Generalization Bounds for Dependent Data using Online-to-Batch Conversion

Sagnik Chatterjee **Manuj Mukherjee** **Alhad Sethi**
Indraprastha Institute of Information Technology, Delhi (IIIT-Delhi)

Abstract

In this work, we give generalization bounds of statistical learning algorithms trained on samples drawn from a dependent data source both in expectation and with high probability, using the Online-to-Batch conversion paradigm. We show that the generalization error of statistical learners in the dependent data setting is equivalent to the generalization error of statistical learners in the i.i.d. setting up to a term that depends on the decay rate of the underlying mixing stochastic process. Our proof techniques involve defining a new notion of stability of online learning algorithms based on Wasserstein distances and employing "near-martingale" concentration bounds for dependent random variables to arrive at appropriate upper bounds for the generalization error of statistical learners trained on dependent data. Finally, we prove that the Exponential Weighted Averages (EWA) algorithm satisfies our new notion of stability and instantiate our bounds using the EWA algorithm.

1 INTRODUCTION

An offline statistical learning algorithm (also referred to as a *batch learner*) A is a randomized map from a set $\mathcal{S}$ of instances drawn from a fixed instance space $\mathcal{Z}$ to a fixed hypothesis space $\mathcal{H}$ associated with A. In the supervised learning setup, the objective is to *learn* a hypothesis function $h \in \mathcal{H}$ from a given set $\mathcal{S}$ of *training* instances using A, such that h performs 'well' on unseen *test* instances with respect to some appropriately chosen loss function. To quantify the performance of the offline learner A, a commonly used metric is the *generalization error* of A. The generalization error[1] of A is the difference of the average loss incurred by the hypothesis returned by A on training and test instances.

Classically, the generalization error of an offline learner A has been characterized in terms of combinatorial complexity measures of the hypothesis space $\mathcal{H}$, such as the VC dimension [Yu94; Mei00]. With the advent of modern over-parameterized models, where the number of tunable hyper-parameters far exceeds the size of the training set, generalization bounds based on combinatorial measures are often vacuous in nature [Zha+17; Zha+21]. In an effort to come up with better measures of the generalization error of offline learners, researchers have proposed algorithm-dependent generalization error bounds, such as bounds due to algorithmic stability [BE02], information-theoretic properties [RZ16; XR17], or bounds that are PAC-Bayesian in nature [Hel+24; Alq24]. Of particular relevance to this work is the recent work of Lugosi et al. [LN23], which applies the Online-to-Batch (OTB) framework of [CBCG04] to derive algorithm-dependent generalization error bounds for offline learners.

We remark at this point that most of the generalization error bounds in the literature operate on the underlying assumption that the training and test samples are drawn i.i.d. from the same (unknown) underlying distribution. In many real-world applications, such as learning from time-series dependent data or Federated Learning setups, the i.i.d. assumption does not hold [Vid13; Ami+22; Xio+22; Iye24; Li+24]. Prior works that have tried to address generalization error bounds in non-i.i.d. settings require restrictive stability assumptions on the offline learner [MR10; Fu+23].

1.1 Main Results

In an effort to address the previous shortcomings, we derive bounds on the generalization error of

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

[1] We will sometimes use the contraction *gen-err* for the sake of brevity.

any offline learner by extending the aforementioned Online-to-Batch paradigm to the non-i.i.d. setting, where we assume that the offline learners are trained on data sampled from a stochastic process that is "mixing" (see Definition 2) to a stationary distribution. Mixing is a fairly natural assumption that captures the notion of non-i.i.d'ness for learning tasks (see for example [Yu94; Mei00; LKS05; MR07; MR10; AD11; Duc+12; KM17; Zha+19; Fu+23]). The Online-to-Batch technique allows one to bound the gen-err of an offline learner via the *regret*[2] of an artificially constructed online learner. This allows us to derive gen-err bounds for any offline learner A, without requiring stability assumptions on A itself (unlike in [MR10; Fu+23]), and instead, the stability assumption gets shifted to the artificial online learner. This indicates that our contributions go beyond simply integrating different paradigms. In fact, to the best of our knowledge, along with the concurrent work [ACN25], we are the first to provide gen-err bounds for offline learners (possibly not conforming to any notion of stability) trained on non-i.i.d. data.

However, there is an apparent issue: As mentioned previously, our setup requires the online learners in the Online-to-Batch technique to satisfy a novel notion of stability, which we call *Wasserstein-stability.*[3] It is not apriori evident if such classes of online learners even exist. We resolve the above issue by proving that the class of Exponentially Weighted Averages (EWA) learners conforms to the notion of Wasserstein stability, a result that may be of independent interest. This result then allows us to instantiate our gen-err bounds via EWA learners. We summarize our results in the following informal theorems. The first result provides a framework to obtain gen-err bounds on offline learners via regret of Wasserstein stable online learners.

Theorem (Informal). *Let A be any offline learner trained on a set of n instances drawn from a suitably mixing random process. Then, the expected gen-err of A is upper bounded by $\frac{1}{n}\mathbb{E}[regret_{\mathcal{L}}] + O\left(\frac{1}{n}\right)$, where $\mathcal{L}$ is any arbitrary Wasserstein-stable online learner. Furthermore, the gen-err of A is upper bounded by $\frac{1}{n}regret_{\mathcal{L}} + O\left(\sqrt{\frac{1}{n}\cdot\log\left(1/\delta\right)}\right)$ w.p. $\geq 1-\delta$, for any $\delta > 0$.*

The next result instantiates the framework in the previous theorem by using the EWA online learner.

Theorem (Informal). *For any distribution P_1 over the hypothesis space $\mathcal{H}$, any $\delta > 0$, and an appropriately chosen constant $C > 0$, w.p. $\geq 1-\delta$, the gen-error of any offline learning algorithm A trained on $S_n = (Z_1,\ldots,Z_n)$ drawn from a suitably mixing process $\{Z_t\}_{t\in\mathbb{N}}$ is*

$$O\left(\frac{\mathrm{D_{KL}}\left(P_{A(S_n)}\,\|\,P_1\right) + C\log n + \sqrt{\log n\log\left(1/\delta\right)}}{\sqrt{n}}\right),$$

where $P_{A(S_n)}$ is the distribution on $\mathcal{H}$ which A outputs.

1.2 Overview of Proof Techniques

The Online-to-Batch framework allows us to bound the generalization error of offline learning algorithms by the sum of the regret of an artificially constructed online learning algorithm and the normalized sum of the expected costs incurred by the online learner. In the i.i.d. setting, the costs incurred by the online learner form a martingale difference sequence [CBCG04; LN23]. This simple but powerful observation allows one to upper bound the sum of expected costs using standard concentration inequalities. However, in the dependent data setting, the expected costs incurred by the online learner no longer form a martingale difference sequence. Consequently, it is no longer straightforward to apply concentration inequalities to bound the generalization error of the offline learning algorithms, as in the i.i.d case of [LN23].

To circumvent this issue, we rewrite the sum of expected costs of the online learning algorithm as the expected cost incurred by this online learner at time step $t+\tau$ with respect to its decision at step t, and a few remainder terms – see Lemma 9. The first term is a so-called "near-martingale" [MR10; AD11; Duc+12]. More precisely, this term consists of a sum of random variables forming a martingale difference sequence and an additional expectation term that can be bounded using the mixing coefficients of the random process from which the training data is drawn.

Finally, the remainder terms also consist of differences of expectations of costs incurred by the online learner at one time step with respect to either its output at a different time step or by the output of the offline learner. In order to bound these terms, we impose a doubly-Lipschitz condition on the loss function, boundedness of the observation and hypothesis spaces of the learning problem, and *Wasserstein stability* (see Definition 5) of the online learner. The choice of a Wasserstein distance based stability criterion is motivated by its dual form (see Lemma 3), which is used to bound the difference in expected cost incurred by the online learner in successive steps. This additional stability criterion comes for free as we show that the canonical EWA online learner used to instantiate our bound happens to be Wasserstein-stable – see Theorem 3 and Corollary 14.

[2] Regret of an online learner is defined as the difference between the total cost incurred by the online learner and any fixed offline learner. See Eq. (5) for a precise definition.

[3] See Definition 5 for a formal treatment.

1.3 Organization

Section 2 compares and contrasts our work with a vast body of literature related to deriving generalization bounds for statistical learning algorithms. In particular, a detailed comparison with the concurrent work [ACN25] is provided in Section 2.1. Section 3 introduces the mathematical prerequisites and sets up the problem. Section 4 develops the framework for obtaining generalization error upper bounds for statistical learning algorithms trained on data drawn from mixing processes. Following this, Section 5 shows that the EWA learner conforms to the notion of stability introduced in this work and uses this fact to instantiate the generalization error bounds of Section 4. The paper concludes with Section 6, which summarizes the contributions of this work and provides a few avenues for future work. Due to space constraints and better readability, proofs of certain technical results have been relegated to the Appendices.

2 Related Works

In learning with non-i.i.d. data, previous works have broadly focused on two approaches, via uniform convergence over complexity measures of the hypothesis space [Yu94; Mei00; MR08], or via data dependent bounds on mixing processes [MR10; Zha+19; Fu+23]. As argued before, the first class of bounds can be vacuous in overparameterized regimes, whereas the second approach suffered from the restrictive stability assumption on the offline learner.

As argued previously, we get rid of the stability assumption on offline learners by employing the Online-to-Batch framework. The Online-to-Batch framework was introduced in the seminal work of [CBCG04], which was later extended to give sharp generalization bounds for constrained linear classes based on complexity measures of the hypothesis classes [KT08]. Recently, [LN23] gave a framework that consolidates a vast quantity of the Online-to-Batch literature under its umbrella to obtain generalization bounds for offline learning algorithms trained on i.i.d. data.

Our work also introduces a new notion of algorithmic stability for online learners, which we call Wasserstein stability. Our definition of stability is incomparable to the more popular notions of algorithmic stability in the literature such as uniform stability [BE02; MR10; Fu+23] for offline learners, or the stability notion for online learners introduced by [AD11]. Our definition of stability is similar to the notion of *one-step differential stability* introduced by [Abe+19]. Unlike the notions of stability referenced above, which are incompatible with differential privacy (DP), our notion of Wasserstein-stability holds promise for bridging the two seemingly disparate fields.

2.1 Concurrent Work

Concurrent with our work, Abélès et al. [ACN25] independently addressed the same problem in a draft updated on the arXiv in June 2024.[4] Both works derive generalization bounds of the same order, i.e., $\frac{\text{regret}}{n} + O(\frac{1}{\sqrt{n}})$, but involve different assumptions, and hence vastly different techniques, and also differ in the formulation of the Online-to-Batch framework.

Our work uses the standard Online-to-Batch framework, whereas [ACN25] uses a variant of the Online-to-Batch framework, using a delayed variant of the online learning problem where the learner starts seeing the cost of its choices only after a finite number of plays. This delayed Online-to-Batch framework allows [ACN25] to avoid the stability requirement on the online learner, which is imposed by us. However, as noted earlier, this stability assumption essentially doesn't limit our choice of potential online learners, as the canonical EWA algorithm turns out to be stable.

To circumvent the issue of having non-i.i.d. data, both works require assumptions on the mixing properties of the random process. We use a standard variant of the β and the ϕ mixing assumptions (see, for example, Section 2 of [AD11]) on the random process from which the dataset is drawn. On the other hand, [ACN25] uses a much stronger mixing assumption, applied directly on the associated loss function – see Assumption 1 of [ACN25]. Our mixing assumption being weaker, imposes additional technical steps on our proofs – see Lemmas 10 and 11. However, we require an additional Lipschitz assumption on the loss function which is not required by [ACN25] thanks to their already strong mixing assumption on the loss function. Further, both works require the loss function to be bounded. While we state the boundedness assumption explicitly, [ACN25] uses it implicitly through the use of Hoeffding's Lemma in the proof of [ACN25, Lemma 2].

As a final note, our bounds seem to be more favorable when the KL divergence is high since in our bounds, we incur an additive term arising due to the specific definition of our mixing process, while in [ACN25], they incur a multiplicative factor on the KL divergence term which arises due to their use of the delayed-Online-to-Batch technique.

[4] Our work was first uploaded to the arXiv in May 2024 – see [CMS24].

3 NOTATION AND PRELIMINARIES

We will sometimes interchangeably denote the expectation of a random variable X w.r.t. a distribution P, i.e., $\mathbb{E}_P[X]$ as $\langle P, X\rangle$.

3.1 Information-theoretic Inequalities

Let P and Q be two distributions on the same probability space $(\Omega, \mathcal{F})$, having densities p and q respectively with respect to an underlying measure μ. Then the total variation distance $d_{\mathrm{TV}}(\cdot,\cdot)$ is defined w.r.t. P and Q as

$$\begin{aligned} d_{\mathrm{TV}}(P,Q) &:= \sup_{A\in\mathcal{F}} |P(A) - Q(A)| \\ &= \frac{1}{2}\int_{\omega\in\Omega} |p(\omega) - q(\omega)|\, d\mu(\omega). \end{aligned}$$

The Kullback–Leibler (KL) divergence between P and Q, denoted by $\mathrm{D}_{\mathrm{KL}}(P \,\|\, Q)$ is defined as

$$\mathrm{D}_{\mathrm{KL}}(P \,\|\, Q) = \mathop{\mathbb{E}}_{X\sim P}\left[\ln\left(\frac{dP}{dQ}(X)\right)\right].$$

The relation between total variation distance and KL divergence is captured by Pinsker's inequality.

Lemma 1 (Pinsker's Inequality)**.** *Every P, Q on $(\Omega, \mathcal{F})$ satisfies $d_{\mathrm{TV}}(P,Q) \le \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(P \,\|\, Q)}$.*

The KL divergence can also be expressed as a variational form.

Lemma 2. *Let X be a real-valued integrable random variable. Then for every $\lambda \in \mathbb{R}$,*

$$\log \mathop{\mathbb{E}}_P\left[e^{\lambda(X - \mathbb{E}_P(X))}\right] = \sup_{Q\ll P}\left[\lambda\langle Q-P, X\rangle - \mathrm{D}_{\mathrm{KL}}(Q \,\|\, P)\right].$$

See Theorem 4.19 and Corollary 4.14 of [BLM13] for a proof of Lemma 1 and Lemma 2, respectively.

3.2 Mixing Processes

Consider a random process $(Z_t)_{t\in\mathbb{N}}$ with the probability distribution $\mathcal{P}$. Let $\mathcal{F}_t = \sigma(Z_1, \ldots, Z_t)$ denotes the smallest sigma-algebra generated by the set $\{Z_s\}_{s\in[t]}$. We denote by $\mathcal{P}^t_{[s]} = \mathcal{P}^t(\cdot \mid \mathcal{F}_s)$ the conditional probability distribution of Z_t given the sigma-algebra $\mathcal{F}_s$. In this work, we consider the case where the distribution of Z_t converges (w.r.t. two different notions of convergence) to a stationary distribution $\mathcal{D}$ as $t \to \infty$, as defined below.

Definition 1 (β and ϕ coefficients)**.** The β and ϕ mixing coefficients for the distribution $\mathcal{P}$ are defined as

$$\beta(k) := \sup_{t\in\mathbb{N}}\left\{2 \mathop{\mathbb{E}}_{\mathcal{P}_{[t]}}\left[d_{\mathrm{TV}}\left(\mathcal{P}^{t+k}_{[t]}, \mathcal{D}\right)\right]\right\}, \tag{1}$$

$$\phi(k) := \sup_{t\in\mathbb{N}, B\in\mathcal{F}_t}\left\{2\, d_{\mathrm{TV}}\left(\mathcal{P}^{t+k}(\cdot|B), \mathcal{D}\right)\right\}, \tag{2}$$

where the supremum in the definition of $\phi(k)$ is over elements of $\mathcal{F}_t$ having non-zero measure, and $\mathcal{P}_{[t]}$ is the joint distribution of $Z_1, \ldots, Z_t$.

Definition 2 (β and ϕ mixing)**.** A stochastic process $\{Z_t\}_{t\in\mathbb{N}}$ is β-mixing (or ϕ-mixing) if its distribution $\mathcal{P}$ satisfies $\lim_{k\to\infty}\beta(k) = 0$ (respectively, if $\lim_{k\to\infty}\phi(k) = 0$).

Remark 1. It is trivial to see that for i.i.d. random processes, with $\mathcal{D}$ being the per-letter marginal, the mixing coefficients satisfy $\beta(k) = \phi(k) = 0$ for all $k \ge 1$. Hence, i.i.d. processes are both β and ϕ mixing.

Definition 3 (Geometric ϕ-mixing)**.** Let $K, r > 0$. A stochastic process is geometrically ϕ-mixing with rate K if $\phi(k) \le K \cdot \exp\{-k^r\}$, for all $k > 0$.

We note at this point that there are no practical approaches to finding the decay rate of an unknown mixing process or even determining whether a stochastic process is mixing [Yu94; Mei00] unless other properties (such as Gaussianity or Markovity) of the mixing process are known beforehand. There are, however, known examples of stochastic processes that are exponentially mixing (see [Mok88; Mei00] for examples).

3.3 Wasserstein distances

Let $(\mathcal{X}, d)$ be any Polish space[5], and let P and Q be any pair of probability measures on $\mathcal{X}$. We denote by $\Pi(\mathrm{P},\mathrm{Q})$ the set of joint measures on $\mathcal{X}$ whose marginals are respectively P and Q. The *Wasserstein distance* of order one between P and Q is defined as

$$W(\mathrm{P},\mathrm{Q}) \triangleq \inf_{\pi\in\Pi(\mathrm{P},\mathrm{Q})}\int_{\mathcal{X}} d(x,y)d\pi(x,y). \tag{3}$$

We now state the *Kantorovich-Rubinstein duality formula* for Wasserstein distances of order one as given in Remark 6.5 of [Vil08].

Lemma 3. *Let P and Q be any pair of probability measures on a Polish space $(\mathcal{X}, d)$. Then,*

$$W(P,Q) = \sup_{\substack{\phi:\mathcal{X}\to\mathbb{R} \\ \phi \text{ is 1-Lipschitz}}}\left\{\int_{\mathcal{X}} \phi\, dP - \int_{\mathcal{X}} \phi\, dQ\right\}.$$

As an immediate consequence of Lemma 3, we have the following corollary.

Corollary 4. *Let P and Q be any pair of probability measures on a Polish space $(\mathcal{X}, d)$. Let $\phi : \mathcal{X} \to \mathbb{R}$ be G-Lipschitz. Then, $W(P,Q) \ge \frac{1}{G}\left[\int_{\mathcal{X}} \phi\, dP - \int_{\mathcal{X}} \phi\, dQ\right]$.*

[5] A complete metric space is *Polish* if it has a countable dense subset.

Next, we present an inequality that relates Wasserstein distances of order 1 to the Total Variation distance.[6]

Lemma 5. *Let P and Q be two probability measures on a Polish space $(\mathcal{X}, d)$. If the diameter of the underlying metric space is bounded by $M \geq 0$, then $W(P,Q) \leq M \cdot d_{\mathrm{TV}}(P,Q)$.*

3.4 Problem Setup

Consider a measurable instance space $\mathcal{Z}$. Let the set $S_n = (Z_1, \ldots, Z_n)$ denoted as a *training set*, be a tuple of n random variables (not necessarily independent), drawn from some random process $Z_1, Z_2, \cdots$ over $\mathcal{Z}$, with a probability distribution $\mathcal{P}$ which mixes to a stationary distribution $\mathcal{D}$, as defined in Definition 2.

Assumption A. We assume that $\mathcal{Z}$ is equipped with a metric $\|\cdot\|_{\mathcal{Z}}$, and its diameter is $R_{\mathcal{Z}}$.

A *learning algorithm* $A : \mathcal{Z}^n \mapsto \mathcal{H}$ maps (in a randomized fashion) any such n tuple to an element $H^* = A(S_n)$ in a measurable set $\mathcal{H}$, where $\mathcal{H}$ is known as the *hypothesis class*. The performance of the learning algorithm A is measured with respect to a loss function $\ell : \mathcal{H} \times \mathcal{Z} \mapsto \mathbb{R}_+$. We now state our assumptions with respect to the hypothesis space $\mathcal{H}$ and the loss function ℓ below.

Assumption B (Assumptions on the Hypothesis Space $\mathcal{H}$)**.** We assume that the space $\mathcal{H}$ is equipped with the metric $\|\cdot\|_{\mathcal{H}}$, and $\mathcal{H}$ is *Polish* with respect to $\|\cdot\|_{\mathcal{H}}$. Moreover, the diameter of $\mathcal{H}$ is $R_{\mathcal{H}}$.

We denote by $\Delta_{\mathcal{H}}$ the set of all distributions over $\mathcal{H}$.

Assumption C (Doubly Lipschitz Loss)**.** Let $\ell : \mathcal{H} \times \mathcal{Z} \mapsto \mathbb{R}_+$, be a loss function. We assume that ℓ is *doubly Lipschitz*. More precisely, ℓ is *$G_{\mathcal{H}}$-Lipschitz* w.r.t the first argument h, and it is $G_{\mathcal{Z}}$-Lipschitz w.r.t. the second argument z. Since ℓ is doubly-Lipschitz and both of its inputs are from bounded domains, the range of ℓ is also bounded, as shown in the following lemma.

Lemma 6. *For any $(h,z) \in \mathcal{H} \times \mathcal{Z}$, the loss function ℓ satisfies $|\ell(h,z)| \leq B_\ell$, where $B_\ell := \inf_{(h',z') \in \mathcal{H} \times \mathcal{Z}} |\ell(h',z')| + G_{\mathcal{H}} R_{\mathcal{H}} + G_{\mathcal{Z}} R_{\mathcal{Z}}$.*

The *training error* of the learning algorithm A is the cumulative loss conceded by A over its training set S_n: $\sum_{i=1}^n \ell(H^*, Z_i)$, where $H^* = A(S_n)$. The *test error* of A is defined as the expected loss of the learning algorithm over any instance from the instance space: $\underset{Z' \sim \mathcal{D}}{\mathbb{E}}[\ell(H^*, Z')]$.

Definition 4 (Generalization Error of Offline Learners)**.** The *overfitting error* of A is defined as the difference between the test error and the mean training error of A, as

$$\mathrm{gen}(A, S_n) = \underset{Z' \sim \mathcal{D}}{\mathbb{E}}[\ell(H^*, Z')] - \frac{1}{n} \sum_{t=1}^{n} \ell(H^*, Z_t).$$

The generalization error of a fixed statistical learning algorithm A is defined as

$$\overline{\mathrm{gen}}(A, S_n) := \mathbb{E}_{H \sim \mathrm{P}_{H^*}}[\mathrm{gen}(H, S_n) \mid S_n],$$

where $\mathrm{P}_{H^*} := \mathrm{P}_{A(S_n)}$, i.e., the conditional distribution of the output H^* produced by A given training set S_n.

3.5 The Online-to-Batch framework

An interesting paradigm for evaluating the generalization ability of statistical learning algorithms studied in the literature is Online-to-Batch conversions [CBCG04]. In the Online-to-Batch setting, a connection is established between the performance of batch learners (on unknown instances) and the performance of online learning algorithms (on known instances). We first introduce the online learning setting and subsequently describe the Online-to-Batch paradigm.

3.5.1 Overview of Online Learning

The online learning setting can be modeled as the following two-player game, henceforth referred to as the Online-to-Batch game, between a learning algorithm $\mathcal{L}$ and an adversary. The learner $\mathcal{L}$ has sequential access to a stream of data generated from an arbitrary source, and at every time step t, based on decisions taken up to the $t-1^{\mathrm{th}}$ time-step, the learner tries to predict the correct label of the next data point and incurs a loss that is decided by an adversary. The goal of the online learning setup is to minimize some notion of *regret*, i.e., the loss incurred at the t^{th} time step by the online learner should be reasonably close to the loss incurred by the best possible offline learner that has access to the data points up to the $(t-1)^{\mathrm{th}}$ time-step.

An example of an online learning game is the Hedging algorithm or the Exponential-Weighted Average (EWA) algorithm [LW94; FS97; Vov98]. In EWA, we first fix a data-free prior distribution $\mathcal{P}_1 \in \Delta_{\mathcal{H}}$ and a learning rate parameter $\eta > 0$. At every iteration $t > 0$, we perform the following updates:

$$\mathcal{P}_{t+1} := \underset{\mathcal{P} \in \Delta_{\mathcal{H}}}{argmin} \left\{ \langle \mathcal{P}, c_t \rangle + \frac{1}{\eta} \mathrm{D}_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{P}_t\right) \right\}. \quad (4)$$

3.5.2 The Online-to-Batch Game

In the Online-to-Batch conversion game, we assume that the instances $Z_t, t \in [n]$ in the offline setting are provided to the online learner $\mathcal{L}_n$. We now describe the generalization game from Lugosi and Neu [LN23], played over n rounds below.

[6] See Particular Case 6.16 of [Vil08].

Game 1 (The Online-to-Batch Generalization game)**.**

▷ At the t^{th} iteration,

1. The online learning algorithm $\mathcal{L}_n$ chooses a distribution $\mathrm{P}_t \in \Delta_{\mathcal{H}}$ over the hypothesis space, with knowledge of only $Z_1, \ldots, Z_{t-1}$.
2. The adversary picks a cost function $c_t : \mathcal{H} \to \mathbb{R}$ for each hypothesis $h \in \mathcal{H}$ as
$$c_t(h) := \ell(h, Z_t) - \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(h, Z')] \,.$$
3. The online learning algorithm $\mathcal{L}_n$ incurs a cost $\langle \mathrm{P}_t, c_t \rangle := \mathbb{E}_{\mathrm{P}_t}[c_t(H_t)]$.
4. The adversary reveals to the online learner the sample Z_t. Now the online learner can compute the cost function.

Recall that $\mathrm{P}_{H^*} := \mathrm{P}_{A(S_n)}$. Then the regret of the learning algorithm $\mathcal{L}_n$ with respect to the comparator distribution P_{H^*} over the hypothesis space is:

$$\mathrm{regret}_{\mathcal{L}_n, A}(P_{H^*}) := \sum_{t=1}^{n} \langle \mathrm{P}_t - \mathrm{P}_{H^*}, c_t \rangle, \tag{5}$$

Henceforth, we shall drop the comparator distribution P_{H^*} from the notation of regret for the sake of brevity. We now present the following technical lemma that bounds the cost function from Game 1.

Lemma 7. *For any fixed instance of Z_t, and any $h_1, h_2 \in \mathcal{H}$, the cost function $c_t(\cdot)$ picked by the adversary in the generalization game satisfies $|c_t(h_1) - c_t(h_2)| \le 2G_{\mathcal{H}} R_{\mathcal{H}}$. On the other hand, for any fixed $h \in \mathcal{H}$, and any t, t' and a fixed realization of $Z_t, Z_{t'}$, we have $|c_t(h) - c_{t'}(h)| \le G_{\mathcal{Z}} R_{\mathcal{Z}}$.*

In order to bound the generalization error of A using the regret of the online learner $\mathcal{L}_n$ for Game 1, we shall require the $\mathcal{L}_n$ to be *Wasserstein-stable* as defined below.

Definition 5 (Wasserstein Stable)**.** Given a non-increasing sequence $\kappa(t), t \ge 1$, an online learning algorithm is said to be $\kappa(t)$-Wasserstein-stable if for any $t \in [n]$, the online learner $\mathcal{L}_n$ satisfies

$$W(\mathrm{P}_t, \mathrm{P}_{t+1}) \le \kappa(t). \tag{6}$$

We refer to $\kappa(t)$ as the stability parameter at round t.

The following technical lemma allows us to bound the sum of differences in expected costs between Wasserstein-stable online learning algorithms for Game 1 and the offline comparator.

Lemma 8. *For any Wasserstein-stable online learning algorithm $\mathcal{L}_n$ for Game 1 and any $\tau = o(n)$, the following bound holds with probability one.*

$$\sum_{t=1}^{n} \left[\mathop{\mathbb{E}}_{H \sim P_t}[c_{t+\tau}(H)] - \mathop{\mathbb{E}}_{H \sim P_{H^*}}[c_{t+\tau}(H)] \right]$$
$$\le \mathit{regret}_{\mathcal{L}_n, A} + 2G_{\mathcal{H}} \tau \sum_{t=1}^{n} \kappa(t) + 4\tau G_{\mathcal{H}} R_{\mathcal{H}}.$$

4 GENERALIZATION ERROR BOUNDS FOR MIXING PROCESSES

In this section, we state and prove our main results on the generalization error of statistical learning algorithms trained on training samples drawn from a mixing process. Our first result is an upper bound on the expected generalization error in terms of the expected regret of an online learner $\mathcal{L}_n$ for Game 1.

Theorem 1 (Expected generalization error)**.** *For any arbitrary Wasserstein-stable online learner $\mathcal{L}_n$ for Game 1, and any $\tau = o(n)$, the expected generalization error $\mathbb{E}[\overline{\mathrm{gen}}(A, S_n)]$ of the learning algorithm A with input $S_n = (Z_1, \ldots, Z_n)$ drawn from the mixing random process $\{Z_t\}_{t \in \mathbb{N}}$ is upper bounded by*

$$\frac{1}{n}\mathbb{E}\big[\mathit{regret}_{\mathcal{L}_n, A}\big] + \frac{2\tau G_{\mathcal{H}}}{n} \left(\sum_{t=1}^{n} \kappa(t) + 2R_{\mathcal{H}} \right)$$
$$+ \frac{\tau G_{\mathcal{Z}} R_{\mathcal{Z}}}{n} + B_\ell \cdot \beta(\tau + 1).$$

The next theorem complements Theorem 1 by providing a high probability upper bound on the generalization error of the learning algorithm A in terms of the regret of any learner $\mathcal{L}_n$ for Game 1.

Theorem 2 (Generalization Error)**.** *For any arbitrary Wasserstein-stable online learner $\mathcal{L}_n$ for Game 1, and any $\tau = o(n)$, $\delta > 0$, the generalization error $\overline{\mathrm{gen}}(A, S_n)$ of the learning algorithm A with input $S_n = (Z_1, \ldots, Z_n)$ drawn from the mixing random process $\{Z_t\}_{t \in \mathbb{N}}$ is upper bounded, with probability at least $1 - \delta$, by*

$$\frac{\mathit{regret}_{\mathcal{L}_n, A}}{n} + \frac{2\tau G_{\mathcal{H}}}{n} \left(\sum_{t=1}^{n} \kappa(t) + 2R_{\mathcal{H}} \right)$$
$$+ \frac{\tau G_{\mathcal{Z}} R_{\mathcal{Z}}}{n} + 2G_{\mathcal{H}} R_{\mathcal{H}} \sqrt{\frac{2\tau \log(\tau/\delta)}{n}} + B_\ell \cdot \phi(\tau + 1).$$

In order to establish Theorem 1 and Theorem 2, we first prove the following intermediate lemma which relates the generalization error of the offline learner A to the regret of the online learner $\mathcal{L}_n$.

Lemma 9. *For any $\tau = o(n)$, and any Wasserstein-stable online learner $\mathcal{L}_n$ for Game 1, with probability one, $\overline{\text{gen}}(A, S_n)$ is at most*

$$\frac{regret_{\mathcal{L}_n,A}}{n} + \frac{\tau}{n}\left[2G_{\mathcal{H}}\sum_{t=1}^{n}\kappa(t) + 4G_{\mathcal{H}}R_{\mathcal{H}} + G_{\mathcal{Z}}R_{\mathcal{Z}}\right] - \frac{1}{n}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim P_t}[c_{t+\tau}(H)].$$

To complete the proofs of Theorem 1 and Theorem 2, we upper bound the final term in Lemma 9, respectively, in expectation and with high probability. This is accomplished in the following pair of lemmas which rearranges the terms in $\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)]$ as sums of random variables forming a martingale difference sequence, and a remainder term which can be bounded using the mixing coefficients for the random process $\{Z_t\}_{t\in\mathbb{N}}$.

Lemma 10. *For any $\tau = o(n)$,*

$$-\sum_{t=1}^{n}\mathop{\mathbb{E}}_{\mathcal{P}}\left[\mathop{\mathbb{E}}_{H\sim P_t}[c_{t+\tau}(H)]\right] \le n\cdot B_\ell\cdot\beta(\tau+1).$$

Lemma 11. *With probability at least $1-\delta$, for any $\tau = o(n), \delta > 0$,*

$$-\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim P_t}[c_{t+\tau}(H)] \le 2G_{\mathcal{H}}R_{\mathcal{H}}\sqrt{2n\tau\log(\tau\delta)} + n\cdot B_\ell\cdot\phi(\tau+1).$$

We can now complete the proofs of Theorems 1 and 2.

Proofs of Theorems 1 and 2. Theorem 1 now follows by plugging in Lemma 10 in Lemma 9. Similarly, Theorem 2 follows by plugging Lemma 11 in Lemma 9. ∎

5 APPLICATIONS: GENERALIZATION ERROR BOUNDS FOR EWA

In this section, we obtain generalization bounds for data drawn from mixing processes by using the EWA learner (see Eq. (4)) as our online learning strategy. First, we prove that the EWA Learner is Wasserstein-stable.

Theorem 3. *Under Assumptions A, B and C, the Exponentially Weighted Averages (EWA) algorithm is a Wasserstein-stable online learner (see Definition 5), where the stability parameter $\kappa(t) \le \eta G_{\mathcal{H}}R_{\mathcal{H}}^2$ for all $t\in[n]$, and $\eta > 0$ is the learning rate of the EWA learner.*

To prove Theorem 3, we require the following lemmas.

Lemma 12 (EWA Minimizer). *Let $\eta > 0$ be a learning rate. Then, Eq. (4) is minimized by a distribution $\mathcal{P}^* \ll \mathcal{P}_t$ s.t. $\mathcal{P}^*$ satisfies the following equality:*

$$\langle\mathcal{P}^*, c_t\rangle + \frac{1}{\eta}\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t) = -\frac{1}{\eta}\log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta c_t(H)}\right]\right). \tag{7}$$

We can now complete the proof of Theorem 3.

Proof of Theorem 3. Applying Jensen's inequality and using the concavity of log in the RHS of Eq. (7), we get:

$$\langle\mathcal{P}^*, c_t\rangle + \frac{1}{\eta}\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t) \le \langle\mathcal{P}_t, c_t\rangle.$$

Rearranging terms, we obtain:

$$\frac{1}{\eta}\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t) \le \langle\mathcal{P}_t - \mathcal{P}^*, c_t\rangle. \tag{8}$$

From Lemma 7, we have that c_t is $2G_{\mathcal{H}}$-Lipschitz. Hence, $|\langle\mathcal{P}_t - \mathcal{P}^*, c_t\rangle| \le 2G_{\mathcal{H}}W(\mathcal{P}^*, \mathcal{P}_t)$. Therefore, we can rewrite Eq. (8) as follows:

$$\frac{1}{\eta}\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t) \le 2G_{\mathcal{H}}W(\mathcal{P}^*, \mathcal{P}_t). \tag{9}$$

Recall that the hypothesis space is bounded by $R_{\mathcal{H}}$ by Assumption B. We can therefore upper bound $W(\mathcal{P}^*, \mathcal{P}_t)$ using Lemma 1 and Lemma 5 as:

$$W(\mathcal{P}^*, \mathcal{P}_t) \le R_{\mathcal{H}}\sqrt{\frac{\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t)}{2}}. \tag{10}$$

Plugging Eq. (10) into Eq. (9), we have

$$\frac{1}{\eta}\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t) \le G_{\mathcal{H}}R_{\mathcal{H}}\sqrt{2\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t)}. \tag{11}$$

Thus, as KL divergence is non-negative, we have:

$$\sqrt{\mathrm{D}_{\mathrm{KL}}(\mathcal{P}^*\,\|\,\mathcal{P}_t)} \le \sqrt{2}\eta G_{\mathcal{H}}R_{\mathcal{H}}. \tag{12}$$

Plugging Eq. (12) back into Eq. (10) gives us the desired upper bound for $W(\mathcal{P}^*, \mathcal{P}_t)$.

$$W(\mathcal{P}^*, \mathcal{P}_t) \le \eta G_{\mathcal{H}}R_{\mathcal{H}}^2.$$

Now setting $\mathcal{P}^* = \mathcal{P}_{t+1}$ implies that $\kappa(t) \le \eta G_{\mathcal{H}}R_{\mathcal{H}}^2$, $\forall t\in[n]$. ∎

We now instantiate Theorem 2 by picking the EWA algorithm as our online learning strategy in Game 1.

Theorem 4 (Generalization Bound via EWA). *Let $\{Z_t\}_{t\in\mathbb{N}}$ be a geometric ϕ-mixing process with rate $K > 0$, and $r > 1$. Then for any $P_1 \in \Delta_{\mathcal{H}}$, and any $n > 1$, the generalization error $\overline{\text{gen}}(A, S_n)$ of any learning* *algorithm A trained on $S_n = (Z_1, \ldots, Z_n)$ drawn from the mixing process $\{Z_t\}_{t\in\mathbb{N}}$ is upper bounded by*

$$\frac{\mathrm{D_{KL}}\left(P_{A(S_n)} \,\|\, P_1\right)}{n \cdot \eta} + \frac{\eta \cdot B_\ell^2 + 4\eta G_\mathcal{H}^2 R_\mathcal{H}^2 \log n}{2} + \frac{4G_\mathcal{H} R_\mathcal{H} \log n}{n} + \frac{G_\mathcal{Z} R_\mathcal{Z} \log n}{n} + 2G_\mathcal{H} R_\mathcal{H} \sqrt{\frac{2\log n \log\left(\log n/\delta\right)}{n}} + \frac{B_\ell \cdot K}{n},$$

with probability at least $1-\delta$, for any $\delta > 0$.

To prove Theorem 4, we first state the following lemma[7] to upper bound the regret of the EWA learner.

Lemma 13 (Regret of EWA)**.** *For any $P_1 \in \Delta_\mathcal{H}$ and any comparator $P^* \in \Delta_\mathcal{H}$, the regret of EWA satisfies for any learning rate $\eta > 0$ is at most*

$$\mathit{regret}_{EWA}(P^*) \le \frac{\mathrm{D_{KL}}\left(P^* \,\|\, P_1\right) ||P_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n} \|c_t\|_\infty^2.$$

Proof of Theorem 4. From Lemma 7 and Lemma 13, we have for any $P_1 \in \Delta_\mathcal{H}$ and any comparator $P^* \in \Delta_\mathcal{H}$, the regret of EWA satisfies for any learning rate $\eta > 0$ is at most

$$\mathrm{regret_{EWA}}(P^*) \le \frac{\mathrm{D_{KL}}\left(P^* \,\|\, P_1\right) ||P_1)}{\eta} + \frac{n\cdot\eta\cdot B_\ell^2}{2}. \tag{13}$$

Let $\tau = \lceil \log n \rceil - 1$. Therefore, from Definition 3, we have using $r > 1$,

$$\phi(\tau+1) \le K \cdot \exp\{-(\tau+1)^r\} \le \frac{K}{n}. \tag{14}$$

Plugging Eq. (13) and Eq. (14) into Theorem 2 gives us the required bound. ■

We note here that Theorem 4 gives generalization bounds in terms of the learning rate η of the EWA learner. While this is not a problem in and of itself, optimizing the generalization bound would require choosing a learning rate that is data-dependent due to the $\mathrm{D_{KL}}\left(P_{A(S_n)} \,\|\, P_1\right)$ term, something that is not possible according to Game 1. Fortunately, we can obtain data-independent generalization bounds for an appropriate choice of learning rate, as stated in the following corollary of Theorem 4.

Corollary 14. *Let $\{Z_t\}_{t\in\mathbb{N}}$ be a geometric ϕ-mixing process with rate $K > 0$, and $r > 1$. Then for any $P_1 \in \Delta_\mathcal{H}$, the generalization error $\overline{\mathrm{gen}}\,(A, S_n)$ of any learning algorithm A trained on $S_n = (Z_1, \ldots, Z_n)$ drawn from the mixing process $\{Z_t\}_{t\in\mathbb{N}}$ is upper bounded by*

$$\frac{\mathrm{D_{KL}}\left(P_{A(S_n)} \,\|\, P_1\right)}{\sqrt{n}} + \frac{B_\ell^2 + 4G_\mathcal{H}^2 R_\mathcal{H}^2 \log n}{2\sqrt{n}} + \frac{4G_\mathcal{H} R_\mathcal{H} \log n}{n} + \frac{G_\mathcal{Z} R_\mathcal{Z} \log n}{n} + 2G_\mathcal{H} R_\mathcal{H} \sqrt{\frac{2\log n \log\left(\log n/\delta\right)}{n}} + \frac{B_\ell \cdot K}{n},$$

with probability at least $1-\delta$, for any $\delta > 0$.

6 DISCUSSION AND FUTURE WORK

In this work, we extend the Online-to-Batch framework to give generalization bounds for statistical learning algorithms that are trained on data sampled from mixing processes. An immediate avenue of future work is to obtain generalization bounds for non-i.i.d. data in various settings by using different choices of online learners to instantiate our framework as presented in Theorem 1 and Theorem 2.

To compensate for considering the weaker non-i.i.d. assumption on our dataset, our Online-to-Batch framework requires online learners that are Wasserstein-stable. As noted earlier, the notion of Wasserstein-stability is quite similar to the differential privacy inspired notion of stability proposed by Abernethy et al. [Abe+19]. Abernethy et al. [Abe+19] used their notion of stability to develop regret bounds for a variety of online learning problems, such as follow-the-perturbed-leader algorithms, thereby cementing a connection between differentially private learning algorithms and online learners. In this light, we ask the following question - can we use the techniques introduced in this paper to analyze generalization error bounds for differentially private learners, especially in the non-i.i.d. setting?

Algorithmic stability and its relation to tight generalization bounds have been recently studied by Gastpar et al. [Gas+24a; Gas+24b] in the i.i.d. setting. Due to the dearth of notions of algorithmic stability in the non-i.i.d setting (see Section 2), our Wasserstein stability criteria is a potential candidate to extend the results of [Gas+24b; Gas+24a] to the non-i.i.d. setting.

Acknowledgements

The authors thank the anonymous reviewers for their useful comments.

Note

The authors of this work are listed in alphabetical order.

References

[Abe+19] J. Abernethy, Y. H. Jung, C. Lee, A. McMillan, and A. Tewari. “Online learning via the differential privacy lens”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cit. on pp. 3, 8).

[7] See Appendix A.1 of [LN23] for a detailed proof.

[ACN25] B. Abélès, E. Clerico, and G. Neu. "Generalization bounds for mixing processes via delayed online-to-PAC conversions". In: *36th International Conference on Algorithmic Learning Theory*. 2025. eprint: https://arxiv.org/abs/2406.12600v1. URL: https://openreview.net/forum?id=FsI3wb61G0 (cit. on pp. 2, 3).

[AD11] A. Agarwal and J. C. Duchi. "The Generalization Ability of Online Algorithms for Dependent Data". In: *IEEE Transactions on Information Theory* 59 (2011), pp. 573–587. DOI: 10.1109/TIT.2012.2212414 (cit. on pp. 2, 3).

[Alq24] P. Alquier. "User-friendly Introduction to PAC-Bayes Bounds". In: *Foundations and Trends® in Machine Learning* 17.2 (2024), pp. 174–303. ISSN: 1935-8237. DOI: 10.1561/2200000100. URL: http://dx.doi.org/10.1561/2200000100 (cit. on p. 1).

[Ami+22] S. Amiri, A. Belloum, E. Nalisnick, S. Klous, and L. Gommans. "On the impact of non-IID data on the performance and fairness of differentially private federated learning". In: *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. 2022, pp. 52–58. DOI: 10.1109/DSN-W54100.2022.00018 (cit. on p. 1).

[BE02] O. Bousquet and A. Elisseeff. "Stability and generalization". In: *J. Mach. Learn. Res.* 2 (Mar. 2002), 499–526. ISSN: 1532-4435. DOI: 10.1162/153244302760200704 (cit. on pp. 1, 3).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Feb. 2013. ISBN: 9780199535255. DOI: 10.1093/acprof:oso/9780199535255.001.0001. URL: http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001 (cit. on p. 4).

[CBCG04] N. Cesa-Bianchi, A. Conconi, and C. Gentile. "On the generalization ability of on-line learning algorithms". In: *IEEE Transactions on Information Theory* 50.9 (2004), pp. 2050–2057 (cit. on pp. 1–3, 5).

[CMS24] S. Chatterjee, M. Mukherjee, and A. Sethi. *Generalization Bounds for Dependent Data using Online-to-Batch Conversion*. 2024. arXiv: 2405.13666v1 [cs.LG] (cit. on p. 3).

[Duc+12] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan. "Ergodic Mirror Descent". In: *SIAM Journal on Optimization* 22.4 (2012), pp. 1549–1578. DOI: 10.1137/110836043 (cit. on p. 2).

[FS97] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: https://www.sciencedirect.com/science/article/pii/S002200009791504X (cit. on p. 5).

[Fu+23] S. Fu, Y. Lei, Q. Cao, X. Tian, and D. Tao. "Sharper Bounds for Uniformly Stable Algorithms with Stationary Mixing Process". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=8E5Yazboyh (cit. on pp. 1–3).

[Gas+24a] M. Gastpar, I. Nachum, J. Shafer, and T. Weinberger. "Fantastic Generalization Measures are Nowhere to be Found". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=NkmJotfL42 (cit. on p. 8).

[Gas+24b] M. Gastpar, I. Nachum, J. Shafer, and T. Weinberger. *Which Algorithms Have Tight Generalization Bounds?* 2024. arXiv: 2410.01969 [cs.LG]. URL: https://arxiv.org/abs/2410.01969 (cit. on p. 8).

[Hel+24] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky. *Generalization Bounds: Perspectives from Information Theory and PAC-Bayes*. 2024. arXiv: 2309.04381 [cs.LG]. URL: https://arxiv.org/abs/2309.04381 (cit. on p. 1).

[Iye24] V. N. Iyer. *A review on different techniques used to combat the non-IID and heterogeneous nature of data in FL*. 2024. arXiv: 2401.00809 [cs.LG] (cit. on p. 1).

[Kle07] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer London, 2007. ISBN: 9781848000483. URL: https://books.google.co.in/books?id=tcm3y5UJxDsC (cit. on p. 15).

[KM17] V. Kuznetsov and M. Mohri. "Generalization bounds for non-stationary mixing processes". In: *Machine Learning* 106.1 (2017), pp. 93–117 (cit. on p. 2).

[KT08] S. M. Kakade and A. Tewari. “On the Generalization Ability of Online Strongly Convex Programming Algorithms”. In: *Neural Information Processing Systems*. 2008 (cit. on p. 3).

[Li+24] Y. Li, S. Wang, C.-Y. Chi, and T. Q. S. Quek. “Differentially Private Federated Clustering Over Non-IID Data”. In: *IEEE Internet of Things Journal* 11.4 (2024), pp. 6705–6721. DOI: 10.1109/JIOT.2023.3312852 (cit. on p. 1).

[LKS05] A. C. Lozano, S. Kulkarni, and R. E. Schapire. “Convergence and Consistency of Regularized Boosting Algorithms with Stationary B-Mixing Observations”. In: *Advances in Neural Information Processing Systems*. Ed. by Y. Weiss, B. Schölkopf, and J. Platt. Vol. 18. MIT Press, 2005 (cit. on p. 2).

[LN23] G. Lugosi and G. Neu. “Online-to-PAC Conversions: Generalization Bounds via Regret Analysis”. In: *ArXiv* abs/2305.19674 (2023) (cit. on pp. 1–3, 5, 8).

[LW94] N. Littlestone and M. Warmuth. “The Weighted Majority Algorithm”. In: *Information and Computation* 108.2 (1994), pp. 212–261. ISSN: 0890-5401. DOI: https://doi.org/10.1006/inco.1994.1009. URL: https://www.sciencedirect.com/science/article/pii/S0890540184710091 (cit. on p. 5).

[Mei00] R. Meir. “Nonparametric time series prediction through adaptive model selection”. In: *Machine learning* 39 (2000), pp. 5–34 (cit. on pp. 1–4).

[Mok88] A. Mokkadem. “Mixing properties of ARMA processes”. In: *Stochastic Processes and their Applications* 29.2 (1988), pp. 309–315. ISSN: 0304-4149. DOI: https://doi.org/10.1016/0304-4149(88)90045-2 (cit. on p. 4).

[MR07] M. Mohri and A. Rostamizadeh. “Stability Bounds for Non-i.i.d. Processes”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc., 2007 (cit. on p. 2).

[MR08] M. Mohri and A. Rostamizadeh. “Rademacher Complexity Bounds for Non-I.I.D. Processes”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc., 2008. URL: https://proceedings.neurips.cc/paper_files/paper/2008/file/7eacb532570ff6858afd2723755ff790-Paper.pdf (cit. on p. 3).

[MR10] M. Mohri and A. Rostamizadeh. “Stability Bounds for Stationary ϕ-mixing and β-mixing Processes”. In: *J. Mach. Learn. Res.* 11 (Mar. 2010), 789–814. ISSN: 1532-4435 (cit. on pp. 1–3).

[MU05] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005 (cit. on p. 15).

[RZ16] D. Russo and J. Zou. “Controlling Bias in Adaptive Data Analysis Using Information Theory”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Gretton and C. C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, May 2016, pp. 1232–1240 (cit. on p. 1).

[Vid13] M. Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013 (cit. on p. 1).

[Vil08] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008 (cit. on pp. 4, 5).

[Vov98] V Vovk. “A Game of Prediction with Expert Advice”. In: *Journal of Computer and System Sciences* 56.2 (1998), pp. 153–173. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1556. URL: https://www.sciencedirect.com/science/article/pii/S0022000097915567 (cit. on p. 5).

[Xio+22] Z. Xiong, Z. Cai, D. Takabi, and W. Li. “Privacy Threat and Defense for Federated Learning With Non-i.i.d. Data in AIoT”. In: *IEEE Transactions on Industrial Informatics* 18.2 (2022), pp. 1310–1321. DOI: 10.1109/TII.2021.3073925 (cit. on p. 1).

[XR17] A. Xu and M. Raginsky. “Information-theoretic analysis of generalization capability of learning algorithms”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 1).

[Yu94] B. Yu. "Rates of convergence for empirical processes of stationary mixing sequences". In: *The Annals of Probability* (1994), pp. 94–116 (cit. on pp. 1–4).

[Zha+17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy8gdB9xx (cit. on p. 1).

[Zha+19] R. R. Zhang, X. Liu, Y. Wang, and L. Wang. "McDiarmid-Type Inequalities for Graph-Dependent Variables and Stability Bounds". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 2, 3).

[Zha+21] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. "Understanding deep learning (still) requires rethinking generalization". In: *Commun. ACM* 64.3 (Feb. 2021), 107–115. ISSN: 0001-0782. DOI: 10.1145/3446776. URL: https://doi.org/10.1145/3446776 (cit. on p. 1).

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **[Yes]**
   Justification: See Section 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **[Not Applicable]**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **[Not Applicable]**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **[Yes]**
   (b) Complete proofs of all theoretical results. **[Yes]**
   Justification: See Appendix A in the Appendix for proofs omitted from the main paper.
   (c) Clear explanations of any assumptions. **[Yes]**
   Justification: See discussions in Sections 1.2 and 2 and discussions following Assumptions A, B, and C.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **[Not Applicable]**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **[Not Applicable]**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **[Not Applicable]**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **[Not Applicable]**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **[Not Applicable]**
   (b) The license information of the assets, if applicable. **[Not Applicable]**
   (c) New assets either in the supplemental material or as a URL, if applicable. **[Not Applicable]**
   (d) Information about consent from data providers/curators. **[Not Applicable]**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **[Not Applicable]**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **[Not Applicable]**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **[Not Applicable]**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **[Not Applicable]**

A APPENDIX: OMITTED PROOFS

A.1 PRELIMINARIES AND NOTATION

Corollary 4. *Let P and Q be any pair of probability measures on a Polish space $(\mathcal{X}, d)$. Let $\phi : \mathcal{X} \to \mathbb{R}$ be G-Lipschitz. Then, $W(P, Q) \geq \frac{1}{G}\left[\int_{\mathcal{X}} \phi \, dP - \int_{\mathcal{X}} \phi \, dQ\right]$.*

Proof. Define a new function $\psi : \mathcal{X} \to \mathbb{R}$ as $\psi(x) = \frac{1}{G}\phi(x)$, and note that ψ is 1-Lipschitz by definition. The result now follows using Lemma 3. ■

Lemma 6. *For any $(h, z) \in \mathcal{H} \times \mathcal{Z}$, the loss function ℓ satisfies $|\ell(h, z)| \leq B_\ell$, where $B_\ell := \inf_{(h', z') \in \mathcal{H} \times \mathcal{Z}} |\ell(h', z')| + G_{\mathcal{H}} R_{\mathcal{H}} + G_{\mathcal{Z}} R_{\mathcal{Z}}$.*

Proof of Lemma 6. Fix any $(h', z') \in \mathcal{H} \times \mathcal{Z}$. On applying triangle inequality, we have $\Big| |\ell(h, z)| - |\ell(h', z')| \Big| \leq |\ell(h, z) - \ell(h', z')|$. Now, noting that ℓ is doubly-Lipschitz, and $\mathcal{H}$ and $\mathcal{Z}$ have diameters $R_{\mathcal{H}}$ and $R_{\mathcal{Z}}$ respectively, we have

$$\begin{aligned} |\ell(h, z) - \ell(h', z')| &= |\ell(h, z) - \ell(h, z') + \ell(h, z') - \ell(h', z')| \\ &\leq |\ell(h, z) - \ell(h, z')| + |\ell(h, z') - \ell(h', z')| \\ &\leq G_{\mathcal{Z}} R_{\mathcal{Z}} + G_{\mathcal{H}} R_{\mathcal{H}}. \end{aligned}$$

Therefore, we have $|\ell(h, z)| \leq |\ell(h', z')| + G_{\mathcal{Z}} R_{\mathcal{Z}} + G_{\mathcal{H}} R_{\mathcal{H}}$. The lemma is then proved by taking an infimum over $(h', z') \in \mathcal{H} \times \mathcal{Z}$. ■

Lemma 7. *For any fixed instance of Z_t, and any $h_1, h_2 \in \mathcal{H}$, the cost function $c_t(\cdot)$ picked by the adversary in the generalization game satisfies $|c_t(h_1) - c_t(h_2)| \leq 2G_{\mathcal{H}} R_{\mathcal{H}}$. On the other hand, for any fixed $h \in \mathcal{H}$, and any t, t' and a fixed realization of $Z_t, Z_{t'}$, we have $|c_t(h) - c_{t'}(h)| \leq G_{\mathcal{Z}} R_{\mathcal{Z}}$.*

Proof. Let $h_1, h_2 \in \mathcal{H}$. Then, for any fixed instance of Z_t, we have

$$\begin{aligned} |c_t(h_1) - c_t(h_2)| &= |\ell(h_1, Z_t) - \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(h_1, Z')] - \ell(h_2, Z_t) + \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(h_2, Z')]\,| \\ &\overset{(a)}{\leq} |\ell(h_1, Z_t) - \ell(h_2, Z_t)| + |\mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(h_1, Z')] - \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(h_2, Z')]\,| \\ &\overset{(b)}{\leq} G_{\mathcal{H}} \|h_1 - h_2\|_{\mathcal{H}} + \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[|\ell(h_1, Z') - \ell(h_2, Z')|] \\ &\overset{(c)}{\leq} 2G_{\mathcal{H}} \|h_1 - h_2\|_{\mathcal{H}} \\ &\overset{(d)}{\leq} 2G_{\mathcal{H}} R_{\mathcal{H}}, \end{aligned}$$

where (a), (b) use the triangle inequality, and (b) and (c) use the fact that $\ell(\cdot, \cdot)$ is $G_{\mathcal{H}}$-Lipschitz in the first argument, and (d) uses the bounded diameter of the hypothesis space $\mathcal{H}$.

On the other hand, for any $h \in \mathcal{H}$ and any t, t' and any instance of $Z_t, Z_{t'}$,

$$\begin{aligned} |c_t(h) - c_{t'}(h)| &= |\ell(h, Z_t) - \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(h, Z')] - \ell(h, Z_{t'}) + \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(h, Z')]\,| \\ &\overset{(a)}{\leq} G_{\mathcal{Z}} \|Z_t - Z_{t'}\|_{\mathcal{Z}} \\ &\overset{(b)}{\leq} G_{\mathcal{Z}} R_{\mathcal{Z}}, \end{aligned}$$

where (a) follows from the fact that $\ell(\cdot, \cdot)$ is $G_{\mathcal{Z}}$-Lipschitz in the second argument, and (b) follows by noting that the diameter of $\mathcal{Z}$ is $R_{\mathcal{Z}}$. ■

Lemma 8. *For any Wasserstein-stable online learning algorithm $\mathcal{L}_n$ for Game 1 and any $\tau = o(n)$, the following bound holds with probability one.*

$$\sum_{t=1}^{n} \left[\mathop{\mathbb{E}}_{H \sim P_t} [c_{t+\tau}(H)] - \mathop{\mathbb{E}}_{H \sim P_{H^*}} [c_{t+\tau}(H)] \right]$$
$$\leq \mathit{regret}_{\mathcal{L}_n, A} + 2G_{\mathcal{H}} \tau \sum_{t=1}^{n} \kappa(t) + 4\tau G_{\mathcal{H}} R_{\mathcal{H}}.$$

Proof. We first rearrange the terms of the summation in the LHS of the hypothesis of Lemma 8, and proceed to bound each term individually as follows.

$$\begin{aligned}
&\sum_{t=1}^{n} \left[\mathop{\mathbb{E}}_{H \sim \mathrm{P}_t} [c_{t+\tau}(H)] - \mathop{\mathbb{E}}_{H \sim \mathrm{P}_{H^*}} [c_{t+\tau}(H)] \right] \\
&\quad = \underbrace{\sum_{t=1}^{n} \left[\mathop{\mathbb{E}}_{H \sim \mathrm{P}_t} [c_t(H)] - \mathop{\mathbb{E}}_{H \sim \mathrm{P}_{H^*}} [c_t(H)] \right]}_{T_1} + \underbrace{\sum_{t=1}^{n-\tau} \left[\mathop{\mathbb{E}}_{H \sim \mathrm{P}_t} [c_{t+\tau}(H)] - \mathop{\mathbb{E}}_{H \sim \mathrm{P}_{t+\tau}} [c_{t+\tau}(H)] \right]}_{T_2} \\
&\qquad + \underbrace{\sum_{t=n-\tau+1}^{n} \mathop{\mathbb{E}}_{H \sim \mathrm{P}_t} [c_{t+\tau}(H)] - \sum_{t=n+1}^{n+\tau} \mathop{\mathbb{E}}_{H \sim \mathrm{P}_{H^*}} [c_t(H)]}_{T_3} + \underbrace{\sum_{t=1}^{\tau} \mathop{\mathbb{E}}_{H \sim \mathrm{P}_{H^*}} [c_t(H)] - \sum_{t=1}^{\tau} \mathop{\mathbb{E}}_{H \sim \mathrm{P}_t} [c_t(H)]}_{T_4}.
\end{aligned} \tag{15}$$

By definition, we have $T_1 \leq \mathrm{regret}_{\mathcal{L}_n, A}$. Next,

$$\begin{aligned}
T_2 &\overset{(a)}{\leq} \sum_{t=1}^{n-\tau} 2G_{\mathcal{H}} W(\mathrm{P}_t, \mathrm{P}_{H_{t+\tau}}) \overset{(b)}{\leq} \sum_{t=1}^{n-\tau} 2G_{\mathcal{H}} \sum_{r=0}^{\tau-1} W(\mathrm{P}_{t+r}, \mathrm{P}_{t+r+1}) \\
&\overset{(c)}{\leq} \sum_{t=1}^{n-\tau} 2G_{\mathcal{H}} \sum_{r=0}^{\tau-1} \kappa(t+r) \overset{(d)}{\leq} 2G_{\mathcal{H}} \tau \sum_{t=1}^{n-\tau} \kappa(t) \\
&\overset{(e)}{\leq} 2G_{\mathcal{H}} \tau \sum_{t=1}^{n} \kappa(t),
\end{aligned}$$

where (a) follows from Corollary 4 and the fact that c_t is $2G_{\mathcal{H}}$-Lipschitz (see Lemma 7), (b) follows using the triangle inequality, (c) uses the stability assumption of the learner $\mathcal{L}_n$, (d) uses the fact that $\kappa(\tau)$ is non-increasing, and (e) uses the non-negativity of $\kappa(\tau)$ which follows from Eq. (6).

Next, we bound T_3 as follows.

$$\begin{aligned}
\sum_{t=n-\tau+1}^{n} \mathop{\mathbb{E}}_{H \sim \mathrm{P}_t} [c_{t+\tau}(H)] - \sum_{t=n+1}^{n+\tau} \mathop{\mathbb{E}}_{H \sim \mathrm{P}_{H^*}} [c_t(H)] &= \sum_{t=n+1}^{n+\tau} \left[\mathop{\mathbb{E}}_{H \sim \mathrm{P}_{H_{t-\tau}}} [c_t(H)] - \mathop{\mathbb{E}}_{H \sim \mathrm{P}_{H^*}} [c_t(H)] \right] \\
&= \sum_{t=n+1}^{n+\tau} \mathop{\mathbb{E}}_{\substack{H_1 \sim \mathrm{P}_{H_{t-\tau}} \\ H_2 \sim \mathrm{P}_{H^*}}} [c_t(H_1) - c_t(H_2)] \\
&\leq \sum_{t=n+1}^{n+\tau} 2G_{\mathcal{H}} R_{\mathcal{H}} = 2\tau G_{\mathcal{H}} R_{\mathcal{H}},
\end{aligned}$$

where the penultimate step uses the fact that c_t is $2G_{\mathcal{H}}$-Lipshitz via Lemma 7, and that the diameter of $\mathcal{H}$ is $R_{\mathcal{H}}$. Similarly, one can bound $T_4 \leq 2\tau G_{\mathcal{H}} R_{\mathcal{H}}$. Plugging in all the bounds in Eq. (15), we have the result. ■

A.2 GENERALIZATION ERROR BOUNDS FOR MIXING PROCESSES

Lemma 9. *For any $\tau = o(n)$, and any Wasserstein-stable online learner $\mathcal{L}_n$ for Game 1, with probability one, $\overline{\mathrm{gen}}(A, S_n)$ is at most*

$$\frac{regret_{\mathcal{L}_n,A}}{n} + \frac{\tau}{n}\left[2G_{\mathcal{H}}\sum_{t=1}^{n}\kappa(t) + 4G_{\mathcal{H}}R_{\mathcal{H}} + G_{\mathcal{Z}}R_{\mathcal{Z}}\right] \\ - \frac{1}{n}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)]\,.$$

Proof.

$$\begin{aligned}
\overline{\mathrm{gen}}(A,S_n) &:= \mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}[\mathrm{gen}(H,S_n)\mid S_n] \\
&= \mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}\left[\mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}[\ell(H,Z')] - \frac{1}{n}\sum_{t=1}^{n}\ell(H,Z_t)\mid S_n\right] = -\frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}[c_t(H)] \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)] - \mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}[c_{t+\tau}(H)]\right] - \frac{1}{n}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)] \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\left[\mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}[c_{t+\tau}(H)] - \mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}[c_t(H)]\right] \\
&\overset{(a)}{\leq} \frac{1}{n}\mathrm{regret}_{\mathcal{L}_n,A} + \frac{\tau}{n}\left[2G_{\mathcal{H}}\sum_{t=1}^{n}\kappa(t) + 4G_{\mathcal{H}}R_{\mathcal{H}}\right] - \frac{1}{n}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)] \\
&\quad + \frac{1}{n}\sum_{t=1}^{\tau}\left[\mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}[c_{n+t}(H)] - \mathop{\mathbb{E}}_{H\sim \mathrm{P}_{H^*}}[c_t(H)]\right] \\
&\overset{(b)}{\leq} \frac{1}{n}\mathrm{regret}_{\mathcal{L}_n,A} + \frac{\tau}{n}\left[2G_{\mathcal{H}}\sum_{t=1}^{n}\kappa(t) + 4G_{\mathcal{H}}R_{\mathcal{H}} + G_{\mathcal{Z}}R_{\mathcal{Z}}\right] - \frac{1}{n}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)]\,,
\end{aligned}$$

where (a) uses Lemma 8 and (b) uses Lemma 7. ■

Lemma 10. *For any $\tau = o(n)$,*

$$-\sum_{t=1}^{n}\mathop{\mathbb{E}}_{\mathcal{P}}\left[\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)]\right] \leq n\cdot B_\ell\cdot\beta(\tau+1).$$

Proof. First, we rearrange the terms in $\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)]$ as follows. Consider the indices $a\in\{1,\dots,\tau\}$, and $b\in\{1,\dots,i_a\}$, where $i_a := \min\{b' : (b'-1)\tau + a \leq n\}$, and note that $i_a \leq \lceil n/\tau\rceil$ for any $1\leq a\leq\tau$. Now, let $\mathcal{F}_{(b-1)\tau+a-1} := \sigma(Z_1,\dots,Z_{(b-1)\tau+a-1})$, and define $X_b^a := -\mathop{\mathbb{E}}_{H\sim \mathrm{P}_{(b-1)\tau+a}}[c_{b\tau+a}(H)]$. We rewrite the term $Y := -\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim \mathrm{P}_t}[c_{t+\tau}(H)]$ as follows:

$$Y = \sum_{a=1}^{\tau}\underbrace{\sum_{b=1}^{i_a}\left(X_b^a - \mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a\mid\mathcal{F}_{(b-1)\tau+a-1}\right]\right)}_{\mathcal{M}_a} + \sum_{a=1}^{\tau}\sum_{b=1}^{i_a}\left(\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a\mid\mathcal{F}_{(b-1)\tau+a-1}\right]\right). \tag{16}$$

Firstly, note that $\mathop{\mathbb{E}}_{\mathcal{P}}[\mathcal{M}_a] = 0$ for all $1\leq a\leq\tau$. Therefore,

$$\mathop{\mathbb{E}}_{\mathcal{P}}[Y] = \sum_{a=1}^{\tau}\sum_{b=1}^{i_a}\mathop{\mathbb{E}}_{\mathcal{P}}\left[\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a\mid\mathcal{F}_{(b-1)\tau+a-1}\right]\right]. \tag{17}$$

Now, let $\mathrm{p}_{[s]}^{t}$ and d be the densities with respect to some measure μ.[8] Then, the second term in Eq. (16) can be rewritten as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a \mid \mathcal{F}_{(b-1)\tau+a-1}\right] &= \mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[-\mathop{\mathbb{E}}_{\mathrm{P}_{(b-1)\tau+a}}\left[c_{b\tau+a}(H)\right] \;\middle|\; \mathcal{F}_{(b-1)\tau+a-1}\right] \\
&\leq \mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[\mathop{\mathbb{E}}_{\mathrm{P}_{(b-1)\tau+a}}\left[\left|c_{b\tau+a}(H)\right|\right] \;\middle|\; \mathcal{F}_{(b-1)\tau+a-1}\right] \\
&\overset{(a)}{\leq} \mathop{\mathbb{E}}_{\mathrm{P}_{[(b-1)\tau+a-1]}^{(b-1)\tau+a}}\left[\int_{\mathcal{Z}} \ell(H,Z)\cdot\left|\mathrm{p}_{[(b-1)\tau+a-1]}^{b\tau+a}-\mathrm{d}\right| d\mu\right] \overset{(b)}{\leq} 2\cdot B_\ell\cdot d_{\mathrm{TV}}\left(\mathrm{P}_{[(b-1)\tau+a-1]}^{b\tau+a},\mathcal{D}\right),
\end{aligned} \tag{18}$$

where (a) follows via Fubini's Theorem (Theorem 14.19 of [Kle07]) and by noting the fact that the distribution $\mathrm{P}_{(b-1)\tau+a}$ returned by the online-learner is independent of $Z_{b\tau+a}$ conditioned on $Z_1,\ldots,Z_{(b-1)\tau+a-1}$, and (b) uses Lemma 6. Then, via Definition 1, we have

$$\mathop{\mathbb{E}}_{\mathcal{P}}\left[\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a \mid \mathcal{F}_{(b-1)\tau+a-1}\right]\right] \leq B_\ell\cdot\beta(\tau+1).$$

Plugging the above in (17) completes the proof. ■

Consider any probability space $(\Omega,\mathcal{F},\mathbb{P})$, and let $(\mathcal{F}_t)_{t\in\mathbb{N}}$ be a filtration.[9] Then, a random process $(M_t)_{t\in\mathbb{N}}$ adapted to the filtration $\mathcal{F}_t$ is said to be a *martingle difference sequence* if $\mathbb{E}[|M_t|]<\infty$ and $\mathbb{E}[M_t|\mathcal{F}_{t-1}]=0$ almost surely. We state below the Azuma-Hoeffding inequality, which bounds the probability of the sum of the first T terms of a martingale difference sequence exceeding some constant.

Lemma 15 (Azuma-Hoeffding inequality)**.** *Let $(M_t)_{t\in\mathbb{N}}$ be a martingale difference sequence with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. Let there be constants $c_t\in(0,\infty)$ such that $\forall t\geq 1$, almost surely $|M_t|\leq c_t$. Then, $\forall\gamma>0$,* $\Pr\left(\sum_{t=1}^{T} M_t\geq\gamma\right)\leq\exp\left(-\frac{\gamma^2}{2\sum_{t=1}^{T}c_t^2}\right)$.

See Theorem 13.4 of [MU05][10] for a proof of Lemma 15.

Lemma 11. *With probability at least $1-\delta$, for any $\tau=o(n),\delta>0$,*

$$\begin{aligned}
-\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim P_t}\left[c_{t+\tau}(H)\right] \leq& 2G_{\mathcal{H}}R_{\mathcal{H}}\sqrt{2n\tau\log\left(\tau\delta\right)} \\
&+ n\cdot B_\ell\cdot\phi(\tau+1).
\end{aligned}$$

Proof. Let $\mathcal{F}_{(b-1)\tau+a-1} := \sigma(Z_1,\ldots,Z_{(b-1)\tau+a-1})$. The, exactly as in the proof of Lemma 10, we define $X_b^a := -\mathop{\mathbb{E}}_{H\sim\mathrm{P}_{(b-1)\tau+a}}\left[c_{b\tau+a}(H)\right]$ and $Y := -\sum_{t=1}^{n}\mathop{\mathbb{E}}_{H\sim\mathrm{P}_t}\left[c_{t+\tau}(H)\right]$, and observe that

$$Y=\sum_{a=1}^{\tau}\underbrace{\sum_{b=1}^{i_a}\left(X_b^a-\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a \mid \mathcal{F}_{(b-1)\tau+a-1}\right]\right)}_{\mathcal{M}_a}+\sum_{a=1}^{\tau}\sum_{b=1}^{i_a}\left(\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a \mid \mathcal{F}_{(b-1)\tau+a-1}\right]\right). \tag{19}$$

Now, note that for each $1\leq a\leq\tau$, the term $\mathcal{M}_a$ is a sum of random variables forming a martingale difference sequence w.r.t. the filtration $\mathcal{F}_{(b-1)\tau+a-1}, 1\leq b\leq i_a$. Furthermore, we also observe that $\left|X_b^a-\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a \mid \mathcal{F}_{(b-1)\tau+a-1}\right]\right|\leq 2G_{\mathcal{H}}R_{\mathcal{H}}$ via Lemma 7. From Lemma 15, using the fact that $i_a\leq\lceil\frac{n}{\tau}\rceil$, we can now obtain for any $\gamma>0$

$$\Pr\left[\mathcal{M}_a\geq\gamma\right]\leq\exp\left(-\frac{\tau\gamma^2}{8(n+\tau)G_{\mathcal{H}}^2R_{\mathcal{H}}^2}\right). \tag{20}$$

[8] For example, μ can be chosen as $\mathrm{P}_{[s]}^{t}+\mathcal{D}$.

[9] See Chapter 9 of [Kle07] for a detailed exposition on filtrations and random processes adapted to a filtration.

[10] Theorem 3.4 of [MU05] states the two-sided version of the inequality, i.e., for $\Pr(|\sum_{t=1}^{T}M_t|\geq\beta)$, and hence there is an additional factor of 2 in the RHS. Furthermore, Theorem 3.4 of [MU05] states the inequality for martingales and not martingale difference sequences, but the modification of the proof is straightforward.

Again, as in the proof of Lemma 10, let $\mathrm{p}^{t}_{[s]}$ and d be the densities with respect to some measure μ. Then, by applying Definition 1 in Eq. (18) we get

$$\mathop{\mathbb{E}}_{\mathcal{P}^{b\tau+a}}\left[X_b^a \mid \mathcal{F}_{(b-1)\tau+a-1}\right] \leq B_\ell \cdot \phi(\tau+1). \tag{21}$$

Combining Eq. (20) and Eq. (21), we can now write

$$\Pr\left[Y > \gamma + n \cdot B_\ell \cdot \phi(\tau+1)\right] \overset{(a)}{\leq} \sum_{a=1}^{\tau} \Pr\left[\mathcal{M}_a \geq \gamma/\tau\right] \leq \tau \cdot \exp\left(-\frac{\gamma^2}{8(n+\tau)\tau G_{\mathcal{H}}^2 R_{\mathcal{H}}^2}\right) \tag{22}$$

where (a) follows from a union-bound argument. Therefore, by setting $\gamma = O\left(\sqrt{n \cdot \tau \log\left(\tau/\delta\right)}\right)$, we have $Y \leq 2G_{\mathcal{H}} R_{\mathcal{H}} \sqrt{2n\tau \log\left(\tau/\delta\right)} + n \cdot B_\ell \cdot \phi(\tau+1)$ with probability $\geq 1-\delta$. ■

A.3 APPLICATIONS: GENERALIZATION ERROR BOUNDS FOR EWA

Lemma 12 (EWA Minimizer)**.** *Let $\eta > 0$ be a learning rate. Then, Eq. (4) is minimized by a distribution $\mathcal{P}^* \ll \mathcal{P}_t$ s.t. $\mathcal{P}^*$ satisfies the following equality:*

$$\langle \mathcal{P}^*, c_t \rangle + \frac{1}{\eta} \mathrm{D_{KL}}\left(\mathcal{P}^* \,\|\, \mathcal{P}_t\right) = -\frac{1}{\eta} \log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta c_t(H)}\right]\right). \tag{7}$$

Proof. Seting $\mathcal{P}$ to $\mathcal{P}_t$, Q to $\mathcal{P}$, and Z to $c_t(H)$ in Lemma 2, we obtain for all $\mathcal{P} \ll \mathcal{P}_t$: $\log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta(c_t(H)-\langle\mathcal{P}_t,c_t\rangle)}\right]\right) \geq -\eta\left(\langle P, c_t\rangle - \langle \mathcal{P}_t, c_t\rangle\right) - \mathrm{D_{KL}}\left(\mathcal{P} \,\|\, \mathcal{P}_t\right)$. Rearranging the terms and dividing by η, we get the following:

$$\begin{aligned}
\langle \mathcal{P}, c_t \rangle + \frac{1}{\eta}\mathrm{D_{KL}}\left(\mathcal{P} \,\|\, \mathcal{P}_t\right) &\geq \langle \mathcal{P}_t, c_t\rangle - \frac{1}{\eta}\log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta(c_t(H)-\langle\mathcal{P}_t,c_t\rangle)}\right]\right) \\
&= \langle \mathcal{P}_t, c_t\rangle - \frac{1}{\eta}\log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta c_t(H)} \cdot e^{\eta\langle\mathcal{P}_t,c_t\rangle}\right]\right) \\
&= \langle \mathcal{P}_t, c_t\rangle - \frac{1}{\eta}\log\left(e^{\eta\langle\mathcal{P}_t,c_t\rangle} \cdot \mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta c_t(H)}\right]\right) \\
&= \langle \mathcal{P}_t, c_t\rangle - \langle \mathcal{P}_t, c_t\rangle - \frac{1}{\eta}\log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta c_t(H)}\right]\right) \\
&= -\frac{1}{\eta}\log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta c_t(H)}\right]\right).
\end{aligned}$$

This implies that

$$\langle \mathcal{P}, c_t \rangle + \frac{1}{\eta}\mathrm{D_{KL}}\left(\mathcal{P} \,\|\, \mathcal{P}_t\right) \geq -\frac{1}{\eta}\log\left(\mathbb{E}_{H\sim\mathcal{P}_t}\left[e^{-\eta c_t(H)}\right]\right). \tag{23}$$

Since the above inequality holds for all $\mathcal{P} \ll \mathcal{P}_t$, we now show that for a distribution $\mathcal{P}'$ s.t. $\mathcal{P}' \ll \mathcal{P}_t$, equality is attained when the following condition holds.

$$\frac{d\mathcal{P}'}{d\mathcal{P}_t}(H) = \frac{e^{-\eta c_t(H)}}{\int_{\mathcal{H}} e^{-\eta c_t(H')} d\mathcal{P}_t(H')}. \tag{24}$$

If we plug $\mathcal{P}'$ from Eq. (24) into Eq. (23), the LHS of Eq. (23) becomes

$$\begin{aligned}
\langle \mathcal{P}', c_t \rangle + \frac{1}{\eta} \mathrm{D}_{\mathrm{KL}}\left(\mathcal{P}' \,\|\, \mathcal{P}_t\right) &= \langle \mathcal{P}', c_t \rangle + \frac{1}{\eta} \int_{\mathcal{H}} \log \left(\frac{d\mathcal{P}'}{d\mathcal{P}_t}(H) \right) d\mathcal{P}'(H) \\
&= \langle \mathcal{P}', c_t \rangle + \frac{1}{\eta} \int_{\mathcal{H}} \log \left(\frac{e^{-\eta c_t(H)}}{\int_{\mathcal{H}} e^{-\eta c_t(H')} d\mathcal{P}_t(H')} \right) d\mathcal{P}'(H) \\
&= \langle \mathcal{P}', c_t \rangle - \int_{\mathcal{H}} c_t(H) d\mathcal{P}'(H) - \frac{1}{\eta} \int_{\mathcal{H}} \log \left(\int_{\mathcal{H}} e^{-\eta c_t(H')} d\mathcal{P}_t(H') \right) d\mathcal{P}'(H) \\
&= -\frac{1}{\eta} \log \left(\int_{\mathcal{H}} e^{-\eta c_t(H')} d\mathcal{P}_t(H') \right) \underbrace{\int_{\mathcal{H}} d\mathcal{P}'(H)}_{=1} \\
&= -\frac{1}{\eta} \log \Big(\mathbb{E}_{H \sim \mathcal{P}_t} \left[e^{-\eta c_t(H)} \right] \Big).
\end{aligned}$$

Hence, the minimizer for Eq. (23) exists and is attained when Eq. (24) holds. ■

Corollary 14. *Let $\{Z_t\}_{t \in \mathbb{N}}$ be a geometric ϕ-mixing process with rate $K > 0$, and $r > 1$. Then for any $P_1 \in \Delta_{\mathcal{H}}$, the generalization error $\overline{\mathrm{gen}}\,(A, S_n)$ of any learning algorithm A trained on $S_n = (Z_1, \ldots, Z_n)$ drawn from the mixing process $\{Z_t\}_{t \in \mathbb{N}}$ is upper bounded by*

$$\begin{aligned}
&\frac{\mathrm{D}_{\mathrm{KL}}\left(P_{A(S_n)} \,\|\, P_1\right)}{\sqrt{n}} + \frac{B_\ell^2 + 4G_{\mathcal{H}}^2 R_{\mathcal{H}}^2 \log n}{2\sqrt{n}} + \frac{4G_{\mathcal{H}} R_{\mathcal{H}} \log n}{n} \\
&+ \frac{G_{\mathcal{Z}} R_{\mathcal{Z}} \log n}{n} + 2G_{\mathcal{H}} R_{\mathcal{H}} \sqrt{\frac{2 \log n \log\left(\log n / \delta\right)}{n}} + \frac{B_\ell \cdot K}{n},
\end{aligned}$$

with probability at least $1 - \delta$, for any $\delta > 0$.

Proof. Recall that the R.H.S. in Theorem 4 can be upper-bounded by

$$\frac{\mathrm{D}_{\mathrm{KL}}\left(P_{A(S_n)} \,\|\, P_1\right)}{n \cdot \eta} + \frac{\eta \cdot B_\ell^2}{2} + 2\eta\tau G_{\mathcal{H}}^2 R_{\mathcal{H}}^2. \tag{25}$$

Suppose, for any constant $C > 0$, we set the optimal learning rate

$$\eta_{\mathrm{opt}} = \sqrt{\frac{C \cdot \mathrm{D}_{\mathrm{KL}}\left(P_{A(S_n)} \,\|\, P_1\right)}{n}}.$$

Then for $\eta < \frac{1}{\sqrt{n}}$, $\mathrm{D}_{\mathrm{KL}}\left(P_{A(S_n)} \,\|\, P_1\right) < \frac{1}{C}$, and the R.H.S. in Theorem 4 is at most $O\left(\frac{1}{\sqrt{Cn}}\right)$.

Hence, it is sufficient to analyze the case when $\eta_{\mathrm{opt}} \geq \frac{1}{\sqrt{n}}$ and obtain generalization bounds which are not data-dependent. Plugging $\eta_{\mathrm{opt}} \geq \frac{1}{\sqrt{n}}$ in Theorem 4 gives us the desired bounds. ■