Learning Theory and Algorithms for Forecasting Non-stationary Time Series

Authored by:

Mehryar Mohri
Vitaly Kuznetsov

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. We use our learning bounds to devise new algorithms for non-stationary time series forecasting for which we report some preliminary experimental results.

1 Paper Body

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 only asymptotic. 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 -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 -mixing stationary scenario. A similar result was proven by Shalizi and Kontorovitch [2013] for mixtures of -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. 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 ’-mixing and -mixing stationary processes. Steinwart and Christmann [2009] proved fast learning rates for regularized algorithms with ?-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, 1

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 - and ’-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 and state space 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] also give generalization bounds in terms of discrepancy. However, unlike the result of Kuznetsov and Mohri [2014], 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 devise new algorithms for non-stationary time series forecasting that benefit from our datadependent 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 nonstationarity 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. In Section 4, we give data-dependent learning bounds based on the empirical discrepancy. These results, combined with a novel analysis of kernel-based hypotheses for time series forecasting (Appendix B), are used to devise new forecasting algorithms in Section 5. In Appendix C, we report the results of preliminary experiments using these algorithms.

2

Preliminaries

We consider the following general time series prediction setting where the learner receives a realization (X1 , Y1), . . . , (XT , YT) of some stochastic process, with (Xt , Yt) 2 Z = X ? Y. The objective of the learner is to select out of a specified family H a hypothesis h : X ! Y that achieves a small generalization error E[L(h(XT +1), YT +1)—Z1 , . . . , ZT] conditioned on observed data, where L : Y ? Y ! [0, 1) is a given loss function. The path-dependent generalization error that we consider in this work is a finer measure of the generalization ability than the averaged generalization error E[L(h(XT +1), YT +1)] = E[E[L(h(XT +1), YT +1)—Z1 , . . . , ZT]] since it only takes into consideration the realized history of the stochastic process and does not average over the set of all possible histories. The results that we present in this paper also apply to the setting where the time parameter t can take non-integer values and prediction lag is an arbitrary number l 0. That is, the error is defined by E[L(h(XT +l), YT +l)—Z1 , . . . , ZT] but for notational simplicity we set l = 1. 2

Our setup covers a larger number of scenarios commonly used in practice. The case X = Y p p corresponds to a large class of autoregressive models. Taking X = [1 p=1 Y leads to growing memory models which, in particular, include state space models. More generally, X may contain both the history of the process {Yt } and some additional side information.

To simplify the notation, in the rest of the paper, we will use the shorter notation f (z) = L(h(x), y), for any z = (x, y) 2 Z and introduce the family F = {(x, y) ! L(h(x), y) : h 2 H} containing such functions f . We will assume a bounded loss function, that is —f — ? M for all f 2 F for some M 2 R+ . Finally, we will use the shorthand Zba to denote a sequence of random variables Za , Za+1 , . . . , Zb . The key quantity of interest in the analysis of generalization is the following supremum of the empirical process defined as follows: (ZT1)

= sup f 2F

E[f (ZT +1)—ZT1]

T X
!
qt f (Zt) ,
t=1
(1)

where q1 , . . . , qT are real numbers, which in the standard learning scenarios are chosen to be uniform. In our general setting, different Zt s 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 ZT +1 . The generalization bounds that we present below are for an arbitrary sequence q = (q1 , . . . qT) which, in particular, covers the case of uniform weights. Remarkably, our bounds do not even require the non-negativity of q. Our generalization bounds are expressed in terms of data-dependent measures of sequential complexity such as expected sequential covering number or sequential Rademacher complexity [Rakhlin et al., 2010]. We give a brief overview of the notion of sequential covering number and refer the reader to the aforementioned reference for further details. We adopt the following definition of a complete binary tree: a Z-valued complete binary tree z is a sequence (z1 , . . . , zT) of T mappings zt : {?1}t 1 ! Z, t 2 [1, T]. A path in the tree is = (1 , . . . , T 1). To simplify the notation we will write zt ( ) instead of zt (1 , . . . , t 1), even though zt depends only on the first t 1 elements of . The following definition generalizes the classical notion of covering numbers to sequential setting. A set V of R-valued trees of depth T is a sequential ?-cover (with respect to q-weighted ‘p norm) of a function class G on a tree z of depth T if for all g 2 G and all 2 {?}T , there is v 2 V such that
T X
vt ( )
g(zt ( ))
t=1
p
! p1
? kqkq 1 ?,

where k ? kq is the dual norm. The (sequential) covering number Np (?, G, z) of a function class G on a given tree z is defined to be the size of the minimal sequential cover. The maximal covering number is then taken to be Np (?, G) = supz Np (?, G, z). One can check that in the case of uniform weights this definition coincides with the standard definition of sequential covering numbers. Note that this is a purely combinatorial notion of complexity which ignores the distribution of the process in the given learning problem. Data-dependent sequential covering numbers can be defined as follows. Given a stochastic process distributed according to the distribution p with pt (?—zt1 1) denoting the conditional distribution at time t, we sample a Z ? Z-valued tree of depth T according to the following procedure. Draw two independent samples Z1 , Z10 from p1 : in the left child of the root draw Z2 , Z20 according to p2 (?—Z1) and in the right child according to p2 (?—Z20). More generally, for a node that can be reached by a path (1 , . . . , t), we draw Zt , Zt0 according to pt (?—S1 (1), . . . , St 1 (t 1)), where St (1) = Zt and St (1) = Zt0 . Let z denote the tree formed using Zt s and define the expected covering number to be Ez?T (p) [Np (?, G, z)], where T (p) denotes the distribution of z. In a similar manner, one can define other measures of complexity such as sequential Rademacher complexity and the Littlestone dimension [Rakhlin et al., 2015] as well as their data-dependent counterparts [Rakhlin et al., 2011]. 3

The final ingredient needed for expressing our learning guarantees is the notion of discrepancy between target distribution and the distribution of the sample: ? ? T X = sup E[f (ZT +1)—ZT1] qt E[f (Zt)—Zt1 1] . (2) f 2F

t=1

The discrepancy is a natural measure of the non-stationarity of the stochastic process Z with respect to both the loss function L and the hypothesis set H. In particular, note that if the process Z is i.i.d., then we simply have = 0 provided that qt s form a probability distribution. It is also possible to give bounds on in terms of other natural distances between distribution. For instance, Pinsker?s inequality yields r ? ? PT PT t 1 T ? M PT +1 (?—Z1) ? 12 D PT +1 (?—ZT1) k t=1 qt Pt (?—Zt1 1) , t=1 qt Pt (?—Z1) TV

where k ? kTV is the total variation distance and D(? k ?) the relative entropy, Pt+1 (?—Zt1) the condiPT tional distribution of Zt+1 , and t=1 qt Pt (?—Zt1 1) the mixture of the sample marginals. Alternatively, if the target distribution at lag l, P = PT +l is a stationary distribution of an asymptotically stationary process Z [Agarwal and Duchi, 2013, Kuznetsov and Mohri, 2014], then for qt = 1/T we have ?

T MX kP T t=1

Pt+l (?—Zt

1)kTV

?

(l),

where (l) = sups supz [kP Pl+s (?—zs 1)kTV] is the coefficient of asymptotic stationarity. The process is asymptotically stationary if liml!1 (l) = 0. However, the most important property of the discrepancy is that, as shown later in Section 4, it can be estimated from data under some additional mild assumptions. [Kuznetsov and Mohri, 2014] also give generalization bounds 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 (ZT1). Given a sequence of random variables ZT1 we say that T Z0 1 is a decoupled tangent sequence if Zt0 is distributed according to P(?—Zt1 1) and is independent of Z1 na and Gin?e, t . It is always possible to construct such a sequence of random variables [De la Pe? 1999]. The next theorem is the main result of this section. Theorem 1. Let ZT1 be a sequence of random variables distributed according to p. Fix ? ¿ 2? ¿ 0. Then, the following holds: ? ? ? ? (? 2?)2 T P (Z1) ? ? E N1 (?, F, v) exp . 2M 2 kqk22 v?T (p) Proof. The first step is to observe that, since the difference of the suprema is upper bounded by the supremum of the difference, it suffices to bound the probability of the following event (!) T X sup qt (E[f (Zt)—Zt1 1] f (Zt)) ? . f 2F

By Markov?s inequality, for any P

sup f 2F

? exp(

T X t=1

t=1

¿ 0, the following inequality holds: ! !

qt (E[f (Zt)—Zt1 1] ”

?) E exp

sup f 2F

f (Zt))

T X t=1

4

?

qt (E[f (Zt)—Zt1 1]

!!#

f (Zt))

.

T

Since Z0 1 is a tangent sequence the following equalities hold: E[f (Zt)—Zt1 1] = E[f (Zt0)—Zt1 1] = E[f (Zt0)—ZT1]. Using these equalities and Jensen?s inequality, we obtain the following: ? T ? ? X E exp sup qt E[f (Zt)—Zt1 1] f (Zt) f 2F t=1

? T ? hX = E exp sup E qt f (Zt0) f 2F

?

? E exp

?

sup

t=1

T X

f 2F t=1

qt f (Zt0)

f (Zt) —ZT1 ?

f (Zt)

i?

, T

where the last expectation is taken over the joint measure of ZT1 and Z0 1 . Applying Lemma 5 (Appendix A), we can further bound this expectation by ? ? T ? ?? X 0 E E exp sup f (zt ( )) t qt f (z t ( )) 0 (z,z)?T (p)

? ? =

E 0

(z,z)?T (p)

?
f 2F t=1
E exp
E
z?T (p)
?
?
sup
T X
f 2F t=1
0 t qt f (z t ( )) +
sup
? T X 0 E 0 E exp 2 sup + t qt f (z t ( ))
1 2 (z,z)
?
?
?
f 2F t=1
? q f (z ( )) t t t
T X
f 2F t=1 1 E0 E 2 (z,z)
? T X E exp 2 sup , t qt f (zt ( ))
?
?
exp 2 sup
T X
f 2F t=1
? t qt f (zt ( ))
f 2F t=1

where for the second inequality we used Young?s inequality and for the last equality we used symmetry. Given z let C denote the minimal ?-cover with respect to the q-weighted ‘1 -norm of F on z. Then, the following bound holds sup

T X
f 2F t=1
t qt f (zt (
)) ? max c2C
T X
t q t ct (
) + ?.
t=1

By the monotonicity of the exponential function, ? ? ? ? ? T T X X E exp 2 sup ? exp(2 ?) E exp 2 max t qt f (zt ( )) c2C

f 2F t=1
? exp(2 ?)
X

c2C

?

?

E exp 2

t=1

T X t=1

Since ct ( ) depends only on 1 , . . . , T 1 , by Hoeffding?s bound, ? ? X ? ? ? TX1 ? ? ? T E exp 2 q c ( ) = E exp 2 q c ( ) E exp 2 t t t t t t t t=1

?

?

? E exp 2

T

t=1

T X1

t q t ct (

t=1

?) exp(2

? t q t ct ( )

? . t q t ct ( )

? q c ( ) T T T

v?T (p)

Optimizing over

completes the proof.

An immediate consequence of Theorem 1 is the following result. 5

1

2 2 qT M 2)

and iterating this inequality and using the union bound, we obtain the following: ? ? T ? X P sup qt (E[f (Zt)—Zt1 1] f (Zt)) ? ? E [N1 (?, G, v)] exp (? 2?)+2 f 2F t=1

T 1

2

? M 2 kqk22 .

Corollary 2. For any

¿ 0, with probability at least 1

E[f (ZT +1)—ZT1] ?

T X

qt f (Zt) +

t=1

, for all f 2 F and all ? ¿ 0, r Ev?T (P) [N1 (?, G, v)] + 2? + M kqk2 2 log .

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. Using chaining techniques bounds, Theorem 1 and Corollary 2 can be further improved and we will present these results in the full version of this paper. While Rakhlin et al. [2015] give high probability bounds for a different quantity than the quantity of interest in time series prediction, ! T X t 1 sup qt (E[f (Zt)—Z1] f (Zt)) , (3) f 2F t=1

their analysis of this quantity can also be used in our context to derive high probability bounds for (ZT1) . However, this approach results in bounds that are in terms of purely combinatorial notions such as maximal sequential covering numbers N1 (?, 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 Z1 be drawn from a uniform distribution on {0, 1} and Zt ? p(?—Zt 1) with p(?—y) being a distribution over {0, 1} such that p(x—y) = 2/3 if x = y and 1/3 otherwise. Let G be defined by G = {g(x) = 1x ? : ? 2 [0, 1]}. Then, one can check that Ev?T (P) [N1 (?, G, v)] = 2, while N1 (?, G) 2T . The data-dependent bounds of Theorem 1 and Corollary 2 highlight the fact that the task of time series prediction lies in between the familiar i.i.d. scenario and adversarial on-line learning setting. However, the key component of our learning guarantees is the discrepancy term . 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 4. 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. In the drifting scenario, with ZT1 being a sequence of independent random variables, our discrepancy measure coincides with the one used and studied in [Mohri and Mu?noz Medina, 2012]. Convergence can also be established in asymptotically stationary and stationary mixing cases. However, as we show in Section 4, the most important advantage of our bounds is that the discrepancy measure we use can be estimated from data.

4

Estimating Discrepancy

In Section 3, we showed that the discrepancy is crucial for forecasting non-stationary time series. In particular, if we could select a distribution q over the sample ZT1 that would minimize the discrepancy and use it to weight training points, then we would have a better learning guarantee for an algorithm trained on this weighted sample. In some special cases, the discrepancy can be computed analytically. However, in general, we do not have access to the distribution of ZT1 and hence we need to estimate the discrepancy from the data. Furthermore, in practice, we never observe ZT +1 and it is not possible to estimate without some further assumptions. One natural assumption is that the distribution Pt of Zt does not change drastically with t on average. Under this assumption the last s observations ZTT s+1 are effectively drawn from the distribution close to PT +1 . More precisely, we can write ? ? T T X 1 X t 1 t 1 ? sup E[f (Zt)—Z1] qt E[f (Zt)—Z1] s f 2F t=1 t=T

+ sup f 2F

?

s+1

1 s

E[f (ZT +1)—ZT1]

T X

t=T

s+1

?

E[f (Zt)—Zt1 1]

.

We will assume that the second term, denoted by 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 6

this setting. Observe that ? sup E[ZT +1 —ZT1]

E[f (Zr)—Zr1

f 2F

1

T ? X ?] ? sup E[f (Zt+1)—Zt1] t=r f 2F

?M for all r = T

T X t=r

kPt+1 (?—Zt1)

? E[f (Zt)—Zt1 1]

Pt (?—Zt1 1)kTV ,

s + 1, . . . , T . Therefore, we must have ? ? s+1 1 X sup E[ZT +1 —ZT1] E[f (Zt)—Zt1] ? M , s ? s 2 f 2F t=T

s+1

1

= supt kPt+1 (?—Zt1)

where Pt (?—Zt1 1)kTV . Barve and Long [1996] showed that [VC-dim(H)] 3 is a lower bound on the generalization error in the setting of binary classification where ZT1 is a sequence of independent but not identically distributed random variables (drifting). This setting is a special case of the more general scenario that we are considering. The following result shows that we can estimate the first term in the upper bound on . Theorem 3. Let ZT1 be a sequence of random variables. Then, for any ¿ 0, with probability at least 1 , the following holds for all ? ¿ 0: ! ! T T X X t 1 sup (pt qt) E[f (Zt)—Z1] ? sup (pt qt)f (Zt) + B, f 2F

t=1

q pk2 2 log

where B = 2? + M kq the last s points.

f 2F

Ez?T (p) [N1 (?,G,z)]

t=1

and where p is the uniform distribution over

The proof of this result is given in Appendix A. Theorem 1 and Theorem 3 combined with the union bound yield the following result. Corollary 4. Let ZT1 be a sequence of random variables. Then, for any ¿ 0, with probability at least 1 , the following holds for all f 2 F and all ? ¿ 0: E[f (ZT +1)—ZT1] ? T X t=1

qt f (Zt) + e +

where e = supf 2F
5
Algorithms
?P
T t=1 (pt
s
? + 4? + M kqk2 + kq
? qt)f (Zt) .
pk2
q ? 2 log
2 Ev?T (p) [N1 (?,G,z)]
,

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. We present the full analysis of this setting in Appendix B including novel bounds on the sequential Rademacher complexity. The learning bounds 1 can be generalized q ? of Theorem ? to hold uniformly 1 over q at the price of an additional term in O kq uk1 log2 log2 kq uk1 . We prove this result in Theorem 8 (Appendix B). Suppose L is the squared loss and H = {x ! w ? (x) : kwkH ? ?}, where : X ! H is a feature mapping from X to a Hilbert space H. By Lemma 6 (Appendix B), we can bound the complexity term in our generalization bounds by ? ? ?r O (log3 T) p + (log3 T)kq uk1 , T where K is a PDS kernel associated with H such that supx K(x, x) ? r and u is the uniform distribution over the sample. Then, we can formulate a joint optimization problem over both q and w based on the learning guarantee of Theorem 8, which holds uniformly over all q: ?X T T X min qt (w ? (xt) yt)2 + 1 dt qt + 2 kwk2H + 3 kq uk1 . (4) 0?q?1,w
t=1
t=1
7

PT Here, we have upper bounded the empirical discrepancy term by t=1 dt qt with each dt defined PT by supw0 ?? — s=1 ps (w0 ? (xs) ys)2 (w0 ? (xt) yt)2 —. Each dt can be precomputed using DC-programming. 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]. Note that problem (4) is not jointly convex in q and w. However, using the dual problem associated to w yields the following equivalent problem, it can be rewritten as follows: min
0?q?1
?
max ?
n
2
T X ?2
?T K? + 2

t
qt
t=1
T 2? Y
o
+
1 (d?q)
+
3 kq
uk1 ,
(5)

where d = (d1 , . . . , dT)T , K is the kernel matrix and Y = (y1 , . . . , yT)T . We use the change of variables rt = 1/qt and further upper bound 3 kq uk1 by 03 kr T 2 uk2 , which follows from —qt ut — = —qt ut (rt T)— and H?older?s inequality. Then, this yields the following optimization problem: min r2D
?
max ?
n
2
T X
?T K? + 2
rt ?t2
t=1
o T ? Y + 2
1
T X dt t=1
+
rt
3 kr
T 2 uk22 ,
(6)

where D = {r : rt 1, t 2 [1, T]}. The optimization problem (6) is convex since D is a convex set, the first term in (6) is convex as a maximum of convex (linear) functions of r. This problem can be solved using standard descent methods, where, at each iteration, we solve a standard QP in ?, which admits a closed-form solution. Parameters 1 , 2 , and 3 are selected through cross-validation. An alternative simpler algorithm based on the data-dependent bounds of Corollary 4 consists of first finding a distribution q minimizing the (regularized) discrepancy and then using that to find a hypothesis minimizing the (regularized) weighted empirical risk. This leads to the following twostage procedure. First, we find a solution q? of the following convex optimization problem: min q 0
?
sup
w0 ??

T ?X
(pt
t=1
qt)(w0 ? (xt)
? yt)2 +
1 kq
uk1 ,
(7)

where 1 and ? are parameters that can be selected via cross-validation. Our generalization bounds hold for arbitrary weights q but we restrict them to being positive sequences. Note that other regularization terms such as kqk22 and kq pk22 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 w0 . Alternatively, one can upper bound the suprePT mum by t=1 qt dt and then solve the resulting optimization problem. The solution q? of (7) is then used to solve the following (weighted) kernel ridge regression problem: min w

?X T t=1
qt? (w ? (xt)
yt)2 +
2 2 kwkH

Note that, in order to guarantee the convexity of this problem, we require q?

6
(8)
. 0.

Conclusion

We presented a general theoretical analysis of learning in the broad scenario of non-stationary nonmixing processes, the realistic setting for a variety of applications. We discussed in detail several algorithms benefitting from the learning guarantees presented. Our theory can also provide a finer analysis of several existing algorithms and help devise alternative principled learning algorithms. Acknowledgments This work was partly funded by NSF IIS-1117591 and CCF-1535987, and the NSERC PGS D3. 8

2 References

T. M. Adams and A. B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. The Annals of Probability, 38(4):1345?1367, 2010. A. Agarwal and J. Duchi. The generalization ability of online algorithms for dependent data. Information Theory, IEEE Transactions on, 59(1):573?587, 2013. P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. Technical Report 2010-39, Centre de Recherche en Economie et Statistique, 2010. P. Alquier, X. Li, and O. Wintenberger.

Prediction of time series by statistical learning: general losses and fast rates. Dependence Modelling, 1:65?93, 2014. D. Andrews. First order autoregressive processes and strong mixing. Cowles Foundation Discussion Papers 664, Cowles Foundation for Research in Economics, Yale University, 1983. R. Baillie. Long memory processes and fractional integration in econometrics. Journal of Econometrics, 73 (1):5?59, 1996. R. D. Barve and P. M. Long. On the complexity of learning from drifting distributions. In COLT, 1996. P. Berti and P. Rigo. A Glivenko-Cantelli theorem for exchangeable random variables. Statistics & Probability Letters, 32(4):385 ? 391, 1997. T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. J Econometrics, 1986. G. E. P. Box and G. Jenkins. Time Series Analysis, Forecasting and Control. Holden-Day, Incorporated, 1990. P. J. Brockwell and R. A. Davis. Time Series: Theory and Methods. Springer-Verlag, New York, 1986. V. H. De la Pe?na and E. Gin?e. Decoupling: from dependence to independence: randomly stopped processes, U-statistics and processes, martingales and beyond. Probability and its applications. Springer, NY, 1999. P. Doukhan. Mixing: properties and examples. Lecture notes in statistics. Springer-Verlag, New York, 1994. R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4):987?1007, 1982. J. D. Hamilton. Time series analysis. Princeton, 1994. V. Kuznetsov and M. Mohri. Generalization bounds for time series prediction with non-stationary processes. In ALT, 2014. A. C. Lozano, S. R. Kulkarni, and R. E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary -mixing observations. In NIPS, pages 819?826, 2006. R. Meir. Nonparametric time series prediction through adaptive model selection. Machine Learning, pages 5?34, 2000. D. Modha and E. Masry. Memory-universal prediction of stationary random processes. Information Theory, IEEE Transactions on, 44(1):117?133, Jan 1998. M. Mohri and A. Mu?noz Medina. New analysis and algorithm for learning with drifting distributions. In ALT, 2012. M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In NIPS, 2009. M. Mohri and A. Rostamizadeh. Stability bounds for stationary '-mixing and -mixing processes. Journal of Machine Learning Research, 11:789?814, 2010. V. Pestov. Predictive PAC learnability: A paradigm for learning from exchangeable input data. In GRC, 2010. A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In NIPS, 2010. A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In NIPS, 2011. A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. Probability Theory and Related Fields, 2015. C. Shalizi and A. Kontorovitch. Predictive PAC learning and process decompositions. In NIPS, 2013. I. Steinwart and A. Christmann. Fast learning from non-i.i.d. observations. In NIPS, 2009. P. D. Tao and L. T. H. 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. B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability, 22(1):94?116, 1994.

9