ and let

$$J_n = \inf\{t \ge 1 : d_{t,n} \le b_n\}. \quad (20)$$

Then, for $t \ge J_n$

$$d_{t,n} \le d_{J_n,n} + d_{t,J_n} \le b_n + d_{t,J_n}, \quad (21)$$

by a triangle inequality. Therefore, we can set $\Delta_n = b_n$ and $M_{t,j} = d_{t,j}$ for $t \ge j$ and $M_{t,j} = 1$ for $t < j$.

To achieve the most obvious behaviour of Λ_n, we choose $g_n(x) = \mathbb{I}[x \le b_n]$ as smoothing function. Then we can guarantee that $\Delta_n + \Lambda_n \le 2b_n$ if $J_n < n$ and for $b_n \to 0$, we only need to show the existence of k_n and $m_n \to \infty$ such that $\frac{m_n}{\log n} \to \infty$ and $\mathbb{P}[E^c_{k_n,m_n}] \to 0$. Now we consider a few different classes of processes and analyze the behaviour of $\mathbb{P}[E^c_{k_n,m_n}]$ for the defined M-bound. We repeatedly use that $\mathbb{P}[E^c_{k,m}] = \mathbb{P}[A_k] + \mathbb{P}[B_{k,m}]$ for $A_k = \{J_n > k\}$ and $B_{k,m} = \left\{J_n \le k \wedge \sum_{t=J_n}^{n} \mathbb{I}[d_{t,J_n} \le b_n] < m\right\}$. Hence, we can consider the two events separately if needed. The first two examples were already covered by the convergent case, but we still mention them for illustrative purposes.

I.i.d. As noticed above, in this case $d_{i,j} = 0$ for all i, j. This means that J_n always equals to 1 and we can guarantee that $\mathbb{P}[E^c_{k,m}] = 0$ for $k = 1$ and $m = n - 1$.

Complete dependence. Let $\mathbf{z}_1$ be a random variable and $\mathbf{z}_t = \mathbf{z}_1$ for $t > 1$. Then after the first step, the conditional distributions are just delta measures concentrated on the previous point and we always get $J_n = 2$, so that we obtain $\mathbb{P}[E^c_{k,m}] = 0$ for $k = 2$ and $m = n - 2$.

Periodic sequences. Consider a periodic deterministic sequence with a fixed period $T \in \mathbb{N}$, like the one obtained by observing the trajectory of a pendulum. Because of periodicity, we know that every conditional distribution occurs at least once within each cycle, therefore, $J_n \le T$ is guaranteed and, hence, $\mathbb{P}[E^c_{k,m}] = 0$ for $k = T$ and $m = \lfloor \frac{n}{T} \rfloor - 1$.

Discrete state Markov chains. For discrete state Markov chains the bounds on the probability of $E^c_{k,m}$ are deeply connected to the notion of recurrence times. For a state $s \in S$ let T_s be the recurrence time to this state: $T_s = \inf\{t > 1 : \mathbf{z}_t = s | \mathbf{z}_1 = s\} - 1$. Then the following connection holds, as shown in the supplementary material.

$$\mathbb{P}[B_{k,m}] \le |S| m \max_s \mathbb{P}\Big[T_s > \lfloor \frac{n-k}{m} \rfloor\Big]. \quad (22)$$

Therefore, the bound can be devised from the concentration properties of the recurrence times. For the

other part of $E^c_{k,m}$ we can show that

$$\mathbb{P}\left[A_k\right] \leq |S| \max_s \mathbb{P}\left[F_s > k\right], \tag{23}$$

where $F_s = \inf\{t \geq 1 : \mathbf{z}_t = s\}$ are the first passage times, which also play an important role in the theory of Markov chains, because they reflect how fast the chain explores its state space. In combination,

$$\mathbb{P}\left[E^c_{k,m}\right] \leq |S| \max_s \mathbb{P}\left[F_s > k\right] + |S| m \max_s \mathbb{P}\Big[T_s > \lfloor \frac{n-k}{m} \rfloor\Big]. \tag{24}$$

As an example of an obtainable rate, we apply Markov's inequality to (22),

$$\mathbb{P}\left[B_{k,m}\right] \leq \frac{|S| m^2}{n-k} \max_s \mathbb{E}\left[T_s\right]. \tag{25}$$

One of the basic results from the theory of finite-state Markov chains tells us that some particular state s can be either recurrent or transient, depending on whenever $\mathbb{E}\left[T_s\right]$ is finite or not. If all $\mathbb{E}\left[T_s\right]$ are finite, then all we need is m growing slower than $\sqrt{n-k}$. This offers a nice connection of the recurrence properties of Markov chains to their learnability.

Dynamical systems. Let $(\mathcal{Z}, \Sigma, \mu, F)$ be a dynamical system, where Σ is a σ-algebra on $\mathcal{Z}$, μ is some measure on $(\mathcal{Z}, \Sigma)$ and $F : \mathcal{Z} \to \mathcal{Z}$ is a measure-preserving transformation, meaning that for any set $A \in \Sigma$ we have $\mu(F^{-1}(A)) = \mu(A)$. The evolution of a system is as follows: first $\mathbf{z}_1 \sim \mu$ is sampled and then any subsequent point is obtained through the iteration $\mathbf{z}_{t+1} = F(\mathbf{z}_t) = F^t(\mathbf{z}_1)$. Consequently, $d_{i,j} = \sup_{h \in \mathcal{H}} |\ell(h, F(\mathbf{z}_i)) - \ell(h, F(\mathbf{z}_j))|$. We assume $d_{i,j} \leq \lambda(\mathbf{z}_i, \mathbf{z}_j)$ for some metric λ on $\mathcal{Z}$. Let $C_j = \{z \in \mathcal{Z} : \lambda(z, \mathbf{z}_j) \leq b_n\}$ be a ball around $\mathbf{z}_j$ with radius b_n, then $\mathbb{P}\big[E^c_{k,m}\big]$ is controlled by the first passage times and the recurrence times to the sets C_j (analogously to the discrete Markov chain case). Formally, the recurrence time from a point $z \in \mathcal{Z}$ to a set C is defined as $T(z, C) = \inf\{t \geq 1 : F^t(z) \in C\}$. Then, the first passage time to the set is defined as $F(C) = T(\mathbf{z}_1, C)$ and the recurrence time to a set from itself is $T(C) = \operatorname{ess\,sup}_{z \in C} T(z, C)$. Similarly to the Markov chain case, the following bound holds

$$\mathbb{P}\left[E^c_{k,m}\right] \leq \mathbb{P}\left[F(C_n) > k\right] + k \max_{1 \leq j \leq k} \mathbb{P}\Big[T(C_j) > \lfloor \frac{n-j}{m} \rfloor\Big]. \tag{26, 27}$$

Poincaré's theorem, e.g. [Katok and Hasselblatt, 1997], tells us that any of the sets C_j will be visited infinitely often. A quantitative characterization of the behaviour of the recurrence times for dynamical systems can be found, for example, in [Barreira, 2008].

General stationary processes. To relate the setting to the existing work in the nonparametric prediction, assume that the process is stationary and ergodic and $d_{i,j} \leq \lambda(\mathbf{z}_{i-q+1:i}, \mathbf{z}_{j-q+1:j})$ for some integer q and metric λ on $\mathcal{Z}^q$. For $\bar{z} \in \mathcal{Z}^q$ let $C(\bar{z}) = \{\bar{y} \in \mathcal{Z}^q : \lambda(\bar{y}, \bar{z}) \leq b_n\}$. Along the lines of the previous examples, define $F(C) = \inf\{t \geq 1 : \mathbf{z}_t \in C\}$ as a first passage time to a set C. Then we have

$$\mathbb{P}\left[E^c_{k,m}\right] \leq \mathbb{P}\left[(F(C(\mathbf{z}_n))) > k\right] + k \max_{1 \leq j \leq k} \mathbb{P}\left[\sum_{t=k+1}^{n} \mathbb{I}\left[\mathbf{z}_t \in C(\mathbf{z}_j)\right] < m\right]. \tag{28}$$

In case of mixing processes, it is possible to determine the rate of recurrence for the second term. More concretely, it can be shown that $\sum_{t=k+1}^{n} \mathbb{I}\left[\mathbf{z}_t \in C(\mathbf{z}_j)\right] \approx \sum_{t=k+1}^{n} \mathbb{P}\left[\mathbf{z}_t \in C(\mathbf{z}_j)\right] \geq \inf_{\bar{z}}(n-k)\mathbb{P}\left[\mathbf{z}_t \in C(\bar{z})\right]$, see for example [Caires and Ferreira, 2005]. Therefore, for mixing processes m can be chosen proportionally to n.

Distribution drift. Bartlett [1992] introduced the setting of distributional drift: there is a deterministic sequence of distributions $D_1, \ldots, D_{n+1}$ and samples are drawn independently from the corresponding distribution: $\mathbf{z}_i \sim D_i$. Therefore, any conditional expectations is the expectation with respect to the marginal distribution of a point and we have the following expression for the discrepancies:

$$d_{i,j} = \sup_{h \in \mathcal{H}} \left| \mathbb{E}_{\mathbf{z} \sim D_{i+1}}\left[\ell(h, \mathbf{z})\right] - \mathbb{E}_{\mathbf{z} \sim D_{j+1}}\left[\ell(h, \mathbf{z})\right] \right|. \tag{29}$$

Since in the distribution drift scenario the samples are independent, the values of J_n and $\sum_{t=J_n}^{n} \mathbb{I}\left[d_{t,J_n} \leq b_n\right]$ in the definitions of $E_{k,m}$ are deterministic. Hence, we can ensure that $\mathbb{P}\big[E^c_{k,m}\big] = 0$ by trivially setting $k = J_n$ and $m = \sum_{t=J_n}^{n} \mathbb{I}\left[d_{t,J_n} \leq b_n\right]$.

7 Conclusion

We presented the first general study of the learnability of stochastic processes with respect to the conditional risk. We highlighted the central role of the pairwise discrepancies between conditional distributions and proved two theorems that establish criteria for the learnability of many classes of stochastic processes. Our results suggests that it will be beneficial to look at how existing practical methods and models relate to the discrepancies, thereby obtaining a better understanding of their effectiveness.

Acknowledgments

This work was in parts funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036.

References

Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

Pierre Alquier, Xiaoyin Li, and Olivier Wintenberger. Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 1: 65–93, 2013.

Luis Barreira. *Dimension and recurrence in hyperbolic dynamics*. Springer, 2008.

Peter L. Bartlett. Learning with a slowly changing distribution. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 243–252. ACM, 1992.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Conference on Neural Information Processing Systems (NIPS)*, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Patrizia Berti and Pietro Rigo. Asymptotic predictive inference with exchangeable data. *Submitted, currently available at http://www-dimat. unipv. it/rigo/smps2016. pdf*, 2016.

Patrizia Berti, A Mattei, and Pietro Rigo. Uniform convergence of empirical and predictive measures. *Atti del Seminario Matematico e Fisico dell'Universita' di Modena*, 50(2):465–478, 2002.

George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Sofia Caires and Jose A. Ferreira. On the nonparametric prediction of conditionally stationary sequences. *Statistical inference for stochastic processes*, 8(2):151–184, 2005.

Martin Casdagli. Nonlinear prediction of chaotic time series. *Physica D: Nonlinear Phenomena*, 35(3):335–356, 1989.

J. Doyne Farmer and John J. Sidorowich. Predicting chaotic time series. *Physical review letters*, 59(8): 845, 1987.

László Györfi, Wolfgang Härdle, Pascal Sarda, and Philippe Vieu. *Nonparametric curve estimation from time series*, volume 60. Springer-Verlag Berlin, 1989.

László Györfi, Gusztáv Morvai, and Sidney J Yakowitz. Limits to consistent on-line forecasting for ergodic time series. *IEEE Transactions on Information Theory*, 44(2):886–892, 1998.

Bruce E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(03):726–748, 2008.

Anatole Katok and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems*, volume 54. Cambridge University Press, 1997.

Azadeh Khaleghi and Daniil Ryabko. Nonparametric multiple change point estimation in highly dependent time series. *Theoretical Computer Science*, 620: 119–133, 2016.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on very large data bases*, volume 30, pages 180–191. VLDB Endowment, 2004.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for time series prediction with non-stationary processes. In *Algorithmic Learning Theory (ALT)*, pages 260–274. Springer, 2014.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Conference on Neural Information Processing Systems (NIPS)*, pages 541–549, 2015.

Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and online learning. In *Workshop on Computational Learning Theory (COLT)*, pages 1190–1213, 2016.

Oliver Linton and Alessio Sancetta. Consistent estimation of a general nonparametric regression function in time series. *Journal of Econometrics*, 152(1):70–78, 2009.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Workshop on Computational Learning Theory (COLT)*, 2009.

Daniel J. McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Time series forecasting: model evaluation and selection using nonparametric risk bounds. *arXiv preprint arXiv:1212.0463*, 2012.

Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, 2000.

Dharmendra S. Modha and Elias Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42(6):2133–2145, 1996.

Dharmendra S. Modha and Elias Masry. Memory-universal prediction of stationary random processes. *IEEE Transactions on Information Theory*, 44(1): 117–133, 1998.

Mehryar Mohri and Andres Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory (ALT)*, pages 124–138. Springer, 2012.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. http://www.cs.nyu.edu/mohri/pub/niidj.pdf, 2013. (Oct 10, corrected version of [JMLR (11), 2010]).

Gusztáv Morvai. Guessing the output of a stationary binary time series. In *Foundations of Statistical Inference*, pages 207–215. Springer, 2003.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

Andrew Nobel. Consistent estimation of a dynamical map. In *Nonlinear dynamics and statistics*, pages 267–280. Springer, 2001.

Anastasia Pentina and Christoph H. Lampert. Active task selection for multi-task learning. *arXiv preprint arXiv:1602.06518*, 2016.

Vladimir Pestov. Predictive PAC learnability: A paradigm for learning from exchangeable input data. In *IEEE International Conference on Granular Computing (GrC)*, pages 387–391, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, pages 1–43, 2014.

Cosma Shalizi and Aryeh Kontorovitch. Predictive PAC learning and process decompositions. In *Conference on Neural Information Processing Systems (NIPS)*, pages 1619–1627, 2013.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

Ingo Steinwart and Marian Anghel. Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *The Annals of Statistics*, pages 841–875, 2009.

Alexander Tartakovsky, Igor Nikiforov, and Michèle Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

Vladimir Naumovich Vapnik and Alexey Chervonenkis. *Theory of pattern recognition*. Nauka, 1974.

David Williams. *Probability with martingales*. Cambridge University Press, 1991.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *arXiv preprint arXiv:1404.1356*, 2014.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, pages 94–116, 1994.

Alexander Zimin and Christoph H. Lampert. Conditional risk minimization for stochastic processes. *arXiv preprint arXiv:1510.02706*, 2015.

Vladimir Mikhailovich Zolotarev. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28(2):264–287, 1983.