Generalization bounds for mixing processes via delayed online-to-PAC conversions

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the generalization error of statistical learning algorithms in a non-*i.i.d.* setting, where the training data is sampled from a stationary mixing process. We develop an analytic framework for this scenario based on a reduction to online learning with delayed feedback. In particular, we show that the existence of an online learning algorithm with bounded regret (against a fixed statistical learning algorithm in a specially constructed game of online learning with delayed feedback) implies low generalization error of said statistical learning method even if the data sequence is sampled from a mixing time series. The rates demonstrate a trade-off between the amount of delay in the online learning game and the degree of dependence between consecutive data points, with near-optimal rates recovered in a number of well-studied settings when the delay is tuned appropriately as a function of the mixing time of the process.

1 Introduction

In machine learning, generalization denotes the ability of a model to infer patterns from a dataset of training examples and apply them to analyze previously unseen data (Shalev-Shwartz and Ben-David, 2014). The gap in accuracy between the model's predictions on new data and those on the training set is usually referred to as *generalization error*. Providing upper bounds on this quantity is a central goal in statistical learning theory. Classically, bounds based on notions of complexity (*e.g.*, VC dimension and Rademacher complexity) for the model's hypothesis space were used to provide uniform worst-case guarantees (see Bousquet et al., 2004; Vapnik, 2013; Shalev-Shwartz and Ben-David, 2014). However, results of this kind are often too loose to be applied to the most common machine learning over-parameterised models, such as deep neural networks (Zhang et al., 2021). As a consequence, several approaches have been proposed to obtain algorithm-dependent generalization bounds, which can adapt to the problem and be much tighter in practice than their uniform counterparts. Often, the underlying idea is that if the algorithm's output does not have a too strong dependence on the specific input dataset used for the training, then the model should not be prone to overfitting, and so generalize well. Examples of results that build onto these ideas are stability bounds, information-theoretic bounds, and PAC-Bayesian bounds (see, *e.g.*, Bousquet and Elisseeff, 2002; Russo and Zou, 2020; Hellström et al., 2023; Alquier, 2024).

Most results in the literature focus on the *i.i.d.* setting, where the training dataset is made of independent draws from some underlying data distribution. However, for several applications, this assumption is far from realistic. For instance, it excludes the case where observations received by the learner have some inherent temporal dependence, as it is the case for stock prices, daily energy consumption, or sensor data from physical environments (Ariyo et al., 2014; Takeda et al., 2016). This calls for the development of theory for addressing non-*i.i.d.* data. A common approach in the extant literature is to consider a class of non-*i.i.d.* data-generating processes usually referred to as stationary β-mixing

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

or φ-mixing processes. This assumption, together with a "blocking" trick introduced by Yu (1994), has led to a few results in the literature: Meir (2000), Mohri and Rostamizadeh (2008), Shalizi and Kontorovich (2013), and Wolfer and Kontorovich (2019) provided uniform worst-case generalization bounds, Steinwart and Christmann (2009) and Agarwal and Duchi (2012) discussed excess risk bound (comparing the algorithm's output with the best possible hypothesis), while Mohri and Rostamizadeh (2010) gave bounds based on a stability analysis (in the sense of Bousquet and Elisseeff, 2002).

Here, we propose propose results for the non-*i.i.d.* setting in the form of PAC-Bayesian bounds (Guedj, 2019; Alquier, 2024): high probability upper bounds on the expected generalization error of randomized learning algorithms. We achieve this by combining the "blocking" argument by Yu (1994) to manage the concentration of sums of correlated random variables, with the recent online-to-PAC conversion technique recently proposed by Lugosi and Neu (2023). Using their framework we show a new way to obtain generalization bounds for stationary dependent processes that satisfy a certain "short-memory" property (intuitively meaning that data points that are closer in time are more heavily dependent on each other). Our assumption slightly differs from β-mixing in the sense that we only need it to hold for a specific class of bounded loss functions. Among other results, this allows us to prove PAC-Bayesian generalization bounds for mixing processes. This complements previous work on such bounds that have only considered mild relaxations of the *i.i.d.* condition such as assuming that the data has a martingale structure (see, *e.g.*, Seldin et al., 2012; Chugg et al., 2023; Haddouche and Guedj, 2023). Notable exceptions are the works of Alquier and Wintenberger (2012), Alquier et al. (2013), and Eringis et al. (2022, 2024), who provided generalization bounds for a sequential prediction setting where both the data-generating process and the hypothesis class used for prediction are stable dynamical systems. Their results are proved under some very specific conditions on these systems, and their guarantees involve unspecified problem-dependent constants that may be large. In contrast, our bounds hold under general, simple-to-verify conditions and feature explicit constants.

The rest of the paper is organized as follows. In Section 2 we properly define the generalization error of a statistical learning algorithm for both *i.i.d.* and non-*i.i.d.* cases, and state our main assumption on the data dependence. Our main contribution lies in Section 3, where after recalling the results for the *i.i.d.* setting we show how to adapt this to stationary mixing processes. In Section 4 we provide concrete results of the bounds we can obtain through the online-to-PAC conversion. Finally in Section 5 we extend our results to the setting where the hypothesis class itself may consist of dynamical systems.

Notation. For a distribution over hypotheses $P \in \Delta_W$ and bounded function $f : \mathcal{W} \to \mathbb{R}$ we write $\langle P, f \rangle$ to refer to the expectation of $\mathbb{E}_{W\sim P}[f(W)]$. We denote $\mathcal{D}_{KL}(P||Q) = \mathbb{E}_{X\sim P}\left[\ln\left(\frac{P(X)}{Q(X)}\right)\right]$ to refer to the Kullback-Leibler divergence. We use $||.||$ to denote a norm on the Banach space $\mathcal{Q}$ of the finite signed measures, and $||.||_*$ the corresponding dual norm on the dual space $\mathcal{Q}^*$ of measurable functions f on $\mathcal{W}$ such that $||f||_* = \sup_{Q\in\mathcal{Q}:||Q||\leq 1}\langle Q, f\rangle$.

2 Preliminaries

The classical statistical learning framework usually considers a dataset $S_n = (Z_1, ..., Z_n)$, made of n *i.i.d.* elements drawn from a distribution μ over a measurable instance space $\mathcal{Z}$. Often, one can think of each Z_i as a feature-label pair (X_i, Y_i). Furthermore, we are given a measurable class $\mathcal{W}$ of hypotheses and a loss function $\ell : \mathcal{W}\times\mathcal{Z} \to \mathbb{R}_+$, with $\ell(w, z)$ measuring the quality of the hypothesis $w \in \mathcal{W}$ on the data instance $z \in \mathcal{Z}$. For any given hypothesis $w \in \mathcal{W}$, two key objects of interest are the *training error* $\widehat{\mathcal{L}}(w, S_n) = \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$ and the *test error* $\mathcal{L}(w) = \mathbb{E}_{Z'\sim\mu}[\ell(w, Z')]$, where the random element Z' has the same distribution as Z_i and is independent of S_n.

A learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ maps the training sample to an hypothesis in $\mathcal{W}$. More generally, we will focus on randomized learning algorithms, returning a probability distribution $P_{W_n|S_n} \in \Delta_{\mathcal{W}}$ over $\mathcal{W}$, conditionally on S_n (deterministic algorithms can be recovered as special cases, whose the outputs are Dirac distributions). The ultimate goal of the learner is to minimize the test error. Yet, this quantity cannot be computed without knowledge of the data generating distribution μ. In practice, one typically relies on the training error in order to gauge the quality of the algorithm. For an algorithm $\mathcal{A} : S_n \mapsto P_{W_n|S_n}$, we define the *generalization error* as the expected gap between training and test error:

$$\mathrm{Gen}(\mathcal{A}, S_n) = \mathbb{E}\left[\mathcal{L}(W_n) - \widehat{\mathcal{L}}(W_n, S_n)\middle| S_n\right].$$

The expectation in the above expression integrates over the randomness in the output of the algorithm $W_n \sim P_{W_n|S_n}$, conditionally on the sample S_n. We remark that the test error is *not* equal to the mean of the training error, due to the dependence of W_n on the training data.

We extend the previous setting by considering the case where the data have an intrinsic temporally ordered structure, and come in the form of a stationary process $(Z_t)_{t\in\mathbb{N}^*} \sim \nu$. Formally, we assume that the joint marginal distribution of any block $(Z_t, Z_{t-1}, \ldots, Z_{t-i})$ is the same as the distribution of $(Z_{t+j}, Z_{t+j-1}, \ldots, Z_{t+j-i})$ for any t, i and j, but the data points are not necessarily independent of each other. In particular, the marginal distribution of Z_t is constant and is denoted by μ. Thus, it is natural to continue to use the definition of the test loss and generalization error given above, although with the understanding that μ now refers to the marginal distribution of an independent copy of Z_1, a sample point from a stationary non-*i.i.d.* process. We remark here that other notions of the test loss may also be considered, and the framework that we propose can be extended to most natural definitions with little work (but potentially large notational overhead). In Section 5, we provide such an extension for a more general setting where the hypotheses themselves are allowed to have memory and the process may not be as strongly stationary as our assumption above requires.

In order to obtain generalization results we need to have some control on how strong the dependencies between different datapoints are allowed to be. To this regard, we consider the following assumption.

Assumption 1. *There exists a non-increasing sequence $(\phi_d)_{d\in\mathbb{N}^*}$ of non-negative real numbers such that, for all $w \in \mathcal{W}$ and all $t \in \mathbb{N}^*$:*

$$\mathbb{E}\left[\mathcal{L}(w) - \ell(w, Z_t)\middle|\mathcal{F}_{t-d}\right] \leq \phi_d\,,$$

where $\mathcal{L}(w) = \mathbb{E}_{Z'\sim\mu}[\ell(w, Z')]$, with Z' being independent on the process $(Z_t)_{t\in\mathbb{N}^*}$ and having as distribution the stationary marginal μ of the Z_t.

The intuition behind this assumption is that the loss associated with the observations Z_t becomes almost independent of the past after d steps, enabling us to treat each sequence of the form $(Z_t, Z_{t+d}, \ldots, Z_{t+(n-t)d})$ as an approximately *i.i.d.* sequence. Note that this assumption differs from the usual β-mixing assumption which requires the distribution of $Z_t|\mathcal{F}_{t-d}$ to be close to the marginal distribution μ for all t, in terms of total variation distance. Our assumption is somewhat weaker in the sense that it only requires the expected losses under these distributions to be close, and only a one-sided inequality is required. It is easy to verify that our assumption is satisfied if the process is β-mixing in the usual sense and the losses are bounded in $[0, 1]$.

3 Proving generalization bounds via online learning

Online learning focuses on algorithms that aim to improve performance incrementally as new information becomes available, often without any underlying assumption on how data are generated. The online learner's performance is typically measured leveraging the idea of regret. This involves introducing a cost function for the problem and defining the regret as the difference between the cumulative cost of the online learner and that of a fixed comparator. We refer to the monographs Cesa-Bianchi and Lugosi, 2006 and Orabona, 2019 for comprehensive overviews on online learning and regret analysis. Recently, Lugosi and Neu (2023) established a connection between upper bounds on the regret and generalization bounds, showing that the existence of a strategy with a bounded regret in a specially designed online game translates into a generalization bound, via a technique dubbed *online-to-PAC conversion*. Their focus is on the *i.i.d.* setting, where the training dataset is made of independent draws. Here, we show that this framework can naturally be extended beyond the *i.i.d.* assumption.

In what follows, we briefly review the setup of Lugosi and Neu (2023) in Section 3.1 and then describe our new extension of their model to the non-*i.i.d.* case in Section 3.2. In particular, we prove a high-probability bound for the generalization error of any statistical learning algorithm learnt with a stationary mixing process verifying Assumption 1.

3.1 Online-to-PAC conversions for *i.i.d.* data

Lugosi and Neu (2023) have recently established a framework to obtain generalization bounds via a reduction to online learning. Their technique allows to recover several classic PAC-Bayesian

results, and provide a range of generalizations thereof. The main idea of Lugosi and Neu (2023) is to introduce an online learning game called the *generalization game*, where the following steps are repeated for a sequence of rounds $t = 1, 2, \ldots, n$:

- the online learner picks a distribution $P_t \in \Delta_{\mathcal{W}}$;
- the adversary selects the cost function $c_t : w \mapsto \ell(w, Z_t) - \mathcal{L}(w)$;
- the online learner incurs the cost $\langle P_t, c_t \rangle = \mathbb{E}_{W \sim P_t}[c_t(W)]$;
- Z_t is revealed to the learner.

The learner can adopt any strategy to pick P_t, but they can only rely on past knowledge to make their prediction. Explicitly, if $\mathcal{F}_t$ denotes the sigma-algebra generated by $Z_1, ..., Z_t$, then P_t has to be $\mathcal{F}_{t-1}$-measurable. We also emphasize that in this setup the online learner is allowed to know the loss function ℓ and the distribution μ of the data points Z_t, and therefore by revealing the value of Z_t, the online learner may compute the entire cost function c_t.

We define the *regret* of the online learner against the possibly data-dependent *comparator* $P^* \in \Delta_{\mathcal{W}}$ as $\text{Regret}(P^*) = \sum_{t=1}^n \langle P_t - P^*, c_t \rangle$. Now, denote as $P_{W_n|S_n}$ the distribution produced by the supervised learning algorithm. With this notation, the generalization error can be written as $\text{Gen}(\mathcal{A}, S_n) = -\frac{1}{n}\sum_{t=1}^n \langle P_{W_n|S_n}, c_t \rangle$. By adding and subtracting the quantity $M_n = -\frac{1}{n}\sum_{t=1}^n \langle P_t, c_t \rangle$ we get the following decomposition.

Theorem 1 (Theorem 1 in Lugosi and Neu, 2023; see appendix A.1)**.** *With the notation introduced above,*

$$\text{Gen}(\mathcal{A}, S_n) = \frac{\text{Regret}_n(P_{W_n|S_n})}{n} + M_n \,. \tag{1}$$

The first of these terms correspond to the *regret* of the online learner against a fixed *comparator strategy* that picks $P_{W_n|S_n}$ at each step. The second term is a martingale and can be bounded in high probability with standard concentration tools. Indeed, since P_t is chosen before Z_t is revealed, one can easily check that $\mathbb{E}[\langle P_t, c_t \rangle | \mathcal{F}_{t-1}] = 0$. Thus, to prove a bound on the generalization error of the statistical learning algorithm, it is enough to find an online learning algorithm with bounded regret against $P_{W_n|S_n}$ in the generalization game.

As a concrete application of the above, the following generalization bound is obtained when picking the classic exponential weighted average (EWA) algorithm (Vovk, 1990; Littlestone and Warmuth, 1994; Freund and Schapire, 1997) as online strategy, and plugging its regret bound into (1).

Theorem 2 (Corollary 6 in Lugosi and Neu, 2023)**.** *Suppose that $\ell(w, z) \in [0, 1]$ for all w, z. Then, for any $P_1 \in \Delta_{\mathcal{W}}$ and $\eta > 0$, with probability at least $1 - \delta$ on the draw of S_n, uniformly on every learning algorithm $\mathcal{A} : S_n \mapsto P_{W_n|S_n}$, we have*

$$\text{Gen}(\mathcal{A}, S_n) \leq \frac{\mathcal{D}_{KL}(P_{W_n|S_n}||P_1)}{\eta n} + \frac{\eta}{2} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \,.$$

Proof. We can bound each term of (1) separately. A data-dependent bound for the regret term is obtained via a direct application of the regret analysis of EWA which brings the term $\frac{\mathcal{D}_{KL}(P_{W_n|S_n}||P_1)}{\eta n} + \frac{\eta}{2}$ (see Appendix B.1). The term $\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}$ results from bounding the martingale M_n via an application of Hoeffding–Azuma inequality. □

Note that the first term in the above bound is data-dependent due to the presence of $P_{W_n|S_n}$, and thus optimizing it requires a data-dependent choice of η, which is not allowed by Theorem 2. However, via a union bound argument it is possible to get a bound in the form

$$\text{Gen}(\mathcal{A}, S_n) = \mathcal{O}\left(\sqrt{\frac{\mathcal{D}_{KL}(P_{W_n|S_n}||P_1)}{n}} + \sqrt{\frac{1}{n}\log\left(\frac{\log n}{\delta}\right)}\right),$$

For the details, we refer to the proof of Corollary 5 of Lugosi and Neu (2023), which recovers a classical PAC-Bayes bound of McAllester (1998).

3.2 Online-to-PAC conversions for non-*i.i.d.* data

In what follows, we will drop the *i.i.d.* assumption for the data, and instead consider non-*i.i.d.* sequences satisfying Assumption 1. For this setting we define the following variant of the generalization game.

Definition 1 (Generalization game with delay)**.** *The generalization game with delay $d \in \mathbb{N}^*$ is an online learning game where the following steps are repeated for a sequence of rounds $t = 1, ..., n$:*

- *the online learner picks a distribution $P_t \in \Delta_{\mathcal{W}}$;*
- *the adversary selects the cost function $c_t : w \mapsto \ell(w, Z_t) - \mathcal{L}(w)$;*
- *the online learner incurs the cost $\langle P_t, c_t \rangle = \mathbb{E}_{W \sim P_t}[c_t(W)]$;*
- *if $t \geq d$, Z_{t-d+1} (and thus c_{t-d+1}) is revealed to the learner.*

The main difference between our version of the generalization game and the standard one of Lugosi and Neu (2023) is the introduction of a *delay* on the online learning algorithm's decisions. Specifically, we will force the online learner to only take information into account up to time $t - d$ when picking their action P_t. Clearly, setting $d = 1$ recovers the original version of the generalization game with no delay.

It is easy to see that the regret decomposition of Theorem 1 still remains valid in the current setting. The purpose of introducing the delay is to be able to make sure that the term $M_n = -\frac{1}{n}\sum_{t=1}^{n} \langle P_t, c_t \rangle$ is small. The lemma below states that the increments of M_n behave similarly to a martingale-difference sequence, thanks to the introduction of the delay.

Lemma 1. *Fix $d \in [\![1, n]\!]$. Under assumption 1, it holds for all $t \in [\![1, n]\!]$:*

$$\mathbb{E}[\langle -P_t, c_t \rangle | \mathcal{F}_{t-d}] \leq \phi_d \,.$$

where P_t and c_t are defined as in 1.

Proof. Since P_t is $\mathcal{F}_{t-d}$-measurable we have $\mathbb{E}[\langle -P_t, c_t \rangle | \mathcal{F}_{t-d}] = \langle P_t, \mathbb{E}[-c_t | \mathcal{F}_{t-d}] \rangle \leq \phi_d$, where the last step uses Assumption 1. □

Thus, by following the decomposition of Theorem 1, we are left with the problem of bounding the regret of the delayed online learning algorithm against $P_{W_n|S_n}$, denoted as $\mathrm{Regret}_{d,n}(P_{W_n|S_n}) = \sum_{t=1}^{n} \left\langle P_t - P_{W_n|S_n}, c_t \right\rangle$. The following proposition states a simple and clean bound that one can immediately derive from these insights.

Proposition 1 (Bound in expectation)**.** *Consider $(Z_t)_{t \in \mathbb{N}^*}$ satisfying Assumption 1 and suppose there exists a d-delayed online learning algorithm with regret bounded by $\mathrm{Regret}_{d,n}(P^*)$ against any comparator P^*. Then, the expected generalization of $\mathcal{A}$ is bounded as*

$$\mathbb{E}\left[\mathrm{Gen}(\mathcal{A}, S_n)\right] \leq \frac{\mathbb{E}\left[\mathrm{Regret}_{d,n}(P_{W_n|S_n})\right]}{n} + \phi_d \,.$$

Proof. By Theorem 1, it holds that $\mathbb{E}[\mathrm{Gen}(\mathcal{A}, S_n)] = \frac{\mathbb{E}[\mathrm{Regret}_{d,n}(P_{W_n|S_n})]}{n} + \mathbb{E}[M_n]$, where the regret is for a strategy P_t in the delayed generalization game. Hence, by Lemma 1

$$\mathbb{E}[M_n] = \mathbb{E}\left[-\frac{1}{n}\sum_{t=1}^{n} \langle P_t, c_t \rangle\right] = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\langle -P_t, c_t \rangle] = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{E}[\langle -P_t, c_t \rangle | \mathcal{F}_{t-d}]\right] \leq \phi_d \,,$$

which proves the claim. □

The above result holds in expectation over the training sample. We now provide a high-probability guarantee on the generalization error.

Theorem 3 (Bound in probability)**.** *Assume that $(Z_t)_{t \in \mathbb{N}^*}$ satisfies Assumption 1 and consider a d-delayed online learning algorithm with regret bounded by $R_{d,n}(P^*)$ against any comparator P^*. Then, for any $\delta > 0$, it holds with probability $1 - \delta$ on the draw of S_n, uniformly for all $\mathcal{A}$,*

$$\mathrm{Gen}(\mathcal{A}, S_n) \leq \frac{R_{d,n}(P_{W_n|S_n})}{n} + \phi_d + \sqrt{\frac{2d \log\left(\frac{d}{\delta}\right)}{n}}.$$

The proof of this claim follows directly from combining the decomposition of Theorem 1 with a standard concentration result for mixing processes that we state below.

Lemma 2. *Fix $d \in [\![1, n]\!]$ and consider $(Z_t)_{t\in\mathbb{N}^*}$ satisfying Assumption 1. Consider the generalization game of Definition 1. Then, for any $\delta > 0$, the following bound is satisfied with probability at least $1 - \delta$:*

$$M_n \leq \phi_d + \sqrt{\frac{2d \log\left(\frac{d}{\delta}\right)}{n}}.$$

The proof is based on a classic "blocking" technique due to Yu (1994). For the sake of completeness, we provide a proof in Appendix A.2.

4 New generalization bounds for non-*i.i.d.* data

The dependence on the delay d for the bounds that we presented in the previous section is non-trivial. Indeed, if on the one hand increasing the delay will reduce the magnitude of ϕ_d, on the other hand the regret of the online learner will grow with d. There is hence a trade-off between these two terms appearing in our bounds. In what follows, we derive some concrete generalization bounds from Theorem 3, under a number of different choices of the online learning algorithm. For concreteness, we will consider two types of mixing assumptions, but stress that the approach can be applied to any process that satisfies Assumption 1.

4.1 Regret bounds for delayed online learning

From Theorem 3, we can obtain a generalization bound using our framework if we have a regret bound for a delayed online algorithm. This is a well-known problem in the area of online learning (see, *e.g.*, Weinberger and Ordentlich, 2002; Joulani et al., 2013). In the following, we will leverage the following simple trick that allows us to extend the regret bounds of any online learning algorithm to its delayed counterpart, provided that the regret bound respects some specific assumptions.

Lemma 3 (Weinberger and Ordentlich, 2002)**.** *Consider any online algorithm whose regret satisfies $\mathrm{Regret}_n(P^*) \leq R(n)$ for any comparator P^*, where R is a non-decreasing real-valued function such that $y \mapsto yR(x/y)$ is a concave function of y for any fixed x. Then, for any $d \geq 1$ there exists an online learning algorithm with delay d such that, for any comparator P^*,*

$$\mathrm{Regret}_{d,n}(P^*) \leq dR\left(n/d\right) .$$

The proof idea is closely related to the blocking trick of Yu (1994), with an algorithmic construction that runs one instance of the base method for each index $i = 1, 2, \ldots, d$, with the i-th instance being responsible for the regret in rounds $i, i + d, i + 2d, \ldots$ (more details are provided in Appendix B.3). For most of the regret bounds that we consider, the function R takes the form $R(n) = O(\sqrt{n})$, so that the first term in the generalization bound is typically of order $\sqrt{d/n}$. Since this term matches the bound on M_n in Lemma 2, in this case the final generalization bound behaves effectively as if the sample size was n/d instead of n.

4.2 Geometric and algebraic mixing

The following definition gives two concrete examples of mixing processes that satisfy Assumption 1 with different choices of ϕ_d, and are commonly considered in the related literature (see, e.g., Mohri and Rostamizadeh, 2010, Levin and Peres, 2017).

Definition 2. *We say that a stationary process $(Z_t)_{t\in\mathbb{N}^*}$ satisfying Assumption 1 is:*

- geometrically mixing*, if $\phi_d = Ce^{-\frac{d}{\tau}}$, for some positive τ and C;*
- algebraically mixing*, if $\phi_d = Cd^{-r}$, for some positive r and C.*

Instantiating the bound of Theorem 3 to these two cases yields the following two corollaries.

Corollary 1. *Assume $(Z_t)_{t\in\mathbb{N}^*}$ is a geometrically mixing process with constants $\tau, C > 0$. Consider a d-delayed online learning algorithm with regret bounded by $R_{d,n}(P^*)$ for all comparators P^*.*

Then, setting $d = \lceil \tau \log n \rceil$, for any $\delta > 0$, with probability at least $1 - \delta$ we have that, uniformly for any algorithm $\mathcal{A}$,

$$\mathrm{Gen}(\mathcal{A}, S_n) \leq \frac{R_{d,n}(P_{W_n|S_n})}{n} + \frac{C}{n} + \sqrt{\frac{2\left(\tau \log n + 1\right) \log\left(\frac{\tau \log n + 1}{\delta}\right)}{n}}\,.$$

Up to a term linear in τ and some logarithmic factors, the above states that under the geometric mixing the same rates are achievable as in the *i.i.d.* setting. Roughly speaking, this amounts to saying that the effective sample size is a factor τ smaller than the original number of samples n, as long as generalization is concerned.

Corollary 2. *Assume $(Z_t)_{t \in \mathbb{N}^*}$ is an algebraic mixing process with constants $r, C > 0$. Consider a d-delayed online learning algorithm with regret bounded by $R_{d,n}(P^*)$ against any comparator P^*. Then, setting $d = \left(C^2 n\right)^{1/(1+2r)}$, for any $\delta > 0$, with probability at least $1 - \delta$ we have that, uniformly for any algorithm $\mathcal{A}$,*

$$\mathrm{Gen}(\mathcal{A}, S_n) \leq \frac{R_{d,n}(P_{W_n|S_n})}{n} + C\left(1 + \sqrt{\log(d/\delta)}\right) n^{-\frac{2r}{2(1+2r)}}\,.$$

This result suggests that the rates achievable for algebraically mixing processes are qualitatively much slower than what one can get for *i.i.d.* or geometrically mixing data sequences (although the rates do eventually approach $1/\sqrt{n}$ as r goes to infinity).

4.3 Multiplicative weights with delay

We start our discussion on possible online strategies by focusing on the classic exponential weighted average (EWA) algorithm (Vovk, 1990; Littlestone and Warmuth, 1994; Freund and Schapire, 1997). We fix a data-free prior $P_1 \in \Delta_{\mathcal{W}}$ and a learning rate parameter $\eta > 0$. We consider the updates

$$P_{t+1} = \operatorname*{arg\,min}_{P \in \Delta_{\mathcal{W}}} \left\{ \langle P, c_t \rangle + \frac{1}{\eta} \mathcal{D}_{KL}(P||P_t) \right\},$$

Combining the standard regret bound of EWA (see Appendix B.1) with Lemma 3 and Corollary 1 yields the result that follows.

Corollary 3. *Suppose that $(Z_t)_{t \in \mathbb{N}^*}$ is a geometric mixing process with constants $\tau, C > 0$. Suppose that $\ell(w, z) \in [0, 1]$ for all w, z. Then, for any $P_1 \in \Delta_{\mathcal{W}}$ and any $\delta > 0$, with probability at least $1 - \delta$, uniformly on any learning algorithm $\mathcal{A}$ we have*

$$\mathrm{Gen}(\mathcal{A}, S_n) \leq \frac{\mathcal{D}_{KL}(P^*||P_1)(\tau \log n + 1)}{\eta n} + \frac{\eta}{2} + \frac{C}{n} + \sqrt{\frac{2\left(\tau \log n + 1\right) \log\left(\frac{\tau \log n + 1}{\delta}\right)}{n}}\,.$$

This results suggests that when considering geometric mixing processes, by applying a union bound over a well-chosen range of η we recover the PAC-Bayes bound of McAllester (1998) up to a $O(\sqrt{\tau \log n})$ factor. A similar result can be derived from Corollary 2 for algebraically mixing processes, leading to a bound typically scaling as $n^{-2r/(2(1+2r))}$.

4.4 Follow the regularized leader with delay

In this subsection we extend the common class of online learning algorithms known as follow the regularized leader (FTRL, see *e.g.*, Abernethy and Rakhlin, 2009; Orabona, 2019) to the problem of learning with delay. FTRL algorithms are defined using a convex regularization function $h : \Delta_{\mathcal{W}} \to \mathbb{R}$. We restrict ourselves to the set of proper, lower semi-continuous and α-strongly convex functions with respect to a norm $||.||$ (and its respective dual norm $||.||_*$) defined on the set of signed finite measures on $\mathcal{W}$ (see Appendix B.2 for more details). The online procedure (without delay) of the FTRL algorithm is as follows:

$$P_{t+1} = \operatorname*{arg\,min}_{P \in \Delta_{\mathcal{W}}} \left\{ \sum_{s=1}^{t} \langle P, c_s \rangle + \frac{1}{\eta} h(P) \right\}.$$

The existence of the minimum is guaranteed by the compactness of $\Delta_{\mathcal{W}}$ under $\|\cdot\|$, and its uniqueness is ensured by the strong convexity of h. Combining the analysis of FTRL (see Appendix B.2) with Lemma 3 and Corollary 1 yields the following result.

Corollary 4. *Suppose that $(Z_t)_{t\in\mathbb{N}^*}$ is a geometric mixing process with constants $\tau, C > 0$. Suppose that $\ell(w, z) \in [0, 1]$ for all w, z. Assume there exists $B > 0$ such that for all t, $||c_t||_* \leq B$. Then, for any $P_1 \in \Delta_{\mathcal{W}}$, for any $\delta > 0$ with probability at least $1 - \delta$ on the draw of S_n, uniformly for all $\mathcal{A}$,*

$$\mathrm{Gen}(\mathcal{A}, S_n) \leq \frac{(h(P^*) - h(P_1))\,(\tau \log n + 1)}{\eta n} + \frac{\eta B^2}{2\alpha} + \frac{C}{n} + \sqrt{\frac{2\,(\tau \log n + 1) \log\left(\frac{\tau \log n + 1}{\delta}\right)}{n}}.$$

This generalization bound is similar to the bound of Theorem 9 of Lugosi and Neu (2023) up to a $O(\sqrt{\tau \log n})$ factor, when applying a union-bound argument over an appropriate grid of learning-rates η. In particular, this result recovers PAC-Bayesian bounds like those of Corollary 3 when choosing $h = \mathcal{D}_{\mathrm{KL}}(\cdot\|P_1)$. We refer to Section 3.2 in Lugosi and Neu (2023) for more discussion on such bounds. As before, a similar result can be stated for algebraically mixing processes, with the leading terms approaching zero at rate of $n^{-2r/2(1+2r)}$ instead of $n^{-1/2}$.

5 Generalization bounds for dynamic hypotheses

Finally, inspired by the works of Eringis et al. (2022, 2024), we extend our framework to accommodate loss functions ℓ that rely not only on the last data point Z_t, but on the entire data sequence $\overline{Z}_t = (Z_t, Z_{t-1}, \ldots, Z_1)$. Formally, we will consider loss functions of the form $\ell : \mathcal{W} \times \mathcal{Z}^* \to \mathbb{R}_+$[1] and write $\ell(w, \overline{z}_t)$ to denote the loss associated with hypothesis $w \in \mathcal{W}$ on sequence $\overline{z}_t \in \mathcal{Z}^t$. This consideration extends the learning problem to class of dynamical predictors such as Kalman filters, autoregressive models, or recurrent neural networks (RNNs), broadly used in time-series forecasting (Ariyo et al., 2014; Takeda et al., 2016). Specifically, if we think of $z_t = (x_t, y_t)$ as a data-pair of context and observation, in time-series prediction we usually not only rely on the context x_t but also on the past sequence of contexts and observations $(x_{t-1}, y_{t-1}, \ldots, x_1, y_1)$. As an example, consider $\ell(w, z_t, \ldots, z_1) = \frac{1}{2}(y_t - h_w(x_t, z_{t-1}, \ldots, z_1))^2$ where $h \in \mathcal{H}$ is a function class parameterized by $\mathcal{W}$. For this type of loss function a natural definition of the test error is:

$$\widetilde{\mathcal{L}}(w) = \lim_{n\to\infty} \mathbb{E}[\ell(w, Z'_t, Z'_{t-1}, ..., Z'_{t-n})],$$

where $\overline{Z}'_t = (Z'_t, Z'_{t-1}, \ldots)$ is a semi-infinite random sequence drawn from the same stationary process that has generated the data $\overline{Z}_t$. We consider the following assumption.

Assumption 2. *For a given process $(Z_t)_{t\in\mathbb{Z}}$ with joint-distribution ν over $\mathcal{Z}^{\mathbb{Z}}$ and same marginals μ over $\mathcal{Z}$, there exists a non-increasing sequence $(\phi_d)_{d\in\mathbb{N}^*}$ of non-negative real numbers such that the following holds for all $w \in \mathcal{W}$, for all $t \in \mathbb{N}^*$:*

$$\mathbb{E}\left[\ell(w, Z_t, \ldots, Z_1) - \widetilde{\mathcal{L}}(w)\middle|\mathcal{F}_{t-d}\right] \leq \phi_d.$$

This is a generalization of Assumption 1 in the sense that taking $\ell(w, Z_t, \ldots, Z_1) = \ell(w, Z_t)$ simply amounts to requiring the same mixing condition as before. For our online-to-PAC conversion we consider the same framework as in Definition 1, except that now the cost function is defined as

$$c_t : w \mapsto \ell(w, Z_t, \ldots, Z_1) - \widetilde{\mathcal{L}}(w)\,.$$

Then it easy to check that result of Lemma 2 still holds for this specific cost, and we can thus extend all the results of Section 4. For concreteness, we state the following adaptation of Theorem 3 below.

Theorem 4. *Assume $(Z_t)_{t\in\mathbb{Z}}$ which satisfies Assumption 2 and consider a d-delayed online learning algorithm with regret bounded by $R_{d,n}(P^*)$ against any comparator P^*. Then, for any $\delta > 0$, it holds with probability $1 - \delta$:*

$$\mathrm{Gen}(\mathcal{A}, S_n) \leq \frac{R_{d,n}(P_{W_n|S_n})}{n} + \phi_d + \sqrt{\frac{2d \log\left(\frac{d}{\delta}\right)}{n}}.$$

[1]Here, $\mathcal{Z}^*$ denotes the disjoint union $\mathcal{Z}^* = \sqcup_{t\in\mathbb{N}}\mathcal{Z}^t$.

To see that Assumption 2 can be verified and the resulting bounds can be meaningfully applied, consider the following concrete assumptions about the hypothesis class, the loss function, and the data generating process. The first assumption says that for any given hypothesis, the influence of past data points on the associated loss vanishes with time (*i.e.*, the hypothesis forgets the old data points at a controlled rate).

Assumption 3. *There exists a decreasing sequence $(B_d)_{d\in\mathbb{N}^*}$ of non-negative real numbers such that for any two sequences $\overline{z}_t = (z_t, \dots, z_i)$ and $\overline{z}'_t = (z'_t, \dots, z'_j)$ of possibly different lengths that satisfy $z_k = z'_k$ for all $k \in t, \dots, t-d+1$, we have $|\ell(w, \overline{z}_t) - \ell(w, \overline{z}'_t)| \le B_d$, for all $w \in \mathcal{W}$.*

This condition can be verified for stable dynamical systems like autoregressive models, certain classes of RNNs, or sequential predictors that have bounded memory by design (see Eringis et al., 2022, 2024). The next assumption is a refinement of Assumption 1, adapted to the case where the loss function acts on blocks of d data points $\overline{z}_{t-d+1:t} = (z_t, z_{t-1}, \dots, z_{t-d+1})$.

Assumption 4. *Let $\overline{Z}_t = (Z_t, \dots, Z_1)$ be a sequence of data points and let $\overline{Z}'_t = (Z'_t, \dots, Z'_0, \dots)$ be an independent copy of the same process. Then, there exists a decreasing sequence $(\beta_d)_{d\in\mathbb{N}^*}$ non-negative real numbers such that the following is satisfied for all hypotheses $w \in \mathcal{W}$ and all $d \in \mathbb{N}^*$:*

$$\mathbb{E}\left[\ell(w, \overline{Z}'_{t-d+1:t}) - \ell(w, \overline{Z}_{t-d+1:t}) \middle| \mathcal{F}_{t-2d}\right] \le \beta_d\,.$$

This assumption can be verified whenever the loss function is bounded and the joint distribution of the data block $\overline{Z}_{t-d+1:t}$ satisfies a β-mixing assumption. In more detail, this latter condition amounts to requiring that the conditional distribution of each data block given a block that trails d steps behind is close to the marginal distribution in total variation distance, up to an additive term of β_d. The following proposition shows that these two simple conditions together imply that Assumption 2 holds, and that thus the bound of Theorem 4 can be meaningfully instantiated for bounded-memory hypothesis classes deployed on mixing processes.

Proposition 2. *Suppose that the loss function satisfies Assumption 3 and the data distribution satisfies Assumption 4. Then Assumption 2 is satisfied with $\phi_d = 2B_{d/2} + \beta_{d/2}$.*

6 Conclusion

We have developed a general framework for deriving generalization bounds for non-i.i.d. processes under a general mixing assumption, via an extension of the online-to-PAC-conversion framework of Lugosi and Neu (2023). Among other results, this approach has allowed us to prove PAC-Bayesian generalization bounds for such data in a clean and transparent way, and even study classes of dynamic hypotheses under a simple bounded-memory condition. These results provide a clean and tight alternative to the results of (Alquier and Wintenberger, 2012; Eringis et al., 2022). The generality of our approach further demonstrates the power of the Online-to-PAC scheme of Lugosi and Neu (2023), and in particular our results provide further evidence that this framework is particularly promising for developing techniques for generalization in non-i.i.d. settings. We hope that flexibility of our framework will find further uses and enables more rapid progress in the area.

References

Abernethy, J. and Rakhlin, A. (2009). Beating the adaptive bandit with high probability. *IEEE Information Theory and Applications Workshop*.

Agarwal, A. and Duchi, J. (2012). The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1).

Alquier, P. (2024). User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2).

Alquier, P., Li, X., and Wintenberger, O. (2013). Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 1(1).

Alquier, P. and Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3).

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Stock price prediction using the ARIMA model. *UKSim-AMSS International Conference on Computer Modelling and Simulation*.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). *Introduction to Statistical Learning Theory*. Springer.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Chugg, B., Wang, H., and Ramdas, A. (2023). A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *Journal of Machine Learning Research*, 24(372).

Eringis, D., Leth, J., Tan, Z., Wisniewski, R., and Petreczky, M. (2022). PAC-Bayesian-like error bound for a class of linear time-invariant stochastic state-space models. *arXiv:2212.14838*.

Eringis, D., Leth, J., Tan, Z., Wisniewski, R., and Petreczky, M. (2024). PAC-Bayes generalisation bounds for dynamical systems including stable rnns. *AAAI Conference on Artificial Intelligence*.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *Second congress of the French Mathematical Society*.

Haddouche, M. and Guedj, B. (2023). PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2023(4).

Hellström, F., Durisi, G., Guedj, B., and Raginsky, M. (2023). Generalization bounds: Perspectives from information theory and PAC-Bayes. *arXiv:2309.04381*.

Joulani, P., Gyorgy, A., and Szepesvári, C. (2013). Online learning under delayed feedback. *ICML*.

Levin, D. and Peres, Y. (2017). *Markov chains and mixing times*. American Mathematical Soc.

Littlestone, N. and Warmuth, M. (1994). The weighted majority algorithm. *Information and computation*, 108(2).

Lugosi, G. and Neu, G. (2023). Online-to-PAC conversions: Generalization bounds via regret analysis. *arXiv:2305.19674*.

McAllester, D. A. (1998). Some PAC-Bayesian theorems. *COLT*.

Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39.

Mohri, M. and Rostamizadeh, A. (2008). Rademacher complexity bounds for non-i.i.d. processes. *NeurIPS*.

Mohri, M. and Rostamizadeh, A. (2010). Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(26).

Orabona, F. (2019). A modern introduction to online learning. *arXiv:1912.13213*.

Russo, D. and Zou, J. (2020). How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1).

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. (2012). PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12).

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press.

Shalizi, C. and Kontorovich, A. (2013). Predictive PAC learning and process decompositions. *NeurIPS*.

Steinwart, I. and Christmann, A. (2009). Fast learning from non-i.i.d. observations. *NeurIPS*.

Takeda, H., Tamura, Y., and Sato, S. (2016). Using the ensemble Kalman filter for electricity load forecasting and analysis. *Energy*, 104.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vovk, V. (1990). Aggregating strategies. *COLT*.

Weinberger, M. and Ordentlich, E. (2002). On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7).

Wolfer, G. and Kontorovich, A. (2019). Minimax learning of ergodic Markov chains. *ALT*.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1).

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3).

A Omitted proofs

A.1 The proof of Theorem 1

Let $(P_t)_{t=1}^n \in \Delta_{\mathcal{W}}^n$ be the predictions of an online learner playing the *generalization game*. Then

$$\begin{aligned}
\mathrm{Gen}(\mathcal{A}, S_n) &= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\ell_t(W_n) - \mathcal{L}(W_n)|S_n] \\
&= -\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[c_t(W_n)|S_n] \\
&= -\frac{1}{n}\sum_{t=1}^{n} \langle P_{W_n|S_n}, c_t\rangle \\
&= \frac{1}{n}\sum_{t=1}^{n} \langle P_t - P_{W_n|S_n}, c_t\rangle - \frac{1}{n}\sum_{t=1}^{n}\langle P_t, c_t\rangle \\
&= \frac{\mathrm{Regret}_n(P_{W_n|S_n})}{n} + M_n.
\end{aligned}$$

A.2 The proof of Lemma 2

Assume $n = Kd$ for simplicity:

$$\begin{aligned}
M_n &= -\frac{1}{n}\sum_{t=1}^{n}\langle P_t, c_t\rangle \\
&= \frac{1}{dK}\sum_{i=1}^{d}\sum_{t=1}^{K}\langle -P_{i+d(t-1)}, c_{i+d(t-1)}\rangle
\end{aligned}$$

We denote $X_t^{(i)} = \langle -P_{i+d(t-1)}, c_{i+d(t-1)}\rangle$ and we want to bound in high-probability the term $\frac{1}{K}\sum_{t=1}^{K} X_t^{(i)}$. Let also denote $\mathcal{F}_t^{(i)} = \mathcal{F}_{i+d(t-1)}$. Then for $i \in [\![1, d]\!]$, we can write using Chernoff's technique that for all $\lambda > 0$ it holds:

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{K}\sum_{t=1}^{K} X_t^{(i)} \geq u\right) &\leq \frac{\mathbb{E}\left[e^{\frac{\lambda}{K}\sum_{t=1}^{K} X_t^{(i)}}\right]}{e^{\lambda u}} \\
&\leq \mathbb{E}\left[e^{\frac{\lambda}{K}\sum_{t=1}^{K-1} X_t^{(i)}}\mathbb{E}\left[e^{\frac{\lambda}{K}X_K^{(i)}}\middle|\mathcal{F}_{K-1}^{(i)}\right]\right]e^{-\lambda u}.
\end{aligned}$$

Now remark that:

$$\mathbb{E}\left[e^{\frac{\lambda}{K}X_K^{(i)}}\middle|\mathcal{F}_{K-1}^{(i)}\right] = \mathbb{E}\left[e^{\frac{\lambda}{K}(X_K^{(i)} - \mathbb{E}[X_K^{(i)}|F_{K-1}^{(i)}])}\middle|F_{K-1}^{(i)}\right]e^{\frac{\lambda}{K}\mathbb{E}[X_K^{(i)}|F_{K-1}^{(i)}]}.$$

If we denote $Z = X_K^{(i)} - \mathbb{E}[X_K^{(i)}|F_{K-1}^{(i)}]$ then $|Z| \leq 2$ and $\mathbb{E}[Z|F_{K-1}^{(i)}] = 0$ so via Hoeffding's lemma:

$$\mathbb{E}[e^{\frac{\lambda}{K}Z}] \leq e^{\frac{\lambda^2}{2K^2}}.$$

Now by construction of the P_t and because of Lemma 1 it follows that for all i, $\mathbb{E}[X_K^{(i)}|F_{K-1}^{(i)}] \leq \phi_d$. Repeating the same reasoning for each term of the sum yields:

$$\mathbb{P}\left(\frac{1}{K}\sum_{t=1}^{K} X_t^{(i)} \geq u\right) \leq e^{\frac{\lambda^2}{2K}}e^{\lambda\phi_d}e^{-\lambda u}.$$

Optimzing with $\lambda = K(u - \phi_d)$ and taking $\delta = e^{-\frac{K(u-\phi)^2}{2}}$ it finally holds for any $\delta > 0$, with probability $1 - \frac{\delta}{d}$:

$$\frac{1}{K}\sum_{t=1}^{K} X_t^{(i)} \leq \phi_d + \sqrt{\frac{2\log\left(\frac{d}{\delta}\right)}{K}}.$$

Thus applying a union bound we have with probability $1-\delta$:

$$M_n \leq \phi_d + \sqrt{\frac{2\log\left(\frac{d}{\delta}\right)}{K}},$$

which concludes the proof. □

A.3 Proof of Proposition 2

Suppose without loss of generality that d is even and define $d' = d/2$. For the proof, let $\overline{Z}'_n$ be a semi-infinite sequence drawn independently from the same process as $\overline{Z}_n$. Then, we have

$$\begin{aligned}
\tilde{\mathcal{L}}(w) &= \lim_{n\to\infty} \mathbb{E}[\ell(w, Z'_t, Z'_{t-1}, ..., Z'_{t-n})] \\
&\leq \mathbb{E}[\ell(w, Z'_t, Z'_{t-1}, \ldots, Z'_{t-d'})] + B_{d'} \\
&\leq \mathbb{E}\left[\ell(w, Z_t, Z_{t-1}, \ldots, Z_{t-d'})|\,\mathcal{F}_{t-2d'}\right] + B_{d'} + \beta_{d'} \\
&\leq \mathbb{E}\left[\ell(w, Z_t, Z_{t-1}, \ldots, Z_{t-d'}, \ldots, Z_1)|\,\mathcal{F}_{t-2d'}\right] + 2B_{d'} + \beta_{d'} \\
&\leq \mathbb{E}\left[\ell(w, Z_t, Z_{t-1}, \ldots, Z_1)|\,\mathcal{F}_{t-2d'}\right] + 2B_{d'} + \beta_{d'},
\end{aligned}$$

where we used Assumption 3 in the first inequality, Assumption 4 in the second one, and Assumption 3 again in the last step. This proves the statement. □

B Online Learning Tools and Results

B.1 Regret Bound for EWA

Recalling EWA updates we have:

$$P_{t+1} = \operatorname*{arg\,min}_{P\in\Delta_{\mathcal{W}}} \left\{ \langle P, c_t\rangle + \frac{1}{\eta}\mathcal{D}_{KL}(P||P_t) \right\},$$

where $\eta > 0$ is a learning-rate parameter. The minimizer can be shown to exist and satisfies:

$$\frac{\mathrm{d}P_{t+1}}{\mathrm{d}P_t}(w) = \frac{e^{-\eta c_t(w)}}{\int_{\mathcal{W}} e^{-\eta c_t(w')}\mathrm{d}P_t(w')},$$

and the following result holds.

Proposition 3. *For any prior $P_1 \in \Delta_{\mathcal{W}}$ and any comparator $P^* \in \Delta_{\mathcal{W}}$ the regret of EWA simultaneously satisfies for $\eta > 0$:*

$$\mathrm{Regret}(P^*) \leq \frac{\mathcal{D}_{KL}(P^*||P_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n} ||c_t||^2_\infty.$$

We refer the reader to Appendix A.1 of Lugosi and Neu (2023) for a complete proof of the result above.

B.2 Regret Bound for FTRL

We say that h is $\alpha-$strongly convex if the following inequality is satisfied for all $P, P' \in \Delta_{\mathcal{W}}$ and all $\lambda \in [0,1]$:

$$h(\lambda P + (1-\lambda)P') \leq \lambda h(P) + (1-\lambda)h(P') - \frac{\alpha\lambda(1-\lambda)}{2}||P-P'||^2.$$

Recalling the FTRL updates:

$$P_{t+1} = \operatorname*{arg\,min}_{P\in\Delta_{\mathcal{W}}} \left\{ \sum_{s=1}^{t}\langle P, c_s\rangle + \frac{1}{\eta}h(P) \right\},$$

the following results holds.

Proposition 4. *For any prior $P_1 \in \Delta_{\mathcal{W}}$ and any comparator $P^* \in \Delta_{\mathcal{W}}$ the regret of FTRL simultaneously satisfies for $\eta > 0$:*

$$\mathrm{Regret}_n(P^*) \leq \frac{h(P^*) - h(P_1)}{\eta} + \frac{\eta}{2\alpha} \sum_{t=1}^{n} ||c_t||_*^2.$$

We refer the reader to Appendix A.3 of Lugosi and Neu (2023) for a complete proof of the results above.

B.3 Details about the reduction of Weinberger and Ordentlich (2002)

For concretenes we formally present how to turn any online learning algorithm into its delayed version. For sake of convenience, assume $n = Kd$. We denote $\tilde{c}_t^{(i)} = c_{i+d(t-1)}$ (for instance $\tilde{c}_1^{(1)} = c_1$ is the cost revealed at time $d+1$). Then we create d instances of horizon time K of the online learning as follows, for $i = 1, \ldots, d$:

- We initialize $\tilde{P}_1^{(i)} = P_0$,
- for each block i of length K we update for $t = 1, \ldots, K$:

$$\tilde{P}_{t+1}^{(i)} = \mathrm{OL}_{\text{update}}\left((\tilde{c}_s^{(i)})_{s=1}^{t}\right).$$

Here $\mathrm{OL}_{\text{update}}$ refers to the update function of the online learning algorithm we consider which can possibly depend of the whole history of cost functions (e.g., in the case of the FTRL update).

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim that we present a new framework adapted from Lugosi and Neu, 2023 to prove generalization bounds in non-*i.i.d* setting. We present it in Section 3and we provide PAC-Bayesian bounds in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main result of the paper lies in Section 3.2 and is carefully explained. Regarding Section 4 where most of the results are presented we give all the technical results and references in the AppendixB.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: paper does not include experiments requiring code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper does not include experiments requiring code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The contribution is mainly theoretical so we do not discuss these issues in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We do not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.