Annals of Mathematics and Artificial Intelligence

Discrepancy-Based Theory and Algorithms for Forecasting Non-Stationary Time Series

Vitaly Kuznetsov · Mehryar Mohri

Received: January 20, 2019

Abstract We present data-dependent learning bounds for the general scenario of non-stationary non-mixing stochastic processes. Our learning guarantees are expressed in terms of a data-dependent measure of sequential complexity and a discrepancy measure that can be estimated from data under some mild assumptions. Our learning bounds guide the design of new algorithms for non-stationary time series forecasting for which we report several favorable experimental results.

Keywords time series · forecasting · non-stationary · non-mixing · generalization bounds · discrepancy · expected sequential covering numbers · sequential Rademacher complexity

1 Introduction

Time series forecasting plays a crucial role in a number of domains ranging from weather forecasting and earthquake prediction to applications in economics and finance. The classical statistical approaches to time series analysis are based on generative models such as the autoregressive moving average (ARMA) models, or their integrated versions (ARIMA) and several other extensions [Engle, 1982; Bollerslev, 1986; Brockwell and Davis, 1986; Box and Jenkins, 1990; Hamilton, 1994]. Most of these models rely on strong assumptions about the noise terms, often assumed to be i.i.d. random variables sampled from a Gaussian distribution, and the guarantees provided in their support are often only asymptotic.

Vitaly Kuznetsov (Corresponding author)
Google Research, 76 Ninth Avenue, New York, NY 10011.
E-mail: vitaly@cims.nyu.edu

Mehryar Mohri
Google Research and Courant Institute, 251 Mercer Street, New York, NY 10012.
E-mail: mohri@cims.nyu.edu

An alternative non-parametric approach to time series analysis consists of extending the standard i.i.d. statistical learning theory framework to that of stochastic processes. In much of this work, the process is assumed to be stationary and suitably mixing [Doukhan, 1994]. Early work along this approach consisted of the VC-dimension bounds for binary classification given by Yu [1994] under the assumption of stationarity and β-mixing. Under the same assumptions, Meir [2000] presented bounds in terms of covering numbers for regression losses and Mohri and Rostamizadeh [2009] proved general data-dependent Rademacher complexity learning bounds. Vidyasagar [1997] showed that PAC learning algorithms in the i.i.d. setting preserve their PAC learning property in the β-mixing stationary scenario. A similar result was proven by Shalizi and Kontorovich [2013] for mixtures of β-mixing processes and by Berti and Rigo [1997] and Pestov [2010] for exchangeable random variables. Alquier and Wintenberger [2010] and Alquier et al. [2014] also established PAC-Bayesian learning guarantees under weak dependence and stationarity. Chen and Wu [2018] provide concentration results for linear time series.

A number of algorithm-dependent bounds have also been derived for the stationary mixing setting. Lozano et al. [2006] studied the convergence of regularized boosting. Mohri and Rostamizadeh [2010] gave data-dependent generalization bounds for stable algorithms for φ-mixing and β-mixing stationary processes. Steinwart and Christmann [2009] proved fast learning rates for regularized algorithms with α-mixing stationary sequences and Modha and Masry [1998] gave guarantees for certain classes of models under the same assumptions.

However, stationarity and mixing are often not valid assumptions. For example, even for Markov chains, which are among the most widely used types of stochastic processes in applications, stationarity does not hold unless the Markov chain is started with an equilibrium distribution. Similarly, long memory models such as ARFIMA, may not be mixing or mixing may be arbitrarily slow [Baillie, 1996]. In fact, it is possible to construct first order autoregressive processes that are not mixing [Andrews, 1983]. Additionally, the mixing assumption is defined only in terms of the distribution of the underlying stochastic process and ignores the loss function and the hypothesis set used. This suggests that mixing may not be the right property to characterize learning in the setting of stochastic processes.

A number of attempts have been made to relax the assumptions of stationarity and mixing. Adams and Nobel [2010] proved asymptotic guarantees for stationary ergodic sequences. Agarwal and Duchi [2013] gave generalization bounds for asymptotically stationary (mixing) processes in the case of stable on-line learning algorithms. Kuznetsov and Mohri [2014] established learning guarantees for fully non-stationary β- and φ-mixing processes.

In this paper, we consider the general case of non-stationary non-mixing processes. We are not aware of any prior work providing generalization bounds in this setting. In fact, our bounds appear to be novel even when the process is stationary (but not mixing). The learning guarantees that we present hold for both bounded and unbounded memory models. Deriving generalization bounds for unbounded memory models even in the stationary mixing case was an open question prior to our work [Meir, 2000]. Our guarantees cover the majority of approaches used in practice, including various autoregressive models.

The key ingredients of our generalization bounds are a data-dependent measure of sequential complexity (*expected sequential covering number* or *sequential Rademacher complexity* [Rakhlin et al., 2010]) and a measure of *discrepancy* between the sample and target distributions. Kuznetsov and Mohri [2014, 2017a] also give generalization bounds in terms of discrepancy. However, unlike these result, our analysis does not require any mixing assumptions which are hard to verify in practice. More importantly, under some additional mild assumption, the discrepancy measure that we propose can be estimated from data, which leads to data-dependent learning guarantees for non-stationary non-mixing case.

We use the theory that we develop to devise new algorithms for non-stationary time series forecasting that benefit from our data-dependent guarantees. The parameters of generative models such as ARIMA are typically estimated via the maximum likelihood technique, which often leads to non-convex optimization problems. In contrast, our objective is convex and leads to an optimization problem with a unique global solution that can be found efficiently. Another issue with standard generative models is that they address non-stationarity in the data via a *differencing* transformation which does not always lead to a stationary process. In contrast, we address the problem of non-stationarity in a principled way using our learning guarantees.

The rest of this paper is organized as follows. The formal definition of the time series forecasting learning scenario as well as that of several key concepts is given in Section 2. In Section 3, we introduce and prove our new generalization bounds. Section 4 provides an analysis in the special case of kernel-based hypotheses with regression losses. In Section 5, we give data-dependent learning bounds based on the empirical discrepancy. These results are used to devise new forecasting algorithms in Section 6. In Appendix 7, we report the results of preliminary experiments using these algorithms.

2 Preliminaries

We consider a general time series prediction scenario where the learner receives a realization $Z_1, \dots, Z_T$ of some stochastic process, where, for any $t \in [T]$, $Z_t = (X_t, Y_t)$ is in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{X}$ an input space and $\mathcal{Y}$ an output space. We will often use the shorthand $\mathbf{Z}_n^m$ to denote a sequence of random variables $Z_n, Z_{n+1}, \dots, Z_m$. Given a loss function $L \colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ and a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$, the objective of the learner is to select a predictor $h \colon \mathcal{X} \to \mathcal{Y}$ in $\mathcal{H}$ that achieves a small *path-dependent* generalization error $\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T)$, that is the generalization error conditioned on the data observed:

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) = \mathbb{E}[L(h(X_{T+1}), Y_{T+1}) | Z_1, \dots, Z_T]. \tag{1}$$

The path-dependent generalization error that we consider in this work is a finer measure than the *averaged* generalization error

$$\mathcal{L}_{T+1}(h) = \mathbb{E}[L(h(X_{T+1}), Y_{T+1})] = \mathbb{E}[\mathbb{E}[L(h(X_{T+1}), Y_{T+1})|Z_1, \ldots, Z_T]], \quad (2)$$

since it takes into consideration the specific realization of the stochastic process and does not average over all possible histories. The results presented in this paper apply as well to the setting where the time parameter t can take non-integer values and where the prediction lag is an arbitrary value $l \geq 0$. Thus, the error can be defined more generally by $\mathbb{E}[L(h(X_{T+l}), Y_{T+l})|Z_1, \ldots, Z_T]$, but for notational simplicity we set $l = 1$.

Our setup covers a larger number of scenarios commonly used in practice. The case $\mathcal{X} = \mathcal{Y}^p$ corresponds to a large class of autoregressive models. Taking $\mathcal{X} = \bigcup_{p=1}^{\infty} \mathcal{Y}^p$ leads to growing memory models which, in particular, include state space models. More generally, $\mathcal{X}$ may contain both the history of the process $\{Y_t\}$ and some additional side information. Note that the output space $\mathcal{Y}$ may also be high-dimensional. This covers both the case where we seek to forecast a multi-variate or high-dimensional time series, and that of multi-step forecasting.

We denote by $\mathcal{F}$ the family of loss functions associated to $\mathcal{H}$: $\mathcal{F} = \{(x, y) \to L(h(x), y) : h \in \mathcal{H}\}$. For any $z = (x, y) \in \mathcal{Z}$, $L(h(x), y)$ can thus be replaced by $f(z)$ for some $f \in \mathcal{F}$. We will assume a bounded loss function, that is $|f| \leq M$ for all $f \in \mathcal{F}$ for some $M \in \mathbb{R}_+$.

The key quantity of interest in the analysis of generalization is the following supremum of the empirical process defined as follows:

$$\Phi(\mathbf{Z}_1^T) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] - \sum_{t=1}^{T} q_t f(Z_t) \right), \quad (3)$$

where $q_1, \ldots, q_T$ are real numbers, which in the standard learning scenarios are chosen to be all equal to $\frac{1}{T}$. In our general setting, different Z_ts may follow different distributions, thus distinct weights could be assigned to the errors made on different sample points, depending on their relevance to forecasting the future Z_{T+1}. The generalization bounds that we present below are for an arbitrary sequence $\mathbf{q} = (q_1, \ldots q_T)$ which, in particular, covers the case of uniform weights. Remarkably, our bounds do not even require the non-negativity of $\mathbf{q}$.

The two key ingredients of our analysis are the notions of sequential complexity [Rakhlin et al., 2010] and that of discrepancy measure between target and source distributions. In the next two sections, we give a detailed description of these notions.

2.1 Sequential Complexities

Our generalization bounds are expressed in terms of data-dependent measures of sequential complexity such as expected sequential covering number or se-

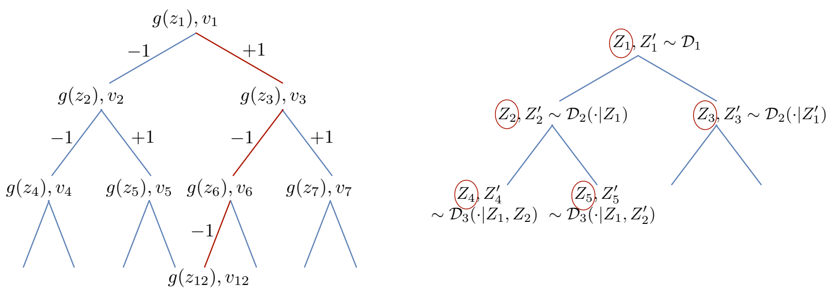
$g(z_1), v_1$
-1
$+1$
$g(z_2), v_2$
$g(z_3), v_3$
-1
$+1$
-1
$+1$
$g(z_4), v_4$
$g(z_5), v_5$
$g(z_6), v_6$
$g(z_7), v_7$
-1
$g(z_{12}), v_{12}$
$Z_1, Z'_1 \sim \mathcal{D}_1$
$Z_2, Z'_2 \sim \mathcal{D}_2(\cdot|Z_1)$
$Z_3, Z'_3 \sim \mathcal{D}_2(\cdot|Z'_1)$
$Z_4, Z'_4 \sim \mathcal{D}_3(\cdot|Z_1, Z_2)$
$Z_5, Z'_5 \sim \mathcal{D}_3(\cdot|Z_1, Z'_2)$

Fig. 1 Weighted sequential cover and sequential covering numbers.

quential Rademacher complexity [Rakhlin et al., 2010], which we review in this section.

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 $z_t\colon \{\pm 1\}^{t-1} \to \mathcal{Z}$, $t \in [1, T]$. A path in the tree is $\sigma = (\sigma_1, \ldots, \sigma_{T-1}) \in \{\pm 1\}^{T-1}$. 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 sequential setting. A set $\mathcal{V}$ 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 $\mathbf{v} \in \mathcal{V}$ such that

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

where $\|\cdot\|_q$ is the dual norm associated to $\|\cdot\|_p$, that is $\frac{1}{p} + \frac{1}{q} = 1$. 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 $\mathcal{N}_p(\alpha, \mathcal{G}) = \sup_{\mathbf{z}} \mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z})$. 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 $\mathbf{p}_2(\cdot|Z'_1)$. 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 $\mathcal{T}$ the distribution of trees $\mathbf{z}$ thereby defined. Then, the *expected covering number* is defined as $\mathbb{E}_{\mathbf{z} \sim \mathcal{T}}[\mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z})]$.

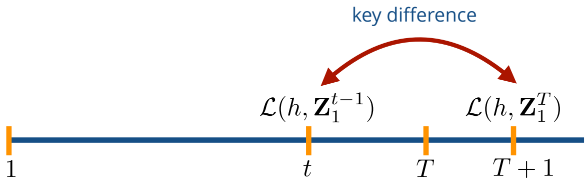
key difference
$\mathcal{L}(h, \mathbf{Z}_1^{t-1})$
$\mathcal{L}(h, \mathbf{Z}_1^{T})$
1
t
T
$T+1$

Fig. 2 Key difference in the definition of discrepancy.

Figure 1 illustrates these notions. For i.i.d. sequences, the notion of expected sequential covering numbers exactly coincides with that of expected covering numbers from classical statistical learning theory.

The sequential Rademacher complexity $\mathfrak{R}_T^{\mathrm{seq}}(\mathcal{G})$ of a function class $\mathcal{G}$ is defined by

$$\mathfrak{R}_T^{\mathrm{seq}}(\mathcal{G}) = \sup_{\mathbf{z}} \mathbb{E}\left[\sup_{g \in \mathcal{G}} \sum_{t=1}^{T} \sigma_t q_t g(z_t(\boldsymbol{\sigma}))\right], \tag{4}$$

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 [Rakhlin et al., 2010, 2011, 2015a,b]. Similarly, one can also define the notion of *distribution-dependent* sequential Rademacher complexity as well as other notions of sequential complexity such as *Littlestone dimension* and *sequential metric entropy* that have been shown to characterize learning in the on-line learning scenario. For further details, we refer the reader to [Littlestone, 1987; Rakhlin et al., 2010, 2011, 2015a,b].

2.2 Discrepancy Measure

The final ingredient needed for expressing our learning guarantees is the notion of *discrepancy* between target distribution and the distribution of the sample:

$$\mathrm{disc}_T(\mathbf{q}) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] - \sum_{t=1}^{T} q_t \, \mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] \right). \tag{5}$$

In what follows we will omit subscript T and write disc_T to simplify the notation. Figure 2 illustrates the key difference term in the definition of discrepancy.

The discrepancy disc is a natural measure of the non-stationarity of the stochastic process $\mathbf{Z}$ with respect to both the loss function L and the hypothesis set $\mathcal{H}$. 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. To help the reader develop further intuition about discrepancy, we provide several explicit examples below.

Example 1. Consider the case of a time-homogeneous Markov chain on a set $\{0, \ldots, N-1\}$ such that $\mathbf{P}(X_t \equiv (i-1) \bmod N | X_{t-1} = i) = p$ and $\mathbf{P}(X_t \equiv (i+1) \bmod N | X_{t-1} = i) = 1-p$ for some $0 \leq p \leq 1$. This process is non-stationary if it is not started with an equilibrium distribution. Suppose that the set of hypothesis is $\{x \mapsto a(x-1)+b(x+1)\colon a+b=1, a, b \geq 0\}$ and the loss function $L(y, y') = \ell(|y-y'|)$ for some ℓ. It follows that for any (a, b)

$$\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] = p|a-b-1| + (1-p)|a-b+1|$$

and hence $\mathrm{disc}(\mathbf{q}) = 0$ provided $\mathbf{q}$ is a probability distribution. Note that if we chose a larger hypothesis set $\{x \mapsto a(x-1)+b(x+1)\colon a, b \geq 0\}$ then

$$\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] = p|(a+b-1)X_t + a - b| + (1-p)|(a+b+1)X_t + a - b|$$

and in general it may be the case that $\mathrm{disc}(\mathbf{q}) \neq 0$. This highlights an important property of discrepancy: it takes into account not only the underlying distribution of the stochastic process but other components of the learning problem such as the loss function and the hypothesis set that is being used.

Example 2. Let $\epsilon_0, \epsilon_1, \ldots$ be a sequence of i.i.d. random variables such that $\mathbb{P}(\epsilon_t = -1) = p$ and $\mathbb{P}(\epsilon = 1) = 1-p$ for some $0 \leq p \leq 1$. Consider the following stochastic: $Y_t = Y_{t-1} + \epsilon_t Y_0$ for $t \geq 1$ and $X_0 = \epsilon_0$. Observe that this process is not Markov, not stationary and not mixing. Furthermore, when $p \neq \frac{1}{2}$ this process has a stochastic trend. Let $\mathcal{H} = \{x \mapsto x + c\colon c \in [-1, 1]\}$ and consider a loss function L such that $L(y', y) = \ell(y' - y)$ for some ℓ and any y, y'. Observe that discrepancy in this case is given by

$$\sup_{h \in \mathcal{H}} \left(\mathbb{E}_{\epsilon_{T+1}}[\ell(c - \epsilon_{T+1} X_0)] - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\epsilon_t}[\ell(c - \epsilon_t X_0)] \right) = 0.$$

This result combined with Corollary 2 below can be used to provide a learning guarantee for this setting.

Example 3. The next example is concerned with periodic time series. Let $\epsilon_0, \epsilon_1, \ldots$ be a sequence of i.i.d. standard Gaussian random variables and consider a stochastic process $Y_t = \sin(\phi t) + \epsilon_t$. We let $\mathcal{H}$ be a set of offset functions as before[1] $\mathcal{H} = \{x \mapsto x + c : c \in [-1, 1]\}$. If $L(y', y') = (y - y')^2$ then discrepancy is given by

$$\sup_{h \in \mathcal{H}} \left((c - \sin(\phi(T+1))^2 - \sum_{t=1}^{T} q_t (c - \sin(\phi t))^2 \right)$$

and since sin is a periodic function it is possible to choose $\mathbf{q}$ such that above expression is non-positive and $\|\mathbf{q}\|_2 = \frac{1}{\sqrt{T}}$. This result together with Corollary 2 below can be used to provide learning guarantees in this setting and shows that periodic time series can be improperly learned with offset functions.

The weights $\mathbf{q}$ play a crucial role in the learning problem. Consider our Example 1, where transition probability distributions $(p_i, 1 - p_i)$ are different for each state i. Note that choosing $\mathbf{q}$ to be a uniform distribution, in general, leads to a non-zero discrepancy. However, with $q'_t = \mathbf{1}_{X_{t-1} = X_T}$ and $q_t = q'_t / \sum_{t=1}^{T} q'_t$ discrepancy is zero. Note that in fact it is not the only choice

[1] A set of periodic functions would also be sufficient for this example.

that leads to a zero discrepancy in this example and in fact any distribution that is supported on ts for which $X_{t-1} = X_T$ will lead to the same result. However, q_ts based on q'_t have an advantage of providing the largest effective sample.

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

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

where $\|\cdot\|$ is the total variation distance and $D(\cdot \parallel \cdot)$ 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. Note that these upper bounds are often too loose since they are agnostic to the loss function and the hypothesis set that is being used. For our earlier Markov chain example, the support of $\mathbf{P}_{T+1}(\cdot|\mathbf{Z}_1^T)$ is $\{X_T - 1, X_T + 1\}$ while the mixture $\frac{1}{T}\sum_{t=1}^T \mathbf{P}_t(\cdot|\mathbf{Z}_1^{t-1})$ is likely to be supported on the whole set $\{0, \ldots, N-1\}$ which leads to a large total variation distance. Of course, it is possible to choose q_ts so that the mixture is also supported only on $\{X_T - 1, X_T + 1\}$ but that may reduce the effective sample size which is not necessary when working with $\operatorname{disc}(\mathbf{q})$.

However, the most important property of the discrepancy $\operatorname{disc}(\mathbf{q})$ is that, as shown later in Section 5, it can be estimated from data under some additional mild assumptions. Kuznetsov and Mohri [2014] also give generalization bounds based on averaged generalization error for non-stationary mixing processes in terms of a related notion of discrepancy. It is not known if the discrepancy measure used in [Kuznetsov and Mohri, 2014] can be estimated from data.

3 Generalization Bounds

In this section, we prove new generalization bounds for forecasting non-stationary time series. The first step consists of using *decoupled tangent* sequences to establish concentration results for the supremum of the empirical process $\Phi(\mathbf{Z}_1^T)$. Given a sequence of random variables $\mathbf{Z}_1^T$ we say that $\mathbf{Z}_1'^T$ is a decoupled tangent sequence if Z'_t is distributed according to $\mathbb{P}(\cdot|\mathbf{Z}_1^{t-1})$ and is independent of $\mathbf{Z}_t^\infty$. It is always possible to construct such a sequence of random variables [De la Peña and Giné, 1999]. The next theorem is the main result of this section.

Theorem 1 *Let $\mathbf{Z}_1^T$ be a sequence of random variables distributed according to $\mathbf{p}$. Fix $\epsilon > 2\alpha > 0$. Then, the following holds:*

$$\mathbb{P}\Big(\Phi(\mathbf{Z}_1^T) - \operatorname{disc}(\mathbf{q}) \ge \epsilon\Big) \le \mathop{\mathbb{E}}_{\mathbf{v}\sim\mathcal{T}}\big[\mathcal{N}_1(\alpha, \mathcal{F}, \mathbf{v})\big] \exp\Bigg(-\frac{(\epsilon - 2\alpha)^2}{8M^2\|\mathbf{q}\|_2^2}\Bigg).$$

Proof Since the difference of two suprema is upper bounded by the supremum of the difference, by Markov's inequality, the following holds for any $\epsilon > 0$:

$$\begin{aligned}
&\mathbb{P}\Big(\Phi(\mathbf{Z}_1^T) - \operatorname{disc}(\mathbf{q}) \geq \epsilon\Big) \\
&\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t(\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] - f(Z_t))\right) \geq \epsilon\right) \\
&\leq \exp(-\lambda\epsilon)\,\mathbb{E}\left[\exp\left(\lambda \sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t(\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] - f(Z_t))\right)\right)\right].
\end{aligned}$$

Let $\mathbf{Z}_1'^T$ be a decoupled tangent sequence associated to $\mathbf{Z}_1^T$, then, the following equalities hold: $\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] = \mathbb{E}[f(Z_t')|\mathbf{Z}_1^{t-1}] = \mathbb{E}[f(Z_t')|\mathbf{Z}_1^T]$. Using these equalities and Jensen's inequality, we obtain the following:

$$\begin{aligned}
&\mathbb{E}\left[\exp\Big(\lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T} q_t\big(\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] - f(Z_t)\big)\Big)\right] \\
&= \mathbb{E}\left[\exp\Big(\lambda \sup_{f\in\mathcal{F}}\mathbb{E}\Big[\sum_{t=1}^{T} q_t\big(f(Z_t') - f(Z_t)\big)|\mathbf{Z}_1^T\Big]\Big)\right] \\
&\leq \mathbb{E}\left[\exp\Big(\lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T} q_t\big(f(Z_t') - f(Z_t)\big)\Big)\right],
\end{aligned}$$

where the last expectation is taken over the joint measure of $\mathbf{Z}_1^T$ and $\mathbf{Z}_1'^T$. Applying Lemma 2 (Appendix A), we can further bound this expectation by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')\sim\mathcal{T}}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(\lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sigma_t q_t\Big(f(\mathbf{z}_t'(\boldsymbol{\sigma})) - f(\mathbf{z}_t(\boldsymbol{\sigma}))\Big)\right)\right] \\
&\leq \mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')\sim\mathcal{T}}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(\lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sigma_t q_t f(\mathbf{z}_t'(\boldsymbol{\sigma})) + \lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T} -\sigma_t q_t f(\mathbf{z}_t(\boldsymbol{\sigma}))\right)\right] \\
&\leq \frac{1}{2}\mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2\lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sigma_t q_t f(\mathbf{z}_t'(\boldsymbol{\sigma}))\right)\right] \\
&\quad + \frac{1}{2}\mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2\lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sigma_t q_t f(\mathbf{z}_t(\boldsymbol{\sigma}))\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{T}}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2\lambda \sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sigma_t q_t f(\mathbf{z}_t(\boldsymbol{\sigma}))\right)\right],
\end{aligned}$$

where the second inequality holds by the convexity of the exponential function and the last inequality by symmetry. Given $\mathbf{z}$, let C denote the minimal α-cover with respect to the $\mathbf{q}$-weighted ℓ_1-norm of $\mathcal{F}$ on $\mathbf{z}$. Then, the following bound holds

$$\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sigma_t q_t f(\mathbf{z}_t(\boldsymbol{\sigma})) \leq \max_{\mathbf{c}\in C}\sum_{t=1}^{T}\sigma_t q_t \mathbf{c}_t(\boldsymbol{\sigma}) + \alpha.$$

By the monotonicity of the exponential function,

$$\begin{aligned}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2\lambda\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sigma_t q_t f(\mathbf{z}_t(\boldsymbol{\sigma}))\right)\right] &\le \exp(2\lambda\alpha)\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2\lambda\max_{\mathbf{c}\in C}\sum_{t=1}^{T}\sigma_t q_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\right] \\ &\le \exp(2\lambda\alpha)\sum_{\mathbf{c}\in C}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2\lambda\sum_{t=1}^{T}\sigma_t q_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\right].\end{aligned}$$

Since $\mathbf{c}_t(\boldsymbol{\sigma})$ depends only on $\sigma_1,\ldots,\sigma_{T-1}$, by Hoeffding's bound,

$$\begin{aligned}&\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(2\lambda\sum_{t=1}^{T}\sigma_t q_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\right] \\ &= \mathbb{E}\left[\exp\left(2\lambda\sum_{t=1}^{T-1}\sigma_t q_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\mathop{\mathbb{E}}_{\sigma_T}\left[\exp\left(2\lambda\sigma_T q_T \mathbf{c}_T(\boldsymbol{\sigma})\right)\Big|\boldsymbol{\sigma}_1^{T-1}\right]\right] \\ &\le \mathbb{E}\left[\exp\left(2\lambda\sum_{t=1}^{T-1}\sigma_t q_t \mathbf{c}_t(\boldsymbol{\sigma})\right)\exp(2\lambda^2 q_T^2 M^2)\right].\end{aligned}$$

Iterating this inequality and using the union bound, we obtain the following:

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{T} q_t(\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] - f(Z_t)) \ge \epsilon\right) \\ &\le \mathop{\mathbb{E}}_{\mathbf{v}\sim\mathcal{T}}[\mathcal{N}_1(\alpha,\mathcal{G},\mathbf{v})]\exp\left(-\lambda(\epsilon-2\alpha)+2\lambda^2M^2\|\mathbf{q}\|_2^2\right).\end{aligned}$$

Optimizing over $\lambda > 0$ completes the proof. □

An immediate consequence of Theorem 1 is the following result.

Corollary 1 *For any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for all $f\in\mathcal{F}$ and all $\alpha > 0$:*

$$\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] \le \sum_{t=1}^{T} q_t f(Z_t) + \mathrm{disc}(\mathbf{q}) + 2\alpha + M\|\mathbf{q}\|_2\sqrt{8\log\frac{\mathbb{E}_{\mathbf{v}\sim\mathcal{T}}[\mathcal{N}_1(\alpha,\mathcal{F},\mathbf{v})]}{\delta}}.$$

We are not aware of other finite sample bounds in a non-stationary non-mixing case. In fact, our bounds appear to be novel even in the stationary non-mixing case.

While Rakhlin et al. [2015a] give high probability bounds for a different quantity than the quantity of interest in time series prediction,

$$\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t(\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}] - f(Z_t))\right), \tag{6}$$

their analysis of this quantity can also be used in our context to derive high probability bounds for $\Phi(\mathbf{Z}_1^T) - \mathrm{disc}(\mathbf{q})$. However, this approach results in bounds that are in terms of purely combinatorial notions such as maximal sequential covering numbers $\mathcal{N}_1(\alpha,\mathcal{F})$. While at first sight, this may seem as a minor technical detail, the distinction is crucial in the setting of time series prediction. Consider the following example. Let Z_1 be drawn from a uniform distribution on $\{0,1\}$ and $Z_t \sim p(\cdot|Z_{t-1})$ with $p(\cdot|y)$ being a distribution over $\{0,1\}$ such that $p(x|y) = 2/3$ if $x = y$ and $1/3$ otherwise. Let $\mathcal{G}$ be defined by $\mathcal{G} = \{g(x) = \mathbf{1}_{x \geq \theta} \colon \theta \in [0,1]\}$. Then, one can check that $\mathbb{E}_{\mathbf{v} \sim \mathcal{T}}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{v})] = 2$, while $\mathcal{N}_1(\alpha, \mathcal{G}) \geq 2^T$. The data-dependent bounds of Theorem 1 and Corollary 1 highlight the fact that the task of time series prediction lies in-between the familiar i.i.d. scenario and the adversarial on-line learning setting.

However, the key component of our learning guarantees is the discrepancy term $\operatorname{disc}(\mathbf{q})$. Note that in the general non-stationary case, the bounds of Theorem 1 may not converge to zero due to the discrepancy between the target and sample distributions. This is also consistent with the lower bounds of Barve and Long [1996] that we discuss in more detail in Section 5. However, convergence can be established in some special cases. In the i.i.d. case, our bounds reduce to the standard covering numbers learning guarantees. The discrepancy examples of the previous section also show that convergence can be established for various stochastic processes, including non-mixing and non-stationary ones. In the drifting scenario, with $\mathbf{Z}_1^T$ being a sequence of independent random variables, our discrepancy measure and bounds coincide with those studied in [Mohri and Muñoz Medina, 2012]. In the case of ϕ-mixing non-stationary stochastic processes, our results provide tighter learning bounds for path-dependent generalization error than the previous best results in [Kuznetsov and Mohri, 2017a].[2] However, shown in Section 5, the most important advantage of our bounds is that the discrepancy measure we use can be estimated from data.

We now show that expected sequential covering numbers can be upper bounded in terms of the sequential Rademacher complexity. Generalization bounds in terms of the sequential Rademacher complexity are not as tight as bounds in terms the expected sequential covering numbers since the former is a purely combinatorial notion. Nevertheless, the analysis of sequential Rademacher complexity may be simpler for certain hypothesis classes such as for instance that of kernel-based hypotheses, which we study in Section 4. We have the following extension of Sudakov's Minoration Theorem to the setting of sequential complexities.

Theorem 2 *The following bound holds for the sequential Rademacher complexity of $\mathcal{F}$:*

$$\sup_{\alpha > 0} \frac{\alpha}{2} \sqrt{\log \mathcal{N}_2(2\alpha, \mathcal{F})} \leq 3 \sqrt{\frac{\pi}{2} \log T}\, \mathfrak{R}_T^{seq}(\mathcal{F}),$$

whenever $\mathcal{N}_2(2\alpha, \mathcal{F}) < +\infty$.

Proof We consider the following Gaussian-Rademacher sequential complexity:

$$\mathfrak{G}_T^{\mathrm{seq}}(\mathcal{F}, \mathbf{z}) = \mathop{\mathbb{E}}_{\boldsymbol{\gamma}, \boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left(\sum_{t=1}^{T} q_t \sigma_t \gamma_t f(z_t(\boldsymbol{\sigma})) \right) \right], \tag{7}$$

[2] For further details, see the discussion following Corollary 2.

where $\boldsymbol{\sigma}$ is an independent sequence of Rademacher random variables, $\boldsymbol{\gamma}$ is an independent sequence of standard Gaussian random variables and $\mathbf{z}$ is a complete binary tree of depth T with values in $\mathcal{Z}$.

Observe that if V is any α-cover with respect to the $\mathbf{q}$-weighted ℓ_2-norm of $\mathcal{F}$ on $\mathbf{z}$, then, the following holds by independence of $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$:

$$\mathfrak{G}^{\text{seq}}(\mathcal{F},\mathbf{z}) \geq \mathop{\mathbb{E}}_{\boldsymbol{\gamma}}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^{T} q_t\sigma_t\gamma_t\mathbf{v}_t(\boldsymbol{\sigma})\right)\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{\boldsymbol{\gamma}}\left[\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^{T} q_t\sigma_t\gamma_t\mathbf{v}_t(\boldsymbol{\sigma})\right)\right].$$

Observe that V is also 2α-cover with respect to the $\mathbf{q}$-weighted ℓ_2-norm of $\mathcal{F}$ on $\mathbf{z}$. We can obtain a smaller 2α-cover V_0 from V be eliminating $\mathbf{v}$s that are α close to some other $\mathbf{v}' \in V$. Since V is finite, let $V = \{\mathbf{v}^1,\ldots,\mathbf{v}^{|V|}\}$, and for each $\mathbf{v}^i$ we delete $\mathbf{v}^j \in \{\mathbf{v}_{i+1},\ldots,\mathbf{v}^{|V|}\}$ such that

$$\left(\sum_{t=1}^{T}\left(q_t\mathbf{v}_t^i(\boldsymbol{\sigma}) - q_t\mathbf{v}_t^j(\boldsymbol{\sigma})\right)^2\right)^{1/2} \leq \alpha.$$

It is straightforward to verify that V_0 is 2α-cover with respect to the $\mathbf{q}$-weighted ℓ_2-norm of $\mathcal{F}$ on $\mathbf{z}$. Furthermore, it follows that for a fixed $\boldsymbol{\sigma}$, the following holds:

$$\mathop{\mathbb{E}}_{\boldsymbol{\gamma}}\left[\left(\sum_{t=1}^{T} q_t\sigma_t\gamma_t\mathbf{v}_t(\boldsymbol{\sigma}) - \sum_{t=1}^{T} q_t\sigma_t\gamma_t\mathbf{v}'_t(\boldsymbol{\sigma})\right)^2\right] \geq \alpha^2.$$

for any $\mathbf{v}',\mathbf{v} \in V_0$. Let $Z_i, i = 1,\ldots,|V_0|$ be a sequence of independent Gaussian random variables with $\mathbb{E}[Z_i] = 0$ and $\mathbb{E}[Z_i^2] = \alpha^2/2$. Observe that $\mathbb{E}[(Z_i - Z_j)] = \alpha^2$ and hence by Sudakov-Fernique inequality it follows that

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{\boldsymbol{\gamma}}\left[\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^{T} q_t\sigma_t\gamma_t\mathbf{v}_t(\boldsymbol{\sigma})\right)\right] &\geq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{\boldsymbol{\gamma}}\left[\sup_{\mathbf{v}\in V_0}\left(\sum_{t=1}^{T} q_t\sigma_t\gamma_t\mathbf{v}_t(\boldsymbol{\sigma})\right)\right]\\ &\geq \mathbb{E}\left[\max_{i=1,\ldots,|V_0|} Z_i\right]\\ &\geq \frac{\alpha}{2}\sqrt{\log|V_0|},\end{aligned}$$

where the last inequality is the standard result for Gaussian random variables. Therefore, we conclude that $\mathfrak{G}^{\text{seq}}(\mathcal{F},\mathbf{z}) \geq \sup_{\alpha>0}\frac{\alpha}{2}\sqrt{\log\mathcal{N}_2(2\alpha,\mathcal{F},\mathbf{z})}$. On the other hand, by the standard properties of Gaussian complexity [Ledoux and Talagrand, 1991], we can write:

$$\mathfrak{G}_T^{\text{seq}}(\mathcal{F},\mathbf{z}) \leq 3\sqrt{\frac{\pi}{2}\log T}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon},\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t\sigma_t\epsilon_t f(z_t(\boldsymbol{\sigma}))\right)\right],$$

where $\boldsymbol{\epsilon}$ is an independent sequence of Rademacher variables. We re-arrange $\mathbf{z}$ into $\mathbf{z}^{\boldsymbol{\epsilon}}$ so that $z_t(\boldsymbol{\sigma}) = z_t^{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon\sigma})$ for all $\boldsymbol{\sigma} \in \{\pm 1\}^T$ and it follows that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\epsilon},\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t\sigma_t\epsilon_t f(z_t(\boldsymbol{\sigma}))\right)\right] &= \mathop{\mathbb{E}}_{\boldsymbol{\epsilon},\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t\sigma_t\epsilon_t f(z_t^{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon\sigma}))\right)\right] \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\sup_{\mathbf{z}}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t\sigma_t\epsilon_t f(z_t(\boldsymbol{\epsilon\sigma}))\right)\right] \\
&= \sup_{\mathbf{z}}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T} q_t\sigma_t f(z_t(\boldsymbol{\sigma}))\right)\right].
\end{aligned}$$

Therefore, the following inequality holds

$$\sup_{\alpha>0}\frac{\alpha}{2}\sqrt{\log\mathcal{N}_2(2\alpha,\mathcal{F},\mathbf{z})} \le 3\sqrt{\frac{\pi}{2}\log T}\,\mathfrak{R}_T^{\mathrm{seq}}(\mathcal{F})$$

and the proof is completed by taking the supremum with respect to $\mathbf{z}$ of both sides of this inequality. □

Observe that, by definition of the sequential complexities, the following inequalities hold: $\mathbb{E}_{\mathbf{v}\sim\mathcal{T}}[\mathcal{N}_1(\alpha,\mathcal{G},\mathbf{v})] \le \mathbb{E}_{\mathbf{v}\sim\mathcal{T}}[\mathcal{N}_2(\alpha,\mathcal{G},\mathbf{v})] \le \mathcal{N}_2(\alpha,\mathcal{G})$. Thus, setting $\alpha = \|\mathbf{q}\|_2/2$, applying Corollary 1 and Theorem 2, and using the fact that $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y > 0$, yields the following result.

Corollary 2 *For any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for all $f \in \mathcal{F}$ and all $\alpha > 0$:*

$$\begin{aligned}
&\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] \\
&\le \sum_{t=1}^{T} q_t f(Z_t) + \mathrm{disc}(\mathbf{q}) + \|\mathbf{q}\|_2 + 6M\sqrt{4\pi\log T}\,\mathfrak{R}_T^{seq}(\mathcal{F}) + M\|\mathbf{q}\|_2\sqrt{8\log\frac{1}{\delta}}.
\end{aligned}$$

As already mentioned in Section 2, sequential Rademacher complexity can be further upper bounded in terms of the sequential metric entropy, sequential Littlestone dimension, maximal sequential covering numbers and other combinatorial notions of sequential complexity of $\mathcal{F}$. These notions have been extensively studied in the past [Rakhlin et al., 2015b]. Note that, in [Kuznetsov and Mohri, 2017a], an almost identical bound on $\mathbb{E}[f(Z_{T+s})|\mathbf{Z}_1^T]$ is proven under the assumption that the underlying stochastic process is ϕ-mixing (see Theorem 3 in [Kuznetsov and Mohri, 2017a]). Our result requires neither of these assumptions.

Corollary 1 or Corollary 2 can also be used to derive *oracle inequalities* for the setting that we are considering. Let $f^* = \mathrm{argmin}_{f\in\mathcal{F}}\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T]$ and

$f_0 = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{t=1}^T q_t f(Z_t)$. Then, it follows that

$$\begin{aligned}\mathbb{E}[f_0(Z_{T+1})|\mathbf{Z}_1^T] - \mathbb{E}[f^*(Z_{T+1})|\mathbf{Z}_1^T] &= \mathbb{E}[f_0(Z_{T+1})|\mathbf{Z}_1^T] - \sum_{t=1}^T q_t f_0(Z_t) \\ &+ \sum_{t=1}^T q_t f_0(Z_t) - \sum_{t=1}^T q_t f^*(Z_t) \\ &+ \sum_{t=1}^T q_t f^*(Z_t) - \mathbb{E}[f^*(Z_{T+1})|\mathbf{Z}_1^T] \\ &\le 2\Phi(\mathbf{Z}_1^T).\end{aligned}$$

The following result immediately follows.

Corollary 3 *For any $\delta > 0$, with probability at least $1-\delta$, for all $\alpha > 0$,*

$$\begin{aligned}&\mathbb{E}[f_0(Z_{T+1})|\mathbf{Z}_1^T] \\ &\le \inf_{f \in \mathcal{F}} \mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] + \operatorname{disc}(\mathbf{q}) + \|\mathbf{q}\|_2 + 6M\sqrt{4\pi \log T}\,\mathfrak{R}_T^{seq}(\mathcal{F}) + M\|\mathbf{q}\|_2\sqrt{8\log\frac{1}{\delta}}.\end{aligned}$$

We conclude this section by observing that our results also hold in the case where $q_t = q_t(f, X_{T+1}, Z_t)$, which is a common heuristic used in some algorithms for forecasting non-stationary time series [Lorenz, 1969; Zhao and Giannakis, 2016]. We formalize this result in the following theorem.

Theorem 3 *Let $q \colon \mathcal{F} \times \mathcal{X} \times \mathcal{Z} \mapsto [-B, B]$ and suppose X_{T+1} is $\mathbf{Z}_1^T$-measurable. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$ and all $\alpha > 0$,*

$$\begin{aligned}&\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] \\ &\le \frac{1}{T}\sum_{t=1}^T q(f, X_{T+1}, Z_t) f(Z_t) + \operatorname{disc}(q) + 2\alpha + 2MB\sqrt{8\frac{\log\frac{\mathbb{E}_{\mathbf{v}\sim\mathcal{T}}[\mathcal{N}_1(\alpha,\mathcal{G},\mathbf{v})]}{\delta}}{T}},\end{aligned}$$

where $\operatorname{disc}(q)$ *is defined by*

$$\operatorname{disc}(q) = \sup_{f \in \mathcal{F}}\left(\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] - \sum_{t=1}^T \mathbb{E}_{Z_t}[q(f, X_{T+1}, Z_t) f(Z_t)|\mathbf{Z}_1^{t-1}]\right), \quad (8)$$

and where $\mathcal{G} = \{(x, z) \mapsto q(f, x, z) f(z) \colon f \in \mathcal{F}\}$.

We illustrate this result with some examples. Consider for instance a Gaussian Markov process with $\mathbf{P}_t(\cdot|\mathbf{Z}_1^T)$ being a normal distribution with mean Z_{t-1} and unit variance. Suppose $f(h, z) = \ell(\|h(x) - y\|_2)$ for some function ℓ. We let $q(h, x', (x, y)) = \exp(-\frac{1}{2}\|y - h(x) - x' + h(x')\|_2^2) / \exp\left(-\frac{1}{2}\|x - y\|_2^2\right)$ and observe that for any f:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{Z_t}[q(X_{T+1}, Z_t) f(Z_t) | \mathbf{Z}_1^{t-1}] \\
&= \int \ell(\|y - h(X_t)\|_2) q(h, X_{T+1}, (X_t, y)) \exp\Big(-\frac{1}{2}\|y - X_t\|_2^2\Big) dy \\
&= \int \ell(\|y - h(X_t)\|_2) \exp\Big(-\frac{1}{2}\|y - h(X_t) - X_{T+1} + h(X_{T+1})\|_2^2\Big) dy \\
&= \int \ell(\|x\|_2) \exp\Big(-\frac{1}{2}\|x - X_{T+1} + h(X_{T+1})\|_2^2\Big) dx \\
&= \int \ell(\|y - h(X_{T+1})\|_2) \exp\Big(-\frac{1}{2}\|y - X_{T+1}\|_2^2\Big) dy \\
&= \mathbb{E}[f(Z_{T+1}) | \mathbf{Z}_1^T],
\end{aligned}
$$

which shows that $\operatorname{disc}(q) = 0$ in this case. More generally, if $\mathbf{Z}$ is a time-homogeneous Markov process, then one can use the Radon-Nikodym derivative $\frac{d\mathbf{P}(\cdot|x')}{d\mathbf{P}(\cdot|x)}(y - h(x) + h(x'))$ for q, which will again lead to zero discrepancy. The major obstacle for this approach is that Radon-Nikodym derivatives are typically unknown and one needs to learn them from data via density estimation, which itself can be a difficult task. In Section 4, we investigate an alternative approach to choosing weights $\mathbf{q}$ based on extending the results of Theorem 1 to hold uniformly over weight vectors $\mathbf{q}$.

4 Kernel-Based Hypotheses with Regression Losses

In this section, we present generalization bounds for kernel-based hypotheses with regression losses. Our results in this section are based on the learning guarantee presented in Corollary 2 in terms of the sequential Rademacher complexity of a class. Our first result is a bound on the sequential Rademacher complexity of the kernel-based hypothesis with regression losses.

Lemma 1 *Let $p \geq 1$ and $\mathcal{F} = \{(\mathbf{x}, y) \to (\mathbf{w} \cdot \Psi(\mathbf{x}) - y)^p : \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$ where $\mathbb{H}$ is a Hilbert space and $\Psi: \mathcal{X} \to \mathbb{H}$ a feature map. Assume that the condition $|\mathbf{w} \cdot \mathbf{x} - y| \leq M$ holds for all $(\mathbf{x}, y) \in \mathcal{Z}$ and all $\mathbf{w}$ such that $\|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda$. Then, the following inequalities hold:*

$$
\mathfrak{R}_T^{seq}(\mathcal{F}) \leq pM^{p-1} C_T \mathfrak{R}_T^{seq}(\mathcal{H}) \leq C_T\Big(pM^{p-1}\Lambda r \|\mathbf{q}\|_2\Big), \tag{9}
$$

where K is a PDS kernel associated to $\mathbb{H}$, $\mathcal{H} = \{\mathbf{x} \to \mathbf{w} \cdot \Psi(\mathbf{x}) : \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$, $r = \sup_x K(x, x)$, and $C_T = 8(1 + 4\sqrt{2}\log^{3/2}(eT^2))$.

Proof We begin the proof by setting $q_t f(\mathbf{z}_t(\boldsymbol{\sigma})) = q_t(\mathbf{w} \cdot \Psi(\mathbf{x}_t(\sigma)) - \mathbf{y}_t(\sigma))^2 = \frac{1}{T}(\mathbf{w} \cdot \mathbf{x}'(\boldsymbol{\sigma}) - \mathbf{y}'_t)^2$, where $\mathbf{x}'_t(\boldsymbol{\sigma}) = \sqrt{Tq_t}\Psi(\mathbf{x}_t(\sigma))$ and $\mathbf{y}'_t(\boldsymbol{\sigma}) = \sqrt{Tq_t}\mathbf{y}_t(\boldsymbol{\sigma})$. We let $\mathbf{z}'_t = (\mathbf{x}'_t, \mathbf{y}'_t)$. Then we observe that

$$\begin{aligned}\mathfrak{R}^{\text{seq}}_T(\mathcal{F}) &= \sup_{\mathbf{z}'=(\mathbf{x}',\mathbf{y}')} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}} \frac{1}{T}\sum_{t=1}^{T} \sigma_t(\mathbf{w}\cdot\mathbf{x}'_t(\boldsymbol{\sigma}) - \mathbf{y}'_t(\boldsymbol{\sigma}))^p\right] \\ &= \sup_{\mathbf{z}=(\mathbf{x},\mathbf{y})} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}} \sum_{t=1}^{T} q_t\sigma_t(\mathbf{w}\cdot\mathbf{x}_t(\boldsymbol{\sigma}) - \mathbf{y}_t(\boldsymbol{\sigma}))^p\right].\end{aligned}$$

Since $x \to |x|^p$ is pM^{p-1}-Lipschitz over $[-M, M]$, by Lemma 13 in [Rakhlin et al., 2015a], the following bound holds:

$$\mathfrak{R}^{\text{seq}}_T(\mathcal{F}) \le pM^{p-1}C_T\mathfrak{R}^{\text{seq}}_T(\mathcal{H}'),$$

where $\mathcal{H}' = \{(\mathbf{x}, y) \to \mathbf{w}\cdot\Psi(\mathbf{x}) - y : \|\mathbf{w}\|_{\mathbb{H}} \le \Lambda\}$. Note that Lemma 13 requires that $\mathfrak{R}^{\text{seq}}_T(\mathcal{H}') > 1/T$ which is guaranteed by Khintchine's inequality. By definition of the sequential Rademacher complexity

$$\begin{aligned}\mathfrak{R}^{\text{seq}}_T(\mathcal{H}') &= \sup_{(\mathbf{x},y)} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}} \sum_{t=1}^{T} \sigma_t q_t(\mathbf{w}\cdot\Psi(\mathbf{x}_t(\boldsymbol{\sigma})) - y(\boldsymbol{\sigma}))\right] \\ &= \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}} \sum_{t=1}^{T} \sigma_t q_t \mathbf{w}\cdot\Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\right] + \sup_{y} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \sigma_t q_t y(\boldsymbol{\sigma})\right] = \mathfrak{R}^{\text{seq}}_T(\mathcal{H}),\end{aligned}$$

where for the last equality we used the fact that σ_ts are mean zero random variables and σ_t is independent of $y(\boldsymbol{\sigma}) = y(\sigma_1, \sigma_2, \ldots, \sigma_{t-1})$. This proves the first result. To prove the second bound we observe that the right-hand side can be bounded as follows:

$$\begin{aligned}\sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}} \sum_{t=1}^{T} \sigma_t q_t \mathbf{w}\cdot\Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\right] &\le \Lambda \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left\|\sum_{t=1}^{T} \sigma_t q_t \Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\right\|_{\mathbb{H}} \\ &\le \Lambda \sup_{\mathbf{x}} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left\|\sum_{t=1}^{T} \sigma_t q_t \Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\right\|^2_{\mathbb{H}}} \\ &= \Lambda \sup_{\mathbf{x}} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{t,s=1}^{T} \sigma_t\sigma_s q_t q_s \Psi(\mathbf{x}_t(\boldsymbol{\sigma}))\cdot\Psi(\mathbf{x}_s(\boldsymbol{\sigma}))\right]} \\ &\le \Lambda \sup_{\mathbf{x}} \sqrt{\sum_{t=1}^{T} q_t^2 \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[K(x_t(\boldsymbol{\sigma}), x_t(\boldsymbol{\sigma}))]} \\ &\le \Lambda r\|\mathbf{q}\|_2,\end{aligned}$$

where again we are using the fact that if $s < t$ then

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_t\sigma_s q_t q_s K(x_t(\sigma), x_s(\sigma))] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_t] \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_s q_t q_s K(x_t(\sigma), x_s(\sigma))] = 0$$

by the independence of σ_t from σ_s, as well as $x_t(\sigma) = x_t(\sigma_1, \ldots, \sigma_{t-1})$ and $x_s(\sigma) = x_s(\sigma_1, \ldots, \sigma_s)$. □

Our next result establishes a high-probability learning guarantee for kernel-based hypothesis. Combining Corollary 2 with Lemma 1, yields the following result.

Theorem 4 *Let $p \geq 1$ and $\mathcal{F} = \{(\mathbf{x}, y) \to (\mathbf{w} \cdot \Psi(\mathbf{x}) - y)^p \colon \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$ where $\mathbb{H}$ is a Hilbert space and $\Psi \colon \mathcal{X} \to \mathbb{H}$ a feature map. Assume that the condition $|\mathbf{w} \cdot \mathbf{x} - y| \leq M$ holds for all $(\mathbf{x}, y) \in \mathcal{Z}$ and all $\mathbf{w}$ such that $\|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda$. If $\mathbf{Z}_1^T = (\mathbf{X}_1^T, \mathbf{Y}_1^T)$ is a sequence of random variables then, for any $\delta > 0$, with probability at least $1-\delta$ the following holds for all $h \in \{\mathbf{x} \to \mathbf{w} \cdot \Psi(\mathbf{x}) \colon \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$:*

$$\mathbb{E}[(h(X_{T+1}) - Y_{T+1})^p | \mathbf{Z}_1^T] \leq \sum_{t=1}^T q_t (h(X_t) - Y_t)^p + \mathrm{disc}(\mathbf{q}) + C_T \Lambda r \|\mathbf{q}\|_2 + 2M\|\mathbf{q}\|_2 \sqrt{8 \log \frac{1}{\delta}},$$

where $C_T = 48pM^p\sqrt{4\pi \log T}(1 + 4\sqrt{2}\log^{3/2}(eT^2))$. Thus, for $p = 2$,

$$\mathbb{E}[(h(X_{T+1}) - Y_{T+1})^2 | \mathbf{Z}_1^T] \leq \sum_{t=1}^T q_t (h(X_t) - Y_t)^2 + \mathrm{disc}(\mathbf{q}) + O\Big((\log^2 T)\Lambda r \|\mathbf{q}\|_2\Big).$$

The results in Theorem 4 (as well as Theorem 1) can be extended to hold uniformly over $\mathbf{q}$ and we provide exact statement in Theorem 6 in Appendix A. This result suggests that we should seek to minimize $\sum_{t=1}^T q_t f(Z_t) + \mathrm{disc}(\mathbf{q})$ over $\mathbf{q}$ and $\mathbf{w}$. This insight is used to develop our algorithmic solutions for forecasting non-stationary time series in Section 6.

5 Estimating Discrepancy

In Section 3, we showed that the discrepancy measure $\mathrm{disc}(\mathbf{q})$ is crucial for forecasting non-stationary time series. In particular, 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 reweight training points, then we could achieve a more favorable learning guarantee for an algorithm trained on this reweighted sample. In some special cases, the discrepancy $\mathrm{disc}(\mathbf{q})$ can be computed analytically. However, in general, we do not have access to the distribution of $\mathbf{Z}_1^T$ and hence we need to estimate the discrepancy from data. Furthermore, in practice, we never observe Z_{T+1} and it is not possible to estimate $\mathrm{disc}(\mathbf{q})$ without some further assumptions. One natural assumption is that the distribution $\mathbf{P}_t$ of Z_t does not change drastically with t on average. Under this assumption, the last s observations $\mathbf{Z}_{T-s+1}^T$ are effectively drawn from the distribution close to $\mathbf{P}_{T+1}$. More precisely, we can write

$$\mathrm{disc}(\mathbf{q}) \leq \sup_{f \in \mathcal{F}} \left(\frac{1}{s} \sum_{t=T-s+1}^T \mathbb{E}[f(Z_t) | \mathbf{Z}_1^{t-1}] - \sum_{t=1}^T q_t \, \mathbb{E}[f(Z_t) | \mathbf{Z}_1^{t-1}] \right) + \sup_{f \in \mathcal{F}} \left(\mathbb{E}[f(Z_{T+1}) | \mathbf{Z}_1^T] - \frac{1}{s} \sum_{t=T-s+1}^T \mathbb{E}[f(Z_t) | \mathbf{Z}_1^{t-1}] \right).$$

We will assume that the second term, denoted by disc_s, is sufficiently small and will show that the first term can be estimated from data. But, we first note that our assumption is necessary for learning in this setting. Observe that the following inequalities hold:

$$\begin{aligned}\sup_{f\in\mathcal{F}}\Big(\mathbb{E}[Z_{T+1}|\mathbf{Z}_1^T]-\mathbb{E}[f(Z_r)|\mathbf{Z}_1^{r-1}]\Big) &\le \sum_{t=r}^{T}\sup_{f\in\mathcal{F}}\Big(\mathbb{E}[f(Z_{t+1})|\mathbf{Z}_1^t]-\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}]\Big)\\ &\le M\sum_{t=r}^{T}\|\mathbf{P}_{t+1}(\cdot|\mathbf{Z}_1^t)-\mathbf{P}_t(\cdot|\mathbf{Z}_1^{t-1})\|_{\mathrm{TV}},\end{aligned}$$

for all $r=T-s+1,\ldots,T$. Therefore, we must have

$$\mathrm{disc}_s \le \frac{1}{s}\sum_{t=T-s+1}\sup_{f\in\mathcal{F}}\Big(\mathbb{E}[Z_{T+1}|\mathbf{Z}_1^T]-\mathbb{E}[f(Z_t)|\mathbf{Z}_1^t]\Big)\le\frac{s+1}{2}M\gamma,$$

where $\gamma=\sup_t\|\mathbf{P}_{t+1}(\cdot|\mathbf{Z}_1^t)-\mathbf{P}_t(\cdot|\mathbf{Z}_1^{t-1})\|_{\mathrm{TV}}$. Barve and Long [1996] showed that $[\text{VC-dim}(\mathcal{H})\gamma]^{\frac{1}{3}}$ is a lower bound on the generalization error in the setting of binary classification where $\mathbf{Z}_1^T$ is a sequence of independent but not identically distributed random variables (drifting).

More precisely, Barve and Long [1996] consider the setting in which the learner observes a sequence $(X_1,Y_1),(X_2,Y_2),\ldots$ from $P_1,P_2,\ldots$ respectively. It is shown that there exists a sufficiently small $\epsilon>0$ and a constant $c>0$ such that if $\gamma>c\epsilon^3/d$ then, for any algorithm $\mathcal{A}$ and any t_0, there exists a sequence $P_1,P_2,\ldots$ and $T>t_0$ such that the generalization error of $h_{\mathcal{A}}$ is at least $\epsilon=\Omega((\gamma d)^{\frac{1}{3}})$, where $h_{\mathcal{A}}$ is the hypothesis produced by $\mathcal{A}$ using the sample $(X_1,Y_1),\ldots,(X_T,Y_T)$.

Observe that this setting is the special case of the general setup considered in our work, since in our case the pairs (X_t,Y_t) are not required to be independent.

The following result shows that we can estimate the first term in the upper bound on $\mathrm{disc}(\mathbf{q})$.

Theorem 5 *Let $\mathbf{Z}_1^T$ be a sequence of random variables. Then, for any $\delta>0$, with probability at least $1-\delta$, the following holds for all $\alpha>0$:*

$$\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(p_t-q_t)\,\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}]\right)\le\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(p_t-q_t)f(Z_t)\right)+B,$$

where $B=2\alpha+M\|\mathbf{q}-\mathbf{p}\|_2\sqrt{8\log\frac{\mathbb{E}_{\mathbf{z}\sim\mathcal{T}}[\mathcal{N}_1(\alpha,\mathcal{F},\mathbf{z})]}{\delta}}$ and where $\mathbf{p}$ is the uniform distribution over the last s points.

Proof The first step consists of upper-bounding the difference of suprema by the supremum of the differences:

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(p_t-q_t)\,\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}]\right)-\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(p_t-q_t)f(Z_t)\right)\\ &\le\sup_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(p_t-q_t)(\mathbb{E}[f(Z_t)|\mathbf{Z}_1^{t-1}]-f(Z_t))\right).\end{aligned}$$

Next, arguments similar to those in the proof of Theorem 1 can be applied to complete the proof. □

Theorem 1 and Theorem 5 combined with the union bound yield the following result.

Corollary 4 *Let $\mathbf{Z}_1^T$ be a sequence of random variables. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f \in \mathcal{F}$ and all $\alpha > 0$:*

$$\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] \leq \sum_{t=1}^{T} q_t f(Z_t) + \widetilde{\mathrm{disc}}(\mathbf{q}) + \mathrm{disc}_s + 4\alpha$$
$$+ M\big[\|\mathbf{q}\|_2 + \|\mathbf{q} - \mathbf{p}\|_2\big]\sqrt{8 \log \tfrac{2\,\mathbb{E}_{\mathbf{v}\sim\mathcal{T}}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},$$

where $\widetilde{\mathrm{disc}}(\mathbf{q}) = \sup_{f\in\mathcal{F}}\Big(\sum_{t=1}^{T}(p_t - q_t)f(Z_t)\Big)$ *and* $\mathbf{p}$ *is the uniform distribution over the last* s *points.*

We note that, while we used a uniform prior $\mathbf{p}$ over the last s points to state Theorem 5 and Corollary 4, any other distribution over the sample points can be used as well. Our choice of $\mathbf{p}$ is based on the assumption that the last s points admit a distribution that is the most similar to the future that we are seeking to predict.

In Section 6, we combine these results with Theorem 6 that extends learning guarantees to hold uniformly over $\mathbf{q}$s to derive novel algorithms for non-stationary time series prediction.

6 Algorithms

In this section, we use our learning guarantees to devise algorithms for forecasting non-stationary time series. We consider a broad family of kernel-based hypothesis classes with regression losses which we analyzed in Section 4.

Suppose L is the squared loss and $\mathcal{H} = \{\mathbf{x} \to \mathbf{w} \cdot \Psi(\mathbf{x}) \colon \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$, where $\Psi \colon \mathcal{X} \to \mathbb{H}$ is a feature mapping from $\mathcal{X}$ to a Hilbert space $\mathbb{H}$. Theorem 4, Theorem 6 and Theorem 4 suggest that we should solve the following optimization problem:

$$\min_{\mathbf{q}\in\mathcal{Q},\mathbf{w}} \left\{ \sum_{t=1}^{T} q_t(\mathbf{w} \cdot \Psi(x_t) - y_t)^2 + \widetilde{\mathrm{disc}}(\mathbf{q}) + \lambda_1 \|\mathbf{w}\|_{\mathbb{H}}^2 \right\} \tag{10}$$

where λ_1 is a regularization hyperparameter and $\mathcal{Q}$ is some convex bounded domain. Note that for simplicity, we have omitted $O(\|\mathbf{q}\|)$ terms in (10). One can extend the formulation in (10) to account for these terms as well.

The optimization problem in (10) is quadratic (and hence convex) in $\mathbf{w}$ and convex in $\mathbf{q}$ since $\widetilde{\mathrm{disc}}(\mathbf{q})$ is a convex function of $\mathbf{q}$ as a supremum of linear functions. However, this objective is not jointly convex in $(\mathbf{q}, \mathbf{w})$. Of course, one could apply a block coordinate descent to solve this objective which alternates between optimizing over $\mathbf{q}$ and solving a QP in $\mathbf{w}$. In general, no convergence guarantees can be provided for this algorithm. In addition, each $\mathbf{q}$-step involves finding $\widetilde{\operatorname{disc}}(\mathbf{q})$ which in itself may be a costly procedure. In the sequel we address both of these concerns.

First, let us observe that if $a(\mathbf{w}) = \sum_{t=1}^{T} p_t(\mathbf{w}\cdot\Psi(x_t) - y_t)^2$ with p_t being a uniform distribution on the last s points, then by definition of empirical discrepancy

$$\begin{aligned}\widetilde{\operatorname{disc}}(\mathbf{w}) &= \sup_{\mathbf{w}\leq\Lambda}\left(a(\mathbf{w}) - \sum_{t=1}^{T} q_t(\mathbf{w}\cdot\Psi(x_t) - y_t)^2\right)\\ &= \sup_{\mathbf{w}\leq\Lambda}\left(\sum_{t=1}^{T}(u_t - q_t)a(\mathbf{w}) - \sum_{t=1}^{T} q_t((\mathbf{w}\cdot\Psi(x_t) - y_t)^2 - a(\mathbf{w}))\right)\\ &\leq \sum_{t=1}^{T} q_t d_t + \lambda_2\|\mathbf{q}-\mathbf{v}\|_p\end{aligned}$$

where λ_2 is some constant (a hyperparameter), $\mathbf{v}$ is a prior typically chosen to uniform weights $\mathbf{u}$, $p \geq 1$ and d_ts are instantaneous discrepancies defined by

$$d_t = \sup_{\mathbf{w}\leq\Lambda}\left|a(\mathbf{w}) - (\mathbf{w}\cdot\Psi(x_t) - y_t)^2\right|.$$

Note that d_t can also be defined in terms of windows averages, where $(\mathbf{w}\cdot\Psi(x_t) - y_t)^2$ is replaced with $\frac{1}{2l}\sum_{s=t-l}^{t+l}(\mathbf{w}\cdot\Psi(x_s) - y_s)^2$ for some l.

This leads to the following optimization problem:

$$\min_{\mathbf{q}\in\mathcal{Q},\mathbf{w}}\left\{\sum_{t=1}^{T} q_t(\mathbf{w}\cdot\Psi(x_t) - y_t)^2 + \sum_{t=1}^{T} d_t q_t + \lambda_1\|\mathbf{w}\|_{\mathbb{H}}^2 + \lambda_2\|\mathbf{q}-\mathbf{v}\|_p\right\}. \tag{11}$$

This optimization problem is still not convex but now d_ts can be precomputed before solving (11) which may be considerably more efficient. For instance, the $\mathbf{q}$-step in the block coordinate descent reduces to a simple LP. Below we show how (11) can be turned into a convex optimization problem when $\mathcal{Q} = [0,1]^T$.

6.1 Convex Optimization over $[0,1]^T$ and Dual Problems

In this section, we consider the case when $\mathcal{Q} = [0,1]^T$ and we show how (11) can be turned into a convex optimization problem which then can be solved efficiently. We apply change of variables $r_t = 1/q_t$, which leads to the following optimization problem:

$$\min_{\mathbf{r}\in\mathcal{D},\mathbf{w}}\left\{\sum_{t=1}^{T}\frac{(\mathbf{w}\cdot\Psi(x_t) - y_t)^2 + d_t}{r_t} + \lambda_1\|\mathbf{w}\|_{\mathbb{H}}^2 + \lambda_2\left(\sum_{t=1}^{T}|r_t^{-1} - v_t|^p\right)^{1/p}\right\}, \tag{12}$$

where $\mathcal{D} = \{\mathbf{r}\colon r_t \geq 1\}$. We can remove the $(\cdot)^{1/p}$ on the last term by first turning it into a constraint, raising it to the pth power and then moving it back to the objective by introducing a Lagrange multiplier:

$$\min_{\mathbf{r}\in\mathcal{D},\mathbf{w}} \left\{ \sum_{t=1}^{T} \frac{(\mathbf{w}\cdot\Psi(x_t) - y_t)^2 + d_t}{r_t} + \lambda_1 \|\mathbf{w}\|_{\mathbb{H}}^2 + \lambda_2 \sum_{t=1}^{T} |r_t^{-1} - v_t|^p \right\}. \tag{13}$$

Note that the first two terms in (12) are jointly convex in $(\mathbf{r}, \mathbf{w})$: the first term is a sum of quadratic-over-linear functions which is convex and the second term is a squared norm which is again convex.

The last step is to observe that $|r_t^{-1} - v_t| = |v_t r_t^{-1}|^p |r_t - v_t^{-1}|^p \leq v_t^p |r_t - v_t|^p$ since $r_t^{-1} \leq 1$. Therefore, we have the following optimization problem:

$$\min_{\mathbf{r}\in\mathcal{D},\mathbf{w}} \left\{ \sum_{t=1}^{T} \frac{(\mathbf{w}\cdot\Psi(x_t) - y_t)^2 + d_t}{r_t} + \lambda_1 \|\mathbf{w}\|_{\mathbb{H}}^2 + \lambda_2 \sum_{t=1}^{T} v_t^p |r_t - v_t^{-1}|^p \right\}, \tag{14}$$

which is jointly convex over $(\mathbf{r}, \mathbf{w})$.

For many real-world problems, Ψ is specified implicitly via a PDS kernel K and it is computationally advantageous to consider the dual formulation of (14). Using the dual problem associated to $\mathbf{w}$ (14) can be rewritten as follows:

$$\min_{\mathbf{r}\in\mathcal{D}} \left\{ \max_{\boldsymbol{\alpha}} \left\{ -\lambda_1 \sum_{t=1}^{T} r_t \alpha_t^2 - \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + 2\lambda_1 \boldsymbol{\alpha}^T \mathbf{Y} \right\} + \sum_{t=1}^{T} \frac{d_t}{r_t} + \lambda_2 \sum_{t=1}^{T} v_t^p |r_t - v_t^{-1}|^p \right\}, \tag{15}$$

where $\mathbf{K}$ is the kernel matrix and $\mathbf{Y} = (y_1, \ldots, y_T)^T$. We provide a full derivation of this result in Appendix B.

Both (14) and (15) can be solved using standard descent methods, where, at each iteration, we solve a standard QP in $\boldsymbol{\alpha}$ or $\mathbf{w}$, which admits a closed-form solution.

6.2 Discrepancy Computation

The final ingredient that is needed to solve optimization problems (14) or (15) is an algorithm to find instantaneous discrepancies d_ts. Recall that in general these are defined as

$$\sup_{\mathbf{w}'\leq\Lambda} \left| \sum_{s=1}^{T} p_s \ell(\mathbf{w}'\cdot\Psi(x_s) - y_s) - \ell(\mathbf{w}'\cdot\Psi(x_t) - y_t) \right| \tag{16}$$

where ℓ is some specified loss function. For an arbitrary ℓ this may be a difficult optimization problem. However, observe that if ℓ is a convex function then the objective in (16) is a difference of convex functions so we may use DC-programming to solve this problem. For general loss functions, the DC-programming approach only guarantees convergence to a stationary point. However, for the squared loss, our problem can be cast as an instance of the trust region problem, which can be solved globally using the DCA algorithm of Tao and An [1998].

6.3 Two-Stage Algorithms

An alternative simpler algorithm based on the data-dependent bounds of Corollary 4 consists of first finding a distribution $\mathbf{q}$ minimizing the discrepancy and then using that to find a hypothesis minimizing the (regularized) weighted empirical risk. This leads to the following two-stage procedure. First, we find a solution $\mathbf{q}^*$ of the following convex optimization problem:

$$\min_{\mathbf{q}\geq 0}\left\{\sup_{\mathbf{w}'\leq \Lambda}\left(\sum_{t=1}^{T}(p_t-q_t)(\mathbf{w}'\cdot\Psi(x_t)-y_t)^2\right)\right\}, \tag{17}$$

where Λ is parameter that can be selected via cross-validation (for example using techniques in [Kuznetsov and Mohri, 2016]). Our generalization bounds hold for arbitrary weights $\mathbf{q}$ but we restrict them to being positive sequences. Note that other regularization terms such as $\|\mathbf{q}\|_2^2$ and $\|\mathbf{q}-\mathbf{p}\|_2^2$ from the bound of Corollary 4 can be incorporated in the optimization problem, but we discard them to minimize the number of parameters. This problem can be solved using standard descent optimization methods, where, at each step, we use DC-programming to evaluate the supremum over $\mathbf{w}'$.

The solution $\mathbf{q}^*$ of (17) is then used to solve the following (weighted) kernel ridge regression problem:

$$\min_{\mathbf{w}}\left\{\sum_{t=1}^{T}q_t^*(\mathbf{w}\cdot\Psi(x_t)-y_t)^2+\lambda_2\|\mathbf{w}\|_{\mathbb{H}}^2\right\}. \tag{18}$$

Note that, in order to guarantee the convexity of this problem, we require $\mathbf{q}^*\geq 0$. We note that one can also use ϵ-sensitive loss instead of squared loss for this algorithm.

6.4 Further Extensions and Discussion

We conclude this section with the observation that the learning guarantees that we presented in Section 3 and Section 4 can be used to derive algorithms for many other problems that involve time series data with loss functions and hypothesis set distinct from regression losses and linear hypotheses considered in this section.

For instance, we can choose a hypothesis set $\mathcal{H}$ to be a set of neural networks with pre-specified architecture. In that case, we can apply stochastic gradient descent (SGD) to solve the optimization problem (10). Note that computing $\mathrm{disc}(\mathbf{q})$ in that case is almost identical to solving the Wasserstein GAN optimization problem [Arjovsky et al., 2017]. However, no convergence guarantees are known for this setting. We leave these extensions to future work.

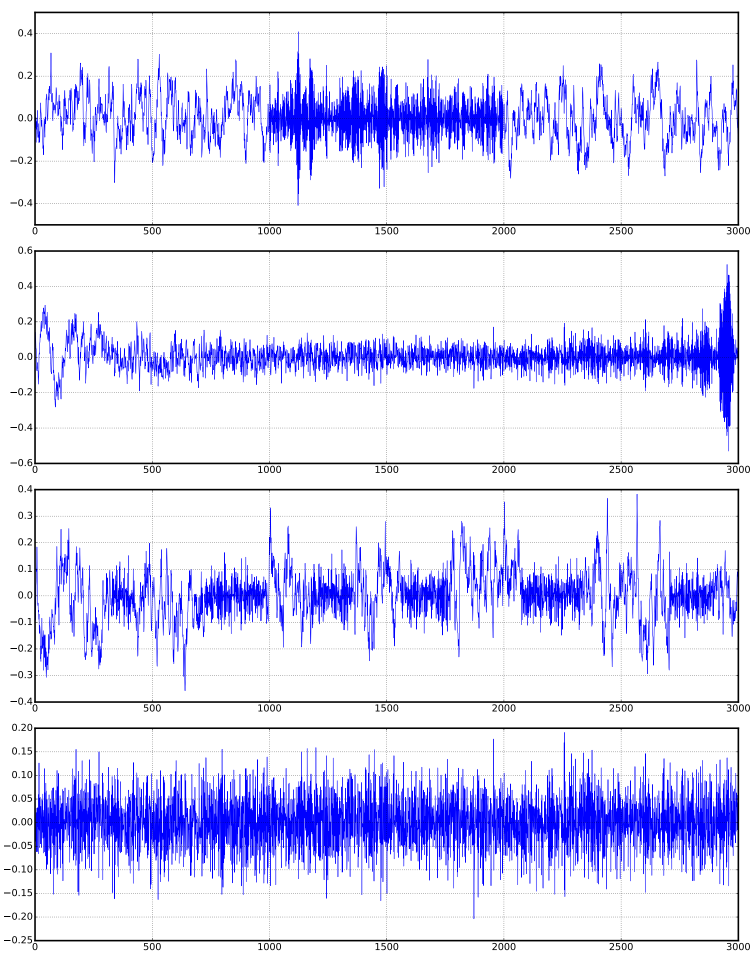

Fig. 3 Synthetic datasets (top to bottom): `ads1`, `ads2`, `ads3`, `ads4`.

7 Experiments

In this section, we present the results of experiments evaluating our algorithmic solutions on a number of synthetic and real-world datasets. In particular, we consider the one-stage algorithm presented in Section 6 which is based on solving the optimization problem in (11). While solving this problem as opposed to (14) may result in a sub-optimal results, this simplification allows us to use an alternating optimization method described in Section 6: for a fixed $\mathbf{q}$, problem (11) is a simple QP over $\mathbf{w}$ and, for a fixed $\mathbf{w}$, the problem reduces to an LP in $\mathbf{q}$. This iterative scheme admits a straightforward implementation using existing QP and LP solvers. In the rest of this section, we will refer to this algorithm as *discrepancy-based forecaster* (DBF).

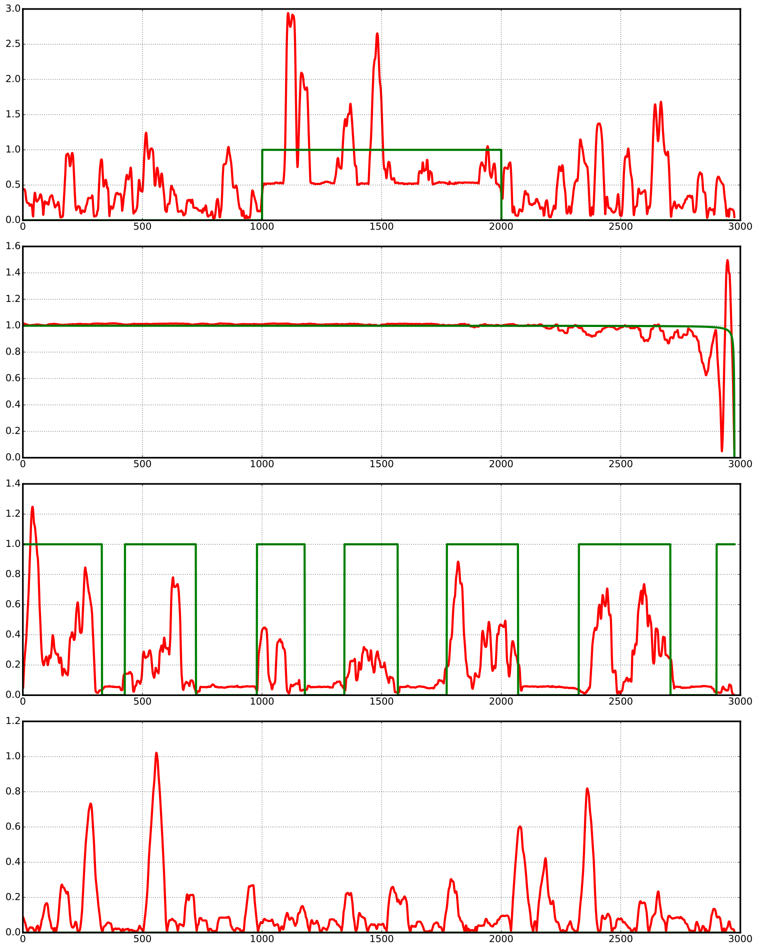

Fig. 4 True (green) and estimated (red) instantaneous discrepancies for synthetic data (top to bottom): `ads1`, `ads2`, `ads3`, `ads4`.

We have chosen ARIMA models as a baseline comparator in our experiments. These models are standard and are commonly used in practice for forecasting non-stationary time series.

We present two sets of experiments: synthetic data experiments (Section 7.1) and real-world data experiments (Section 7.2).

7.1 Experiments with Synthetic Data

In this section, we present the results of our experiments on some synthetic datasets. These experimental results allow us to gain some further understanding of the discrepancy-based approach to forecasting. In particular, they help

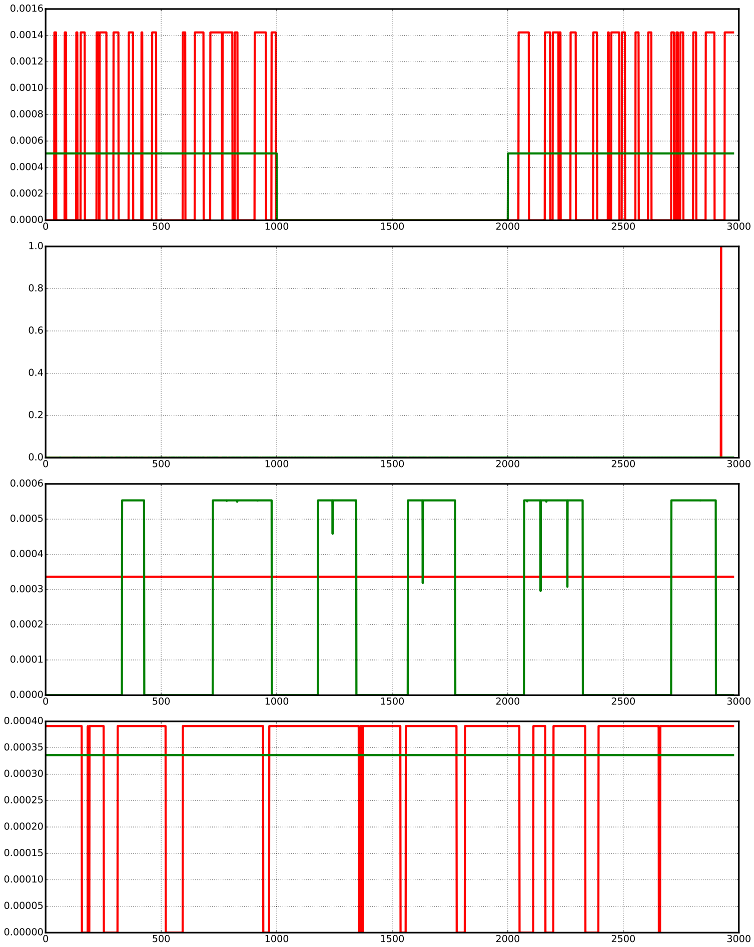

Fig. 5 Weights **q** chosen by DBF when used with true (green) and estimated (red) instantaneous discrepancies for synthetic data (top to bottom): `ads1`, `ads2`, `ads3`, `ads4`.

us study the effects of using estimated instantaneous discrepancies instead of the true ones.

We have used four artificial datasets: `ads1`, `ads2`, `ads3` and `ads4`. These datasets are generated using the following autoregressive processes:

$$\begin{aligned}
&\texttt{ads1:}\ \ Y_t = \alpha_t Y_{t-1} + \epsilon_t, \qquad \alpha_t = -0.9 \text{ if } t \in [1000, 2000] \text{ and } 0.9 \text{ otherwise},\\
&\texttt{ads2:}\ \ Y_t = \alpha_t Y_{t-1} + \epsilon_t, \qquad \alpha_t = 1 - (t/1500),\\
&\texttt{ads3:}\ \ Y_t = \alpha_{i(t)} Y_{t-1} + \epsilon_t, \quad \alpha_1 = -0.5 \text{ and } \alpha_2 = 0.9\\
&\texttt{ads4:}\ \ Y_t = -0.5 Y_{t-1} + \epsilon_t,
\end{aligned}$$

where ϵ_t are independent Gaussian random variables with zero mean and $\sigma = 0.05$. Note that $i(t)$ in the definition of `ads3` is a stochastic process on $\{1, 2\}$,

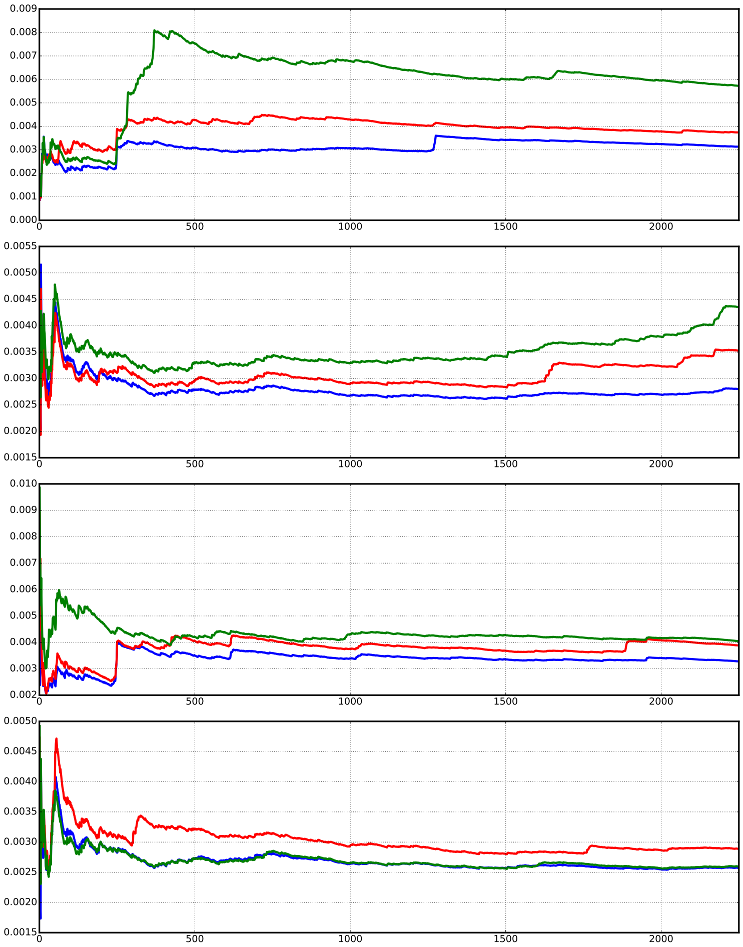

Fig. 6 Running MSE for synthetic data experiments (top to bottom): `ads1`, `ads2`, `ads3`, `ads4`. For each time t on the horizontal axis we plot MSE up to time t of `tDBF` (blue), `eDBF` (red), `ARIMA` (green).

such that $\mathbb{P}(i(t+s) = i | i(t+s-1) = \ldots = i(s) = i, i(s-1) \neq i) = (0.99995)^t$. In other words, if the process $i(t)$ spends exactly τ last time steps in state i, then at the next time step it will stay in i with probability $(0.99995)^\tau$ and will move to a different state with probability $1 - (0.99995)^\tau$.

The first stochastic process (`ads1`) is supposed to model sudden abrupt changes in the data generating mechanism. The scenario in which parameters of the data generating process smoothly drift is modeled by `ads2`. The setting in which the changes can occur at random times is captured by `ads3`. Finally, `ads4` is generated by a stationary random process. See Figure 3.

Table 1 Mean squared error (standard deviation) for synthetic data. `tDBF` is DBF with true instantaneous discrepancies d_t as its input. `eDBF` is DBF with estimated instantaneous discrepancies d_t as its input. The results in bold are statistically significant using one-sided paired t-test at 5% level.

Dataset	`tDBF`	`eDBF`	`ARIMA`
`ads1`	$\mathbf{3.135 \times 10^{-3}}$ $(\mathbf{7.504 \times 10^{-3}})$	3.743×10^{-3} (6.171×10^{-3})	5.723×10^{-3} (10.143×10^{-3})
`ads2`	$\mathbf{2.800 \times 10^{-3}}$ $(\mathbf{3.930 \times 10^{-3}})$	3.530×10^{-3} (6.620×10^{-3})	4.348×10^{-3} (6.770×10^{-3})
`ads3`	$\mathbf{3.282 \times 10^{-3}}$ $(\mathbf{6.417 \times 10^{-3}})$	3.887×10^{-3} (9.277×10^{-3})	4.066×10^{-3} (6.122×10^{-3})
`ads4`	2.573×10^{-3} (3.516×10^{-3})	2.889×10^{-3} (4.262×10^{-3})	2.593×10^{-3} (3.578×10^{-3})

For each dataset, we have generated time series with 3,000 sample points. In all our experiments, we used the following protocol. For each $t \in [750, 775, \ldots, 2995]$, $(y_1, \ldots, y_t)$ is used as a development set and $(y_{t+1}, \ldots, y_{t+25})$ is used as a test set. On the development set, we first train each algorithm with different hyperparameter settings on $(y_1, \ldots, y_{t-25})$ and then select the best performing hyperparameters on $(y_{t-24}, \ldots, y_t)$. This set of hyperparameters is then used for training on the full development set $(y_1, \ldots, y_t)$ and mean squared error (MSE) on $(y_{t+1}, \ldots, y_{t+25})$ averaged over all $t \in [750, 775, \ldots, 2995]$ is reported.

Recall that DBF algorithm requires two regularization hyperparameters λ_1 and λ_2. We optimized these parameters over the following two sets of values for λ_1 and λ_2 respectively: $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and $\{100, 10, 1, 0.1, 0.05, 0.01, 0.001, 0\}$.

ARIMA models have three hyperparameters p, d and q that we also set via the validation procedure described above. Recall that ARIMA(p, d, q) is a generative model defined by the following autoregression:

$$\left(1 - \sum_{i=0}^{p-1} \phi_i \mathfrak{L}^i\right)(1 - \mathfrak{L})^d Y_{t+1} = \left(1 + \sum_{j=1}^{q} \theta_j \mathfrak{L}^j\right)\epsilon_t.$$

where $\mathfrak{L}$ is a lag operator, that is, $\mathfrak{L} Y_t = Y_{t-1}$. Therefore, validation over (p, d, q) is equivalent to validation over different sets of features used to train the model. For instance, $(p, d, q) = (3, 0, 0)$ means that we are using $(y_{t-3}, y_{t-2}, y_{t-1})$ as our features while $(p, d, q) = (2, 1, 0)$ corresponds to $(y_{t-3} - y_{t-2}, y_{t-2} - y_{t-1})$. For DBF, we fix feature vectors to be $(y_{t-3}, y_{t-2}, y_{t-1})$. For ARIMA, we optimize over $p, d, q \in \{0, 1, 2\}^3$. We use the maximum likelihood approach to estimate the unknown parameters of ARIMA models.

Finally, observe that the discrepancy estimation procedure discussed in Section 6 also requires a hyperparameter s representing the length of the most recent history of the stochastic process. We did not make an attempt to optimize this parameter and in all of our experiments we set $s = 20$.

The results of our experiments are presented in Table 1. We have compared DBF with true discrepancies as its input `tDBF`, DBF with estimated discrepancies as its input `eDBF` and ARIMA. In all experiments with non-stationary processes (`ads1`, `ads2`, `ads3`), `tDBF` performs better than both `eDBF` and ARIMA.

Table 2 Real-world datasets statistics

Dataset	URL	Size
`bitcoin`	`https://www.quandl.com/data/BCHARTS/BTCNCNY`	1705
`coffee`	`https://www.quandl.com/data/COM/COFFEE_BRZL`	2205
`eur/jpy`	`https://www.quandl.com/data/ECB/EURJPY`	4425
`jpy/usd`	`http://data.is/269FpLF`	4475
`mso`	`http://data.is/269F3EV`	1235
`silver`	`https://www.quandl.com/data/COM/AG_EIB`	2251
`soy`	`https://www.quandl.com/data/COM/SOYB_1`	2218
`temp`	`http://data.is/1qlX2AN`	3649

Table 3 Mean squared error (standard deviation) for real-world data. The results in bold are statistically significant using one-sided paired t-test at 5% level.

Dataset	DBF	ARIMA
`bitcoin`	$\mathbf{4.400 \times 10^{-3}}$ $(\mathbf{26.500 \times 10^{-3}})$	4.900×10^{-3} (29.990×10^{-3})
`coffee`	$\mathbf{3.080 \times 10^{-3}}$ $(\mathbf{6.570 \times 10^{-3}})$	3.260×10^{-3} (6.390×10^{-3})
`eur/jpy`	$\mathbf{7.100 \times 10^{-5}}$ $(\mathbf{16.900 \times 10^{-5}})$	7.800×10^{-5} (24.200×10^{-5})
`jpy/usd`	$\mathbf{9.770 \times 10^{-1}}$ $(\mathbf{25.893 \times 10^{-1}})$	10.004×10^{-1} (27.531×10^{-1})
`mso`	32.876×10^{0} (55.586×10^{0})	32.193×10^{0} (51.109×10^{0})
`silver`	$\mathbf{7.640 \times 10^{-4}}$ $(\mathbf{46.65 \times 10^{-4}})$	34.180×10^{-4} (158.090×10^{-4})
`soy`	5.071×10^{-2} (9.938×10^{-2})	5.003×10^{-2} (10.097×10^{-2})
`temp`	6.418×10^{0} (9.958×10^{0})	6.461×10^{0} (10.324×10^{0})

Similarly, `eDBF` outperforms ARIMA on the same datasets. These results are statistically significant at 5%-level using one-sided paired t-test. Figure 6 illustrates the dynamics of MSE as a function of time t for all three algorithms on all of the synthetic datasets.

Our results suggest that accurate discrepancy estimation can lead to a significant improvement in performance. We present the results of discrepancy estimation for our experiments in Figure 4. Figure 5 shows the corresponding weights $\mathbf{q}$ chosen by DBF.

7.2 Experiments with Real-World Data

In this section, we present the results of our experiments with real-world datasets. For our experiments, we have chosen eight time series from different domains: currency exchange rates (`bitcoin`, `eur/jpy`, `jpy/usd`), commodity prices (`coffee`, `soy`, `silver`) and meteorology (`mso`, `temp`). Further details of these datasets are summarized in Table 2.

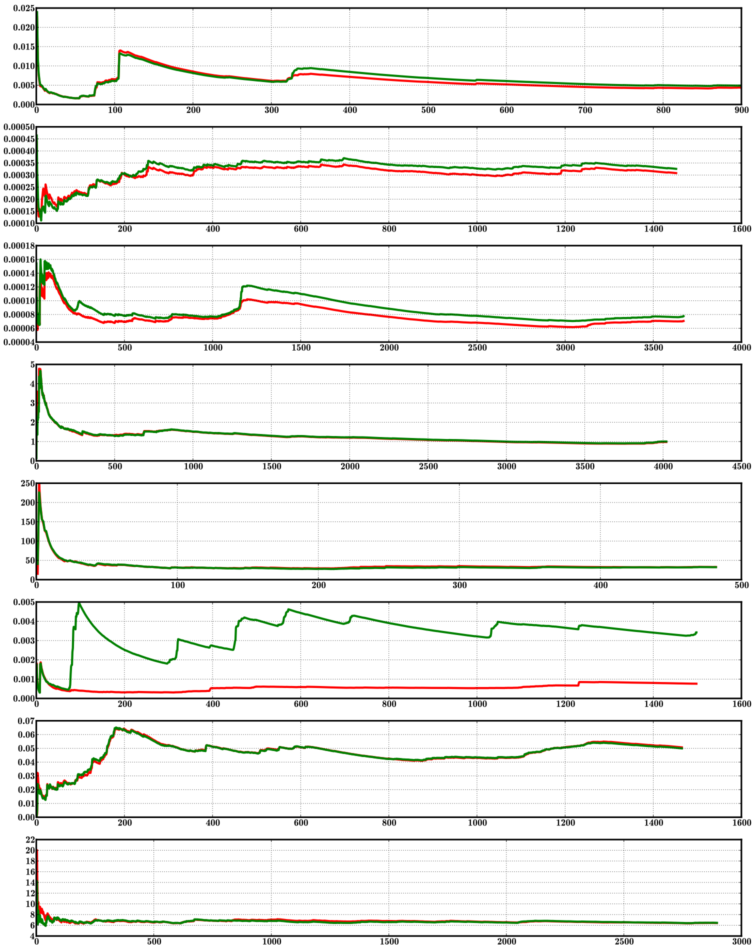

Fig. 7 Running MSE for real-world data experiments (top to bottom): `bitcoin`, `coffee`, `eur/jpy`, `jpy/usd`, `mso`, `silver`, `soy`, `temp`. For each time t on the horizontal axis we plot MSE up to time t of `DBF` (red) and `ARIMA` (green).

In all our experiments, we used the same protocol as in the previous section. In particular, for each $t \in [750, 775, \ldots,$ $(y_1, \ldots, y_t)$ is used as a development set and $(y_{t+1}, \ldots, y_{t+25})$ is used as a test set. On the development set, we first train each algorithm with different hyperparameter settings on $(y_1, \ldots, y_{t-25})$ and then select the best performing hyperparameters on $(y_{t-24}, \ldots, y_t)$. This set of hyperparameters is then used for training on the full development set $(y_1, \ldots, y_t)$ and mean squared error (MSE) over the remaining data points is reported. The range of the hyperparameters for both DBF and ARIMA is also the same as in previous section. Note that since true discrepancies are unknown, we only present the results for DBF with discrepancies estimated from data.

The results of our experiments are summarized in Table 3. Observe that DBF outperforms ARIMA on 5 out of 8 datasets. It should be noted that the error variance is high compared to the mean error. However, the higher variance is likely due to inherent low signal-to-noise ratio in these real world datasets.

Figure 7 illustrates the dynamics of MSE as a function of time t for all three algorithms on all of the synthetic datasets. These results are statistically significant at 5%-level using one-sided paired t-test. There is no statistical difference in the performance of ARIMA and DBF on the rest of the datasets.

Our results suggest that discrepancy-based approach to prediction of non-stationary time series may lead to improved performance compared to other traditional approaches such as ARIMA.

8 Conclusion

We presented a general theoretical analysis of learning in the broad scenario of non-stationary non-mixing processes, the realistic setting for a variety of applications. We discussed in detail several algorithms benefiting from the learning guarantees presented. Our theory can also provide a finer analysis of several existing algorithms and help devise alternative principled learning algorithms.

The key ingredient of our analysis and algorithms is the notion of discrepancy. This is as a fundamental concept for learning with non-stationary stochastic processes and was used in the analysis of several other time series forecasting techniques [Kuznetsov and Mohri, 2017b; Kuznetsov and Mariet, 2019; Zimin and Lampert, 2017; Kuznetsov and Mohri, 2016] following the earlier work in [Kuznetsov and Mohri, 2015].

Acknowledgements This work was partly funded by NSF CCF-1535987, IIS-1618662, and a Google Research Award.

References

Terrence M. Adams and Andrew B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38 (4):1345–1367, 2010.

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

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.

Donald Andrews. First order autoregressive processes and strong mixing. Cowles Foundation Discussion Papers 664, Cowles Foundation for Research in Economics, Yale University, 1983.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, 2017. URL http://arxiv.org/abs/1701.07875.

Richard Baillie. Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73(1):5–59, 1996.

Rakesh D. Barve and Philip M. Long. On the complexity of learning from drifting distributions. In *COLT*, 1996.

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

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J Econometrics*, 1986.

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

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, 1986.

Likai Chen and Wei Biao Wu. Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research*, 18(231): 1–46, 2018.

Victor H. De la Peña and Evarist Giné. *Decoupling: from dependence to independence: randomly stopped processes, U-statistics and processes, martingales and beyond*. Probability and its applications. Springer, NY, 1999.

Paul Doukhan. *Mixing: properties and examples*. Lecture notes in statistics. Springer-Verlag, New York, 1994.

Robert Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

James D. Hamilton. *Time series analysis*. Princeton, 1994.

Vitaly Kuznetsov and Zelda Mariet. Foundations of sequence-to-sequence modeling for time series. In *AISTATS*, 2019.

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

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

Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and on-line learning. In *COLT*, 2016.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017a.

Vitaly Kuznetsov and Mehryar Mohri. Discriminative state space models. In *NIPS*, 2017b.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete. U.S. Government Printing Office, 1991.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 1987.

E. N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26:636–646, 1969.

Aurelie C. Lozano, Sanjeev R. Kulkarni, and Robert E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In *NIPS*, 2006.

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 *ALT*, 2012.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *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.

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

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

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *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 Kontorovich. Predictive PAC learning and process decompositions. In *NIPS*, 2013.

Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. In *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.

Zhizhen Zhao and Dimitrios Giannakis. Analog forecasting with dynamics-adapted kernels. *CoRR*, abs/1412.3831, 2016.

Alexander Zimin and Christoph Lampert. Learning Theory for Conditional Risk Minimization. In *AISTAT*, 2017.

A Proofs

In the construction described in Section 2.1, we denoted by $\mathbf{z}$ the tree defined using Z_ts and denoted by $\mathcal{T}$ the distribution of $\mathbf{z}$. Here, we will also denote by $\mathbf{z}'$ the tree formed by Z'_ts and denote by $\overline{\mathcal{T}}$ the joint distribution of $(\mathbf{z}, \mathbf{z}')$.

Lemma 2 *Let $\mathbf{Z}_1^T$ be a sequence of random variables and let $\mathbf{Z}_1'^T$ be a decoupled tangent sequence. Then, for any measurable function G, the following equality holds:*

$$\mathbb{E}\bigg[G\Big(\sup_{f\in\mathcal{F}}\sum_{t=1}^{T} q_t\big(f(Z'_t)-f(Z_t)\big)\Big)\bigg] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')\sim\overline{\mathcal{T}}}\Big[G\Big(\sup_{f}\sum_{t=1}^{T}\sigma_t q_t\big(f(\mathbf{z}'_t(\boldsymbol{\sigma}))-f(\mathbf{z}_t(\boldsymbol{\sigma}))\big)\Big)\Big]. \quad (19)$$

The result also holds with the absolute value around the sums in (19).

Proof The proof follows an argument invoked in the proof of Theorem 3 of Rakhlin et al. [2011]. We only need to check that every step holds for an arbitrary weight vector $\mathbf{q}$, in lieu of the uniform distribution vector $\mathbf{u}$, and for an arbitrary measurable function G, instead of the identity function. Let $\mathbf{p}$ denote the joint distribution of the random variables Z_ts. Observe that we can write the left-hand side of (19) as follows:

$$\mathbb{E}\Big[G\Big(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\Big)\Big] = \mathop{\mathbb{E}}_{Z_1,Z'_1\sim\mathbf{p}_1}\ \mathop{\mathbb{E}}_{Z_2,Z'_2\sim\mathbf{p}_2(\cdot|Z_1)}\cdots\mathop{\mathbb{E}}_{Z_T,Z'_T\sim\mathbf{p}_T(\cdot|\mathbf{Z}_1^{T-1})}\Big[G\Big(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\Big)\Big],$$

where $\boldsymbol{\sigma} = (1,\ldots,1) \in \{\pm1\}^T$ and $\Sigma(\boldsymbol{\sigma}) = \sum_{t=1}^{T}\sigma_t q_t(f(Z'_t) - f(Z_t))$. Now, by definition of decoupled tangent sequences, the value of the last expression is unchanged if we swap the sign of any σ_{i-1} to -1 since that is equivalent to permuting Z_i and Z'_i. Thus, the last expression is in fact equal to

$$\mathop{\mathbb{E}}_{Z_1,Z'_1\sim\mathbf{p}_1}\ \mathop{\mathbb{E}}_{Z_2,Z'_2\sim\mathbf{p}_2(\cdot|S_1(\sigma_1))}\cdots\mathop{\mathbb{E}}_{Z_T,Z'_T\sim\mathbf{p}_T(\cdot|S_1(\sigma_1),\ldots,S_{T-1}(\sigma_{T-1}))}\Big[G\Big(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\Big)\Big]$$

for any sequence $\boldsymbol{\sigma} \in \{\pm1\}^T$, where $S_t(1) = Z_t$ and Z'_t otherwise. Since this equality holds for any $\boldsymbol{\sigma}$, it also holds for the mean with respect to uniformly distributed $\boldsymbol{\sigma}$. Therefore, the last expression is equal to

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\ \mathop{\mathbb{E}}_{Z_1,Z'_1\sim\mathbf{p}_1}\ \mathop{\mathbb{E}}_{Z_2,Z'_2\sim\mathbf{p}_2(\cdot|S_1(\sigma_1))}\cdots\mathop{\mathbb{E}}_{Z_T,Z'_T\sim\mathbf{p}_T(\cdot|S_1(\sigma_1),\ldots,S_{T-1}(\sigma_{T-1}))}\Big[G\Big(\sup_{f\in\mathcal{F}}\Sigma(\boldsymbol{\sigma})\Big)\Big].$$

This last expectation coincides with the expectation with respect to drawing a random tree $\mathbf{z}$ and its tangent tree $\mathbf{z}'$ according to $\overline{\mathcal{T}}$ and a random path $\boldsymbol{\sigma}$ to follow in that tree. That is, the last expectation is equal to

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\ \mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{z}')\sim\overline{\mathcal{T}}}\Big[G\Big(\sup_{f}\sum_{t=1}^{T}\sigma_t q_t\big(f(\mathbf{z}'_t(\boldsymbol{\sigma}))-f(\mathbf{z}_t(\boldsymbol{\sigma}))\big)\Big)\Big],$$

which concludes the proof. □

Theorem 6 *Let $p \geq 1$ and $\mathcal{F} = \{(\mathbf{x}, y) \to (\mathbf{w}\cdot\Psi(\mathbf{x}) - y)^p : \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$ where $\mathbb{H}$ is a Hilbert space and $\Psi : \mathcal{X} \to \mathbb{H}$ a feature map. Assume that the condition $|\mathbf{w}\cdot\mathbf{x} - y| \leq M$ holds for all $(\mathbf{x}, y) \in \mathcal{Z}$ and all $\mathbf{w}$ such that $\|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda$. Fix $\mathbf{q}^*$. Then, if $\mathbf{Z}_1^T = (\mathbf{X}_1^T, \mathbf{Y}_1^T)$ is a sequence of random variables, for any $\delta > 0$, with probability at least $1-\delta$, the following bound holds for all $h \in \mathcal{H} = \{\mathbf{x} \to \mathbf{w}\cdot\Psi(\mathbf{x}) : \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$ and all $\mathbf{q}$ such that $0 < \|\mathbf{q} - \mathbf{q}^*\|_1 \leq 1$:*

$$\mathbb{E}[(h(X_{T+1}) - Y_{T+1})^p|\mathbf{Z}_1^T] \leq \sum_{t=1}^{T} q_t(h(X_t) - Y_t)^p + \mathrm{disc}(\mathbf{q}) + G(\mathbf{q}) + 4M\|\mathbf{q} - \mathbf{q}^*\|_1$$

where $G(\mathbf{q}) = 4M\Big(\sqrt{8\log\frac{2}{\delta}} + \sqrt{2\log\log_2 2(1-\|\mathbf{q}-\mathbf{q}^*\|_1)^{-1}} + \widetilde{C}_T \Lambda r\Big)\Big(\|\mathbf{q}^*\|_2 + 2\|\mathbf{q}-\mathbf{q}^*\|_1\Big)$ *and* $\widetilde{C}_T = 48pM^p\sqrt{4\pi\log T}(1+4\sqrt{2}\log^{3/2}(eT^2))$*. Thus, for* $p=2$*,*

$$\begin{aligned}\mathbb{E}[(h(X_{T+1})-Y_{T+1})^2|\mathbf{Z}_1^T] \le {} & \sum_{t=1}^T q_t(h(X_t)-Y_t)^2 + \operatorname{disc}(\mathbf{q}) \\ & + O\Bigg(\Lambda r(\log^2 T)\sqrt{\log\log_2 2(1-\|\mathbf{q}-\mathbf{q}^*\|_1)^{-1}}\Big(\|\mathbf{q}^*\|_2 + \|\mathbf{q}-\mathbf{q}^*\|_1\Big)\Bigg).\end{aligned}$$

This result extends Theorem 4 to hold uniformly over $\mathbf{q}$. Similarly, one can prove an analogous extension for Theorem 1. This result suggests that we should try to minimize $\sum_{t=1}^T q_t f(Z_t) + \operatorname{disc}(\mathbf{q})$ over $\mathbf{q}$ and $\mathbf{w}$. This bound is in certain sense analogous to margin bounds: it is the most favorable when there exists a good choice for $\mathbf{q}^*$ and we hope to find $\mathbf{q}$ that is going to be close to this weight vector. These insights are used to develop our algorithmic solutions for forecasting non-stationary time series in Section 6.

Proof Let $(\epsilon_k)_{k=0}^\infty$ and $(\mathbf{q}(k))_{k=0}^\infty$ be infinite sequences specified below. By Theorem 4, the following holds for each k

$$\mathbb{P}\Bigg(\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] > \sum_{t=1}^T q_t(k)f(Z_t) + \Delta(\mathbf{q}(k)) + C(\mathbf{q}(k)) + 4M\|\mathbf{q}\|_2\epsilon_k\Bigg) \le \exp(-\epsilon_k^2),$$

where $\Delta(\mathbf{q}(k))$ denotes the discrepancy computed with respect to the weights $\mathbf{q}(k)$ and $C(\mathbf{q}(k)) = \widetilde{C}_T\|\mathbf{q}(k)\|_2$. Let $\epsilon_k = \epsilon + \sqrt{2\log k}$. Then, by the union bound we can write

$$\begin{aligned}\mathbb{P}\Bigg(\exists k{:}\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] > \sum_{t=1}^T q_t(k)f(Z_t) + \Delta(\mathbf{q}(k)) + C(\mathbf{q}(k)) + 4M\|\mathbf{q}(k)\|_2\epsilon_k\Bigg) &\le \sum_{k=1}^\infty e^{-\epsilon_k^2} \\ &\le \sum_{k=1}^\infty e^{-\epsilon^2-\log k^2} \\ &\le 2e^{-\epsilon^2}.\end{aligned}$$

We choose the sequence $\mathbf{q}(k)$ to satisfy $\|\mathbf{q}(k)-\mathbf{q}^*\|_1 = 1-2^{-k}$. Then, for any $\mathbf{q}$ such that $0 < \|\mathbf{q}-\mathbf{u}\|_1 \le 1$, there exists $k \ge 1$ such that

$$1-\|\mathbf{q}(k)-\mathbf{q}^*\|_1 < 1-\|\mathbf{q}-\mathbf{q}^*\|_1 \le 1-\|\mathbf{q}(k-1)-\mathbf{q}^*\|_1 \le 2(1-\|\mathbf{q}(k)-\mathbf{q}^*\|_1).$$

Thus, the following inequality holds:

$$\sqrt{2\log k} \le \sqrt{2\log\log_2 2(1-\|\mathbf{q}-\mathbf{q}^*\|_1)^{-1}}.$$

Combining this with the observation that the following two inequalities hold:

$$\begin{gathered}\sum_{t=1}^T q_t(k-1)f(Z_t) \le \sum_{t=1}^T q_t f(Z_t) + 2M\|\mathbf{q}-\mathbf{q}^*\|_1 \\ \Delta(\mathbf{q}(k-1)) \le \Delta(\mathbf{q}) + 2M\|\mathbf{q}-\mathbf{q}^*\|_1, \\ \|\mathbf{q}(k-1)\|_2 \le 2\|\mathbf{q}-\mathbf{q}^*\|_1 + \|\mathbf{q}^*\|_2\end{gathered}$$

shows that the event

$$\Bigg\{\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] > \sum_{t=1}^T q_t f(Z_t) + \operatorname{disc}(\mathbf{q}) + G(\mathbf{q}) + 4M\|\mathbf{q}-\mathbf{q}^*\|_1\Bigg\}$$

where $G(\mathbf{q}) = 4M\Big(\epsilon + \sqrt{2\log\log_2 2(1-\|\mathbf{q}-\mathbf{q}^*\|_1)^{-1}} + \widetilde{C}_T\Lambda r\Big)\Big(\|\mathbf{q}^*\|_2 + 2\|\mathbf{q}-\mathbf{q}^*\|_1\Big)$ implies the following one

$$\Bigg\{\mathbb{E}[f(Z_{T+1})|\mathbf{Z}_1^T] > \sum_{t=1}^T q_t(k-1)f(Z_t) + \Delta(\mathbf{q}(k-1)) + C(\mathbf{q}(k-1)) + 4M\|\mathbf{q}(k-1)\|_2\epsilon_{k-1}\Bigg\},$$

which completes the proof. □

B Dual Optimization Problem

In this section, we provide a detailed derivation of the optimization problem in (15) starting with optimization problem in (11). The first step is to appeal to the following chain of equalities:

$$\begin{aligned}
&\min_{\mathbf{w}}\Big\{\sum_{t=1}^{T} q_t(\mathbf{w}\cdot\Psi(x_t)-y_t)^2+\lambda_2\|\mathbf{w}\|_{\mathbb{H}}^2\Big\}\\
&\quad=\min_{\mathbf{w}}\Big\{\sum_{t=1}^{T}(\mathbf{w}\cdot x'_t-y'_t)^2+\lambda_2\|\mathbf{w}\|_{\mathbb{H}}^2\Big\}\\
&\quad=\max_{\boldsymbol{\beta}}\Big\{-\lambda_2\sum_{t=1}^{T}\beta_t^2-\sum_{s,t=1}^{T}\beta_s\beta_t x'_s x'_t+2\lambda_2\sum_{t=1}\beta_t y'_t\Big\}\\
&\quad=\max_{\boldsymbol{\beta}}\Big\{-\lambda_2\sum_{t=1}^{T}\beta_t^2-\sum_{s,t=1}^{T}\beta_s\beta_t\sqrt{q_s}\sqrt{q_t}K_{s,t}+2\lambda_2\sum_{t=1}\beta_t\sqrt{q_t}y_t\Big\}\\
&\quad=\max_{\boldsymbol{\alpha}}\Big\{-\lambda_2\sum_{t=1}^{T}\frac{\alpha_t^2}{q_t}-\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}+2\lambda_2\boldsymbol{\alpha}^T\mathbf{Y}\Big\},
\end{aligned}\tag{20}$$

where the first equality follows by substituting $x'_t=\sqrt{q_t}\Psi(x_t)$ and $y'_t=\sqrt{q_t}y_t$ the second equality uses the dual formulation of the kernel ridge regression problem and the last equality follows from the following change of variables: $\alpha_t=\sqrt{q_t}\beta_t$.

By (20), optimization problem in (11) is equivalent to the following optimization problem

$$\min_{0\leq\mathbf{q}\leq 1}\left\{\max_{\boldsymbol{\alpha}}\Big\{-\lambda_1\sum_{t=1}^{T}\frac{\alpha_t^2}{q_t}-\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}+2\lambda_1\boldsymbol{\alpha}^T\mathbf{Y}\Big\}+(\mathbf{d}\cdot\mathbf{q})+\lambda_2\|\mathbf{q}-\mathbf{u}\|_p\right\}.$$

Next, we apply the change of variables $r_t=1/q_t$ and appeal to the same arguments as were given for the primal problem in Section 6 to arrive at (15).