Forecasting Non-Stationary Time Series: From Theory to Algorithms

Vitaly Kuznetsov
Courant Institute
251 Mercer Street
New York, NY 10012
vitaly@cims.nyu.edu

Mehryar Mohri
Courant Institute & Google Research
251 Mercer Street
New York, NY 10012
mohri@cims.nyu.edu

Abstract

Generalization bounds for time series prediction and other non-i.i.d. learning scenarios that can be found in the machine learning and statistics literature assume that observations come from a (strictly) stationary distribution. The first bounds for completely non-stationary setting were proved in [6]. In this work we present an extension of these results and derive novel algorithms for forecasting non-stationary time series. Our experimental results show that our algorithms significantly outperform standard autoregressive models commonly used in practice.

1 Introduction

Given a sample $((X_1, Y_1), \ldots, (X_m, Y_m))$ of pairs in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, the standard supervised learning task consists of selecting, out of a class of functions H, a hypothesis $h\colon \mathcal{X} \to \mathcal{Y}$ that admits a small expected loss measured using some specified loss function $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. The common assumption in the statistical learning theory and the design of algorithms is that samples are drawn i.i.d. from some unknown distribution and generalization in this scenario has been extensively studied in the past. However, for many problems such as time series forecasting, the i.i.d. assumption is too restrictive and it is important to analyze generalization in the absence of that condition. A variety of relaxations of this i.i.d. setting have been proposed in the machine learning and statistics literature. In particular, the scenario in which observations are drawn from a stationary mixing distribution has become standard and has been adopted by most previous studies [1, 8, 9, 10, 13, 15] and most of the modern time series prediction methods either assume stationarity or attempt to transform the data in order to satisfy this assumption. For a more detailed survey of state-of-the-art results in this area we refer the reader to [6]. However, a wide spectrum of stochastic processes considered in applications, such as for example Markov chains, are in fact non-stationary. In this work, we present generalization bounds under the more realistic assumption of non-stationary data. Furthermore, we argue that under some additional assumptions our generalization bounds lead to novel algorithms for forecasting non-stationary time series.

2 Generalization bounds

To state our main results we will first need to introduce some notation. Suppose we are given a doubly infinite sequence of $\mathcal{Z}$-valued random variables $\{Z_t\}_{t=-\infty}^{\infty}$ jointly distributed according to $\mathbf{P}$. We will write $\mathbf{Z}_a^b$ to denote a vector $(Z_a, Z_{a+1}, \ldots, Z_b)$ where a and b are allowed to take values $-\infty$ and ∞. Similarly, $\mathbf{P}_a^b$ denotes the distribution of $\mathbf{Z}_a^b$. Recall that a sequence of random variables $\mathbf{Z}_{-\infty}^{\infty}$ is (strictly) stationary provided that, for any t and any non-negative integers m and k, $\mathbf{Z}_t^{t+m}$ and $\mathbf{Z}_{t+k}^{t+m+k}$ have the same distribution. We will not assume that the process that we sample from is stationary but we will assume that it is mixing. Following [3], we define β-mixing coefficients for

$\mathbf{P}$ as follows. For each positive integer a, we set $\beta(a) = \sup_t \|\mathbf{P}_{-\infty}^t \otimes \mathbf{P}_{t+a}^\infty - \mathbf{P}_{-\infty}^t \wedge \mathbf{P}_{t+a}^\infty\|_{TV}$, where $\mathbf{P}_{-\infty}^t \wedge \mathbf{P}_{t+a}^\infty$ denotes the joint distribution of $\mathbf{Z}_{-\infty}^t$ and $\mathbf{Z}_{t+a}^\infty$. and $\|\cdot\|_{TV}$ is the total variation distance. Roughly speaking, this means that the future has a sufficiently weak dependence on the distant past. Often, processes that arise naturally in applications are β-mixing. For example, one can show that Markov processes are geometrically β-mixing with $\beta(a) = O(d^{-a})$ for some $d > 1$.

The goal of the learner is to find a hypothesis h that will have a small generalization error in the near future: $\mathcal{L}_{T+s}(h) = \mathbb{E}_{Z_{T+s}}[\ell(h, Z_{T+s})]$, where $\ell(h,z) = L(h(x), y)$ and L is some given loss function. For alternative definitions of the generalization error for time series prediction see [6].

Finally, a key ingredient of the bounds we present is the notion of *discrepancy* between two probability distributions that was used by Mohri and Muñoz Medina [12] to give generalization bounds for sequences of independent (but not identically distributed) random variables. In our setting, discrepancy can be defined as $d(t_1, t_2) = \sup_{h \in H} |\mathcal{L}_{t_1}(h) - \mathcal{L}_{t_2}(h)|$. Discrepancy is a natural measure of the non-stationarity of a stochastic process with respect to the hypothesis class H and a loss function L. For instance, if the process is strictly stationary then $\bar{d}(t_1, t_2) = 0$ for all $t_1, t_2 \in \mathbb{Z}$.

Generalization bounds for non-i.i.d. scenarios that can be found in the machine learning and statistics literature assume that observations come from a (strictly) stationary distribution. The first bounds for completely non-stationary setting were proved in [6]. Here we present an extension of these results and use it to derive novel time series prediction algorithms. Our main result is the following.

Theorem 1. *Let L be a loss function bounded by M and H an arbitrary hypothesis set. For any a and m such that $T = 2am$, partition the given sample $\mathbf{Z}_1^T$ into blocks $2m$ blocks each of size a. Fix any $w_1, \ldots, w_T$, such that $\sum_{t=1}^T w_t = 1$ and $w_t \geq 0$. Then, for any $\delta > 2(m-1)\beta(a)$, with probability $1 - \delta$, the following holds for all hypotheses $h \in H$:*

$$\mathcal{L}_{T+s}(h) \leq \sum_{t=1}^{T} w_t \ell(h, Z_t) + \frac{4}{a} \sum_{j=0}^{2a-1} \mathfrak{R}_j + 2 \sum_{t=1}^{T} w_t d(t, T+s) + c(\sqrt{a}\|\mathbf{w} - \mathbf{u}\|_2 + \tfrac{1}{\sqrt{m}}),$$

where $\mathfrak{R}_j = \frac{1}{m}\mathbb{E}[\sup_{h \in H} \sum_{i=1}^m \sigma_i \ell(h_{2ai+j}, Z_{2ai+j})]$ are Rademacher complexities, $c = 2M\sqrt{\log \frac{2}{\delta'}}$ and $\mathbf{u}$ is the uniform distribution.

The main difference of this result with the result presented in [6] is that w_t is not required to be uniform anymore. Unlike in i.i.d. setting, it is natural to weight the errors of a given hypothesis h differently on different sample points since distances between their distributions and distribution that we are trying to predict may vary. As we shall see below this also leads to new algorithms for time series prediction. The proof of this result follows the same arguments as in [6] which are based on independent block technique of [15]. We omit the proof and refer the reader to [6] for details.

The learning bound of Theorem 1 indicates the challenges faced by the learner when presented with data drawn from a non-stationary stochastic process. In particular, the presence of the third term in the bound shows that generalization in this setting depends on the "degree" of non-stationarity of the underlying process. The dependency in the training instances reduces the effective size of the sample from T to m (if choose uniform weights $w_t = 1/T$).

3 Algorithms

In this section we will show how the bounds of Section 2 can be used to derive novel algorithms for forecasting non-stationary time series. We will assume that that $d_t = d(t, T+s)$ can be computed analytically or has been estimated from the data. Either of these assumptions can naturally arise in applications. For instance, the discrepancy measure d_t can be replaced by an upper bound that, under mild conditions, can be estimated from data [7, 4]. Alternatively, suppose L is a quadratic loss $L(y, y') = (y - y')^2$ and $\mathcal{X} = \mathcal{Y}^d \times \mathcal{X}'$, i.e. our feature vector $x_t = (y_{t-1}, \ldots, y_{t-d}, x'_t)$ consists of the d previous values of the stochastic process that we are trying to predict and a side information x'_t at time t. If we use a set of linear hypothesis $H = \{\mathbf{x} \mapsto \mathbf{h} \cdot \mathbf{x} \colon \|\mathbf{h}\|_2 \leq \Lambda\}$, then we can compute d_t explicitly in terms of the autocovariance function of the underlying stochastic process. In particular, if for simplicity we omit side information then we observe that for any t we can write $\mathbb{E}[(\sum_{j=1}^d h_j Y_{t-j} - Y_t)^2] = \sum_{i,j=0}^d h_i h_j \mathbb{E}[Y_{t-j} Y_{t-i}] = \sum_{i,j=0}^d h_i h_j \rho(t-i, t-j)$, where

$\rho(r, s) = \mathbb{E}[Y_r Y_s]$ and we take $h_0 = -1$. Therefore, we can write

$$d_t = \sup_{h \in H} |\sum_{i,j=0}^{d} h_i h_j (\rho(t-i, t-j) - \rho(T-i, T-j))|.$$

In particular, the last expression implies that if the process is only weakly stationary, i.e. there is a function f such that $\rho(r, s) = f(r - s)$ and $\mathbb{E}Y_r$ is constant as a function of r, then $d_t = 0$ for all t. Note that this result together with Theorem 1 gives strong theoretical guarantees for learning autoregressive processes (AR) with linear hypothesis, since these processes are weakly stationary.

More generally, for linear hypotheses with quadratic loss, d_t is completely determined by the covariance structure of the underlying stochastic process. In particular, $d_t \leq O(\|P_t - P_T\|)$, where $P_t = (\rho(t-i, t-j))_{i,j}$ is $(d+1) \times (d+1)$ matrix. Consider, for instance, a process defined by $Y_{t+1} = aY_t + \epsilon_t$, where ϵ_t's are mean zero independent random variables with $\mathbb{E}[\epsilon_t^2] = \sigma_t$. One can show that the autocovariance function is given by $\rho(t+s, t) = \sigma_t a^s/(1-a^2)$ and the process is not (weakly) stationary unless σ_t is constant. Recall that a standard approach when dealing with non-stationary processes is so called "differencing", i.e. considering $Y_t' = Y_t - Y_{t-1}$ and higher order differences to obtain a stationary processes. This approach, for instance, leads to a celebrated ARIMA model. However, this method will fail for the process Y_t that we have just defined. On the other hand, process Y_t can arise naturally in the applications in which the variance of the stochastic process evolves with time. In summary, in many practical applications, d_t can either be found analytically or estimated from the data and under this assumption we will present our algorithms.

3.1 SRM-style Algorithm

The first algorithm that we present here is a meta-algorithm that is close in spirit to Structural Risk Minimization (SRM) of [14]. The major difference is that now we also need to control the weighted discrepancy term that appears in the bound. Suppose we have access to an infinite nested sequence of hypothesis sets $H_0 \subset H_1 \subset H_2 \ldots$. For each $n \in \mathbb{N}$, we find a hypothesis h_n that optimizes the trade-off between weighted discrepancy and weighted empirical error. More precisely,

$$h_n = \operatorname{argmin}_{h \in H_n} \inf_{\mathbf{w} \in \Delta} \left(\sum_{t=1}^{T} w_t (\ell(h, Z_t) + d_t) + c\|\mathbf{w} - \mathbf{u}\|_2 \right) = \operatorname{argmin}_{h \in H_n} \psi(h),$$

where Δ is a probability simplex and $\psi(h) = \inf_{\mathbf{w} \in \Delta} (\sum_{t=1}^{T} w_t (\ell(h, Z_t) + d_t) + c\|w - u\|_2)$. We set the SRM hypothesis to be h_n that achieves the optimal trade-off between complexity term and discrepancy-risk functional ψ:

$$h_{SRM} = \operatorname{argmin}_{h_n} \left(\psi(h_n) + \frac{4}{a} \sum_{j=0}^{2a-1} \mathfrak{R}_j(H_n) \right).$$

This algorithm directly optimizes the upper bound of Theorem 1, however, it is tractable only in certain special cases and we consider some of these special cases below.

3.2 EM-style Algorithm

Here we consider the case of quadratic loss function L with a set H of linear hypotheses with bounded norm. Recall that Rademacher complexity of such hypothesis set H is bounded above by $\Lambda r/\sqrt{m}$ (see for example [11]). Then Theorem 1 leads to the following optimization problem:

$$\min_{\mathbf{w}, \mathbf{h}} \quad \lambda_1 \|\mathbf{w} - \mathbf{u}\|_2^2 + \lambda_2 \|\mathbf{h}\|_2^2 + \sum_{t=1}^{T} w_t (d_t + (\mathbf{h} \cdot \mathbf{x}_t - y_t)^2)$$

$$\text{subject to:} \quad \sum_{t=1}^{T} w_t = 1 \text{ and } w_t \geq 0, \forall t = 1, \ldots, T$$

where λ_1, λ_2 are parameters to our learning algorithm that can be set via cross-validation. It should be noted that this optimization problem can viewed as a special case of optimization problem considered in Subsection 3.1 with these particular L and H. Note also that this problem is not convex,

Table 1: Stochastic processes for ADS1, ADS2, ADS3 (Z_t i.i.d $N(0, 0.01)$).

ADS1	ADS2	ADS 3
$Y_t = a_t Y_{t-1} + Z_t$ $a_t = 1$ if $t < 1800$ and -1 otherwise	$Y_t = a_t Y_{t-1} + Z_t$ $a_t = 0.9 - 1.8(t/2000)$	$Y_t = a_t Y_{t-1} + (1 - a_t) Y_{t-2} + Z_t$ $a_t = 0.9t/2000$

Table 2: Average L_2 error (st.dev.)

	ADS1	ADS2	ADS3	FX1	FX2
WRA	**0.0099 (0.0155)**	**0.0997 (0.1449)**	**0.1026 (0.1509)**	**0.0072 (0.0102)**	**0.0069 (0.0112)**
ARIMA(q,0,0)	0.1432 (0.2091)	0.4797 (0.6942)	0.2598 (0.3696)	0.0366 (0.0329)	0.0252 (0.0254)
Ratio	14.5	4.8	2.6	5.1	3.7

but we observe that for a fixed $\mathbf{w}$ it is a QP. The same is true in reverse: when $\mathbf{h}$ is fixed, we have a different a QP. This suggests an alternating scheme, similar to EM algorithm, where we alternate between solving QP for $\mathbf{h}$ and keeping $\mathbf{w}$ fixed and vice verse. Of course, this algorithm is not immune to the usual problems that one faces when objective function is non-convex. In particular, there is no guarantee that this algorithm will converge to a global minimum.

3.3 Weighted Ridge Regression (WRA)

In some special cases optimal w_t can be computed explicitly or set to some fixed values according to some natural heuristic. For example, in many applications d_t may increase as t decreases and one can choose an increasing sequence $w_1, \ldots, w_T$ such that $\sum_{t=1}^{T} w_t = 1$ and $w_t \geq 0$. This leads to a simple optimization problem:

$$\min_{\mathbf{h}} \quad \lambda \|\mathbf{h}\|_2^2 + \sum_{t=1}^{T} w_t (\mathbf{h} \cdot \mathbf{x}_t - y_t)^2$$

where λ are parameters that can be set via cross-validation. This is can be viewed as an instance of weighted Ridge Regression, or more generally, QP problem and we can use standard techniques of convex optimization to find the solution.

4 Experiments

We have compared WRA against standard autoregressive model (ARIMA(q,0,0)) that is commonly used in practice. In our experiments we have used a number of artificial (ADS1, ADS2, ADS3) and real (FX1, FX2) datasets. For artificial datasets we have generated time series with 2,000 sample points, trained on the first 1,999 points and tested on the last point. To gain statistically significance, we repeat this procedure $1{,}000$ times. The processes used to generate these time series are summarized in Table 1. We have also used daily foreign exchange rates (12/31/1979 - 12/31/1998) for CAD/USD and FRF/USD pairs (FX1 and FX2 respectively) found in [2] as examples of real life non-stationary time series. Both FX1 and FX2 contain 4,774 points and we train on the first $T - 1$ points and test on the T-th observation, where $T = 250, \ldots, 4{,}774$. Results of our experiments are summarized in Table 2. Observe that results of each experiment are statistically significant using paired t-test and in each case WRA significantly outperforms the standard approach. Moreover, WRA has a better performance on at least 80% of individual runs in each experiment and the average error of ARIMA(q,0,0) is at least two times larger than that of WRA.

5 Conclusion

We presented generalization guarantees for learning in presence of non-stationary stochastic processes in terms of a weighted discrepancy measure that appears as a natural quantity in our general analysis. We show that our results provide learning guarantees for some well-known approaches such as learning autoregressive processes with linear models. We argued that our bounds can guide the design of time series prediction algorithms that would tame non-stationarity in the data by minimizing an upper bound on the discrepancy that can be estimated from the data [7, 4] or computed analytically. Our empirical results show that our algorithm significantly outperform standard autoregression models. In this work we focused here on the problem of time series prediction but the same learning guarantees and algorithms with only minor modifications can be formulated for random fields with more complex temporo-spatial or any other dependence structure.

References

[1] Agarwal, A., Duchi, J.C.: The Generalization Ability of Online Algorithms for Dependent Data. IEEE Transactions on Information Theory. 59(1), 573–587 (2013).

[2] DataMarket: `https://datamarket.com/data/set/237v`

[3] Doukhan, P.: Mixing: Properties and Examples. Lecture Notes in Statistics, vol. 85. Springer Verlag, New York (1989).

[4] Kifer, D., Ben-David, S., & Gehrke, J.: Detecting change in data streams. In: Proceedings of the 30th International Conference on Very Large Data Bases (2004).

[5] Koltchinskii, V., Panchenko, D.: Rademacher processes and bounding the risk of function learning. In High Dimensional Probability II, pp. 443-459, Birkhauser (1999).

[6] Kuznetsov, V., Mohri, M.: Generalization bounds for time series prediction with non-stationary processes. In: Proceedings of The 25th International Conference on Algorithmic Learning Theory (ALT 2014). Bled, Slovenia, October 2014. Springer, Heidelberg, Germany.

[7] Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation: learning bounds and algorithms. In: Proceedings of the Annual Conference on Learning Theory (COLT 2009). Omnipress (2009).

[8] Meir, R.: Nonparametric time series prediction through adaptive model selection. Machine Learning. 39(1), 5–34 (2000).

[9] Mohri, M., Rostamizadeh, A.: Rademacher complexity bounds for non-i.i.d. processes. In: Advances in Neural Information Processing Systems (NIPS 2008), pp. 1097–1104. MIT Press (2009).

[10] Mohri, M., Rostamizadeh, A.: Stability bounds for stationary φ-mixing and β-mixing processes. Journal of Machine Learning. 11 (2010).

[11] Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT press (2012).

[12] Mohri, M., Muñoz Medina, A.: New analysis and algorithm for learning with drifting distributions. In: Bshouty, N., Stoltz, G., Vayatis, N., Zeugmann, T. (eds.) ALT 2012. LNCS, vol. 7568, pp 124–138. Springer, Heidelberg (2012).

[13] Steinwart, I., Christmann, A.: Fast learning from non-i.i.d. observations. In: Bengio, Y., Schurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems (NIPS 2009), pp. 1768–1776. MIT Press (2009).

[14] Vapnik, V.: Statistical Learning Theory. Wiley-Interscience (1998).

[15] Yu, B.: Rates of convergence for empirical processes of stationary mixing sequences. Annals Probability. 22(1), 94–116 (1994).