JMLR: Workshop and Conference Proceedings vol 49:1–24, 2016

Time Series Prediction and Online Learning

Vitaly Kuznetsov VITALY@CIMS.NYU.EDU
Courant Institute, New York

Mehryar Mohri MOHRI@CS.NYU.EDU
Courant Institute and Google Research, New York

Abstract

We present a series of theoretical and algorithmic results for time series prediction leveraging recent advances in the statistical learning analysis of this problem and on-line learning. We prove the first generalization bounds for a hypothesis derived by online-to-batch conversion of the sequence of hypotheses output by an online algorithm, in the general setting of a non-stationary non-mixing stochastic process. Our learning guarantees hold for adapted sequences of hypotheses both for convex and non-convex losses. We further give generalization bounds for sequences of hypotheses that may not be adapted but that admit a stability property. Our learning bounds are given in terms of a discrepancy measure, which we show can be accurately estimated from data under a mild assumption. Our theory enables us to devise a principled solution for the notoriously difficult problem of model section in the time series scenario. It also helps us devise new ensemble methods with favorable theoretical guarantees for forecasting non-stationary time series.

Keywords: time series prediction, on-line learning, generalization bounds, regret minimization, validation, model selection, ensembles, stability, non-stationary, non-mixing.

1. Introduction

Time series appear in a variety of key real-world applications such as signal processing, including audio and video processing; the analysis of natural phenomena such as local weather, global temperature, and earthquakes; the study of economic or financial variables such as stock values, sales amounts, energy demand; and many other similar areas. One of the central problems related to time series analysis is that of forecasting, that is that of predicting the value Y_{T+1}, given past observations $Y_1, \ldots, Y_T$.

Two distinct learning scenarios have been adopted in the past to study the problem of sequential prediction, each leading to a different family of theoretical and algorithmic studies: the statistical learning and the on-line learning scenarios.

The statistical learning scenario assumes that the observations are drawn from some unknown distribution. Within this scenario, early methods to derive a theoretical analysis of the problem have focused on autoregressive generative models, such as the celebrated ARIMA model (Box and Jenkins, 1990) and its many variants. These methods typically require strong distributional assumptions and their guarantees are often only asymptotic. Many of the recent efforts in learning theory focus on generalizing the classical analysis and learning bounds of the i.i.d. setting to scenarios with less restrictive distributional assumptions. *Drifting* or *tracking* scenarios extend the classical setting to non-stationary sequences of independent random variables (Ben-David et al., 1989; Bartlett, 1992; Barve and Long, 1997; Even-Dar et al., 2010; Mohri and Muñoz Medina, 2012). The scenario of learning with dependent variables is another extension of the standard i.i.d. scenario that has been the subject of several recent publications (Yu, 1994; Vidyasagar, 1997; Berti and Rigo, 1997; Modha and Masry, 1998; Meir, 2000; Steinwart and Christmann, 2009; Mohri and Rostamizadeh, 2009; Alquier and Wintenberger, 2010; Pestov, 2010; Mohri and Rostamizadeh, 2010; Shalizi and Kontorovitch, 2013; Alquier et al., 2014; Kuznetsov and Mohri, 2014). In most of this past literature, the underlying stochastic process is assumed to be stationary and mixing. To the best of our knowledge, the only exception is the recent work of Kuznetsov and Mohri (2015), who analyzed the general non-stationary and non-mixing scenario and gave high-probability generalization bounds for this framework.

The on-line learning scenario requires no distributional assumption. In on-line learning, the sequence is revealed one observation at a time and it is often assumed to be generated in an adversarial fashion. The goal of the learner in this scenario is to achieve a regret, that is the difference between the cumulative loss suffered and that of the best expert in hindsight, that grows sub-linearly with time. There is a large body of literature devoted to the study of such problems and the design of algorithms for different variants of this general scenario (Cesa-Bianchi and Lugosi, 2006).

Can we leverage the theory and algorithms developed for these two distinct scenarios to design more accurate solutions for time series prediction? Can we derive generalization guarantees for a hypothesis derived by application of an online-to-batch conversion technique to the sequence of hypotheses output by an online algorithm, in the general setting of a non-stationary non-mixing process? What other benefits can such combinations of the statistical learning and on-line learning tools offer? This paper precisely addresses several of these questions. We present a series of theoretical and algorithmic results combining the benefits of the statistical learning approach to time series prediction with that of on-line learning.

We prove generalization guarantees for predictors derived from regret minimization algorithms in the general scenario of non-stationary non-mixing processes. We are not aware of any prior work that provides a connection between regret minimization and generalization in this general setting. Our results are expressed in terms of a generalization of the discrepancy measure that was introduced in (Kuznetsov and Mohri, 2015), which can be viewed as a natural measure of the degree of non-stationarity of a stochastic process taking into account the loss function and the hypothesis set used. We show that, under some additional mild assumptions, this discrepancy measure can be estimated from data, thereby resulting in fully data-dependent learning guarantees. Our results generalize the previous work of Littlestone (1989) and Cesa-Bianchi et al. (2004) who designed on-line-to-batch conversion techniques that use the hypotheses returned by a regret minimization algorithm to derive a predictor benefitting from generalization guarantees in the setting of i.i.d. samples. Agarwal and Duchi (2013) extended these results to the setting of asymptotically stationary mixing processes for stable on-line learning algorithms under some additional assumptions on the regularity of both the distribution and the loss function. However, stationarity and mixing assumptions often do not hold for the task of time series prediction. For instance, processes that admit a trend or a seasonal component are not stationary. Markov chains are not stationary unless started with an equilibrium distribution. Stochastic processes exhibiting long memory effects, such as fractional Brownian motion, are often slowly or not mixing.

We also highlight the application of our results to two related problems: model selection in time series prediction, and the design of accurate ensembles of time series predictors. Model selection for time series prediction appears to be a difficult task: in contrast with the familiar i.i.d. scenario, in time series analysis, there is no straightforward method for splitting a sample into a training and validation sets. Using the most recent data for validation may result in models that ignore the most recent information. Validating over the most distant past may lead to selecting sub-optimal parameters. Any other split of the sample may result in the destruction of important statistical patterns and correlations across time that may be present in the data. We show that, remarkably, our on-line-to-batch conversions enable us to use the same time series for both training and model selection.

Next, we show that our theory can guide the design of algorithms for learning ensembles of time series predictors via on-line-to-batch conversions. One benefit of this approach is that a battery of existing on-line learning algorithms can be used including those specifically designed for time series prediction (Anava et al., 2013, 2015; Koolen et al., 2015) as well as others capable of dealing with non-stationary sequences (Bousquet and Warmuth, 2001; Cesa-Bianchi et al., 2012; Chaudhuri et al., 2010; Crammer et al., 2010; Herbster and Warmuth, 1998, 2001; Moroshko and Crammer, 2012, 2013; Moroshko et al., 2015).

The rest of this paper is organized as follows. In Section 2, we describe the learning scenario of time series prediction that we consider and introduce some key notation, definitions, and concepts, including a discrepancy measure that plays a central role in our analysis. Section 3 presents our on-line-to-batch conversion techniques and a series of generalization guarantees. We prove the first generalization bounds for a hypothesis derived by online-to-batch conversion of the sequence of hypotheses output by an online algorithm, in the general setting of a non-stationary non-mixing stochastic process. Our learning guarantees hold for adapted sequences of hypotheses both for convex and non-convex losses. We also give generalization bounds for sequences of hypotheses that may not be adapted but that admit a stability property. Our learning bounds are given in terms of the discrepancy measure introduced, which we show can be accurately estimated from data under a mild assumption. In Section 4, we show how our theory can be used to derive principled solutions for model selection and for ensemble learning in the setting of non-stationary non-mixing time series.

2. Learning scenario

Here, we describe the learning scenario we consider and introduce some preliminary definitions and tools.

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. We consider a scenario where the learner receives a sequence $(X_1, Y_1), \ldots, (X_T, Y_T)$ that is the realization of some stochastic process, with $Z_t = (X_t, Y_t) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Note that, quite often in time series prediction, the feature vector at that time t is the collection of the past d output observations, that is $X_t = (Y_{t-1}, \ldots, Y_{t-d})$ for some d. To simplify the notation, we will use the shorthand $\mathbf{z}_a^b$ to denote a sequence $z_a, z_{a+1}, \ldots, z_b$.

The goal of the learner is to select, out of a specified family H, a hypothesis $h\colon \mathcal{X} \to \mathcal{Y}$ that achieves a small *path-dependent* generalization error

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) = \mathbb{E}\left[L(h(X_{T+1}), Y_{T+1}) \mid \mathbf{Z}_1^T\right], \tag{1}$$

where $L\colon \mathcal{Y} \times \mathcal{Y} \to [0, M]$ is a loss function bounded by $M > 0$. To abbreviate the notation, we will often write $L(h, z) = L(h(x), y)$, for any $z = (x, y) \in \mathcal{Z}$. Note that another performance measure commonly used in the time series prediction literature is the *averaged* generalization error $\mathbb{E}[\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T)]$. The path-dependent error that we consider is a finer measure of the generalization ability than the averaged error since it takes into consideration the specific history realized, unlike the averaged error, which is based on all possible trajectories of the stochastic process. Our results can be easily extended to hold for non-integer times t and arbitrary prediction lag $l \geq 0$.

A related learning scenario is that of on-line learning where time series are revealed to the learner one observation at a time and where the goal of the on-line learner is to minimize *regret* after T rounds. In this work, we consider the following general notion of regret for an on-line algorithm $\mathcal{A}$ playing a sequence of hypotheses $\mathbf{h} = (h_1, \ldots, h_T)$

$$\mathrm{Reg}_T = \sum_{t=1}^{T} L(h_t, Z_t) - \inf_{\mathbf{h}^*} \left\{ \sum_{t=1}^{T} L_t(\mathbf{h}^*, Z_t) + \mathcal{R}(\mathbf{h}^*) \right\}, \tag{2}$$

where the infimum is taken over sequences in a (possibly random) subset $H^* \subseteq H^T$ and where $\mathcal{R}$ is a regularization term that controls the complexity of the competitor class H^*. This notion of regret generalizes a number of other existing definitions. For instance, taking $\mathcal{R} = 0$ and H^* to be the set of constant sequences recovers the standard notion of regret. The *dynamic competitor* or *expert tracking* setting correspond to $\mathcal{R} = 0$ and $H^* = H^T$. If we let $H^* = H^T$ with $H \subseteq \mathbb{R}^n$ and $\mathcal{R}(\mathbf{h}) = \lambda_T \sum_{s,t=1}^T h_s \mathbf{K}_{s,t} h_t$ where $\mathbf{K}$ is a positive definite matrix and $\lambda_T \geq 0$ is a regularization parameter, then we retrieve the notion of regret studied in (Herbster and Warmuth, 2001; Koolen et al., 2015). Alternatively, we may require $H^* = \{\mathbf{h} = (h_1, \ldots, h_T) \in H^T \colon \sum_{s,t=1}^T h_s \mathbf{K}_{s,t} h_t \leq C\}$ for some $C > 0$ and set $\lambda_T = 0$. More generally, let $c \in \mathcal{C}$ be a (possibly random) constraints function, then we can define $H^* = \{\mathbf{h} \in H^T \colon c(h_1(X_1), \ldots, h_T(X_T)) < C\}$, which is a generalization of a data-dependent competitor set studied by Rakhlin and Sridharan (2015).

In this work, we give learning guarantees for regret minimization algorithms for forecasting non-stationary non-mixing time series. The key technical tool that we will need for our analysis is the discrepancy measure that quantifies the divergence of the target and sample distributions defined by

$$\mathrm{disc}(\mathbf{q}) = \sup_{\mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big(\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \Big) \right|, \tag{3}$$

where $\mathbf{q} = (q_1, \ldots, q_T)$ is an arbitrary weight vector and where $H_{\mathcal{A}}$ is a set of sequences of hypotheses that the on-line algorithm $\mathcal{A}$ can pick. One possible choice is, for instance, $H_{\mathcal{A}} = \{\mathbf{h} \colon \mathcal{R}(\mathbf{h}) \leq C\}$ for some C. In fact, $H_{\mathcal{A}}$ can be even chosen in data-dependent fashion so long as each $\mathbf{h} \in H_{\mathcal{A}}$ is adapted to the filtration of $\mathbf{Z}_1^T$. The notion of discrepancy that we consider here is a generalization of the definition given by Kuznetsov and Mohri (2015) where the supremum is taken over constant sequences $\mathbf{h}$. The crucial property of the discrepancy considered in that work is that it can be estimated from data. Our generalization also admits this property, as well as a number of other favorable ones. In particular, it can be bounded in terms of other familiar divergences between distributions such as total variation and relative entropy. We provide further details in Appendix B.

3. Theory

In this section, we prove generalization bounds for a hypothesis derived by application of some online-to-batch conversion techniques to the sequence of hypotheses output by an online algorithm, in the general setting of a non-stationary non-mixing stochastic process. We first present learning guarantees for adapted sequences of hypotheses both for convex and non-convex losses (Section 3.1). Next, in Section 3.2, we present generalization bounds for sequences of hypotheses that may not be adapted but that admit a stability property. Our learning bounds are given in terms of the discrepancy measure defined by (3). In Section 3.3, we show that, under a mild assumption, the discrepancy term can be accurately estimated from data.

3.1. Adapted hypothesis sequences

Here, we consider sequences of hypotheses $\mathbf{h} = (h_1, \ldots, h_T)$ adapted to the filtration of $\mathbf{Z}_1^T$, that is, such that h_t is $\mathbf{Z}_1^t$-measurable. This is a natural assumption since the hypothesis output by an on-line algorithm at time t is based on data observed up to time t.

The proof techniques for the first results of this section can be viewed as extensions of those used by Mohri and Muñoz Medina (2012) for the analysis of drifting. The next result is a key lemma used in the proof of our generalization bounds.

Lemma 1 *Let $\mathbf{Z}_1^T$ be any sequence of random variables and let $\mathbf{h} = (h_1, \ldots .h_T)$ be any sequence of hypotheses adapted to the filtration of $\mathbf{Z}_1^T$. Let $\mathbf{q} = (q_1, \ldots, q_T)$ be any weight vector. For any $\delta > 0$, each of the following inequalities holds with probability at least $1 - \delta$:*

$$\sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \mathrm{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}},$$

$$\sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + \mathrm{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

Proof By definition of the path-dependent error, for any $t \in [T]$, $A_t = q_t\big(\mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) - L(h_t, Z_{t+1})\big)$ is a martingale difference:

$$\mathbb{E}\left[A_t | \mathbf{Z}_1^T\right] = q_t\Big(\mathbb{E}\left[L(h_t, Z_{t+1}) | \mathbf{Z}_1^t\right] - \mathbb{E}\left[L(h_t, Z_{t+1}) | \mathbf{Z}_1^t\right]\Big) = 0,$$

since $\mathbf{h}$ is an adapted sequence. Furthermore, since $|A_t| \leq M|q_t|$, by Azuma's inequality, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$\sum_{t=1}^{T} q_t \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

The proof of the first statement can be completed by observing that, by definition of the discrepancy, we can write

$$\sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) + \operatorname{disc}(\mathbf{q}).$$

The second statement follows by symmetry. ■

The next theorem is our main generalization guarantee for on-line-to-batch conversion with bounded convex loss functions.

Theorem 2 *Assume that L is convex and bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let H^* be a set of sequences of hypotheses that are adapted to $\mathbf{Z}_1^T$ and let $h_1, \ldots, h_T$ be a sequence of hypotheses adapted to $\mathbf{Z}_1^T$. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex and let h denote $h = \sum_{t=1}^{T} q_t h_t$. Then, for any $\delta > 0$, each of the following bounds holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \operatorname{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}},$$

$$\begin{aligned}\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \inf_{\mathbf{h}^* \in H^*} &\left\{ \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\} \\ &+ 2 \operatorname{disc}(\mathbf{q}) + \operatorname{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}},\end{aligned}$$

where $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$.

Proof By the convexity of L, we can write $\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T)$. In view of that, by Lemma 1, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \operatorname{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

This proves the first statement of the theorem. To prove the second statement, first observe that, since L is bounded by M, for any $\mathbf{h}^* \in H^*$, the following holds:

$$\begin{aligned}&\sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) - \sum_{t=1}^{T} q_t L(h_t^*, Z_{t+1}) - \mathcal{R}(\mathbf{h}^*) \\ &\leq \sum_{t=1}^{T} \Big(q_t - \frac{1}{T}\Big)(L(h_t, Z_{t+1}) - L(h_t^*, Z_{t+1})) + \frac{1}{T} \sum_{t=1}^{T} (L(h_t, Z_{t+1}) - L(h_t^*, Z_{t+1})) - \mathcal{R}(\mathbf{h}^*) \\ &\leq M\|\mathbf{q} - \mathbf{u}\|_1 + \operatorname{Reg}_T.\end{aligned}$$

Thus, in view of the first statement, for any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$, the following inequalities hold:

$$\begin{aligned}\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \operatorname{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2\sqrt{2\log\frac{2}{\delta}}\\ &\leq \sum_{t=1}^{T} q_t L(h_t^*, Z_{t+1}) + \mathcal{R}(\mathbf{h}^*) + M\|\mathbf{q}-\mathbf{u}\|_1 + \operatorname{Reg}_T + \operatorname{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2\sqrt{2\log\frac{2}{\delta}}.\end{aligned}$$

Now, fix $\epsilon > 0$, and choose $\mathbf{h}^*$ such that

$$\sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \leq \inf_{\mathbf{h}^* \in H^*}\left\{\sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*)\right\} + \epsilon.$$

Using Lemma 1 to bound $\sum_{t=1}^{T} q_t L(h_t^*, Z_{t+1})$ shows that the following inequalities hold with probability at least $1 - \delta$:

$$\begin{aligned}\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) + M\|\mathbf{q}-\mathbf{u}\|_1 + \operatorname{Reg}_T + 2\operatorname{disc}(\mathbf{q})\\ &\quad + 2M\|\mathbf{q}\|_2\sqrt{2\log\frac{2}{\delta}}\\ &\leq \inf_{\mathbf{h}^* \in H^*}\left\{\sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*)\right\} + M\|\mathbf{q}-\mathbf{u}\|_1 + \operatorname{Reg}_T + 2\operatorname{disc}(\mathbf{q})\\ &\quad + 2M\|\mathbf{q}\|_2\sqrt{2\log\frac{2}{\delta}} + \epsilon.\end{aligned}$$

The result follows since this last inequality holds for any $\epsilon > 0$. ■

Theorem 2 establishes an important connection between sequential prediction in the on-line learning framework and time series prediction in the batch setting. In particular, it provides the first generalization bounds for hypotheses obtained by online-to-batch conversion from the output of an regret minimization algorithm in the general setting of time series prediction with non-stationary and non-mixing processes.

These results admit the same flavor as the uniform convergence guarantees of Kuznetsov and Mohri (2015) for forecasting non-stationary time series, which are also expressed in terms of the discrepancy measure. The presence of the discrepancy term in the bound highlights the challenges faced by the learner in this scenario. It suggests that learning is more difficult when the discrepancy term is larger, that is when the time series admits a higher degree of non-stationarity. Note that our proofs are simpler than those presented by Kuznetsov and Mohri (2015) and do not require advanced tools such as sequential complexity measures (Rakhlin et al., 2015b).

When $\mathbf{Z}_1^T$ is an i.i.d. sequence, Theorem 2 recovers the on-line-to-batch guarantees of Cesa-Bianchi et al. (2004), though our results are more general since we adopted a more inclusive notion of regret in (2). Similarly, in the special case of a drifting scenario where $\mathbf{Z}_1^T$ is a sequence of independent random variables, our results coincide with those of Mohri and Muñoz Medina (2012), modulo the extra generality of our definition of regret.

Theorem 2 can be extended to hold for general non-convex bounded functions, which can be useful for problems such as time series classification and anomaly detection where the natural loss is the zero-one loss. The ensemble hypothesis in this case is defined to be

$$h = \operatorname*{argmin}_{h_t} \left(\sum_{s=t+1}^{T-1} q_s L(h_t, Z_s) + \mathrm{pen}(t, \delta) \right), \tag{4}$$

where the penalty term pen is defined by $\mathrm{pen}(t,\delta) = \mathrm{disc}(\mathbf{q}_t^{T-1}) + M\|\mathbf{q}_t^{T-1}\|_2\sqrt{2\log\frac{2(T+1)}{\delta}}$, with $\delta > 0$ the desired confidence parameter. Note that this online-to-batch conversion technique can be useful even with convex losses or whenever convex combinations of elements of $\mathcal{Y}$ are not well-defined, for example for strings or graphs, or when the goal is to choose a single hypothesis, as in the validation setting.

Theorem 3 *Assume that L is bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let H^* be a set of sequences of hypotheses that are adapted to $\mathbf{Z}_1^T$. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex. For any $\delta > 0$, let h be defined by (4), then, each of the following bounds holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \mathrm{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2\sqrt{2\log\frac{2(T+1)}{\delta}},$$

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \inf_{\mathbf{h}^* \in H^*} \left\{ \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\}$$
$$+ 2\,\mathrm{disc}(\mathbf{q}) + \mathrm{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2\sqrt{2\log\frac{2(T+1)}{\delta}},$$

where $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$.

3.2. Stable hypothesis sequences

One of our target applications is model selection for time series prediction. The objective is then to select a good model given N models, each trained on the full sample $\mathbf{Z}_1^T$. One way to do that is to run an online learning algorithm with these models used as experts and then use one of the online-to-batch conversion techniques discussed in the previous section. However, the guarantees presented in the previous section do not hold in this scenario since $\mathbf{h}$ is no longer an adapted sequence.

In this section, we extend our guarantees to this scenario assuming that the algorithms that were used to generate the models were uniformly stable. Let $\mathfrak{A}$ be a collection of a learning algorithms, where each learning algorithm $\mathcal{A} \in \mathfrak{A}$ is defined as a rule that takes as input a sequence $\mathbf{z}_1^T \in \mathcal{Z}^T$ and outputs a hypothesis $\mathcal{A}(\mathbf{z}_1^T) \in H$. In other words, each learning algorithm is a map $\mathcal{A} \colon \mathcal{Z}^T \to H$.

An algorithm $\mathcal{A} \in \mathfrak{A}$ is β-uniformly stable if there exists $\beta > 0$ such that for any $z = (x, y) \in \mathcal{Z}$ and for any two samples $\mathbf{z}$ and $\mathbf{z}'$ that differ by exactly one point, the following condition holds:

$$\big|L(\mathcal{A}(\mathbf{z})(x), y) - L(\mathcal{A}(\mathbf{z}')(x), y)\big| \leq \beta. \tag{5}$$

In what follows, we assume that $\mathfrak{A}$ is a collection of uniformly stable algorithms. This condition is not a strong limitation since many existing learning algorithms have been shown to be stable (Bousquet and Elisseeff, 2002; Mohri and Rostamizadeh, 2010).

Given a sample $\mathbf{Z}_1^T$, we define a set of *stable hypotheses* as follows:

$$\mathcal{H} = \{h \in H\colon \text{ there exists } \mathcal{A} \in \mathfrak{A} \text{ such that } h = \mathcal{A}(\mathbf{Z}_1^T)\}.$$

For each $h = \mathcal{A}(\mathbf{Z}_1^T) \in \mathcal{H}$, we let β_h denote a uniform stability coefficient of $\mathcal{A}$, which is defined by (5). We will also use a shorthand notation β_t to denote β_{h_t}.

The following is an analogue of Lemma 1 in this setting.

Lemma 4 *Let $\mathbf{Z}_1^T$ be any sequence of random variables and let $\mathbf{h} = (h_1, \ldots, h_T) \in \mathcal{H}^T$ be any sequence of stable hypotheses. Let $\mathbf{q} = (q_1, \ldots, q_T)$ be any weight vector. For any $\delta > 0$, each of the following inequalities holds with probability at least $1 - \delta$:*

$$\sum_{t=1}^T q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t L(h_t, Z_{t+1}) + \sum_{t=1}^T q_t \beta_t + \operatorname{disc}(\mathbf{q}) + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}},$$
$$\sum_{t=1}^T q_t L(h_t, Z_{t+1}) \leq \sum_{t=1}^T q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + \sum_{t=1}^T q_t \beta_t + \operatorname{disc}(\mathbf{q}) + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}}.$$

The proof of Lemma 4 is given in Appendix C. This result enables us to extend Theorem 2 and Theorem 3 to the setting of this section.

Theorem 5 *Assume that L is convex and bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let $\mathcal{H}^*$ be a set of sequences of stable hypotheses and let $h_1, \ldots, h_T$ be a sequence of stable hypotheses. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex and let h denote $h = \sum_{t=1}^T q_t h_t$. Then, for any $\delta > 0$, each of the following bounds holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \leq \sum_{t=1}^T q_t L(h_t, Z_{t+1}) + \sum_{t=1}^T q_t \beta_t + \operatorname{disc}(\mathbf{q}) + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}},$$
$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \inf_{\mathbf{h}^* \in \mathcal{H}^*} \left\{ \sum_{t=1}^T q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\} + 2\beta_{\max} + 2\operatorname{disc}(\mathbf{q})$$
$$+ \operatorname{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 4M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}},$$

where $\beta_{\max} = \sup_{h \in \mathcal{H}} \beta_h$ and $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$.

Theorem 6 *Assume that L is bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let $\mathcal{H}^*$ be a set of sequences of stable hypotheses. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex. For any $\delta > 0$, let h be defined by (4), then, each of the following bounds holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \beta_{\max} + \operatorname{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}},$$

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \inf_{\mathbf{h}^* \in \mathcal{H}^*} \left\{ \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\} + 2\beta_{\max}$$

$$+ 2 \operatorname{disc}(\mathbf{q}) + \operatorname{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}},$$

where $\beta_{\max} = \sup_{h \in \mathcal{H}} \beta_h$ and $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$.

3.3. Discrepancy estimation

Our generalization guarantees in Theorem 2, Theorem 3, Theorem 5 and Theorem 6 critically depend on the discrepancy $\operatorname{disc}(\mathbf{q})$. In this section, under some additional mild assumptions, we show that the discrepancy measure admits upper bounds that can be estimated from the input sample.

The challenge in estimating the discrepancy $\operatorname{disc}(\mathbf{q})$ is that it depends on the distribution of Z_{T+1} which we never observe. Our discrepancy measure is the generalization of the discrepancy considered by Kuznetsov and Mohri (2015), thus, a similar approach can be used for its estimation. In particular, we can assume that the distribution of Z_t conditioned on the past history changes slowly with time. Under that assumption, the last s observations $\mathbf{Z}_{T-s+1}^T$ serve as a reasonable proxy for Z_{T+1}. More precisely, we can write

$$\operatorname{disc}(\mathbf{q}) \leq \sup_{\mathbf{h} \in H_{\mathcal{A}}} \left| \frac{1}{s} \sum_{\tau=T-s+1}^{T} \sum_{t=1}^{T} q_t \mathcal{L}_\tau(h_t, \mathbf{Z}_1^{\tau-1}) - \sum_{t=1}^{T} q_t \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^{t-1}) \right|$$

$$+ \sup_{\mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \frac{1}{s} \sum_{\tau=T-s+1}^{T} \sum_{t=1}^{T} q_t \mathcal{L}_\tau(h_t, \mathbf{Z}_1^{\tau-1}) \right|.$$

The first term can be estimated from data as we show in Lemma 9 in Appendix C. For this bound to be meaningful, we need that the second term in the above equation be sufficiently small, which is in fact a necessary condition for learning, even in the space case of a drifting scenario, as shown by Barve and Long (1997). However, the main disadvantage of this approach is that it relies on a parameter s and it is not clear how this parameter can be chosen in a principled way.

Instead, we propose an alternative approach that is based on the implicit assumption that H contains a hypothesis h^* that admits a small path-dependent generalization error, whose prediction can be used as a proxy for Z_{T+1}. We will also assume that the loss function is Lipschitz, that is, $|L(y, y') - L(y, y'')| \leq C\, d(y', y'')$, where $C > 0$ is a constant and d the underlying metric. This assumption holds for a broad family of regression losses commonly used in time series prediction. In fact, d

often coincides with L, for example, $L = d$ when L is the L_2-norm. Observe that the following holds for any h^*:

$$
\begin{aligned}
\operatorname{disc}(\mathbf{q}) \leq & \sup_{\mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \left(\mathbb{E}\left[L(h_t(X_{T+1}), h^*(X_{T+1})) \middle| \mathbf{Z}_1^T \right] - \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \right) \right| \\
& + \sup_{\mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \left(\mathbb{E}\left[L(h_t(X_{T+1}), h^*(X_{T+1})) \middle| \mathbf{Z}_1^T \right] - \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) \right) \right| \\
\leq & \sup_{\mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \left(L(h_t(X_{T+1}), h^*(X_{T+1})) - \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \right) \right| \\
& + \sup_{\mathbf{h} \in H_{\mathcal{A}}} \sum_{t=1}^{T} q_t \, \mathbb{E}\left[d(Z_{T+1}, h^*(X_{T+1})) \middle| \mathbf{Z}_1^T \right] \\
\leq & \sup_{h \in H, \mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \left(L(h_t(X_{T+1}), h(X_{T+1})) - \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \right) \right| + \Delta
\end{aligned} \tag{6}
$$

where $\Delta = \mathbb{E}\left[d(Z_{T+1}, h^*(X_{T+1})) \middle| \mathbf{Z}_1^T \right]$ and we may choose h^* to be the hypothesis that achieves $\inf_{h^*} \mathbb{E}[d(Z_{T+1}, h^*(X_{T+1})) | \mathbf{Z}_1^T]$. The first term in the above bound, that we denote by $\operatorname{disc}_H(\mathbf{q})$ can be estimated from the data as the next lemma shows. The guarantees that we present are in terms of the *expected sequential covering numbers* $\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T}[\mathcal{N}_1(\alpha, \mathcal{F}, \mathbf{z})]$ of the set $\mathcal{F} = \{(z, t) \mapsto L(h_t, z) \colon \mathbf{h} \in H_{\mathcal{A}}\}$ which are natural generalizations to the sequential setting of the standard expected covering numbers (Rakhlin et al., 2015b). Here, $\mathbf{z}$ is a $\mathcal{Z}$-valued complete binary tree of depth T and $\mathcal{Z}_T$ the distribution induced over such trees (see Appendix A). A similar guarantee can be given in terms of sequential Rademacher complexity. A brief review of sequential complexity measures can be found in Appendix A.

Lemma 7 *Let $\mathbf{Z}_1^T$ be a sequence of random variables. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $\alpha > 0$:*

$$
\operatorname{disc}(\mathbf{q}) \leq \widehat{\operatorname{disc}}_H(\mathbf{q}) + \inf_{h^*} \mathbb{E}[d(Z_{T+1}, h^*(X_{T+1})) | \mathbf{Z}_1^T] + 2\alpha + M \|\mathbf{q}\|_2 \sqrt{2 \log \frac{\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},
$$

where $\widehat{\operatorname{disc}}_H(\mathbf{q})$ is the empirical discrepancy defined by

$$
\widehat{\operatorname{disc}}_H(\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \left(L(h_t(X_{T+1}), h(X_{T+1})) - L(h_t, Z_{t+1}) \right) \right|. \tag{7}
$$

The proof of this result can be found in Appendix C. In Section 4, we will show that that there are efficient algorithms for the computation and optimization of this discrepancy $\widehat{\operatorname{disc}}_H(\mathbf{q})$. We can further improve the computational cost by making a stronger implicit assumption that there exists a single h^* that has a small generalization error at each time step t. This assumption is closely related to regret guarantees: the existence of such a hypothesis h^* implies that minimizing regret against such a competitor gives a meaningful learning guarantee. Using the same arguments as in the proof of Lemma 7, one can show that, with high probability, $\mathrm{disc}(\mathbf{q})$ is bounded by $\widehat{\mathrm{disc}}_{H^T}(\mathbf{q}) + \inf_{h^* \in H} \sum_{t=1}^{T} q_t \, \mathbb{E}[d(Z_t, h^*(X_t)) | \mathbf{Z}_1^{t-1}]$ plus a complexity penalty, where

$$\widehat{\mathrm{disc}}_{H^T}(\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in H^T} \left| \sum_{t=1}^{T} q_t \big(L(h_t(X_{T+1}), h(X_{T+1})) - L(h_t(X_{t+1}), h(X_{t+1})) \big) \right|. \quad (8)$$

4. Applications

In this section we present the applications of our theory to two problems: model selection and the design of ensemble solutions in time series prediction.

4.1. Model Selection

The first application that we consider is that of model selection for time series prediction. In this setting, we are given N models that have been trained on the given sample $\mathbf{Z}_1^T$ out of which we wish to select a single best model. In the i.i.d. setting, this problem is typically addressed via cross-validation: part of the sample $\mathbf{Z}_1^T$ is reserved for training, the rest used as a validation set. In that setting, high-probability performance guarantees can be given for this procedure that are close to but weaker than those that hold for structural risk minimization (SRM) (Mohri et al., 2012). Unfortunately, in the setting of time series prediction, it is not immediately clear how this can be accomplished in a principled way. As already mentioned in Section 1, using the most recent data for validation may result in models that ignore the most recent information. Validating over the most distant past may lead to selecting sub-optimal parameters. Any other split of the sample may result in the destruction of important statistical patterns and correlations across time that may be present in the data.

Is it possible to come up with a principled solution for model selection in our general scenario? Our Theorem 6 helps derive a positive response to this question and design an algorithm for this problem. Note that Theorem 6 suggests that, if we could select a distribution $\mathbf{q}$ over the sample $\mathbf{Z}_1^T$ that would minimize the discrepancy $\mathrm{disc}(\mathbf{q})$ and use it to weight training points, then we would have a better learning guarantee for a hypothesis h obtained via on-line-to-batch conversion defined by (4). This leads to the following algorithm:

- choose a weight vector $\mathbf{q}$ in the simplex that minimizes the empirical discrepancy, that is, choose $\mathbf{q}$ as the solution of one of the following two optimization problems: $\min_{\mathbf{q}} \widehat{\mathrm{disc}}_H(\mathbf{q})$ or $\min_{\mathbf{q}} \widehat{\mathrm{disc}}_{H_\mathcal{A}}(\mathbf{q})$. In several important special cases, these problems can be solved efficiently as discussed later.
- use any on-line learning algorithm for prediction with expert advice to generate a sequence of hypotheses $\mathbf{h} \in \mathcal{H}^T$, where $\mathcal{H}$ is a set of N models trained on $\mathbf{Z}_1^T$. Select a single model h according to (4).

Observe that, by definition, the discrepancy is a convex function of $\mathbf{q}$ since the maximum of a set of convex functions is convex and since the composition of a convex function with an affine function is convex. Thus, both $\min_{\mathbf{q}} \widehat{\mathrm{disc}}_H(\mathbf{q})$ and $\min_{\mathbf{q}} \widehat{\mathrm{disc}}_{H_\mathcal{A}}(\mathbf{q})$ are convex optimization problems. Since both objectives admit a subgradient, a standard subgradient descent algorithm can be applied to solve both of these problems. The bottleneck of this procedure, however, is the computation of the gradient which requires solving an optimization problem in the definition of the discrepancy at each iteration of the subgradient procedure.

For convex loss functions and a convex hypothesis set H, this optimization problem can be cast as a difference of convex functions problem (DC-programming problem) which can be solved, for instance, using the DC-algorithm of Tao and An (1998). Furthermore, this algorithm can be shown to be globally optimal in the case of the squared loss with a set of linear hypothesis H, which is the standard setting in time series prediction.

In special case of the squared loss with linear hypotheses $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \colon \|\mathbf{w}\|_2 \leq 1\}$, optimization problem $\min_{\mathbf{q}} \widehat{\mathrm{disc}}_{H_{\mathcal{A}}}(\mathbf{q})$ admits additional structure that can lead to more efficient solutions. Indeed, in that case the objective can be rewritten as follows:

$$\begin{aligned}
\widehat{\mathrm{disc}}_{H_{\mathcal{A}}}(\mathbf{q}) &= \max_{\|\mathbf{w}_t\|_2 \leq 1, \|\mathbf{w}\|_2 \leq 1} \left| \sum_{t=1}^{T} q_t \big\| \mathbf{w}_t \cdot \mathbf{x}_{t+1} - \mathbf{w} \cdot \mathbf{x}_{t+1} \big\|_2^2 - q_t \big\| \mathbf{w}_t \cdot \mathbf{x}_{T+1} - \mathbf{w} \cdot \mathbf{x}_{T+1} \big\|_2^2 \right| \\
&= \max_{\|\mathbf{w}_t\|_2 \leq 1, \|\mathbf{w}\|_2 \leq 1} \left| \sum_{t=1}^{T} (\mathbf{w}_t - \mathbf{w})^\top q_t (\mathbf{x}_{t+1}\mathbf{x}_{t+1}^\top - \mathbf{x}_{T+1}\mathbf{x}_{T+1}^\top)(\mathbf{w}_t - \mathbf{w}) \right| \\
&= \max_{\|\mathbf{w}_t\|_2 \leq 1, \|\mathbf{w}\|_2 \leq 1} \left| (\mathbf{W} - \mathbf{W}')^\top \mathbf{M}(\mathbf{q}) (\mathbf{W} - \mathbf{W}') \right|
\end{aligned}$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_T)$ and $\mathbf{W}' = (\mathbf{w}, \ldots, \mathbf{w})$, and where $\mathbf{M}(\mathbf{q})$ is a block diagonal matrix with block matrices $q_t(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^\top - \mathbf{x}_{T+1}\mathbf{x}_{T+1}^\top)$ on the diagonal. Furthermore, observe that $\mathbf{M}(\mathbf{q})$ can be written as $\mathbf{M}(\mathbf{q}) = \sum_{t=1} q_t \mathbf{M}_t$, where each $\mathbf{M}_t$ is a block diagonal matrix with all its blocks equal to zero except from the tth one which is $q_t(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^\top - \mathbf{x}_{T+1}\mathbf{x}_{T+1}^\top)$. This leads to the following optimization problem:

$$\begin{aligned}
&\min_{\mathbf{q}} \max_{\|\mathbf{w}_t\|_2 \leq 1, \|\mathbf{w}\|_2 \leq 1} \left| (\mathbf{W} - \mathbf{W}')^\top \mathbf{M}(\mathbf{q}) (\mathbf{W} - \mathbf{W}') \right| \\
&\text{subject to} \quad \mathbf{q}^\top \mathbf{1} = 1, \quad \mathbf{q} \geq 0.
\end{aligned} \tag{9}$$

There are multiple different approaches for solving this optimization problem. Observe that this optimization problem is similar to maximum eigenvalue minimization problem:

$$\begin{aligned}
&\min_{\mathbf{q}} \max_{\|\mathbf{V}\|_2 \leq 2\sqrt{T}} \left| \mathbf{V}^\top \mathbf{M}(\mathbf{q}) \mathbf{V} \right| \\
&\text{subject to} \quad \mathbf{q}^\top \mathbf{1} = 1, \quad \mathbf{q} \geq 0.
\end{aligned}$$

In fact, for $T = 1$ these two problems coincide. Maximum eigenvalue minimization problem can be solved using the smooth approximation technique of Nesterov (2007) applied to the objective (Cortes and Mohri, 2014) or by writing it as an equivalent SDP problem which can then be solved in polynomial-time using interior-point methods and several other specific algorithms (Fletcher, 1985; Overton, 1988; Jarre, 1993; Helmberg and Oustry, 2000; Alizadeh, 1995; Cortes and Mohri, 2014). The natural approaches for solving the optimization problem (9) also include the smooth approximation technique of Nesterov (2007) or casting it as an SDP problem. Another approach is based on the relaxation of (9) to maximum eigenvalue minimization. Indeed, observe that

$$\max_{\|\mathbf{w}_t\|_2 \leq 1, \|\mathbf{w}\|_2 \leq 1} \left|(\mathbf{W} - \mathbf{W}')^\top \mathbf{M}(\mathbf{q})(\mathbf{W} - \mathbf{W}')\right| \leq \max_{\|\mathbf{V}\|_2 \leq 2\sqrt{T}} \left|\mathbf{V}^\top \mathbf{M}(\mathbf{q}) \mathbf{V}\right|,$$

where we let $\mathbf{V} = \mathbf{W} - \mathbf{W}'$ and the inequality is a consequence of the fact that, for any $\mathbf{w}, \mathbf{w}_1, \ldots, \mathbf{w}_T$ such that $\|\mathbf{w}_t\|_2 \leq 1$, $\|\mathbf{w}\|_2 \leq 1$ for all t the following bound holds:

$$\|\mathbf{V}\|_2 = \|\mathbf{W} - \mathbf{W}'\|_2 = \sqrt{\sum_{t=1}^T \|\mathbf{w}_t - \mathbf{w}\|_2^2} \leq 2\sqrt{T}.$$

We leave it to the future to give a more detailed analysis of the best specific algorithms for this problem.

In a similar way, we can derive a validation procedure that is based on Theorem 3 for a slightly different setting in which we do not choose among pre-trained models but rather among different hyperparameter vectors $\theta_1, \ldots, \theta_N$. At each round of the execution of an on-line algorithm, one hyperparameter vector θ is chosen as an expert, the corresponding h_θ is trained on data that has been observed so far and is used to make a prediction. The final hyperparameter vector is chosen according to (4).

We conclude this section with the observation that model selection procedures described above can also be applied in the i.i.d. setting in which case we can take $\mathbf{q}$ to be $\mathbf{u}$.

4.2. Ensemble Learning

In this section, we present another application of our theory which is that of learning convex ensembles of time series predictors. Given a hypothesis set H and a sample $\mathbf{Z}_1^T$, that may consist of the models that have been trained on $\mathbf{Z}_1^T$, the goal of the learner is to come up with a convex combination $\sum_{t=1}^T q_t h_t$ for some $\mathbf{h} \in H_{\mathcal{A}}$ and a $\mathbf{q}$ in the simplex. We propose the following two-step procedure:

- run a regret minimization algorithm on $\mathbf{Z}_1^T$ to generate a sequence of hypotheses $\mathbf{h}$.
- select an ensemble hypothesis $h = \sum_{t=1}^T q_t h_t$ where $\mathbf{q}$ is solution of the following convex optimization problem over the simplex:

$$\begin{aligned} \min_{\mathbf{q}} \quad & \widehat{\mathrm{disc}}_H(\mathbf{q}) + \sum_{t=1}^{T-1} q_t L(h_t, Z_{t+1}) \\ \text{subject to} \quad & \|\mathbf{q} - \mathbf{u}\|_1 \leq \lambda_1, \end{aligned} \tag{10}$$

  where $\lambda_1 \geq 0$ is some hyperparameter that can be set via a validation procedure.

Note that (10) directly optimizes the upper bound of Theorem 2. If models in H have been trained on $\mathbf{Z}_1^T$, then an additional linear term $\sum_{t=1}^T q_t \beta_t$ appears in the objective and the overall problem can be handled in exactly the same way. Similarly, $\widehat{\mathrm{disc}}_{H_{\mathcal{A}}}(\mathbf{q})$ may be used in place of $\widehat{\mathrm{disc}}_H(\mathbf{q})$.

As in Section 4.1, the convex optimization problem in (10) can be solved using a standard projected subgradient algorithm where at each iteration a DC-algorithm of Tao and An (1998) is used to compute the discrepancy, if H and L are convex. As before, this DC-algorithm is guaranteed to be optimal if L is the squared loss and H is a set of linear hypothesis. Furthermore, for linear hypotheses with the squared loss and $\widehat{\mathrm{disc}}_{H_A}(\mathbf{q})$ in the objective, the same analysis as in Section 4.1 can be used.

5. Conclusion

Time series prediction is a fundamental learning problem. We presented a series of results exploiting its recent analysis in statistical learning theory in the general scenario of non-stationary non-mixing Kuznetsov and Mohri (2015) and other existing regret-based analysis and guarantees from the broad on-line learning literature. This combination of the benefits of different approaches can lead to a variety of rich problems and solutions in learning theory that we hope this work will promote and stimulate.

Acknowledgments

This work was partly funded by NSF IIS-1117591 and CCF-1535987.

References

A. Agarwal and J.C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.

Farid Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5:13–51, 1995.

Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. Technical Report 2010-39, Centre de Recherche en Economie et Statistique, 2010.

Pierre Alquier, Xiaoyin Li, and Olivier Wintenberger. Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modelling*, 1:65–93, 2014.

Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. In *proceedings of COLT*, 2013.

Oren Anava, Elad Hazan, and Assaf Zeevi. Online time series prediction with missing data. In *proceedings of ICML*, 2015.

Peter L. Bartlett. Learning with a slowly changing distribution. In *proceedings of COLT*, pages 243–252, 1992.

Rakesh D. Barve and Phil M. Long. On the complexity of learning from drifting distributions. *Information and Computation*, 138(2):101–123, 1997.

Shai Ben-David, Gyora M. Benedek, and Yishay Mansour. A parametrization scheme for classifying models of learnability. In *Proceedings of COLT*, pages 285–302, 1989.

Patrizia Berti and Pietro Rigo. A Glivenko-Cantelli theorem for exchangeable random variables. *Statistics & Probability Letters*, 32(4):385 – 391, 1997.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet and Manfred K. Warmuth. Tracking a small set of experts by mixing past posteriors. In *proceedings of COLT*, 2001.

George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, 1990.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inf. Theory*, 50(9), 2004.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In *proceedings of NIPS*, pages 989–997, 2012.

Kamalika Chaudhuri, Yoav Freund, and Daniel Hsu. An online learning-based framework for tracking. In *UAI*, 2010.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.*, 2014.

Koby Crammer, Yishay Mansour, Eyal Even-Dar, and Jennifer Wortman Vaughan. Regret minimization with concept drift. In *proceedings of COLT*, 2010.

Eyal Even-Dar, Yishay Mansour, and Jennifer Wortman. Regret minimization with concept drift. In *Proceedings of COLT*, 2010.

R. Fletcher. On minimizing the maximum eigenvalue of a symmetric matrix. *SIAM J. Control and Opt.*, 23(4):493–513, 1985.

C. Helmberg and F. Oustry. Bundle methods to minimize the maximum eigenvalue function. In *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*. Kluwer Academic Publishers, 2000.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2), 1998.

Mark Herbster and Manfred K. Warmuth. Tracking the best linear predictor. *JMLR*, 2001.

Florian Jarre. An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices. *SIAM J. Control Optim.*, 31(5):1360–1377, 1993.

Wouter M Koolen, Alan Malek, Peter L Bartlett, and Yasin Abbasi. Minimax time series prediction. In *proceedings of NIPS*, 2015.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for time series prediction with non-stationary processes. In *proceedings of ALT*, 2014.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *proceedings of NIPS*, 2015.

Nick Littlestone. From on-line to batch learning. In *Proceedings of COLT*, pages 269–284, 1989.

Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, pages 5–34, 2000.

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

Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In *proceedings of ALT*, 2012.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *proceedings of NIPS*, 2009.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

Edward Moroshko and Koby Crammer. Weighted last-step min-max algorithm with improved sub-logarithmic regret. In *proceedings of ALT*, 2012.

Edward Moroshko and Koby Crammer. A last-step regression algorithm for non-stationary online learning. In *AISTATS*, 2013.

Edward Moroshko, Nina Vaits, and Koby Crammer. Second-order non-stationary online learning for regression. *JMLR*, 2015.

Yurii Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.*, 110:245–259, 2007.

Michael L. Overton. On minimizing the maximum eigenvalue of a symmetric matrix. *SIAM J. Matrix Anal. Appl.*, 9(2), 1988.

Vladimir Pestov. Predictive PAC learnability: A paradigm for learning from exchangeable input data. In *GRC*, 2010.

Alexander Rakhlin and Karthik Sridharan. Hierarchies of relaxations for online prediction problems with evolving constraints. In *proceedings of COLT*, 2015.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *proceedings of NIPS*, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *proceedings of NIPS*, 2011.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 2015a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *JMLR*, 16(1), January 2015b.

Cosma Shalizi and Aryeh Kontorovitch. Predictive PAC learning and process decompositions. In *proceedings of NIPS*, 2013.

Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. In *proceedings of NIPS*, 2009.

Pham Dinh Tao and Le Thi Hoai An. A D.C. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

M. Vidyasagar. *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer-Verlag New York, Inc., 1997.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

Appendix A. Sequential Complexities

The guarantees that we provide in this paper for estimating the discrepancy $\mathrm{disc}(\mathbf{q})$ are expressed in terms of data-dependent measures of sequential complexity such as expected sequential covering number or sequential Rademacher complexity (Rakhlin et al., 2010). We give a brief overview of the notion of sequential covering number and refer the reader to the aforementioned reference for further details. We adopt the following definition of a complete binary tree: a $\mathcal{Z}$-valued complete binary tree $\mathbf{z}$ is a sequence $(z_1, \ldots, z_T)$ of T mappings $z_t \colon \{\pm 1\}^{t-1} \to \mathcal{Z}$, $t \in [T]$. A path in the tree is a sequence $\sigma = (\sigma_1, \ldots, \sigma_{T-1})$, with $\sigma_1, \ldots, \sigma_{T-1} \in \{\pm 1\}$. To simplify the notation we will write $z_t(\boldsymbol{\sigma})$ instead of $z_t(\sigma_1, \ldots, \sigma_{t-1})$, even though z_t depends only on the first $t-1$ elements of $\boldsymbol{\sigma}$. The following definition generalizes the classical notion of covering numbers to the sequential setting. A set V of $\mathbb{R}$-valued trees of depth T is a *sequential α-cover* (with respect to $\mathbf{q}$-weighted ℓ_p norm) of a function class $\mathcal{G}$ on a tree $\mathbf{z}$ of depth T if for all $g \in \mathcal{G}$ and all $\boldsymbol{\sigma} \in \{\pm\}^T$, there is $\mathbf{v} \in V$ such that

$$\left(\sum_{t=1}^{T} \Big| \mathbf{v}_t(\boldsymbol{\sigma}) - g\big(\mathbf{z}_t(\boldsymbol{\sigma})\big) \Big|^p \right)^{\frac{1}{p}} \leq \|\mathbf{q}\|_q^{-1} \alpha,$$

where $\| \cdot \|_q$ is the dual norm. The *(sequential) covering number* $\mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z})$ of a function class $\mathcal{G}$ on a given tree $\mathbf{z}$ is defined to be the size of the minimal sequential cover. The *maximal covering number* is then taken to be $\mathcal{N}_p(\alpha, \mathcal{G}) = \sup_{\mathbf{z}} \mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z})$. One can check that in the case of uniform weights this definition coincides with the standard definition of sequential covering numbers. Note that this is a purely combinatorial notion of complexity which ignores the distribution of the process in the given learning problem.

Data-dependent sequential covering numbers can be defined as follows. Given a stochastic process distributed according to the distribution $\mathbf{p}$ with $\mathbf{p}_t(\cdot|\mathbf{z}_1^{t-1})$ denoting the conditional distribution at time t, we sample a $\mathcal{Z} \times \mathcal{Z}$-valued tree of depth T according to the following procedure. Draw two independent samples Z_1, Z'_1 from $\mathbf{p}_1$: in the left child of the root draw Z_2, Z'_2 according to $\mathbf{p}_2(\cdot|Z_1)$ and in the right child according to $\mathbf{p}_2(\cdot|Z'_2)$. More generally, for a node that can be reached by a path $(\sigma_1, \ldots, \sigma_t)$, we draw Z_t, Z'_t according to $\mathbf{p}_t(\cdot|S_1(\sigma_1), \ldots, S_{t-1}(\sigma_{t-1}))$, where $S_t(1) = Z_t$ and $S_t(-1) = Z'_t$. Let $\mathbf{z}$ denote the tree formed using Z_ts and define the *expected covering number* to be $\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T}\big[\mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z})\big]$, where $\mathcal{Z}_T$ denotes the distribution of $\mathbf{z}$ thereby defined.

One can define similarly other measures of complexity such as sequential Rademacher complexity and the Littlestone dimension (Rakhlin et al., 2015a) as well as their data-dependent counterparts (Rakhlin et al., 2011). One of the main technical tools used in our analysis is the notion of *sequential Rademacher complexity*. Let $\mathcal{G}$ be a set of functions from $\mathcal{Z}$ to $\mathbb{R}$. The sequential Rademacher complexity of a function class $\mathcal{Z}$ is defined as the following:

$$\mathfrak{R}_T^{\text{seq}}(\mathcal{G}) = \sup_{\mathbf{z}} \mathbb{E}\left[\sup_{g \in \mathcal{G}} \sum_{t=1}^{T} \sigma_t \, q_t \, g\big(z_t(\boldsymbol{\sigma})\big)\right], \tag{11}$$

where the supremum is taken over all complete binary trees of depth T with values in $\mathcal{Z}$ and where $\boldsymbol{\sigma}$ is a sequence of Rademacher random variables (i.i.d. uniform random variables taking values in $\{\pm 1\}$).

Appendix B. Discrepancy Measure

One of the key ingredients needed for the derivation of our learning guarantees is the notion of discrepancy between the target distribution and the distribution of the sample that was introduced in Section 2. Our discrepancy measure is a direct extension of the discrepancy measure in (Kuznetsov and Mohri, 2015) to the setting of on-line learning algorithms and it enjoys many of the same favorable properties as its precursor.

One natural interpretation of disc is as a measure of the non-stationarity of the stochastic process $\mathbf{Z}$ with respect to both the loss function L and the hypothesis set H. In particular, note that if the process $\mathbf{Z}$ is i.i.d., then we simply have $\mathrm{disc}(\mathbf{q}) = 0$ provided that q_ts form a probability distribution.

As a more interesting example, consider the case of a Markov chain on a set $\{0, \ldots, N-1\}$ such that $\mathbf{P}(X_t \equiv (i-1) \bmod N | X_{t-1} = i) = p_{i \,(\mathrm{mod}\ 2)}$ and $\mathbf{P}(X_t \equiv (i+1) \bmod N | X_{t-1} = i) = 1 - p_{i\,(\mathrm{mod}\ 2)}$ for some $0 \leq p_0, p_1 \leq 1$. In other words, this is a random walk on $\{0, \ldots, N-1\}$, with transition probability distribution changing depending on the equivalent class of the time step t. This process is non-stationary. Suppose that the set of hypotheses used by an on-line algorithm is $\{x \mapsto a(x-1) + b(x+1) \colon a + b = 1, a, b \geq 0\}$ and the loss function is defined by $L(y, y') = \ell(|y - y'|)$ for some ℓ. It follows that for any (a, b),

$$\mathcal{L}_{s+1}(h_t, \mathbf{Z}_1^s) = p_{s\,(\mathrm{mod}\ 2)} \ell(|a_t - b_t - 1|) + (1 - p_{s\,(\mathrm{mod}\ 2)}) \ell(|a_t - b_t + 1|)$$

and hence $\mathrm{disc}(\mathbf{q}) = 0$ provided that $\mathbf{q}$ is a probability distribution that is supported on odd ts if T is odd or even ts if T is even. Note that if we chose a different hypothesis set, then, in general, we may have $\mathrm{disc}(\mathbf{q}) \neq 0$. This highlights an important property of discrepancy: it takes into account not only the underlying distribution of the stochastic process but also other components of the learning problem such as the loss function and the hypothesis set used. Similar results can be established for weakly stationary stochastic process as well (Kuznetsov and Mohri, 2014).

It is also possible to give bounds on $\operatorname{disc}(\mathbf{q})$ in terms of other natural divergences between distributions. For instance, if $\mathbf{q}$ is a probability distribution, then, by Pinsker's inequality, the following holds:

$$\begin{aligned}\operatorname{disc}(\mathbf{q}) &\leq M \bigg\| \mathbf{P}_{T+1}(\cdot|\mathbf{Z}_1^T) - \sum_{t=1}^{T} q_t \mathbf{P}_t(\cdot|\mathbf{Z}_1^{t-1}) \bigg\|_{\mathrm{TV}} \\ &\leq \tfrac{1}{\sqrt{2}} D^{\frac{1}{2}} \bigg(\mathbf{P}_{T+1}(\cdot|\mathbf{Z}_1^T) \parallel \sum_{t=1}^{T} q_t \mathbf{P}_t(\cdot|\mathbf{Z}_1^{t-1}) \bigg),\end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance, $D(\cdot \parallel \cdot)$ the relative entropy, $\mathbf{P}_{t+1}(\cdot|\mathbf{Z}_1^t)$ the conditional distribution of Z_{t+1}, and $\sum_{t=1}^T q_t \mathbf{P}_t(\cdot|\mathbf{Z}_1^{t-1})$ the mixture of the sample marginals.

However, the most important property of the discrepancy $\operatorname{disc}(\mathbf{q})$ is, as shown later in Lemma 9 and Lemma 7, that it can be accurately estimated from data under some additional mild assumptions.

Appendix C. Proofs

Theorem 3 *Assume that L is bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let H^* be a set of sequences of hypotheses that are adapted to $\mathbf{Z}_1^T$. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex. For any $\delta > 0$, let h be defined by (4), then, each of the following bounds holds with probability at least $1 - \delta$:*

$$\begin{aligned}\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \operatorname{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}}, \\ \mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) &\leq \inf_{\mathbf{h}^* \in H^*} \left\{ \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\} \\ &\qquad + 2 \operatorname{disc}(\mathbf{q}) + \operatorname{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}},\end{aligned}$$

where $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$.

Proof Observe that

$$\begin{aligned}&\min_{0 \leq t \leq T} \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + 2 \operatorname{pen}(t, \delta/2) \\ &\leq \min_{0 \leq t \leq T} \frac{1}{\|\mathbf{q}_{T-t}^T\|_1} \sum_{i=t}^{T-1} q_i \mathcal{L}_{T+1}(h_i, \mathbf{Z}_1^T) + \frac{2}{\|\mathbf{q}_{T-t}^T\|_1} \operatorname{disc}\left(\mathbf{q}_{T-t}^T\right) + 2 \frac{\|\mathbf{q}_{T-t}^T\|_2}{\|\mathbf{q}_{T-t}^T\|_1} M \sqrt{\log \frac{2(T+1)}{\delta}}.\end{aligned}$$

Let M_t and K_t be defined as follows:

$$
\begin{aligned}
M_t &= \sum_{i=t}^{T-1} \widetilde{q}_i \mathcal{L}_{T+1}(h_i, \mathbf{Z}_1^T) + 2\,\mathrm{disc}\left(\widetilde{\mathbf{q}}_{T-t}^T\right) + 2\|\widetilde{\mathbf{q}}_{T-t}^T\|_2\, M\sqrt{\log\frac{2(T+1)}{\delta}}\\
K_t &= \sum_{i=t}^{T-1} \widetilde{q}_i L(h_i, Z_{i+1}) + \mathrm{disc}\left(\widetilde{\mathbf{q}}_{T-t}^T\right) + \|\widetilde{\mathbf{q}}_{T-t}^T\|_2\, M\sqrt{\log\frac{2(T+1)}{\delta}},
\end{aligned}
$$

where $\widetilde{q}_i = q_i / \|\mathbf{q}_{T-t}^T\|_1$. Then, by Lemma 1, it follows that

$$
\mathbb{P}\bigg(\min_{0\le t\le T}\Big(\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + 2\,\mathrm{pen}(t, \delta/2)\Big) \ge \min_{0\le t\le T} K_t\bigg) \le \sum_{t=1}^{T} \mathbb{P}(M_t \ge K_t) \le \frac{\delta}{2}.
$$

Combining this with Lemma 8 yields the first statement of the theorem. The second statement follows from the same arguments as in the proof of Theorem 2. ■

Lemma 8 *Let $\mathbf{Z}_1^T$ be any sequence of random variables and let $h_1, \ldots, h_T$ be any sequence of hypotheses adapted to $\mathbf{Z}_1^T$. Let $\mathbf{q} = (q_1, \ldots, q_T)$ be any weight vector in the simplex. If the loss function L is bounded and h is defined by (4), then, for any $\delta > 0$, each of the following bounds holds with probability at least $1 - \delta$:*

$$
\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \le \min_{0\le t\le T}(\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + 2\,\mathrm{pen}(t, \delta)).
$$

Proof We define

$$
\begin{aligned}
\widetilde{t} &= \operatorname*{argmin}_{0\le t\le T}(G_t + \mathrm{pen}(t, \delta))\\
\widehat{t} &= \operatorname*{argmin}_{0\le t\le T}(S_t + \mathrm{pen}(t, \delta)),
\end{aligned}
$$

where $S_t = \sum_{s=t+1}^{T} q_s L(h_t, Z_s)$ and $G_t = \sum_{s=t}^{T-1} q_s R_{s+1}(h_t, \mathbf{Z}_1^s)$. We also let $\widetilde{h} = h_{\widetilde{t}}$ and $\widetilde{S} = S_{\widetilde{t}}$. Observe that $S_{\widehat{t}} + \mathrm{pen}(\widehat{t}, \delta) \le \widetilde{S} + \mathrm{pen}(\widetilde{t}, \delta)$ and so if we let $A = \{\mathcal{L}_{T+1}(\widehat{h}, \mathbf{Z}_1^T) \ge \mathcal{L}_{T+1}(\widetilde{h}, \mathbf{Z}_1^T) + 2\,\mathrm{pen}(\widetilde{t}, \delta)\}$ then

$$
\mathbb{P}(A) = \mathbb{P}\Big(A, S_{\widehat{t}} + \mathrm{pen}(\widehat{t}, \delta) \le \widetilde{S} + \mathrm{pen}(\widetilde{t}, \delta)\Big) \le \sum_{t=0}^{T-1} \mathbb{P}\Big(A, S_t + \mathrm{pen}(t, \delta) < \widetilde{S} + \mathrm{pen}(\widetilde{t}, \delta)\Big).
$$

Note that $S_t + \mathrm{pen}(t, \delta) < \widetilde{S} + \mathrm{pen}(\widetilde{t}, \delta)$ implies that at least one of the following events must hold:

$$
\begin{aligned}
B_1 &= \{S_t \le \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathrm{pen}(t, \delta)\},\\
B_2 &= \{\widetilde{S} \ge \mathcal{L}_{T+1}(\widetilde{h}, \mathbf{Z}_1^T) + \mathrm{pen}(\widetilde{t}, \delta)\},\\
B_3 &= \{\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) < 2\,\mathrm{pen}(\widetilde{t}, \delta)\}.
\end{aligned}
$$

Therefore,

$$\mathbb{P}\Big(A, S_t + \text{pen}(t,\delta) < \widetilde{S} + \text{pen}(\widetilde{t},\delta)\Big) \leq \mathbb{P}(B_1) + \mathbb{P}(B_2) + \mathbb{P}(B_3, A) = \mathbb{P}(B_1) + \mathbb{P}(B_2)$$

since $\mathbb{P}(B_3, A) = 0$. Then it follows that

$$\mathbb{P}(A) \leq \sum_{t=1}^{T} \mathbb{P}(S_t \leq \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \text{pen}(t,\delta)) + T\sum_{t=1} \mathbb{P}(S_t \geq \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + \text{pen}(t,\delta)).$$

By the special choice of the $\text{pen}(t,\delta)$ it follows that $\mathbb{P}(A) \leq \delta$ and the proof is complete. ■

Lemma 4 *Let $\mathbf{Z}_1^T$ be any sequence of random variables and let $\mathbf{h} = (h_1, \ldots, h_T) \in \mathcal{H}^T$ be any sequence of stable hypotheses. Let $\mathbf{q} = (q_1, \ldots, q_T)$ be any weight vector. For any $\delta > 0$, each of the following inequalities holds with probability at least $1 - \delta$:*

$$\sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \sum_{t=1}^{T} q_t \beta_t + \text{disc}(\mathbf{q}) + 2M\|\mathbf{q}\|_2 \sqrt{2\log\frac{2}{\delta}},$$
$$\sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + \sum_{t=1}^{T} q_t \beta_t + \text{disc}(\mathbf{q}) + 2M\|\mathbf{q}\|_2 \sqrt{2\log\frac{2}{\delta}}.$$

Proof For each t, we let $\mathbf{Z}_t^T$ and $\widetilde{\mathbf{Z}}_t^T$ be independent sequences of random variables drawn from $\mathbf{P}_t^T(\cdot|\mathbf{Z}_1^{t-1})$. Define $\widehat{\mathbf{Z}}(t)$ as the sequence $(Z_1, \ldots, Z_t, \widetilde{Z}_{t+1}, \ldots, \widetilde{Z}_T)$ and observe that for any function $g\colon \mathbb{R}^T \to \mathbb{R}$ and any $s \leq t$ the following holds:

$$\mathbb{E}\Big[g\big(\mathbf{Z}_1^T\big) \mid \mathbf{Z}_1^s\Big] = \mathbb{E}\Big[g\big(\widehat{\mathbf{Z}}(t)\big) \mid \mathbf{Z}_1^s\Big]. \tag{12}$$

Recall that, by definition of $\mathcal{H}^T$, each h_t is a hypothesis $\mathcal{A}_t(\mathbf{Z}_1^T)\colon \mathcal{X} \to \mathcal{Y}$ returned by a stable algorithm $\mathcal{A}_t \in \mathfrak{A}$ trained on $\mathbf{Z}_1^T$.

In view of (12), $A_t = \mathcal{L}_{t+1}(\mathcal{A}_t(\mathbf{Z}_1^T), \mathbf{Z}_1^t) - \mathcal{L}_{t+1}(\mathcal{A}_t(\widehat{\mathbf{Z}}(t)), \mathbf{Z}_1^t)$ forms a martingale sequence. Thus, by Azuma's inequality, with probability at least $1 - \delta/2$, $\sum_{t=1}^T q_t A_t \leq M\|\mathbf{q}\|_2\sqrt{2\log\frac{2}{\delta}}$.

Similarly, $B_t = \mathbb{E}_{Z_{t+1}}[L(\mathcal{A}_t(\widehat{\mathbf{Z}}(t+1)), Z_{t+1})|\mathbf{Z}_1^t] - L(\mathcal{A}_t(\mathbf{Z}_1^T), Z_{t+1})$ is a martingale difference and with probability at least $1 - \delta/2$, $\sum_{t=1}^T q_t B_t \leq M\|\mathbf{q}\|_2\sqrt{2\log\frac{2}{\delta}}$.

By stability, we have $\mathcal{L}_{t+1}(\mathcal{A}_t(\widehat{\mathbf{Z}}(t)), \mathbf{Z}_1^t) - \mathbb{E}_{Z_{t+1}}[L(\mathcal{A}_t(\widehat{\mathbf{Z}}(t+1)), Z_{t+1})|\mathbf{Z}_1^t] \leq \beta_t$. It follows that, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:

$$\sum_{t=1}^{T} q_t \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \sum_{t=1}^{T} q_t \beta_t + 2M\|\mathbf{q}\|_2 \sqrt{2\log\frac{2}{\delta}}.$$

The first statement of the lemma then follows from the definition of the discrepancy. The second statement follows by symmetry. ■

Theorem 5 *Assume that L is convex and bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let $\mathcal{H}^*$ be a set of sequences of stable hypotheses and let $h_1, \ldots, h_T$ be a sequence of stable hypotheses. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex and let h denote $h = \sum_{t=1}^T q_t h_t$. Then, for any $\delta > 0$, each of the following bounds holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \leq \sum_{t=1}^T q_t L(h_t, Z_{t+1}) + \sum_{t=1}^T q_t \beta_t + \operatorname{disc}(\mathbf{q}) + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}},$$

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \inf_{\mathbf{h}^* \in \mathcal{H}^*} \left\{ \sum_{t=1}^T q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\} + 2\beta_{\max} + 2 \operatorname{disc}(\mathbf{q})$$
$$+ \operatorname{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 4M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}},$$

where $\beta_{\max} = \sup_{h \in \mathcal{H}} \beta_h$ and $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$.

Proof The proof of this Theorem is similar to that of Theorem 2 with the only difference that we use Lemma 4 instead of Lemma 1. ■

Theorem 6 *Assume that L is bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let $\mathcal{H}^*$ be a set of sequences of stable hypotheses. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex. For any $\delta > 0$, let h be defined by (4), then, each of the following bounds holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t L(h_t, Z_{t+1}) + \beta_{\max} + \operatorname{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}},$$

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \inf_{\mathbf{h}^* \in \mathcal{H}^*} \left\{ \sum_{t=1}^T q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\} + 2\beta_{\max}$$
$$+ 2 \operatorname{disc}(\mathbf{q}) + \operatorname{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}},$$

where $\beta_{\max} = \sup_{h \in \mathcal{H}} \beta_h$ and $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$.

Proof The proof of this Theorem is analogous to the proof of Theorem 3 with the only difference that we use Lemma 4 instead of Lemma 1. ■

Lemma 9 *Let $\mathbf{Z}_1^T$ be a sequence of random variables. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds for all $\alpha > 0$*

$$\operatorname{disc}(\mathbf{q}) \leq \widehat{\operatorname{disc}}_{\mathcal{Y}}(\mathbf{q}) + s\gamma + 2\alpha + M\|\mathbf{q} - \mathbf{u}_s\|_2 \sqrt{2 \log \frac{\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},$$

where the empirical discrepancy is defined by

$$\widehat{\operatorname{disc}}_{\mathcal{Y}}(\mathbf{q}) = \sup_{\mathbf{h} \in H_{\mathcal{A}}} \left| \frac{1}{s} \sum_{\tau=T-s+1}^T \sum_{t=1}^T q_t L(h_t, Z_\tau) - \sum_{t=1}^T q_t L(h_t, Z_{t+1})) \right|,$$

and $\gamma = \sup_t \|\mathbf{P}_{t+1}(\cdot|\mathbf{Z}_1^t) - \mathbf{P}_t(\cdot|\mathbf{Z}_1^{t-1})\|_{TV}$ *and* $\mathbf{u}_s$ *is the uniform distribution on the last* s *points.*

Proof We observe that

$$\mathrm{disc}_{\mathcal{Y}}(\mathbf{q}) - \widehat{\mathrm{disc}}_{\mathcal{Y}}(\mathbf{q}) \leq \sup_{\mathcal{G}} \left| \sum_{t=1}^{T} (q_t - p_t)(\mathbb{E}[g(Z_t, t)|\mathbf{Z}_1^{t-1}] - g(Z_t, t)) \right|,$$

where $\mathcal{G} = \sup\{(z, s) \mapsto L(h_s, z) \colon \mathbf{h} \in H_{\mathcal{A}}\}$. The conclusion of the lemma follows from Theorem 10. ■

Lemma 7 *Let* $\mathbf{Z}_1^T$ *be a sequence of random variables. Then, for any* $\delta > 0$*, with probability at least* $1 - \delta$*, the following inequality holds for all* $\alpha > 0$*:*

$$\mathrm{disc}(\mathbf{q}) \leq \widehat{\mathrm{disc}}_H(\mathbf{q}) + \inf_{h^*} \mathbb{E}[d(Z_{T+1}, h^*(X_{T+1}))|\mathbf{Z}_1^T] + 2\alpha + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},$$

where $\widehat{\mathrm{disc}}_H(\mathbf{q})$ *is the empirical discrepancy defined by*

$$\widehat{\mathrm{disc}}_H(\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t (L(h_t(X_{T+1}), h(X_{T+1})) - L(h_t, Z_{t+1})) \right|. \quad (13)$$

Proof We observe that

$$\mathrm{disc}_H(\mathbf{q}) - \widehat{\mathrm{disc}}_H(\mathbf{q}) \leq \sup_{\mathcal{G}} \left| \sum_{t=1}^{T} q_t \, \mathbb{E}[g(Z_t, t)|\mathbf{Z}_1^{t-1}] - g(Z_t, t) \right|,$$

where $\mathcal{G} = \sup\{(z, s) \mapsto L(h_s, z) \colon \mathbf{h} \in H_{\mathcal{A}}\}$. The conclusion of the lemma follows from Theorem 10. ■

Theorem 10 *Let* $\mathbf{Z}_1^T$ *be a sequence of random variables distributed according to* $\mathbf{p}$*. Define* $\Psi(\mathbf{Z}_1^T)$ *by* $\Psi(\mathbf{Z}_1^T) = \sup_{f \in \mathcal{F}} \left| \sum_{t=1}^T q_t \, \mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] - f(Z_t) \right|$*. Then, for any* $\epsilon > 2\alpha > 0$*, the following inequality holds:*

$$\mathbb{P}\big(\Psi(\mathbf{Z}_1^T) \geq \epsilon\big) \leq \mathop{\mathbb{E}}_{\mathbf{v} \sim \mathcal{Z}_T} \Big[\mathcal{N}_1(\alpha, \mathcal{F}, \mathbf{v})\Big] \exp\left(-\frac{(\epsilon - 2\alpha)^2}{2M^2\|\mathbf{q}\|_2^2}\right).$$

Theorem 10 is a consequence of Theorem 1 in (Kuznetsov and Mohri, 2015), where a slightly tighter statement is proven by bounding $(\Phi(\mathbf{Z}_1^T) - \Delta)$, with $\Phi(\mathbf{Z}_1^T)$ the supremum of the empirical process and Δ the discrepancy measure defined in (Kuznetsov and Mohri, 2015).