FAST RATES FOR NONSTATIONARY WEIGHTED RISK MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weighted empirical risk minimization is a common approach to prediction under distribution drift. This article studies its out-of-sample prediction error under nonstationarity. We provide a general decomposition of the excess risk into a learning term and an error term associated with distribution drift, and prove oracle inequalities for the learning error under mixing conditions. The learning bound holds uniformly over arbitrary weight classes and accounts for the effective sample size induced by the weight vector, the complexity of the weight and hypothesis classes, and potential data dependence. We also analyze the drift error in the square loss setting and provide explicit bounds for structural-break and smoothly drifting processes. We then combine the learning and drift analysis in (auto-) regression problems with linear models, basis approximations, and neural networks. In all cases, the optimal weights are characterized by a bias–variance trade-off between learning and drift, while minimax-optimal rates (up to logarithmic factors) are recovered when specialized to unweighted and stationary settings.

1 INTRODUCTION

The study of empirical risk minimizers (ERMs) is central to learning theory. In particular, we are interested in the following question: if we learn a hypothesis $\hat{h} \in \mathcal{H}$ from samples $X_1, \dots, X_n$, how well does it perform on an unseen observation X_{n+1}? The setting where samples are independent and identically distributed (iid) has been extensively studied. However, real-world data is often nonstationary and dependent, e.g., financial time series (Cont, 2001; Tsay, 2010) or weather data (Gneiting et al., 2007; Wilks, 2011). A common approach to handling nonstationarity is to introduce weights $w \in \mathbb{R}^n$ to the empirical risk: for some loss function L, define

$$R_n^w(h) = \sum_{t=1}^{n} w_t L(X_t, h), \qquad \hat{h}_w = \arg\min_{h \in \mathcal{H}} R_n^w(h).$$

For example, if the distributions are drifting slowly, one may improve the performance of the minimizer by giving greater weight to more recent observations; see, e.g., Vogel et al. (2020); Mazzetto & Upfal (2023); Jeong et al. (2025). However, when weights concentrate too much on a few observations, the *effective sample size* $n_{\text{eff}}(w) = 1/\|w\|_2^2$ (Kish, 1992) is small and renders the problem noisier, a trade-off that needs to be accounted for.

We want to find $\hat{h}_w$ that is close to the Bayes optimal model $h^*_{P_{n+1}} = \arg\min_h \mathbb{E}[L(X_{n+1}, h)]$. It is thus natural to decompose the excess risk into a learnable and a deterministic component:

$$\underbrace{\mathbb{E}[L(X_{n+1}, \hat{h}_w)] - \mathbb{E}[L(X_{n+1}, h^*_{P_{n+1}})]}_{\text{excess out-of-sample risk}} \lesssim \underbrace{\|\hat{h}_w - h^*_w\|^2_{L_2(P_{n+1})}}_{\text{learning error}} + \underbrace{\|h^*_w - h^*_{P_{n+1}}\|^2_{L_2(P_{n+1})}}_{\text{drift error}},$$

where $h^*_w = \arg\min_h \sum_{t=1}^n w_t \mathbb{E}[L(X_t, h)]$, with both minima taken over all measurable functions.

This decomposition makes the trade-off between learning and drift explicit: larger $n_{\text{eff}}(w)$ can reduce the learning error but may increase drift if the distributions evolve quickly over time. Moreover, the drift error is solely affected by the nonstationarity of the data and the choice of the weights, making it independent of the hypothesis class $\mathcal{H}$. In this work, we study the out-of-sample excess risk of weighted ERM under nonstationarity by analyzing the learning and drift errors.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

CONTRIBUTIONS

Fast rates for the learning error. In Section 3, we state our main result (Theorem 3.2), an oracle inequality for weighted ERM: for any weight class $\mathcal{W}$, with high probability,

$$\forall w \in \mathcal{W}: \quad \|\hat{h}_w - h_w^*\|_{L_2(P_{n+1})} \lesssim r(\|w\|_2),$$

for a rate function r determined by the dependence in the data and the complexities of weight and hypothesis classes, explicitly depending on the effective sample size. In particular:

- *Uniformity over weights:* the guarantee is simultaneous over a generic class $\mathcal{W}$, including any data-dependent choice $\widehat{w} \in \mathcal{W}$.
- *Weight-adaptive models:* the hypothesis class $\mathcal{H}_w$ may depend on the selected weights.
- *Sharp dependence control:* controlling local empirical-risk deviations allows an iid-order leading rate under polynomial mixing, thus improving on existing results even for stationary data and uniform weights.

Section 3.4 gives a matching lower bound for mean estimation that recovers the effective sample size and complexity terms.

Bounds on the drift error. In Section 4, we analyze the drift error for structural break processes and smoothly drifting marginals. For square loss, the drift error depends on how the selected weights interact with changes in the regression function, and we provide explicit bounds for both settings.

Excess-risk analysis for selected models. Section 5 combines learning and drift for linear models, basis expansions, and neural networks. We recover the corresponding iid minimax rates up to logarithmic factors and show how nonstationarity affects the excess risk in the structural break setting: the optimal weight choice must balance the learning and drift errors. We also give a numerical illustration that demonstrates the fast rate under polynomial mixing and the learning-drift trade-off.

RELATED WORK

A large body of literature studies learning under nonstationarity by quantifying distribution drift via the discrepancy, which captures the worst-case expected loss between two time points with respect to the hypothesis class (Mansour et al., 2009). Generalization results leveraging the discrepancy have been derived for independent drifting processes (Yang, 2011; Mohri & Muñoz Medina, 2012; Awasthi et al., 2023; Mazzetto & Upfal, 2023). Such bounds either require restricting the drift rate or require that the discrepancy is sufficiently small over some specified look-back window to be meaningful. Alternatively, Huang & Wang (2025) propose selecting an adaptive look-back window in a nonstationary environment to obtain optimal regret guarantees.

Other authors consider more general settings by jointly relaxing independence and stationarity. Kuznetsov & Mohri (2017) and Hanneke & Yang (2019) derive generalization results for nonstationary mixing processes using the discrepancy; Kuznetsov & Mohri (2016; 2020) impose no restriction on the dependence structure. The latter works also consider weighted ERM, with weights learned by minimizing a discrepancy term. Our decomposition instead isolates drift from the learning problem.

For stationary data, Steinwart & Christmann (2009) and Hang & Steinwart (2014) derive oracle inequalities and fast learning rates for regularized empirical risk minimization for stationary data under mixing assumptions, and Hang & Steinwart (2017) prove a Bernstein-type inequality for geometrically mixing processes and demonstrate learning consequences. Farahmand & Szepesvári (2012) study regularized least-squares regression under exponentially β-mixing dependence and Barrera & Gobet (2021) provide generalization bounds for nonparametric regression under β-mixing, while additionally relaxing the stationarity assumption. Deb & Mukherjee (2024) develop an empirical-process framework for stationary β- and ρ-mixing data and derive localized bounds that imply fast rates in a range of nonparametric problems. Similarly, Alquier et al. (2013) derive PAC-Bayesian oracle inequalities for forecasting and establish fast rates under restrictive dimensionality and dependence conditions. Importantly, Farahmand & Szepesvári (2012); Hang & Steinwart (2014); Barrera & Gobet (2021); Deb & Mukherjee (2024) obtain rates in stationary or unweighted settings that are polynomially worse than $1/n$ for polynomially mixing data, which is not the case for our results.

2 SETUP AND NOTATION

Let $X_1, \dots, X_{t+1}$ be a sequence of random variables with marginal laws $P_1, \dots, P_{t+1}$ and arbitrary dependence. We consider a prediction problem where, at time t, we want to learn a hypothesis h from $X_1, \dots, X_t$ so that the expected out-of-sample loss $\mathbb{E}[L(X_{t+1}, h)]$ is small. This is hopeless without further assumptions, as P_{t+1} can be completely unrelated to $P_1, \dots, P_t$. In most practical scenarios, however, the distributions P_t evolve gradually or have infrequent regime changes. A common approach is to exploit this by considering a weighted empirical risk minimization (ERM) procedure where recent losses are weighted higher.

To be more precise, denote the weighted sample average and its expectation by

$$R_n^w(h) = \sum_{t=1}^{n} w_t L(X_t, h), \qquad R^w(h) = \sum_{t=1}^{n} w_t \mathbb{E}[L(X_t, h)],$$

where $w \in \mathbb{R}^n$ is a weight vector with $\sum_{t=1}^n w_t = 1$. For every such weight vector, write $P_w = \sum_{t=1}^n w_t P_t$. Define the weighted empirical risk minimizer and optimal model as

$$\hat{h}_w = \operatorname*{arg\,min}_{h \in \mathcal{H}_w} R_n^w(h), \qquad h_w^* = \operatorname*{arg\,min}_{h \text{ measurable}} R^w(h),$$

where $\mathcal{H}_w$ is a hypothesis class. We intentionally let $\mathcal{H}_w$ depend on w since we should pick the complexity of the hypothesis class according to the effective sample size $n_{\text{eff}}(w) = 1/\|w\|_2^2$. Moreover, it is unclear which weighting scheme is best for a given problem. The weights are typically themselves selected, for example, through backtesting, so we want to derive bounds that hold uniformly over a class of weights $\mathcal{W}$.

Example 2.1. A few commonly used weight families (Gardner, 2006; Hyndman et al., 2008) are:

(i) Uniform window: $w_i^{(t)} = \mathbb{1}\{i \le t\} \times \begin{cases} 1/s, & i \in \{t-s+1, \dots, t\}, \\ 0, & i \in \{1, \dots, t-s\}. \end{cases}, s \in \{1, \dots, t\}.$

(ii) Exponential smoothing: $w_i^{(t)} = \mathbb{1}\{i \le t\} \dfrac{\exp(-\theta(t-i))}{\sum_{j=1}^t \exp(-\theta(t-j))}, \theta \in (0, \infty).$

The dependence on t is made explicit to emphasize that the weight vectors change over time and only assign non-zero weights on past observations. Our analysis also allows for negative weights, e.g., Brown double exponential weights, discussed in Appendix E.3, provided P_w remains a probability measure and condition (A4) holds.

Our goal is to derive bounds on the out-of-sample excess risk

$$\mathbb{E}[L(X_{t+1}, \hat{h}_w)] - \mathbb{E}[L(X_{t+1}, h_{P_{t+1}}^*)],$$

uniformly in $w \in \mathcal{W}$. Here, $h_{P_{t+1}}^* = \arg\min_h \mathbb{E}[L(X_{t+1}, h)]$ is the Bayes-optimal predictor.

Assumption A1, stated in Section 3.1, and the inequality $(a+b)^2 \le 2(a^2+b^2)$ then imply the following for any fixed $\hat{h}_w$:

$$\underbrace{\mathbb{E}[L(X_{t+1}, \hat{h}_w)] - \mathbb{E}[L(X_{t+1}, h_{P_{t+1}}^*)]}_{\text{excess out-of-sample risk}} \lesssim \underbrace{\|\hat{h}_w - h_w^*\|_{L_2(P_{t+1})}^2}_{\text{learning error}} + \underbrace{\|h_w^* - h_{P_{t+1}}^*\|_{L_2(P_{t+1})}^2}_{\text{drift error}}. \quad (1)$$

Both error terms are measured with respect to the L_2 norm under P_{t+1}. We use the term "learning error" because it combines the estimation error of $\hat{h}_w$ with respect to $\bar{h}_w = \arg\min_{h \in \mathcal{H}_w} R^w(h)$ and the approximation error between $\bar{h}_w$ and h_w^*. The drift error measures the difference between the best hypothesis for the weighted average of past risks and the best hypothesis for the current distribution P_{t+1}. This reflects the nonstationarity of the data and how well the weights w adapt to it, distinguishing our approach from previous works that rely on the discrepancy, which measures the worst-case change in expected loss over the hypothesis class (Mansour et al., 2009).

In practice, the drift error depends on an interaction between the selected weight w and the nonstationarity of the data-generating process. The focus of this section is controlling the learning error uniformly over $w \in \mathcal{W}$. Section 4 gives an analysis of the drift error for concrete drift scenarios.

The bounds depend on several factors. To quantify the dependence structure of the sequence $X_1, X_2, \ldots$, we use two mixing coefficients.

Definition 2.2 (β-mixing)**.** For $k \in \mathbb{N}_0$, let $\mathcal{A}_t = \sigma\{X_j : j \le t\}$ and $\mathcal{B}_t = \sigma\{X_j : j \ge t+k\}$. The β-mixing coefficient at lag $k \ge 0$ is defined as

$$\beta(k) = \sup_{t\ge 1} \beta(\mathcal{A}_t, \mathcal{B}_t) = \sup_{t\ge 1} \frac{1}{2} \sup_{\{A_i\}_i \in \Pi_t, \{B_j\}_j \in \Pi'_{t+k}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|,$$

where Π_t and Π'_{t+k} denote the sets of all finite partitions of $\mathcal{A}_t$ and $\mathcal{B}_t$, respectively.

Definition 2.3 (ρ-mixing)**.** The ρ-mixing coefficient at lag $k \ge 0$ is defined as

$$\rho(k) = \sup_{t\ge 1} \sup_{f,g} \frac{|\mathrm{Cov}(f(X_t), g(X_{t+k}))|}{\sqrt{\mathrm{Var}(f(X_t))\mathrm{Var}(g(X_{t+k}))}},$$

with the inner supremum taken over all functions for which the ratio is well-defined, and its long-run characteristic is defined as $K_\rho = 1 + 2\sum_{k=1}^{\infty} \rho(k)$.

The β-mixing coefficient allows for sharp concentration inequalities via coupling arguments (Berbee, 1979), while the ρ-mixing coefficient is useful to control local fluctuations of the empirical risk around its optimum (e.g., Deb & Mukherjee, 2024). These fluctuations further depend on the complexity of the weight and hypothesis classes. Specifically, denote by $N_1(\epsilon, \mathcal{W}) = N(\epsilon, \mathcal{W}, \|\cdot\|_1)$ the covering number of $\mathcal{W}$ with respect to $\|\cdot\|_1$ and by $N_\infty(\epsilon, \mathcal{H}_w) = N(\epsilon, \mathcal{H}_w, \|\cdot\|_\infty)$ the covering number of $\mathcal{H}_w$ with respect to $\|\cdot\|_\infty$.

When convenient, we write $L_h(\cdot) = L(\cdot, h)$ and $\mathcal{H} = \bigcup_{w\in\mathcal{W}} \mathcal{H}_w$. We frequently use the notation $a \lesssim b$ whenever there is a numerical constant C such that $a \le Cb$. Unless stated otherwise, let $\|w\|$ be the Euclidean norm $\|w\|_2$.

3 Fast rates for the learning error

3.1 Assumptions

We require five key assumptions for Theorem 3.2. Assumptions (A1)–(A2) provide the geometry needed to relate excess risk and compare norms across time, while (A3) enables localization and sharper control of dependence effects. Assumption (A4) is the Bernstein condition that provides the curvature required for fast rates, and (A5) imposes standard boundedness and Lipschitz regularity.

For the weight class $\mathcal{W}$, assume that every P_w is a probability measure and define $\mathcal{P} = \{P_1, \ldots, P_{n+1}\} \cup \{P_w : w \in \mathcal{W}\}$.

(A1) There is $C_L < \infty$ such that for any fixed $h \in \mathcal{H}$ and every $P \in \mathcal{P}$, with $h_P^* \in \arg\min_h \mathbb{E}_P[L(X,h)]$,

$$\mathbb{E}_P[L(X,h)] - \mathbb{E}_P[L(X,h_P^*)] \le C_L \|h - h_P^*\|_{L_2(P)}^2.$$

This assumption is satisfied by many common loss functions, such as squared, Huber, logistic, and zero-one losses, and is only required to hold for deterministic h.

Our main result uses a stronger norm $\|f\|_{L_2(\mathcal{P})} = \sup_{P\in\mathcal{P}} \mathbb{E}_{X\sim P}[f(X)^2]^{1/2}$ to control the learning error. For this norm to be useful, we impose a comparability condition on the set $\mathcal{P}$:

(A2) *There is $C_{\mathcal{P}} < \infty$ such that, for all $h, h' \in \mathcal{H}$,*

$$\sup_{P\in\mathcal{P}} \|h - h'\|_{L_2(P)} \le C_{\mathcal{P}} \inf_{P\in\mathcal{P}} \|h - h'\|_{L_2(P)}.$$

The condition allows us to transfer $L_2(P)$-bounds across time. It is satisfied, for example, if all $P \in \mathcal{P}$ have densities that are uniformly bounded away from zero and infinity with respect to some dominating measure Q—irrespective of $\mathcal{H}$. Much milder conditions usually suffice if the functions in $\mathcal{H}$ are sufficiently regular (see Example 3.1).

The next condition helps to control local deviations in the empirical risk. It is central to our improved rates under polynomial mixing (see Section 3.3).

(A3) *There is a function* $(w, \epsilon) \mapsto C_{w,\infty}(\epsilon) \in [0, \infty)$ *such that, for every* $w \in \mathcal{W}$, $\epsilon > 0$, *and* $h, h' \in \mathcal{H}_w(\epsilon)$,

$$\inf_{P \in \mathcal{P}} \|h - h'\|_{L_2(P)} \geq C_{w,\infty}(\epsilon)\|h - h'\|_\infty,$$

where $\mathcal{H}_w(\epsilon) = \bigcup_{v \in \mathcal{W}: \|v-w\|_1 \leq \epsilon} \mathcal{H}_v$ is the neighborhood class of w with radius ϵ. This formulation allows the hypothesis class to vary locally with the selected weight. Positive $C_{w,\infty}(\epsilon)$ prevents differences between nearby models in the neighborhood from being large in sup-norm while small in L_2; for a fixed hypothesis class, i.e., $\mathcal{H}_w \equiv \mathcal{H}$, it is independent of w or ϵ. A3 always holds with $C_{w,\infty}(\epsilon) = 0$, while $C_{w,\infty}(\epsilon) > 0$ can give sharper rates under dependence (see Section 3.3). It can often be obtained from inverse inequalities, or by direct calculation as in Section 5.

Example 3.1.

- *Scalar class* $\mathcal{H} = \{h(z) = c : |c| \leq B\}$: (A2) and (A3) hold with $C_{\mathcal{P}} = 1$ and $C_{w,\infty}(\epsilon) = 1$.
- *Linear feature class* $\mathcal{H} = \{h(z) = \beta^\top \Phi(z) : \|\beta\|_2 \leq B\}$: if there exist constants $0 < \underline{\lambda} \leq \bar{\lambda} < \infty$ such that for all $P \in \mathcal{P}$,

  $$\underline{\lambda} \leq \lambda_{\min}\big(\mathbb{E}_P[\Phi(Z)\Phi(Z)^\top]\big) \leq \lambda_{\max}\big(\mathbb{E}_P[\Phi(Z)\Phi(Z)^\top]\big) \leq \bar{\lambda},$$

  and $\sup_{z \in \mathcal{Z}} \|\Phi(z)\|_2 \leq 1$, then (A2) and (A3) hold with $C_{\mathcal{P}} = \sqrt{\bar{\lambda}/\underline{\lambda}}$ and $C_{w,\infty}(\epsilon) = \sqrt{\underline{\lambda}}$, where $\lambda_{\min}, \lambda_{\max}$ are the smallest and largest eigenvalues, respectively.

We further impose a margin condition to control the curvature of the excess risk.

(A4) *Let* $P_w = \sum_{t=1}^n w_t P_t$ *be a probability measure for any* $w \in \mathcal{W}$, *and assume that, for any* $w \in \mathcal{W}$ *and* $h \in \mathcal{H}_w$,

$$R^w(h) - R^w(h^*_w) \geq \|h - h^*_w\|^2_{L_2(P_w)}.$$

This is the so-called Bernstein condition (Bartlett & Mendelson, 2006), typically imposed for fast-rate results, but applied to the mixture distribution P_w. It is satisfied, for example, for the square loss (with equality) and a rescaled logistic loss under standard conditions on $\mathcal{H}$. A final simplifying assumption is made to avoid cluttering our main result and arguments:

(A5) The loss function L is uniformly 1-Lipschitz and bounded by 1, and all hypotheses $h \in \mathcal{H}$ and h^*_w are uniformly bounded by 1.

Boundedness and Lipschitzness in sup-norm can be relaxed to milder moment conditions through tedious but standard truncation arguments. The constant 1 can be replaced by a straightforward reparametrization of the problem and bounds.

3.2 Main result

Our main result is a weight-uniform oracle inequality for the learning error. Fix $\delta \in (0, 1/2]$ and let $\mathcal{W}$ be a class of weight vectors. Define

$$C_1 = \sup_{w \in \mathcal{W}} \|w\|_1, \quad C_{\mathcal{W}} = \inf_{w \in \mathcal{W}} \|w\|, \quad B_{\mathcal{W}} = \sup_{w \in \mathcal{W}} \frac{\|w\|_\infty}{\|w\|^2}, \quad m_\beta = \inf\left\{m : \frac{n}{m}\beta(m) \leq \delta\right\}.$$

The first three quantities characterize the weight class: $C_1 = 1$ unless some weights are negative, $C_{\mathcal{W}}^{-2}$ is its maximal effective sample size, and $B_{\mathcal{W}}$ measures weight spikiness. The quantity m_β is the coupling block length used to control the β-mixing error. For each $w \in \mathcal{W}$, define

$$K_w = 4 + \log\left(N_1(\epsilon_{\mathcal{W}}, \mathcal{W}) N_\infty(\epsilon_w, \mathcal{H}_w(\epsilon_{\mathcal{W}}))^2\right),$$

where

$$\epsilon_{\mathcal{W}} = \min\left\{\frac{C_{\mathcal{W}}^3}{64(1 + C_{\mathcal{W}} + C_1 K)}, \frac{C_{\mathcal{W}}}{64(1+K)}\right\}, \qquad \epsilon_w = \frac{\|w\|^2}{32 C_1}.$$

Here, m_β and K_ρ account for potential dependence in the data, while K_w captures the complexity of the weight and hypothesis classes. K is a Lipschitz constant for the rate function r defined below.

Theorem 3.2. *Fix $K \in (0,\infty)$ and assume (A1)–(A5). Let $r\colon [C_{\mathcal{W}}, C_1] \to [0,\infty)$ be increasing and K-Lipschitz, and suppose that for all $w \in \mathcal{W}$,*

$$r(\|w\|)^2 \geq K_w \|w\|^2 \left(C_{\mathcal{P}}^2 K_\rho + m_\beta B_{\mathcal{W}} \min\left\{2, C_{\mathcal{P}} C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} r(\|w\|)\right\}\right), \tag{2}$$

$$r(\|w\|)^2 \geq 4 C_L \inf_{h \in \mathcal{H}_w} \|h - h_w^*\|_\infty^2. \tag{3}$$

Then, with probability at least $1-2\delta$,

$$\forall w \in \mathcal{W}\colon \quad \|\hat{h}_w - h_w^*\|_{L_2(\mathcal{P})}^2 \lesssim R^w(\hat{h}_w) - R^w(h_w^*) \lesssim r(\|w\|)^2 \log^2(1/\delta).$$

Condition (2) reflects the effect of the weight and complexity terms on the estimation error. It is large when the effective sample size $1/\|w\|^2$ is small or when the complexity of the weight and function classes is high. We use the convention $C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} = \infty$ when $C_{w,\infty}(\epsilon_{\mathcal{W}}) = 0$. Condition (3) is an assumption on the approximation error: the rate cannot be lower than the error stemming from approximating h_w^* by an element $h \in \mathcal{H}_w$. The function r is required to be K-Lipschitz only on the interval $[C_{\mathcal{W}}, C_1]$. Since $C_{\mathcal{W}} \geq 1/n$, it is not required to be Lipschitz at zero.

Proof idea. The full proof is deferred to Appendix B and follows a peeling argument with coupling, localization, concentration, and covering. We require two new key techniques:

Weight-dependent localization. For every w, we peel over the weight-sensitive excess-risk shells

$$\mathcal{H}_{k,w} = \left\{h \in \mathcal{H}_w : R^w(h) - R^w(h_w^*) \leq 2^{k+1} r(\|w\|_2)^2\right\}.$$

Thus h_w^*, $\mathcal{H}_w$, the radius, and the normalization all vary with w. The proof jointly discretizes $\mathcal{W}$ and the corresponding localized classes $\mathcal{H}_{k,w}$ while preserving one high-probability event for every w.

Sharp dependence control. Dependence enters in two distinct places. The β-mixing coefficients control the error of coupling the process to independent blocks, whereas K_ρ controls the variance of localized fluctuations. When $C_{w,\infty}(\epsilon_{\mathcal{W}}) > 0$, the local envelope shrinks with the radius, allowing us to recover an iid-order leading term when $K_\rho < \infty$. We discuss this case in the following section.

Remark 3.3. The quadratic Bernstein condition (A4) can be generalized to $[R^w(h) - R^w(h_w^*)]^\vartheta$, $\vartheta \in [0,1]$, covering additional losses with the usual slower rates as ϑ decreases.

3.3 Explicit rate functions

To interpret Theorem 3.2, suppose that the complexity of $\mathcal{H}_w(\epsilon_{\mathcal{W}})$ grows with the effective sample size $n_{\text{eff}}(w) = \|w\|_2^{-2}$: for some $\alpha \in [0,2)$,

$$\log N_\infty(\epsilon, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \lesssim \|w\|_2^{-\alpha} \log(n/\epsilon), \qquad \log N_1(\epsilon, \mathcal{W}) \lesssim \log(n/\epsilon),$$

and let $C_1, B_{\mathcal{W}}, C_{\mathcal{P}} \lesssim 1$. The exponent α determines how model complexity is allowed to adapt to the amount of information retained by the weights. If $C_{w,\infty}(\epsilon_{\mathcal{W}}) > 0$, the local L_2-to-sup-norm control in (A3) yields the admissible weight-specific rate

$$r(\|w\|_2)^2 \lesssim K_\rho \|w\|_2^{2-\alpha} \log n + m_\beta^2 B_{\mathcal{W}}^2 C_{w,\infty}(\epsilon_{\mathcal{W}})^{-2} \|w\|_2^{4-2\alpha} \log^2 n.$$

The first term is the familiar variance contribution at effective sample size $n_{\text{eff}}(w)$; the second is the cost of coupling the dependent sample to independent blocks. Crucially, the latter is of higher order in $\|w\|_2$. Without local sup-norm control ($C_{w,\infty}(\epsilon_{\mathcal{W}}) = 0$), it instead has order $m_\beta B_{\mathcal{W}} \|w\|_2^{2-\alpha} \log n$. Appendix B.2 gives the precise rate functions and conditions.

To sharpen intuition, consider the stationary unweighted case. Let $\mathcal{W} = \{(1/n, \ldots, 1/n)\}$ and $\log N_\infty(\epsilon, \mathcal{H}) \lesssim \log(n/\epsilon)$, where $\mathcal{H}$ is a fixed parametric class (i.e., $\alpha = 0$), $\|w\|_2^2 = 1/n$, and $B_{\mathcal{W}} = 1$. Fixing δ and taking $K_\rho \lesssim 1$, we obtain

$$r(\|w\|_2)^2 \lesssim (K_\rho + m_\beta) \frac{\log n}{n}.$$

Under exponential mixing, where $m_\beta \lesssim \log n$, we recover the optimal rate $n^{-1} \log^2 n$ up to logarithms. Under polynomial mixing $\beta(m) \asymp m^{-(\gamma-1)}$, $\gamma > 2$, however, $m_\beta \asymp n^{1/\gamma}$ and the rate becomes $n^{-1+1/\gamma} \log n$, which is slower than optimal, and comparable to existing stationary bounds (e.g., Hang & Steinwart, 2014; Deb & Mukherjee, 2024).

With $C_{w,\infty}(\epsilon_{\mathcal{W}}) = C_\infty > 0$, the localized bound above instead yields

$$r(\|w\|_2)^2 \lesssim \frac{K_\rho \log n}{n} + \frac{m_\beta^2 \log^2 n}{n^2 C_\infty^2} \lesssim \frac{\log n}{n} + \frac{n^{2/\gamma} \log^2 n}{n^2 C_\infty^2} \lesssim \frac{\log n}{n},$$

in the polynomial case. For fixed $C_\infty > 0$ and $\gamma > 2$, the second term is higher order, leaving an iid-order leading rate up to logarithmic factors. Thus, even in the special case of unweighted ERM with stationary data, Theorem 3.2 improves on the state of the art for polynomially mixing data. In nonparametric settings, where $h_w^* \notin \mathcal{H}$ and $\alpha > 0$, the rates are necessarily slower, and one must carefully balance approximation and estimation errors; see Section 5 for examples.

3.4 Lower bound for mean estimation

To the best of our knowledge, there are no generic lower bounds for nonstationary dependent learning problems, as the optimal rate depends crucially on interactions among the concrete learning problem, the drift model, and the dependence structure. In particular, it is generally unclear what the worst-case dependence structure is within a given budget of m_β or K_ρ for a given problem. Nevertheless, we can still recover the quantities in our upper bound for concrete models, as we illustrate with the following nonstationary block-dependent process.

For a mean sequence $\mu = (\mu_1, \dots, \mu_n)$, denote its weighted oracle by $\mu_w^* = \sum_{i=1}^n w_i \mu_i$. Partition the sample into K consecutive blocks I_k of size M and each I_k into B subblocks $J_{k,b}$ of size m. For $i \in J_{k,b}$, let $Y_i = \mu_k + \xi_{k,b}$, with $\xi_{k,b} \sim N(0, I_d)$ *iid*. Under squared Euclidean loss, $\mu_{w^{(k)}}^* = \mu_k$ and the weighted excess risk is $\|\widehat{\mu}_{w^{(k)}} - \mu_k\|_2^2$. Moreover, $\|w^{(k)}\|_2^2 = M^{-1}$ and $n_{\text{eff}}(w^{(k)}) = M$, whereas repeated observations within a subblock carry the information in one Gaussian vector. Thus each block provides only $B = M/m$ independent observations from $N(\mu_k, I_d)$.

Proposition 3.4. *In the above model, there exist a nonstationary parameter class $\Theta_{\text{ns}} \subset \mathcal{H}^K$, $\mathcal{H} = [-1,1]^d$, and a class $\mathcal{W}$ of block-uniform weights satisfying $\|w\|_2^2 = M^{-1}$ such that, if $d + \log K \le Bd$, then*

$$\inf_{(\widehat{\mu}_w)} \sup_{\mu \in \Theta_{\text{ns}}} \mathbb{E}_\mu \left[\sup_{w \in \mathcal{W}} \|\widehat{\mu}_w - \mu_w^*\|_2^2 \right] \gtrsim K_\rho \|w\|_2^2 \log[N_1(1/2, \mathcal{W}) N_\infty(1/2, \mathcal{H})] \,.$$

Thus the lower bound recovers the effective sample size and complexity terms of (2) up to constants and logarithmic factors. The proof is given in Appendix B.3.

4 Bounds on the drift error

A feature of our risk decomposition (1) is that the drift error depends purely on the interaction between the marginals P_t and the weight w. In particular, it is unrelated to the hypothesis class $\mathcal{H}_w$, the dependence structure, or the realization of the sample. This makes the unlearnable part of the problem transparent, but also implies that a full analysis of out-of-sample risk is highly problem-specific. As an illustrative example, we briefly outline how such an analysis could look for the square loss.

Let $X_t = (Y_t, Z_t) \in [-B, B] \times \mathbb{R}^p, t \ge 1$, and define $L(X_t, h) = (Y_t - h(Z_t))^2$. This also covers autoregressive problems in which Z_t contains lagged values of Y_t. The optimal hypothesis under distribution P_t is the conditional expectation $h_{P_t}^*(z) = \mathbb{E}[Y_t | Z_t = z]$. The following lemma is useful for analyzing the drift error under specific structural assumptions.

Lemma 4.1. *Assume that the features Z_i, $i \ge 1$, have densities p_{Z_i} with respect to a common measure Q, $w \ge 0$, and define $C_p = \sup_{i \in \{1,\dots,t\}, z \in \mathcal{Z}} p_{Z_i}(z) / \sum_{j=1}^t w_j p_{Z_j}(z)$. Then*

$$\|h_w^* - h_{P_{t+1}}^*\|_{L_2(P_{t+1})} \le C_p \sum_{i=1}^t w_i \|h_{P_i}^* - h_{P_{t+1}}^*\|_{L_2(P_{t+1})}.$$

The proof involves simple calculations and is given in Appendix B.5.

Example 4.2 (Smooth drift)**.** Suppose the oracle path varies smoothly in the sense that $\|h_{P_s}^* - h_{P_t}^*\|_{L_2(P_t)} \le \kappa |t-s|^\nu$ for some $\nu, \kappa > 0$. In this setting, putting more weight on recent observations

seems appropriate. For example, the exponential smoothing weights from Example 2.1 (ii) yield

$$\|h^*_w - h^*_{P_{t+1}}\|_{L_2(P_{t+1})} \lesssim C_p\kappa\|w\|^{-2\nu} \lesssim C_p\kappa(1-e^{-\theta})^{-\nu}.$$

Thus, a smaller θ assigns more mass to older observations and increases temporal bias, while smaller κ and smoother variation of the oracle hypothesis allow more dispersed weights.

Example 4.3 (Structural breaks)**.** Let $\tau_1, \tau_2, \cdots \in \mathbb{N}$ be break points and suppose that $P_i = P_j$ for all $i, j \in (\tau_k, \tau_{k+1}], k \geq 1$. This corresponds to a piecewise-stationary process. Consider the uniform weights from Example 2.1 (i) with window size s. If $s \leq t - \tau_k$ and $t < \tau_{k+1}$, the window intersects only a single segment, and therefore $\|h^*_w - h^*_{P_{t+1}}\|_{L_2(P_{t+1})} = 0$. Now suppose the window intersects two segments with end points τ_k, τ_{k+1}, and define the current-regime age $d = t - \tau_k$ and the fraction of earlier-regime observations in the window $\pi_s = (1 - d/s)_+$. Then

$$\|h^*_w - h^*_{P_{t+1}}\|_{L_2(P_{t+1})} \lesssim C_p\pi_s\|h^*_{P_{\tau_k}} - h^*_{P_{\tau_{k+1}}}\|_{L_2(P_{\tau_{k+1}})},$$

and similarly for windows intersecting multiple segments. A "good" weight w must therefore balance the learning error incurred through Theorem 3.2 with the magnitude of recent structural breaks.

5 Excess risk in selected applications

We started the paper by separating the excess risk into a learning term and a deterministic drift term in (1). Section 3 gave a detailed analysis of the learning error and its sharpness, even under slow dependence, while Section 4 analyzed the drift term for concrete drift models in the square loss setting. We now put these pieces together for selected model classes: we first state the learning rate for a generic weight class $\mathcal{W}$, then recover the familiar iid minimax rate up to logarithmic factors, and finally combine the learning term with the drift analysis under structural breaks to derive concrete excess-risk bounds under optimized weight choices.

We begin with an excess out-of-sample guarantee that isolates the remaining effect of the drift from the learning error. Theorem 3.2 and (1) imply the following corollary.

Corollary 5.1. *Under the conditions of Theorem 3.2, with probability at least* $1 - 2\delta$*, for all* $w \in \mathcal{W}$*,*

$$\mathbb{E}[L(X_{n+1}, \hat{h}_w)] - \mathbb{E}[L(X_{n+1}, h^*_{P_{n+1}})] \lesssim r(\|w\|)^2 \log^2(1/\delta) + \|h^*_w - h^*_{P_{n+1}}\|^2_{L_2(P_{n+1})}.$$

The corresponding cumulative excess-risk guarantee is stated as Theorem B.4 in Appendix B.7.

Models. We now specialize Corollary 5.1 to three model classes. Throughout, $Y_t = h^*_t(Z_t) + \eta_t$ is bounded with $\sup_{z\in\mathcal{Z}} \|z\|_2 \leq 1$, $\mathbb{E}[\eta_t \mid Z_t] = 0$, and we use square loss. For simplicity, assume $C_1, B_{\mathcal{W}}, C_{\mathcal{P}}, K_\rho \lesssim 1$ and exponential mixing $\beta(m) \lesssim m\exp(-\gamma m)$, $\gamma > 0$, and $w \geq 0$. Table 1 states the weight-dependent learning rates and the optimal iid excess risk in columns 2 and 3, respectively, for the model classes discussed below. We refer to Appendix C for technical details.

Linear models. Let $\mathcal{H} = \{z \mapsto \beta^\top z : \|\beta\|_2 \leq B\}$ and $h^*_t(z) = \beta^{*\top}_t z$, where $\|\beta^*_t\|_2 \leq B$ and Z_t is uniform on $[0,1]^p$, with uniformly well-conditioned covariance matrix.

Basis expansions. Let Z_t be uniform on $[0,1]$, let h^*_t be 1-Lipschitz, and use step functions on $q(w)$ equal-width bins, with $q(w)$ adapted to the effective sample size.

ReLU networks. Let Z_t be uniform on $[0,1]^p$ and h^*_t be a-smooth. We assume that the ReLU classes have ν_w neurons, ℓ_w layers, parameter bound $b_w \asymp \nu_w$, and architecture size $\nu_w\ell_w \asymp \|w\|_2^{-p/(2a+p)}$ (Nagler & Langer, 2026; Ou & Bölcskei, 2024). We abbreviate $\varrho = 2a/(2a+p)$ and $\zeta = \max\{2a/p, 2\}$.

Table 1: Learning and excess-risk rates; s^* is the optimized post-break window.

Class	Learning error	Excess risk iid	Excess risk at s^*
Linear	$p\|w\|_2^2 \log n$	$p\log n/n$	$(p\log n/d)\big(1 - p\log n/(4C_p^2\Delta^2 d)\big)$
Basis	$\|w\|_2^{4/3}\log n$	$n^{-2/3}\log n$	$d^{-2/3}\log n\big(1 - \log n/(9C_p^2\Delta^2 d^{2/3})\big)$
ReLU	$\|w\|_2^{2\varrho}\log^\zeta n$	$n^{-\varrho}\log^\zeta n$	$d^{-\varrho}\log^\zeta n\big(1 - \varrho^2\log^\zeta n/(4C_p^2\Delta^2 d^\varrho)\big)$

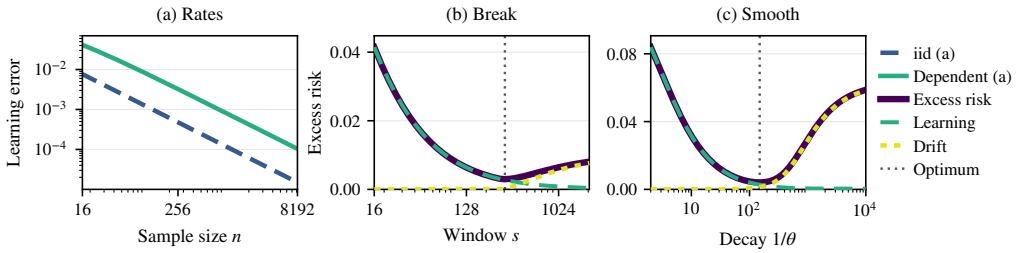

Figure 1: Learning error (a) and excess risk under a break (b) and smooth drift (c); see Appendix D.

Excess risk under structural breaks. To see how drift changes the weights, consider a structural break and uniform windows of length s. Then $n_{\text{eff}}(w) = s$, so each leading learning rate has the form cs^{-b}. Here c contains the model-specific complexity and logarithmic factors, while b is the exponent governing how quickly the learning error decreases with the effective sample size. For the linear, basis, and ReLU models in Table 1, respectively, $(c, b) = (p \log n, 1)$, $(\log n, 2/3)$, and $(\log^{\zeta} n, \varrho)$. If the current regime contains d observations and the oracle changed by Δ, a window extending beyond that regime assigns mass $\pi_s = (1 - d/s)_+$ to pre-break data. By Corollary 5.1 and Example 4.3,

$$\underbrace{\mathbb{E}[L(X_{t+1}, \hat{h}_w) - L(X_{t+1}, h^*_{P_{t+1}})]}_{\text{excess risk}} \lesssim \underbrace{cs^{-b}}_{\text{learning}} + \underbrace{C_p^2 \Delta^2 \pi_s^2}_{\text{drift}},$$

where C_p is the density-ratio factor from Lemma 4.1 that accounts for covariate shift. Choosing the window length balances learning and drift: a longer window reduces estimation error, but including observations from before the break introduces bias. Using d observations from the current regime avoids drift and gives error cd^{-b}, regardless of the series length. When the break is small, including older observations can improve this bound because the drift error grows quadratically with their fraction π_s. Appendix C.4 derives the optimizer s^* and its excess-risk bound.

Simulations. Figure 1 illustrates both trade-offs in the scalar location model $Y_t = \mu_t + \eta_t$. Here η_t is a stationary renewal-block process whose β-coefficients decay only polynomially, but since $C_\infty = 1$, Theorem 3.2 still yields an iid-order learning error, as demonstrated in panel (a). Panels (b)–(c) consider nonstationary scenarios. A structural break favors a window near the current-regime age, whereas smooth drift favors an intermediate exponential-decay horizon. Both panels show that the best weights balance lower learning error from longer averaging against greater drift.

6 Discussion

This article separates the excess out-of-sample risk into a learning and a drift error. We provide fast-rate guarantees for the learning error of weighted empirical risk minimization under nonstationary mixing processes. The results hold uniformly over arbitrary weight classes and are essentially optimal in unweighted regimes. We also give an analysis of the drift error for the square loss. The results are then combined to derive excess-risk bounds under structural breaks for specific models.

A more detailed analysis of the drift error requires nontrivial assumptions. For example, Kuznetsov & Mohri (2020); Awasthi et al. (2023) control nonstationarity by requiring the discrepancy between the recent past and near future to be negligible, so that the drift error can be estimated from the observed data. Such an assumption leads to more concrete risk bounds and algorithms, also within our framework. Our decomposition takes a complementary perspective: it maintains an explicit drift error that captures the unknowable aspects of the underlying data-generating process. An important task for future research is to develop realistic yet general nonstationary models allowing for small drift errors.

Moreover, both error terms depend heavily on the practical choice of w. In applications, weights are often selected by backtesting, drift-detection procedures, or adaptive forecasting rules. Our theory connects quite naturally to online learning algorithms for updating weights dynamically over time. The uniform bounds provide a starting point for analyzing such data-dependent weighting schemes; the remaining challenge is to relate the selected weights to the prediction-relevant drift term.

AI USE STATEMENT

In this work, we used generative AI tools to assist with generating synthetic data for the simple simulations and with proving mathematical claims, primarily the rates for the optimal structural-break window s^* and the derivations for the weight analysis in Appendix E. We did not use generative AI tools for other tasks requiring disclosure, and the remaining categories are not applicable to this work. Additionally, we used generative AI tools to aid and polish the writing. We reviewed all AI-assisted text, code, and mathematical derivations for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We include the code needed to reproduce the simulation study in the supplementary material. The code generates all simulated data and figures presented in the paper.

REFERENCES

Pierre Alquier, Xiaoyin Li, and Olivier Wintenberger. Prediction of time series by statistical learning: General losses and fast rates. *Dependence Modeling*, 1, 12 2013. doi: 10.2478/demo-2013-0004.

Pranjal Awasthi, Corinna Cortes, and Christopher Mohri. Theory and algorithm for batch distribution drift problems. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9826–9851. PMLR, 25–27 Apr 2023.

David Barrera and Emmanuel Gobet. Generalization bounds for nonparametric regression with β-mixing samples. *arXiv preprint arXiv:2108.00997*, 2021.

Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

H. C.P̃. Berbee. *Random Walks with Stationary Increments and Renewal Theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1979.

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

Nabarun Deb and Debarghya Mukherjee. Trade-off between dependence and complexity for non-parametric learning – an empirical process approach. *arXiv preprint arXiv:2401.08978*, 2024.

Amir-massoud Farahmand and Csaba Szepesvári. Regularized least-squares regression: Learning from a β-mixing sequence. *Journal of Statistical Planning and Inference*, 142(2):493–505, 2012. doi: 10.1016/j.jspi.2011.08.007.

Everette S. Gardner. Exponential smoothing: The state of the art—part ii. *International Journal of Forecasting*, 22(4):637–666, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.005.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Hanyuan Hang and Ingo Steinwart. Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127:184–199, 2014. doi: 10.1016/j.jmva.2014.02.012.

Hanyuan Hang and Ingo Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2):708–743, 2017. doi: 10.1214/16-AOS1465.

Steve Hanneke and Liu Yang. Statistical learning under nonstationary mixing processes. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1678–1686. PMLR, 16–18 Apr 2019.

Chengpiao Huang and Kaizheng Wang. A stability principle for learning under nonstationarity. *Operations Research*, 73(6):3044–3064, 2025. doi: 10.1287/opre.2024.0766.

Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. Springer, Berlin / Heidelberg, 2008. ISBN 978-3-540-71916-8. doi: 10.1007/978-3-540-71918-2. eISBN: 978-3-540-71918-2.

Yujin Jeong, Ramesh Johari, Dominik Rothenhäusler, and Emily Fox. Optimal empirical risk minimization under temporal distribution shifts. *arXiv preprint arXiv.2507.13287*, 07 2025. doi: 10.48550/arXiv.2507.13287.

Leslie Kish. Weighting for unequal pi. *Journal of Official Statistics*, 8(2):183, 1992.

Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and online learning. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1190–1213, 2016.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017. ISSN 1573-0565. doi: 10.1007/978-3-319-11662-4\string_19.

Vladimir Kuznetsov and Mehryar Mohri. Discrepancy-based theory and algorithms for forecasting non-stationary time series. *Annals of Mathematics and Artificial Intelligence*, 88:367–399, 2020. doi: 10.1007/s10472-019-09683-1.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory (COLT)*, pp. 19–30, 2009.

Alessio Mazzetto and Eli Upfal. An adaptive algorithm for learning with unknown distribution drift. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory (ALT)*, volume 7568 of *Lecture Notes in Computer Science*, pp. 124–138. Springer, 2012. doi: 10.1007/978-3-642-34106-9_13.

Thomas Nagler and Sophie Langer. Optimal neural network approximation of smooth compositional functions on sets with low intrinsic dimension. *arXiv preprint arXiv:2602.03539*, 2026.

Weigutian Ou and Helmut Bölcskei. Covering numbers for deep relu networks with applications to function approximation and nonparametric regression, 2024.

Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80 of *Probability Theory and Stochastic Modelling*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 978-3-662-54322-1. doi: 10.1007/978-3-662-54323-8.

Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. In *Advances in Neural Information Processing Systems 22 (NeurIPS 2009)*, pp. 1768–1776. Curran Associates, Inc., 2009.

Ruey S. Tsay. *Analysis of Financial Time Series*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2010.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Robin Vogel, Mastane Achab, Stéphan Clémençon, and Charles Tillier. Weighted empirical risk minimization: Sample selection bias correction based on importance sampling. *CoRR*, abs/2002.05145, 2020.

Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, Amsterdam, 3rd edition, 2011.

Liu Yang. Active learning with a drifting distribution. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

A PRELIMINARY RESULTS

A.1 COUPLING AND CONCENTRATION FOR β-MIXING PROCESSES

Let $X_1, \dots, X_n$ be a sequence of random variables. We divide this sequence into alternating blocks of size $m \in \mathbb{N}$, assuming w.l.o.g. that n is a multiple of $2m$. By maximal coupling (e.g., Rio, 2017, Theorem 5.1), there exist random vectors $U_j^* = (X^*_{(j-1)m+1}, \dots, X^*_{jm}) \in \mathcal{X}^m$ such that

- $U_j = (X_{(j-1)m+1}, \dots, X_{jm}) \stackrel{d}{=} U_j^*$ for every $j = 1, \dots, n/m$,
- each of the sequences $(U^*_{2j})_{j=1,\dots,n/2m}$ and $(U^*_{2j-1})_{j=1,\dots,n/2m}$ is independent,
- $\Pr(\exists j \colon U_j \neq U_j^*) \leq \frac{n}{m}\beta(m)$.

Now define the coupled empirical process $\tilde{\mathbb{G}}^{w,m}$ as $\mathbb{G}_n^w$ with all X_j replaced by X_j^*. The following result is immediate from this definition and the third bullet above.

Lemma A.1. *For any class of weights $\mathcal{W}$ and any class of functions $\mathcal{F}$,*

$$\mathbb{P}\{\exists w \in \mathcal{W}, f \in \mathcal{F} \colon \mathbb{G}_n^w(f) \neq \tilde{\mathbb{G}}_n^{w,m}(f)\} \leq \frac{n}{m}\beta(m).$$

We further get the following concentration result for the coupled process.

Lemma A.2. *Assume $\max_{1\leq i\leq n} |f(X_i)| \leq b$ a.s. and $\|f\|^2_{L_2(\mathcal{P})} \leq v$. Then, for any $s > 0$,*

$$\mathbb{P}(|\tilde{\mathbb{G}}_n^{w,m} f| > s) \leq 4 \exp\left(-\frac{s^2}{8v\|w\|^2 K_\rho + 3mb\|w\|_\infty s}\right).$$

Proof. Define $S_i^* = f(X_i^*)$ and A_j^w by

$$A_j^w = \sum_{i=1}^m w_{(j-1)m+i}(S^*_{(j-1)m+i} - \mathbb{E}[S^*_{(j-1)m+i}]),$$

so that

$$|\tilde{\mathbb{G}}_n^{w,m} f| = \left|\sum_{j=1}^{n/m} A_j^w\right| \leq \left|\sum_{j=1}^{n/2m} A_{2j}^w\right| + \left|\sum_{j=1}^{n/2m} A_{2j-1}^w\right|.$$

The random variables in each of the sequences $(A_{2j})_{j=1}^{n/2m}$ and $(A_{2j-1})_{j=1}^{n/2m}$ are independent, so we can apply Bernstein's inequality after showing that (a) each block is uniformly bounded and (b) the sum of the variances of the blocks is bounded. For any $j = 1, \dots, n/m$,

$$\left|A_j^w\right| \leq 2mb\|w\|_\infty.$$

Further, denoting $\mathcal{I}_j = \{(j-1)m + 1, \dots, jm\}$ and using

$$\max_i \operatorname{Var}[S_i^*] \leq \|f\|^2_{L_2(\mathcal{P})} \leq v,$$

we get

$$\begin{aligned}\mathrm{Var}[A_j^w] = \mathbb{E}[(A_j^w)^2] &= \sum_{i,k\in\mathcal{I}_j} w_i w_k \,\mathrm{Cov}(S_i^* - \mathbb{E}S_i^*, S_k^* - \mathbb{E}S_k^*) \\ &\le \sum_{i,k\in\mathcal{I}_j} |w_i w_k| \rho(|k-i|) \sqrt{\mathrm{Var}[S_i^*]\mathrm{Var}[S_k^*]} \\ &\le v\left(\rho(0)\sum_{i\in\mathcal{I}_j} w_i^2 + 2\sum_{u=1}^{m-1}\rho(u)\sum_{i\in\mathcal{I}_j}|w_i w_{i+u}|\right) \\ &\le v\left(\sum_{i\in\mathcal{I}_j} w_i^2 + 2\sum_{u=1}^{m-1}\rho(u)\sum_{i\in\mathcal{I}_j} w_i^2\right) \\ &\le vK_\rho \sum_{i\in\mathcal{I}_j} w_i^2,\end{aligned}$$

where we used the Cauchy–Schwarz inequality in the penultimate step. Summing over the even blocks gives

$$\sum_{j=1}^{n/2m} \mathbb{E}[(A_{2j}^w)^2] \le vK_\rho \sum_{j=1}^{n/2m}\sum_{i\in\mathcal{I}_{2j}} w_i^2 \le v\|w\|^2 K_\rho,$$

so Bernstein's inequality for independent random variables yields

$$\mathbb{P}\left(\left|\sum_{j=1}^{n/2m} A_{2j}^w\right| > s\right) \le 2\exp\left(-\frac{s^2}{2v\|w\|^2K_\rho + \frac{4}{3}mb\|w\|_\infty s}\right).$$

Since the same variance bound and argument hold for the odd blocks, we finally get

$$\begin{aligned}\mathbb{P}(|\tilde{\mathbb{G}}_n^{w,m} f| > s) &\le \mathbb{P}\left(\left|\sum_{j=1}^{n/2m} A_{2j}^w\right| > s/2\right) + \mathbb{P}\left(\left|\sum_{j=1}^{n/2m} A_{2j-1}^w\right| > s/2\right) \\ &\le 4\exp\left(-\frac{s^2}{8v\|w\|^2K_\rho + 3mb\|w\|_\infty s}\right),\end{aligned}$$

as claimed. □

A.2 Weight and hypothesis discretization

Lemma A.3. *For any $w, w' \in \mathcal{W}$ satisfying the conditions of Theorem 3.2,*

$$|r(\|w\|)^{-2} - r(\|w'\|)^{-2}| \le \frac{2K\|w - w'\|_1}{C_{\mathcal{W}}^3}.$$

Proof. First note that (2) implies

$$r(\|w\|)^2 \ge \|w\|^2 C_{\mathcal{P}}^2 K_\rho \ge \|w\|^2 \ge C_{\mathcal{W}}^2.$$

Set $u_1 = \|w\|$ and $u_2 = \|w'\|$ and observe that

$$|r(u_1)^{-2} - r(u_2)^{-2}| = \frac{|r(u_1) - r(u_2)|(r(u_1) + r(u_2))}{r(u_1)^2 r(u_2)^2} \le \frac{2|r(u_1) - r(u_2)|}{C_{\mathcal{W}}^3} \le \frac{2K|u_1 - u_2|}{C_{\mathcal{W}}^3}.$$

Now the claim follows from $|\|w\| - \|w'\|| \le \|w - w'\| \le \|w - w'\|_1$. □

Lemma A.4. *For any $(h, h'), (\tilde{h}, \tilde{h}') \in \mathcal{H} \times \mathcal{H}$ and $w, \tilde{w} \in \mathcal{W}$ satisfying the conditions of Theorem 3.2,*

$$\begin{aligned}&|r(\|w\|)^{-2}\mathbb{G}_n^w(L_h - L_{h'}) - r(\|\tilde{w}\|)^{-2}\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})| \\ &\le 2C_1\|\tilde{w}\|^{-2}(\|h - \tilde{h}\|_\infty + \|h' - \tilde{h}'\|_\infty) + 8C_{\mathcal{W}}^{-3}(1 + C_{\mathcal{W}} + C_1 K)\|w - \tilde{w}\|_1.\end{aligned}$$

Proof. By the triangle inequality

$$\begin{aligned}
&|r(\|w\|)^{-2}\mathbb{G}_n^w(L_h - L_{h'}) - r(\|\tilde{w}\|)^{-2}\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})| \\
&= |r(\|w\|)^{-2}[\mathbb{G}_n^w(L_h - L_{h'}) - \mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})] + [r(\|w\|)^{-2} - r(\|\tilde{w}\|)^{-2}]\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| \\
&\leq r(\|w\|)^{-2}|\mathbb{G}_n^w(L_h - L_{h'}) - \mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| + |r(\|w\|)^{-2} - r(\|\tilde{w}\|)^{-2}|\,|\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})|.
\end{aligned}$$

We begin by analyzing the first term. Observe that

$$\begin{aligned}
|\mathbb{G}_n^w(L_h) - \mathbb{G}_n^{\tilde{w}}(L_h)| &= \left|\sum_{t=1}^n (w_t - \tilde{w}_t)(\mathbb{E}[L(X_t, h)] - L(X_t, h))\right| \\
&\leq 2\sum_{t=1}^n |w_t - \tilde{w}_t| \\
&= 2\|w - \tilde{w}\|_1.
\end{aligned}$$

It further holds $r(\|w\|) \geq \|w\| \geq C_{\mathcal{W}}$. Thus

$$\begin{aligned}
&r(\|w\|)^{-2}|\mathbb{G}_n^w(L_h - L_{h'}) - \mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| \\
&\leq C_{\mathcal{W}}^{-2}|(R^w - R_n^w)(h) - (R^w - R_n^w)(h') - (R^{\tilde{w}} - R_n^{\tilde{w}})(h) + (R^{\tilde{w}} - R_n^{\tilde{w}})(h')| \\
&\leq C_{\mathcal{W}}^{-2}\Big(|(R^w - R_n^w)(h) - (R^{\tilde{w}} - R_n^{\tilde{w}})(h)| + |(R^w - R_n^w)(h') - (R^{\tilde{w}} - R_n^{\tilde{w}})(h')|\Big) \\
&\leq 4C_{\mathcal{W}}^{-2}\|w - \tilde{w}\|_1.
\end{aligned}$$

We now bound the second term by Lemma A.3:

$$\begin{aligned}
&|r(\|w\|)^{-2} - r(\|\tilde{w}\|)^{-2}||\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| \\
&\leq 2KC_{\mathcal{W}}^{-3}\|w - \tilde{w}\|_1|\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| \\
&\leq 2KC_{\mathcal{W}}^{-3}\|w - \tilde{w}\|_1 \sum_{t=1}^n |\tilde{w}_t||L(X_t, h) - L(X_t, h') + \mathbb{E}[L(X_t, h) - L(X_t, h')]| \\
&\leq 8KC_{\mathcal{W}}^{-3}C_1\|w - \tilde{w}\|_1.
\end{aligned}$$

It remains to replace $\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})$ by $\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})$. We have

$$|\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'}) - \mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})| \leq |\mathbb{G}_n^{\tilde{w}}(L_h - L_{\tilde{h}})| + |\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}'} - L_{h'})|,$$

and note that

$$\begin{aligned}
|\mathbb{G}_n^{\tilde{w}}(L_h - L_{\tilde{h}})| &\leq |(R^{\tilde{w}} - R_n^{\tilde{w}})(h) - (R^{\tilde{w}} - R_n^{\tilde{w}})(\tilde{h})| \\
&\leq \sum_{t=1}^n |\tilde{w}_t||L(X_t, h) - L(X_t, \tilde{h}) + \mathbb{E}[L(X_t, h) - L(X_t, \tilde{h})]| \\
&\leq \sum_{t=1}^n |\tilde{w}_t||L(X_t, h) - L(X_t, \tilde{h})| + \mathbb{E}|L(X_t, h) - L(X_t, \tilde{h})| \\
&\leq \sum_{t=1}^n |\tilde{w}_t|(\|h - \tilde{h}\|_\infty + \|h - \tilde{h}\|_\infty) \\
&\leq 2C_1\|h - \tilde{h}\|_\infty.
\end{aligned}$$

Because the same argument holds for $|\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}'} - L_{h'})|$, we have

$$|\mathbb{G}_n^{\tilde{w}}(L_h - L_{\tilde{h}})| + |\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}'} - L_{h'})| \leq 2C_1(\|h - \tilde{h}\|_\infty + \|h' - \tilde{h}'\|_\infty),$$

which, using $r(\|\tilde{w}\|)^{-2} \leq \|\tilde{w}\|^{-2}$, yields

$$r(\|\tilde{w}\|)^{-2}|\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'}) - \mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})| \leq 2C_1\|\tilde{w}\|^{-2}(\|h - \tilde{h}\|_\infty + \|h' - \tilde{h}'\|_\infty).$$

The claim follows after summing the bounds and noting that

$$4C_{\mathcal{W}}^{-2} + 8KC_{\mathcal{W}}^{-3}C_1 \leq 8C_{\mathcal{W}}^{-3}(1 + C_{\mathcal{W}} + C_1K)$$

and multiplying by $\|w - \tilde{w}\|_1$ gives the claim. □

B PROOFS OF MAIN RESULTS

B.1 PROOF OF THEOREM 3.2

For any $w \in \mathcal{W}$, define $\bar{h}_w \in \mathcal{H}_w$ as the closest approximator of h^*_w. By (3), $\|\bar{h}_w - h^*_w\|^2_\infty \le r(\|w\|)^2/(4C_L)$. Invoking (A1), we get

$$\begin{aligned} R^w(\hat{h}_w) - R^w(h^*_w) &\le R^w(\hat{h}_w) - R^w(\bar{h}_w) + R^w(\bar{h}_w) - R^w(h^*_w) \\ &\le R^w(\hat{h}_w) - R^w(\bar{h}_w) + C_L\|\bar{h}_w - h^*_w\|^2_{L_2(\mathcal{P})} \\ &\le R^w(\hat{h}_w) - R^w(\bar{h}_w) + r(\|w\|)^2/4. \end{aligned}$$

Since $R^w_n(\hat{h}_w) \le R^w_n(\bar{h}_w)$, we have

$$\begin{aligned} R^w(\hat{h}_w) - R^w(\bar{h}_w) &= (R^w - R^w_n)(\hat{h}_w) - (R^w - R^w_n)(\bar{h}_w) + R^w_n(\hat{h}_w) - R^w_n(\bar{h}_w) \\ &\le (R^w - R^w_n)(\hat{h}_w) - (R^w - R^w_n)(\bar{h}_w) \\ &= \mathbb{G}^w_n(L_{\hat{h}_w} - L_{\bar{h}_w}), \end{aligned}$$

where $L_h = L(\cdot, h)$. We now want to bound $\mathbb{P}\{\sup_w r(\|w\|)^{-2}(R^w(\hat{h}_w) - R^w(h^*_w)) > 2^M\}$. The previous two displays imply

$$r(\|w\|)^{-2}(R^w(\hat{h}_w) - R^w(h^*_w)) \le r(\|w\|)^{-2}\mathbb{G}^w_n(L_{\hat{h}_w} - L_{\bar{h}_w}) + 1/4.$$

For every $k \ge 1$ and $w \in \mathcal{W}$, define the sets $H_{k,w} = \{h \in \mathcal{H}_w \colon R^w(h) - R^w(h^*_w) \le 2^{k+1} r(\|w\|)^2\}$. Then

$$\begin{aligned} &\mathbb{P}\left(\sup_{w\in\mathcal{W}} r(\|w\|)^{-2}(R^w(\hat{h}_w) - R^w(h^*_w)) > 2^M\right) \\ &= \mathbb{P}\left(\bigcup_{k=M}^{\infty}\left\{2^k < \sup_{w\in\mathcal{W}} r(\|w\|)^{-2}(R^w(\hat{h}_w) - R^w(h^*_w)) \le 2^{k+1}\right\}\right) \\ &= \mathbb{P}\left(\bigcup_{k=M}^{\infty}\left\{\sup_{w\in\mathcal{W}} r(\|w\|)^{-2}(R^w(\hat{h}_w) - R^w(h^*_w))\mathbb{1}_{\hat{h}_w\in H_{k,w}} > 2^k - 1/4\right\}\right) \\ &\le \mathbb{P}\left(\bigcup_{k=M}^{\infty}\left\{\sup_{w\in\mathcal{W}} r(\|w\|)^{-2}\mathbb{G}^w_n(L_{\hat{h}_w} - L_{\bar{h}_w})\mathbb{1}_{\hat{h}_w\in H_{k,w}} > 2^k - 1/4\right\}\right) \\ &\le \mathbb{P}\left(\bigcup_{k=M}^{\infty}\left\{\sup_{w\in\mathcal{W}}\sup_{h\in H_{k,w}} r(\|w\|)^{-2}|\mathbb{G}^w_n(L_h - L_{\bar{h}_w})| > 2^k - 1/4\right\}\right) \\ &\le \mathbb{P}\left(\bigcup_{k=M}^{\infty}\left\{\sup_{w\in\mathcal{W}}\sup_{h\in H_{k,w}} r(\|w\|)^{-2}|\tilde{\mathbb{G}}^{w,m_\beta}_n(L_h - L_{\bar{h}_w})| > 2^k - 1/4\right\}\right) + \frac{n}{m_\beta}\beta(m_\beta) \\ &\le \sum_{k=M}^{\infty}\mathbb{P}\left(\sup_{w\in\mathcal{W}}\sup_{h\in H_{k,w}} r(\|w\|)^{-2}|\tilde{\mathbb{G}}^{w,m_\beta}_n(L_h - L_{\bar{h}_w})| > 2^k - 1/4\right) + \frac{n}{m_\beta}\beta(m_\beta), \end{aligned}$$

where the last two inequalities follow from Lemma A.1 and the union bound, respectively. By definition of m_β, $(n/m_\beta)\beta(m_\beta) \le \delta$. Further, for any $h \in H_{k,w}$, since $h, \bar{h}_w \in \mathcal{H}_w$, (A2), the triangle inequality under P_w, (A4), and (3) imply, for $k \ge M$,

$$\begin{aligned} \|h - \bar{h}_w\|_{L_2(\mathcal{P})} &\le C_{\mathcal{P}}\|h - \bar{h}_w\|_{L_2(P_w)} \\ &\le C_{\mathcal{P}}\big(\|h - h^*_w\|_{L_2(P_w)} + \|\bar{h}_w - h^*_w\|_{L_2(P_w)}\big) \\ &\le C_{\mathcal{P}}\big([R^w(h) - R^w(h^*_w)]^{1/2} + r(\|w\|)\big) \\ &\le C_{\mathcal{P}}(2^{(k+1)/2} + 1)r(\|w\|) \le C_{\mathcal{P}}2^{k/2+1}r(\|w\|). \end{aligned}$$

Thus, defining $H'_{k,w} = \{(h,h') \in \mathcal{H}_w \times \mathcal{H}_w \colon \|h-h'\|_{L_2(\mathcal{P})} \le C_{\mathcal{P}} 2^{k/2+1} r(\|w\|)\}$ and substituting above yields

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}} r(\|w\|)^{-2}(R^w(\hat{h}_w) - R^w(h^*_w)) > 2^M\right) \\
&\le \sum_{k=M}^{\infty} \mathbb{P}\left(\sup_{w\in\mathcal{W}} \sup_{(h,h')\in H'_{k,w}} r(\|w\|)^{-2} |\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^k - 1/4\right) + \delta.
\end{aligned}$$

Let

$$\overline{H}'_{k,w} = \left\{(h,h') \in (\mathcal{H}_w(\epsilon_{\mathcal{W}}))^2 : \|h-h'\|_{L_2(\mathcal{P})} \le C_{\mathcal{P}} 2^{(k+3)/2} r(\|w\|)\right\}. \tag{4}$$

A covering argument (Lemma B.1 below) shows that for every $k \ge M$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W},(h,h')\in H'_{k,w}} r(\|w\|)^{-2} |\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^k - \frac{1}{4}\right) \\
&\le N_1(\epsilon_{\mathcal{W}}, \mathcal{W}) \sup_{w\in\mathcal{W}} \left\{ N_\infty^2(\epsilon_w, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \sup_{(h,h')\in\overline{H}'_{k,w}} \mathbb{P}\left(r(\|w\|)^{-2} |\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1}\right) \right\} \\
&\le \sup_{w\in\mathcal{W}} e^{K_w} \sup_{(h,h')\in\overline{H}'_{k,w}} \mathbb{P}\left(r(\|w\|)^{-2} |\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1}\right).
\end{aligned}$$

Substituting in the previous display, we get

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w} r(\|w\|)^{-2}(R^w(\hat{h}_w) - R^w(h^*_w)) > 2^M\right) \\
&\le \sum_{k=M}^{\infty} \sup_{w\in\mathcal{W}} e^{K_w} \sup_{(h,h')\in\overline{H}'_{k,w}} \mathbb{P}\left(r(\|w\|)^{-2} |\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1}\right) + \delta \\
&\le 2\delta,
\end{aligned}$$

where the last inequality follows from Lemma B.2 upon choosing $M = \lceil 2\log_2(96 + 24\log(34/\delta))\rceil$. This concludes the proof of the main theorem. We now state and prove the intermediate lemmas.

Lemma B.1. *Suppose the conditions of Theorem 3.2 hold and let* $H'_{k,w} = \{(h,h') \in \mathcal{H}_w \times \mathcal{H}_w \colon \|h-h'\|_{L_2(\mathcal{P})} \le C_{\mathcal{P}} 2^{k/2+1} r(\|w\|)\}$. *Let* $\overline{H}'_{k,w}$ *be the union class defined in* (4)*. Then, for any* $k \ge 1$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W},(h,h')\in H'_{k,w}} r(\|w\|)^{-2} |\mathbb{G}_n^{w}(L_h - L_{h'})| > 2^k - \frac{1}{4}\right) \\
&\le N_1(\epsilon_{\mathcal{W}}, \mathcal{W}) \sup_{w\in\mathcal{W}} \left\{ N_\infty^2(\epsilon_w, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \sup_{(h,h')\in\overline{H}'_{k,w}} \mathbb{P}\left(r(\|w\|)^{-2} |\mathbb{G}_n^{w}(L_h - L_{h'})| > 2^{k-1}\right) \right\}.
\end{aligned}$$

Proof. By Lemma A.4, for $h = \tilde{h}$ and $h' = \tilde{h}'$:

$$r(\|w\|)^{-2} |\mathbb{G}_n^w(L_h - L_{h'})| \le r(\|\tilde{w}\|)^{-2} |\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| + 8C_{\mathcal{W}}^{-3}(1 + C_{\mathcal{W}} + KC_1)\|w - \tilde{w}\|_1.$$

Let $\mathcal{M}(\epsilon_{\mathcal{W}})$ be an $\epsilon_{\mathcal{W}}$-net of $\mathcal{W}$ with respect to the $\|\cdot\|_1$-norm such that $|\mathcal{M}(\epsilon_{\mathcal{W}})| \le N_1(\epsilon_{\mathcal{W}}, \mathcal{W})$. In particular, for any $w \in \mathcal{W}$, there is $\tilde{w} \in \mathcal{M}(\epsilon_{\mathcal{W}})$ such that $\|w - \tilde{w}\|_1 \le \epsilon_{\mathcal{W}}$. Since $\epsilon_{\mathcal{W}} \le C_{\mathcal{W}}^3/[64(1+C_{\mathcal{W}}+C_1K)]$, the weight-discretization error is at most $1/8$. Moreover, $\mathcal{H}_w \subseteq \mathcal{H}_{\tilde{w}}(\epsilon_{\mathcal{W}})$. Lipschitzness of r and the definition of $\epsilon_{\mathcal{W}}$ give

$$r(\|w\|) \le r(\|\tilde{w}\|) + K\epsilon_{\mathcal{W}} \le \frac{65}{64} r(\|\tilde{w}\|) \le \sqrt{2}\, r(\|\tilde{w}\|),$$

because $K\epsilon_{\mathcal{W}} \le C_{\mathcal{W}}K/[64(1+K)] \le r(\|\tilde{w}\|)/64$. Hence every $H'_{k,w}$ assigned to $\tilde{w}$ is contained in

$$\left\{(h,h') \in (\mathcal{H}_{\tilde{w}}(\epsilon_{\mathcal{W}}))^2 : \|h-h'\|_{L_2(\mathcal{P})} \le C_{\mathcal{P}} 2^{(k+3)/2} r(\|\tilde{w}\|)\right\}.$$

$$\mathbb{P}\left(\sup_{w\in\mathcal{W},(h,h')\in H'_{k,w}} r(\|w\|)^{-2}|\mathbb{G}_n^w(L_h - L_{h'})| > 2^k - \frac{1}{4}\right)$$
$$\leq \mathbb{P}\left(\sup_{\substack{\tilde{w}\in\mathcal{M}(\epsilon_{\mathcal{W}}),\\ (h,h')\in\bigcup_{v\in\mathcal{W}:\,\|v-\tilde{w}\|_1\leq\epsilon_{\mathcal{W}}} H'_{k,v}}} r(\|\tilde{w}\|)^{-2}|\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| > 2^k - \frac{3}{8}\right).$$

Setting $\tilde{w} = w$ in Lemma A.4 gives

$$r(\|\tilde{w}\|)^{-2}|\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| \leq r(\|\tilde{w}\|)^{-2}|\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})| + 2C_1\|\tilde{w}\|^{-2}(\|h-\tilde{h}\|_\infty + \|h'-\tilde{h}'\|_\infty).$$

For every $\tilde{w} \in \mathcal{M}(\epsilon_{\mathcal{W}})$, cover each component of $\mathcal{H}_{\tilde{w}}(\epsilon_{\mathcal{W}})$ at radius $\epsilon_{\tilde{w}} = \|\tilde{w}\|^2/(32C_1)$ in sup-norm. Taking the Cartesian product of two such covers gives a cover of the product class for the metric $d_\oplus((h,h'),(\tilde{h},\tilde{h}')) = \|h-\tilde{h}\|_\infty + \|h'-\tilde{h}'\|_\infty$ at radius $2\epsilon_{\tilde{w}} = \|\tilde{w}\|^2/(16C_1)$ with at most $N_\infty(\epsilon_{\tilde{w}}, \mathcal{H}_{\tilde{w}}(\epsilon_{\mathcal{W}}))^2$ cells. Retain only cells intersecting $\bigcup_{v\in\mathcal{W}:\,\|v-\tilde{w}\|_1\leq\epsilon_{\mathcal{W}}} H'_{k,v}$ and choose one representative from that intersection. These representatives belong to the localized union and hence, by the preceding argument, to $\overline{H}'_{k,\tilde{w}}$. Any point in a retained cell is within distance $2\cdot 2\epsilon_{\tilde{w}}$ of its representative. Consequently, the discretization error is at most $2C_1\|\tilde{w}\|^{-2}(4\epsilon_{\tilde{w}}) = 1/4$. Denoting the resulting set of representatives by $\mathcal{N}'_{k,\tilde{w}}$, the last display and a union bound give

$$\mathbb{P}\left(\sup_{\substack{\tilde{w}\in\mathcal{M}(\epsilon_{\mathcal{W}}),\\ (h,h')\in\bigcup_{v:\,\|v-\tilde{w}\|_1\leq\epsilon_{\mathcal{W}}} H'_{k,v}}} r(\|\tilde{w}\|)^{-2}|\mathbb{G}_n^{\tilde{w}}(L_h - L_{h'})| > 2^k - \frac{3}{8}\right)$$
$$\leq \mathbb{P}\left(\sup_{\tilde{w}\in\mathcal{M}(\epsilon_{\mathcal{W}}),\,(\tilde{h},\tilde{h}')\in\mathcal{N}'_{k,\tilde{w}}} r(\|\tilde{w}\|)^{-2}|\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})| > 2^k - \frac{5}{8}\right)$$
$$\leq \sum_{\tilde{w}\in\mathcal{M}(\epsilon_{\mathcal{W}})}\sum_{(\tilde{h},\tilde{h}')\in\mathcal{N}'_{k,\tilde{w}}} \mathbb{P}\left(r(\|\tilde{w}\|)^{-2}|\mathbb{G}_n^{\tilde{w}}(L_{\tilde{h}} - L_{\tilde{h}'})| > 2^{k-1}\right)$$
$$\leq N_1(\epsilon_{\mathcal{W}},\mathcal{W}) \sup_{w\in\mathcal{W}}\left\{N_\infty^2(\epsilon_w, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \sup_{(h,h')\in\overline{H}'_{k,w}} \mathbb{P}\left(r(\|w\|)^{-2}|\mathbb{G}_n^w(L_h - L_{h'})| > 2^{k-1}\right)\right\},$$

which completes the proof. □

Lemma B.2. *Let the conditions of Theorem 3.2 hold and $\delta \in (0,1)$ be arbitrary. Then, for $M = \lceil 2\log_2(96 + 24\log(34/\delta))\rceil$,*

$$\sum_{k=M}^{\infty}\sup_{w\in\mathcal{W}} e^{K_w} \sup_{\substack{h,h'\in\mathcal{H}_w(\epsilon_{\mathcal{W}}):\\ \|h-h'\|^2_{L_2(\mathcal{P})}\leq C_{\mathcal{P}}^2 2^{k+3} r(\|w\|)^2}} \mathbb{P}\Big(|\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1} r(\|w\|)^2\Big) \leq \delta.$$

Proof. Let $w \in \mathcal{W}$ be arbitrary and $h, h' \in \mathcal{H}_w(\epsilon_{\mathcal{W}})$ be such that $\|h-h'\|^2_{L_2(\mathcal{P})} \leq C_{\mathcal{P}}^2 2^{k+3} r^2$, where here and below we abbreviate $r = r(\|w\|)$. Then, by assumption,

$$|L_h(X_i) - L_{h'}(X_i)| \leq |h(X_i) - h'(X_i)| \leq \|h-h'\|_\infty$$
$$\leq C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1}\|h-h'\|_{L_2(\mathcal{P})} \leq C_{\mathcal{P}} C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} 2^{(k+3)/2} r,$$

and trivially $|L_h(X_i) - L_{h'}(X_i)| \leq 2$. Similarly,

$$\mathrm{Var}[L_h(X_i) - L_{h'}(X_i)] \leq \|L_h - L_{h'}\|^2_{L_2(\mathcal{P})} \leq \|h-h'\|^2_{L_2(\mathcal{P})} \leq C_{\mathcal{P}}^2 2^{k+3} r^2.$$

We apply Lemma A.2 with

$$v = C_{\mathcal{P}}^2 2^{k+3} r^2, \qquad b = \min\{2,\, C_{\mathcal{P}} C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} 2^{(k+3)/2} r\}, \qquad s = 2^{k-1} r^2,$$

to obtain

$$\begin{aligned}&\mathbb{P}\Big(|\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1}r^2\Big)\\ &\leq 4\exp\left(-\frac{2^{2(k-1)}r^2}{8C_{\mathcal{P}}^2 2^{k+3}\|w\|^2 K_\rho + 3m_\beta\|w\|_\infty \min\{2, C_{\mathcal{P}}C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1}2^{(k+3)/2}r\}\,2^{k-1}}\right).\end{aligned} \tag{5}$$

Case: $C_{w,\infty}(\epsilon_{\mathcal{W}}) = 0$. Then $b = 2$, and (5) and (2) imply

$$\mathbb{P}\Big(|\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1}r^2\Big) \leq 4e^{-K_w 2^k/256}.$$

Because $e^{-K_w 2^M/256} \leq \frac{1}{2}$ for our choice of M and $K_w \geq 4$, we can bound:

$$\begin{aligned}\sum_{k=M}^{\infty} e^{-K_w 2^k/256} &= \sum_{j=0}^{\infty}(e^{-K_w 2^M/256})^{2^j}\\ &\leq \sum_{j=0}^{\infty}(e^{-K_w 2^M/256})^{j+1}\\ &= \frac{e^{-K_w 2^M/256}}{1 - e^{-K_w 2^M/256}} \leq 2e^{-K_w 2^M/256}.\end{aligned}$$

Since $2^M \geq 256 + 64\log(8/\delta)$, we obtain

$$e^{K_w}\sum_{k=M}^{\infty} 4e^{-K_w 2^k/256} \leq 8e^{K_w - K_w 2^M/256} \leq 8e^{-(K_w/4)\log(8/\delta)} \leq 8e^{-\log(8/\delta)} = \delta.$$

Case: $C_{w,\infty}(\epsilon_{\mathcal{W}}) > 0$. We distinguish two sub-cases. Let k_0 be the largest integer such that

$$C_{\mathcal{P}}^2 2^{k+3}\|w\|^2 K_\rho \geq m_\beta\|w\|_\infty C_{\mathcal{P}}C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1}2^{(k+3)/2}r2^{k-1}.$$

For $k \leq k_0$, the same argument as above yields

$$\mathbb{P}\left(|\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1}r^2\right) \leq 4e^{-K_w 2^k/512},$$

and, since $2^M \geq 512 + 128\log(16/\delta)$,

$$e^{K_w}\sum_{k=M}^{\infty} 4e^{-K_w 2^k/512} \leq 8e^{K_w - K_w 2^M/512} \leq 8e^{-(K_w/4)\log(16/\delta)} \leq 8e^{-\log(16/\delta)} = \delta/2.$$

For $k \geq k_0$, we use $\min\{2, C_{\mathcal{P}}C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1}2^{(k+3)/2}r\} \leq \min\{2, C_{\mathcal{P}}C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1}r\}2^{(k+3)/2}$ and (2) to upper-bound (5) by

$$4\exp\left(-\frac{K_w}{16}\frac{2^{2(k-1)}}{2^{(k+3)/2+k-1}}\right) \leq 4e^{-K_w 2^{k/2}/96}.$$

Using a similar geometric series argument, we obtain

$$e^{K_w}\sum_{k=M}^{\infty} 4e^{-K_w 2^{k/2}/96} \leq 16e^{K_w}e^{-K_w 2^{M/2}/96}.$$

Because $2^M \geq (96 + 24\log(34/\delta))^2$, we get

$$16e^{K_w}e^{-K_w 2^{M/2}/96} \leq 16e^{-(K_w/4)\log(34/\delta)} \leq 16e^{-\log(34/\delta)} = \delta/2.$$

Combining the two sub-cases, we have

$$\begin{aligned}
& e^{K_w} \sum_{k=M}^{\infty} \mathbb{P}\left(|\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1} r^2\right) \\
\leq \quad & e^{K_w} \sum_{k=M}^{\max\{M,k_0\}-1} \mathbb{P}\left(|\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1} r^2\right) \\
& + e^{K_w} \sum_{k=\max\{M,k_0\}}^{\infty} \mathbb{P}\left(|\tilde{\mathbb{G}}_n^{w,m_\beta}(L_h - L_{h'})| > 2^{k-1} r^2\right) \\
\leq\ & e^{K_w} \sum_{k=M}^{\infty} 4e^{-K_w 2^k/512} + e^{K_w} \sum_{k=M}^{\infty} 4e^{-K_w 2^{k/2}/96} \\
\leq\ & \delta,
\end{aligned}$$

This completes the proof.

□

B.2 Explicit rate functions

Proposition B.3. *Suppose $C_1, B_{\mathcal{W}} \lesssim 1$ and there is $\alpha \in [0, 2)$ such that*

$$\log N_\infty(\epsilon, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \lesssim \|w\|^{-\alpha} \log(n/\epsilon), \qquad \log N_1(\epsilon, \mathcal{W}) \lesssim \log(n/\epsilon). \tag{6}$$

(i) If $C_{\beta,\rho} = C_{\mathcal{P}}^2 K_\rho + m_\beta B_{\mathcal{W}} \leq n$, there is $A \geq 1$ such that condition (2) holds with

$$r(u) = u^{1-\alpha/2} \sqrt{A C_{\beta,\rho} \log n}.$$

(ii) If $C_{\mathcal{P}}^2 K_\rho \leq n$ and $m_\beta B_{\mathcal{W}} C_{\mathcal{P}} C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} \leq n$ for every $w \in \mathcal{W}$, there is $A \geq 1$ such that condition (2) holds for any increasing K-Lipschitz r satisfying, for every $w \in \mathcal{W}$,

$$r(\|w\|) \geq \|w\|^{1-\alpha/2} \sqrt{A C_{\mathcal{P}}^2 K_\rho \log n} + A m_\beta B_{\mathcal{W}} C_{\mathcal{P}} C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} \|w\|^{2-\alpha} \log n.$$

The complexity of $\mathcal{H}_w(\epsilon_{\mathcal{W}})$ may therefore adapt to $n_{\text{eff}}(w) = \|w\|^{-2}$ through α. The assumptions exclude the unrestricted simplex of all weights but cover the structured families in Appendix E.

Proof. Observe that $\|w\| \geq C_{\mathcal{W}} \geq n^{-1/2}$. Thus, there is a large enough constant A such that, choosing $K = A^2 n^2$, we have $K_w \leq A u^{-\alpha} \log n$. We will show that the rate in (i) and the sufficient bound in (ii) satisfy (2); the Lipschitz property in (ii) is imposed explicitly.

(i) Condition (2) is satisfied provided

$$r(u)^2 \geq A u^{2-\alpha} C_{\beta,\rho} \log n,$$

for some large enough constant $A < \infty$. The function $r(u) = u^{1-\alpha/2} \sqrt{A C_{\beta,\rho} \log n}$ satisfies this with equality and has Lipschitz constant bounded by

$$\begin{aligned}
\sup_{u \in [n^{-1/2}, C_1]} |r'(u)| &= (1 - \alpha/2) \sup_{u \in [n^{-1/2}, C_1]} u^{-\alpha/2} \sqrt{A C_{\beta,\rho} \log n} \\
&\leq n^{\alpha/4} (1 - \alpha/2) \sqrt{An \log n} \\
&\leq n^{3/4} \sqrt{A \log n} \\
&\leq An \leq A^2 n^2,
\end{aligned}$$

as desired.

(ii) Condition (2) is satisfied provided, for every $w \in \mathcal{W}$,

$$r(\|w\|)^2 \geq A \|w\|^{2-\alpha} \left(C_{\mathcal{P}}^2 K_\rho + m_\beta B_{\mathcal{W}} C_{\mathcal{P}} C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} r(\|w\|)\right) \log n.$$

The lower bound in (ii) implies this inequality directly.

□

B.3 Proof of the lower bound

Proof of Proposition 3.4. Partition the sample into K consecutive blocks I_k of size M and each I_k into B subblocks $J_{k,b}$ of size m. For $i \in J_{k,b}$, let $Y_i = \mu_k + \xi_{k,b}$, where $\xi_{k,b} \overset{\text{iid}}{\sim} N(0, I_d)$. Let Θ_{ns} contain the nonconstant mean sequences in $\mathcal{H}^K$, $\mathcal{H} = [-1,1]^d$, and set $\mathcal{W} = \{w^{(k)} : w_i^{(k)} = M^{-1}\mathbb{1}\{i \in I_k\}\}$. Under square loss, $\mu^*_{w^{(k)}} = \mu_k$ and the weighted excess risk is $\|\widehat{\mu}_{w^{(k)}} - \mu_k\|_2^2$. Moreover, $\|w^{(k)}\|_2^2 = M^{-1}$ and $n_{\text{eff}}(w^{(k)}) = M$, whereas repeated observations within a subblock carry exactly the information in one Gaussian vector. Thus each block provides only $B = M/m$ independent observations from $N(\mu_k, I_d)$.

A Gilbert–Varshamov packing supplies $c^0, \dots, c^J \in \{-1,1\}^d$ with $\log(J+1) \asymp d$ and pairwise Hamming distance at least $d/8$. Set

$$a = c_0\sqrt{\frac{\log(KJ)}{Bd}}, \qquad \nu^j = ac^j.$$

The condition $d + \log K \le Bd$ ensures that $\nu^j \in \mathcal{H}$ for a sufficiently small universal c_0.

Let $P^{(0)}$ assign mean ν^0 to every block. For $1 \le k \le K$ and $1 \le j \le J$, let $P^{(k,j)}$ assign mean ν^j to I_k and ν^0 to all other blocks. Every alternative has nonconstant means because $K \ge 2$, so it belongs to Θ_{ns}; $P^{(0)}$ is used only as a reference distribution. Only the B independent Gaussian values on I_k change, giving

$$\mathrm{KL}(P^{(k,j)}, P^{(0)}) = \frac{B}{2}\|\nu^j - \nu^0\|_2^2 \lesssim Ba^2d \lesssim c_0^2 \log(KJ).$$

The collections of weighted oracles for distinct alternatives are separated, in the supremum metric over $\mathcal{W}$, by a constant multiple of $a\sqrt{d}$. Fano's inequality therefore gives

$$\inf_{(\widehat{\mu}_w)} \sup_{\mu \in \Theta_{\text{ns}}} \mathbb{E}_\mu\left[\sup_{w \in \mathcal{W}} \|\widehat{\mu}_w - \mu^*_w\|_2^2\right] \gtrsim a^2 d \asymp \frac{d + \log K}{B}.$$

When KJ is bounded, the same conclusion follows by testing two alternatives: their Kullback–Leibler divergence is bounded by a constant multiple of c_0^2, while their squared separation remains of order $a^2 d$.

Finally, the weight vectors have pairwise ℓ_1 distance 2, so $N_1(1/2, \mathcal{W}) = K$, while $\log N_\infty(1/2, \mathcal{H}) \asymp d$. Substituting $B = M/m$ proves the stated bound. For this process, $\rho(\ell) = 1$ for $1 \le \ell < m$ and $\rho(\ell) = 0$ for $\ell \ge m$, hence $K_\rho = 2m - 1$. Together with $\|w\|_2^2 = 1/M$, this gives the claimed dependence scaling. Setting $m = 1$ in the construction and proof yields the independent nonstationary special case. □

B.4 Proof of Example 3.1

First, under square loss and for any distribution P on (Y, Z), let $h^*_P(z) = \mathbb{E}_P[Y \mid Z = z]$. Then, for every measurable h,

$$\begin{aligned}\mathbb{E}_P[(Y - h(Z))^2 - (Y - h^*_P(Z))^2] &= \mathbb{E}_P[(h(Z) - h^*_P(Z))^2] \\ &\quad + 2\,\mathbb{E}_P[(h^*_P(Z) - Y)(h(Z) - h^*_P(Z))] \\ &= \|h - h^*_P\|^2_{L_2(P)},\end{aligned}$$

since $\mathbb{E}_P[h^*_P(Z) - Y \mid Z] = 0$. Applying this with $P = \sum_{t=1}^n w_t P_t$ shows that (A4) holds with equality.

Scalar hypothesis class. Let $\mathcal{H} = \{h(z) = c : c \in [-B, B]\}$. For $h(z) = c$ and $h'(z) = c'$, $\|h - h'\|_{L_2(P)} = |c - c'|$ and $\|h - h'\|_\infty = |c - c'|$ for every $P \in \mathcal{P}$. Therefore

$$\sup_{P \in \mathcal{P}} \|h - h'\|_{L_2(P)} = \inf_{P \in \mathcal{P}} \|h - h'\|_{L_2(P)}, \qquad \inf_{P \in \mathcal{P}} \|h - h'\|_{L_2(P)} = \|h - h'\|_\infty,$$

so (A2) holds with $C_{\mathcal{P}} = 1$, and (A3) with $C_{w,\infty}(\epsilon) = 1$.

Linear feature class. Let $\mathcal{H} = \{h(z) = \beta^\top \Phi(z) : \|\beta\|_2 \le B\}$ and define $\Sigma_P = \mathbb{E}_P[\Phi(Z)\Phi(Z)^\top]$. For $h(z) = \beta^\top \Phi(z)$ and $h'(z) = \beta'^\top \Phi(z)$, writing $\Delta = \beta - \beta'$, we have $\|h - h'\|^2_{L_2(P)} = \Delta^\top \Sigma_P \Delta$. If the Gram matrices are uniformly well-conditioned, i.e.,

$$0 < \underline{\lambda} \le \lambda_{\min}(\Sigma_P) \le \lambda_{\max}(\Sigma_P) \le \bar{\lambda} < \infty \qquad \text{for all } P \in \mathcal{P},$$

then

$$\underline{\lambda}\|\Delta\|_2^2 \le \|h - h'\|^2_{L_2(P)} \le \bar{\lambda}\|\Delta\|_2^2.$$

Taking the supremum and infimum over $P \in \mathcal{P}$ yields

$$\sup_{P\in\mathcal{P}} \|h - h'\|_{L_2(P)} \le \sqrt{\bar{\lambda}/\underline{\lambda}} \inf_{P\in\mathcal{P}} \|h - h'\|_{L_2(P)}.$$

Hence (A2) holds with $C_{\mathcal{P}} = \sqrt{\bar{\lambda}/\underline{\lambda}}$. If moreover $\sup_{z\in\mathcal{Z}} \|\Phi(z)\|_2 \le 1$, then

$$\|h - h'\|_\infty = \sup_{z\in\mathcal{Z}} |\Delta^\top \Phi(z)| \le \|\Delta\|_2,$$

and therefore it is easy to see that (A3) holds with $C_{w,\infty}(\epsilon) = \sqrt{\underline{\lambda}}$:

$$\inf_{P\in\mathcal{P}} \|h - h'\|_{L_2(P)} \ge \sqrt{\underline{\lambda}}\|\Delta\|_2 \ge \sqrt{\underline{\lambda}}\|h - h'\|_\infty.$$

B.5 Proof of Lemma 4.1

Write $p_i = p_{Z_i}$ and $h_i^* = h^*_{P_i}$. Under square loss, the oracle under the weighted past distribution is given pointwise by the conditional mean under the weighted mixture. Hence, for all z such that $\sum_{j=1}^t w_j p_j(z) > 0$,

$$h_w^*(z) = \frac{\sum_{i=1}^t w_i p_i(z) h_i^*(z)}{\sum_{j=1}^t w_j p_j(z)}.$$

For ease of notation, we omit z in the following. By the triangle inequality,

$$\begin{aligned}
\|h_w^* - h^*_{P_{t+1}}\|_{L_2(P_{t+1})} &= \left\| \sum_{i=1}^t \frac{w_i p_i}{\sum_{j=1}^t w_j p_j} \big(h_i^* - h^*_{P_{t+1}}\big) \right\|_{L_2(P_{t+1})} \\
&\le \sum_{i=1}^t \left\| \frac{w_i p_i}{\sum_{j=1}^t w_j p_j} \big(h_i^* - h^*_{P_{t+1}}\big) \right\|_{L_2(P_{t+1})} \\
&\le \sum_{i=1}^t \Big\| C_p w_i \big(h_i^* - h^*_{P_{t+1}}\big) \Big\|_{L_2(P_{t+1})} \\
&= C_p \sum_{i=1}^t w_i \|h_i^* - h^*_{P_{t+1}}\|_{L_2(P_{t+1})}.
\end{aligned}$$

B.6 Proof of Example 4.2

For exponential weights, set $\rho = \exp(-\theta)$ and $k = t - i$. Then

$$\sum_{i=1}^t w_i (t + 1 - i)^\nu = \frac{1-\rho}{1-\rho^t} \sum_{k=0}^{t-1} \rho^k (k+1)^\nu.$$

Since the weights sum to one and $t + 1 - i \le t$, we have $\sum_{i=1}^t w_i (t+1-i)^\nu \le t^\nu$. If $t(1-\rho) \ge 1$, then $1 - \rho^t$ is bounded away from zero, and hence

$$\sum_{i=1}^t w_i (t + 1 - i)^\nu = \frac{1-\rho}{1-\rho^t} \sum_{k=0}^{t-1} \rho^k (k+1)^\nu \lesssim (1-\rho) \sum_{k=0}^{\infty} \rho^k (k+1)^\nu \lesssim (1-\rho)^{-\nu}.$$

Thus, in all cases,

$$\sum_{i=1}^t w_i (t + 1 - i)^\nu \lesssim \min\{t, (1-\rho)^{-1}\}^\nu.$$

Moreover, the computation in Lemma E.4 gives:

$$\|w^{(t)}(\theta)\|^{-2} = \frac{1+\rho}{1-\rho} \cdot \frac{1-\rho^t}{1+\rho^t} \geq \frac{1-\rho^t}{2(1-\rho)} \gtrsim \min\{t, (1-\rho)^{-1}\}.$$

Furthermore, $\|w^{(t)}(\theta)\|^{-2} \leq \frac{2}{1-\rho}$, and consequently,

$$\sum_{i=1}^{t} w_i (t+1-i)^{\nu} \lesssim \|w^{(t)}(\theta)\|^{-2\nu} \lesssim (1-\rho)^{-\nu} = (1-e^{-\theta})^{-\nu}.$$

Combining this with Lemma 4.1 yields

$$\|h^*_w - h^*_{P_{t+1}}\|_{L_2(P_{t+1})} \lesssim C_p \kappa \|w^{(t)}(\theta)\|^{-2\nu} \lesssim C_p \kappa (1-e^{-\theta})^{-\nu}.$$

B.7 Time-uniform Bound

Let $n_0 \in \{1, \dots, n\}$ and suppose that at every time $t = n_0, \dots, n$, we pick a weight $w^{(t)} \in \mathcal{W}$ such that $w_i^{(t)} = 0$ for all $i > t$ and compute the weighted ERM $\hat{h}_{w^{(t)}} = \arg\min_{h \in \mathcal{H}_{w^{(t)}}} R_n^{w^{(t)}}(h)$.

Theorem B.4. *Under the conditions of Theorem 3.2, with probability at least $1-2\delta$, every sequence $w^{(n_0)}, \dots, w^{(n)} \in \mathcal{W}$ satisfies*

$$\sum_{t=n_0}^{n} \mathbb{E}[L(X_{t+1}, \hat{h}_{w^{(t)}}) - L(X_{t+1}, h^*_{P_{t+1}})] \lesssim \sum_{t=n_0}^{n} r(\|w^{(t)}\|)^2 \log^2(1/\delta) + \sum_{t=n_0}^{n} \|h^*_{w^{(t)}} - h^*_{P_{t+1}}\|^2_{L_2(P_{t+1})}.$$

Proof. By Theorem 3.2, with probability at least $1-2\delta$,

$$\sup_{w \in \mathcal{W}} r(\|w\|)^{-2} \|\hat{h}_w - h^*_w\|^2_{L_2(\mathcal{P})} \lesssim \log^2(1/\delta). \tag{7}$$

Thus, on the same event,

$$\begin{aligned}
&\sum_{t=n_0}^{n} \left\{ \mathbb{E}[L(X_{t+1}, \hat{h}_{w^{(t)}})] - \mathbb{E}[L(X_{t+1}, h^*_{P_{t+1}})] \right\} \\
&\lesssim \sum_{t=n_0}^{n} \left\{ \|\hat{h}_{w^{(t)}} - h^*_{w^{(t)}}\|^2_{L_2(P_{t+1})} + \|h^*_{w^{(t)}} - h^*_{P_{t+1}}\|^2_{L_2(P_{t+1})} \right\} \\
&\lesssim \sum_{t=n_0}^{n} \left\{ r(\|w^{(t)}\|)^2 \log^2(1/\delta) + \|h^*_{w^{(t)}} - h^*_{P_{t+1}}\|^2_{L_2(P_{t+1})} \right\},
\end{aligned}$$

where we used (1) in the first inequality and (7) in the second. □

C Derivations for the Applications

C.1 Linear Models

Here $\mathcal{H}_w$ is fixed, so $\mathcal{H}_w(\epsilon_{\mathcal{W}}) = \mathcal{H}$.

For any $h, h' \in \mathcal{H}$ with $h(z) = \beta^\top z$ and $h'(z) = \beta'^\top z$, we have

$$\inf_{P \in \mathcal{P}} \|h - h'\|^2_{L^2(P)} = \inf_{P \in \mathcal{P}} (\beta - \beta')^\top \mathbb{E}_P[ZZ^\top](\beta - \beta') \geq \inf_{P \in \mathcal{P}} \lambda_{\min}(\mathbb{E}_P[ZZ^\top]) \|\beta - \beta'\|^2,$$

and, by the Cauchy–Schwarz inequality,

$$\|\beta - \beta'\|^2 \geq \|\beta - \beta'\|^2 \sup_{z \in \mathcal{Z}} \|z\|^2 \geq \sup_{z \in \mathcal{Z}} |\beta^\top z - \beta'^\top z|^2.$$

Therefore, (A3) is satisfied with $C_{w,\infty}(\epsilon_{\mathcal{W}}) = \inf_{P\in\mathcal{P}} \sqrt{\lambda_{\min}(\mathbb{E}_P[ZZ^\top])}$. For the covering number, apply Lemma E.1 to obtain

$$\log N_\infty(\epsilon, \mathcal{H}) \le \log\left(\frac{3B}{\epsilon}\right)^p \lesssim p\log(1/\epsilon).$$

Thus, (6) holds with $\alpha = 0$. Applying Proposition B.3 (ii) gives

$$r(u) = u\sqrt{AC_{\mathcal{P}}^2 K_\rho p \log n} + AC_{\beta,\infty}u^2 p\log n,$$

where $C_{\beta,\infty} = m_\beta B_{\mathcal{W}} C_{\mathcal{P}} C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1}$. Since $C_{\mathcal{P}}, B_{\mathcal{W}}, K_\rho \lesssim 1$ by assumption, we have

$$r(u)^2 \lesssim pu^2\log n + p^2 m_\beta^2 C_{w,\infty}(\epsilon_{\mathcal{W}})^{-2}u^4\log^2 n.$$

Combining the preceding display with Theorem 3.2 and setting $u = \|w\|$ gives the final rate:

$$\|\hat{h}_w - h_w^*\|_{L_2(\mathcal{P})}^2 \lesssim p\|w\|^2\log n + p^2\|w\|^4 m_\beta^2 C_{w,\infty}(\epsilon_{\mathcal{W}})^{-2}\log^2 n.$$

For unweighted ERM, we have $\|w\|^2 = 1/n$. For any fixed p,

$$\|\hat{h}_w - h_w^*\|_{L_2(\mathcal{P})}^2 \lesssim \frac{p\log n}{n} + \frac{p^2 m_\beta^2\log^2 n}{n^2} \lesssim \frac{p\log n}{n} + \frac{p^2\log^4 n}{n^2} \lesssim \frac{p\log n}{n}.$$

C.2 Basis expansions

We consider the construction from Section 5. Take the dyadically rounded dimension

$$q(w) = 2^{\lceil \log_2 \lceil \|w\|^{-2/3}\rceil\rceil},$$

so $q(w) \asymp \|w\|^{-2/3}$ and the corresponding equal-width partitions are nested. Define $\phi_{j,w}(z) = \mathbb{1}\{z \in [\frac{j-1}{q(w)}, \frac{j}{q(w)})\}, j = 1,\dots,q(w)$. By 1-Lipschitzness,

$$\inf_{h\in\mathcal{H}_w}\|h - h_w^*\|_\infty \le \sup_{j=1,\dots,q(w)}\ \sup_{z\in[\frac{j-1}{q(w)},\frac{j}{q(w)})} |h_w^*((j-1)/q(w)) - h_w^*(z)| \le q(w)^{-1}.$$

Thus, the squared approximation error is bounded by $q(w)^{-2}$. If $\|v - w\|_1, \|v' - w\|_1 \le \epsilon_{\mathcal{W}}$, then $|\|v\| - \|w\||, |\|v'\| - \|w\|| \le \epsilon_{\mathcal{W}} \le \|w\|/64$, and hence $q(v), q(v') \asymp q(w)$. For $h \in \mathcal{H}_v$ and $h' \in \mathcal{H}_{v'}$, their difference is piecewise constant on the common refinement. Because the dyadic partitions are nested, this refinement has $q_{\max} = \max\{q(v), q(v')\} \asymp q(w)$ equal-width intervals. Under the uniform design,

$$\|h - h'\|_{L_2(P)}^2 = \frac{1}{q_{\max}}\sum_{j=1}^{q_{\max}}(h_j - h'_j)^2 \ge \frac{1}{q_{\max}}\|h - h'\|_\infty^2 \gtrsim q(w)^{-1}\|h - h'\|_\infty^2,$$

so $C_{w,\infty}(\epsilon_{\mathcal{W}}) \asymp q(w)^{-1/2}$. By Lemma E.1,

$$\log N_\infty(\epsilon, \mathcal{H}_w) \le \log\left(\frac{3B}{\epsilon}\right)^{q(w)} \lesssim q(w)\log(1/\epsilon).$$

The union bound $N_\infty(\epsilon, \bigcup_j \mathcal{H}_j) \le \sum_j N_\infty(\epsilon, \mathcal{H}_j)$ over the $O(1)$ dyadic dimensions in the local neighborhood therefore gives $\log N_\infty(\epsilon, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \lesssim q(w)\log(n/\epsilon)$. Then $\log N_\infty(\epsilon, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \lesssim \|w\|^{-2/3}\log(n/\epsilon)$, so (6) holds with $\alpha = 2/3$. Applying Proposition B.3 (ii) with $C_{w,\infty}(\epsilon_{\mathcal{W}})^{-1} \lesssim q(w)^{1/2} \lesssim \|w\|^{-1/3}$ and $u = \|w\|$,

$$r(\|w\|)^2 \lesssim \|w\|^{4/3}\log n + \|w\|^2 m_\beta^2\log^2 n.$$

Combining this with $q(w)^{-2} \lesssim \|w\|^{4/3}$ and Theorem 3.2 yields

$$\|\hat{h}_w - h_w^*\|_{L_2(\mathcal{P})}^2 \lesssim \|w\|^{4/3}\log n + \|w\|^2 m_\beta^2\log^2 n.$$

For unweighted ERM, $\|w\|^2 = 1/n$ and thus

$$\|\hat{h}_w - h_w^*\|_{L_2(\mathcal{P})}^2 \lesssim \frac{\log n}{n^{2/3}} + \frac{m_\beta^2\log^2 n}{n} \lesssim \frac{\log n}{n^{2/3}} + \frac{\log^4 n}{n} \lesssim \frac{\log n}{n^{2/3}}.$$

C.3 Neural networks

Theorem 3.1 of Nagler & Langer (2026) shows that for $b_w \gtrsim \nu_w$,

$$\inf_{h \in \mathcal{H}_w} \|h - h_w^*\|_\infty \lesssim (\nu_w \ell_w / \sqrt{\log \ell_w})^{-2a/p}.$$

Further, Theorem 2.1 of Ou & Bölcskei (2024) yields

$$\log N_\infty(\epsilon, \mathcal{H}_w) \lesssim \nu_w^2 \ell_w^2 \log(\nu_w b_w / \epsilon).$$

Noting $\log(\nu_w b_w) \lesssim \log n$, this simplifies to

$$\inf_{h \in \mathcal{H}_w} \|h - h_w^*\|_\infty \lesssim \|w\|^{2a/(2a+p)} \log^{a/p}(n), \quad \log N_\infty(\epsilon, \mathcal{H}_w) \lesssim \|w\|^{-2p/(2a+p)} \log(n/\epsilon).$$

For $\|v - w\|_1 \le \epsilon_{\mathcal{W}}$, the architecture sizes $\nu_v \ell_v$ and $\nu_w \ell_w$ are comparable. Since ν_v and ℓ_v are positive integers of polynomially bounded size, at most $n^{O(1)}$ distinct architecture pairs occur in this local neighborhood. Taking the union of the covers over these distinct architectures therefore adds only $O(\log n)$ to the logarithm of the covering number, which is absorbed by the displayed bound; hence $\log N_\infty(\epsilon, \mathcal{H}_w(\epsilon_{\mathcal{W}})) \lesssim \|w\|^{-2p/(2a+p)} \log(n/\epsilon)$. Use bounded, clipped outputs and $C_{w,\infty}(\epsilon_{\mathcal{W}}) = 0$. Proposition B.3 (i) and Theorem 3.2 yield

$$\|\hat{h}_w - h_w^*\|_{L_2(\mathcal{P})}^2 \lesssim \|w\|^{4a/(2a+p)} \log^{2a/p}(n) + \|w\|^{2-2p/(2a+p)} \log^2 n \lesssim \|w\|^{4a/(2a+p)} \log^\zeta(n),$$

with $\zeta = \max\{2a/p, 2\}$.

C.4 Structural breaks

Let $w(s)$ denote the uniform weights on the most recent s observations. Let τ_k be the most recent break point, so the current regime has age $d = t - \tau_k$, and suppose the window intersects at most the current and the preceding regime. Define the break size by

$$\Delta = \|h^*_{P_{\tau_k}} - h^*_{P_{t+1}}\|_{L_2(P_{t+1})}.$$

Exactly $(s-d)_+$ observations in the window come from before the break. Lemma 4.1 therefore gives

$$\|h^*_{w(s)} - h^*_{P_{t+1}}\|_{L_2(P_{t+1})} \le C_p \Delta \frac{(s-d)_+}{s} = C_p \Delta \left(1 - \frac{d}{s}\right)_+.$$

Since $\|w(s)\|_2^2 = 1/s$, the leading learning terms derived in Appendices C.1–C.3 are cs^{-b}, where

$$(c, b) = (p \log n, 1), \qquad (c, b) = (\log n, 2/3), \qquad (c, b) = (\log^\zeta n, \varrho)$$

for linear models, basis expansions, and ReLU networks, respectively, with $\varrho = 2a/(2a+p)$. Corollary 5.1 thus yields the direct bound

$$\mathbb{E}[L(X_{t+1}, \hat{h}_{w(s)})] - \mathbb{E}[L(X_{t+1}, h^*_{P_{t+1}})] \lesssim \frac{c}{s^b} + C_p^2 \Delta^2 \left(1 - \frac{d}{s}\right)_+^2. \tag{8}$$

For $s \le d$, the drift term vanishes and the learning term decreases with s. Hence $s = d$ is the best window that uses only observations from the current regime. If $C_p \Delta = 0$, the bound is decreasing on the full range and $s = n$ is optimal. Otherwise, to see when older observations improve the bound, consider $s \ge d$ and write

$$\epsilon = 1 - \frac{d}{s}, \qquad s = \frac{d}{1-\epsilon}, \qquad 0 \le \epsilon \le 1 - \frac{d}{n}.$$

The right-hand side of (8) becomes

$$\frac{c}{d^b}(1-\epsilon)^b + C_p^2 \Delta^2 \epsilon^2.$$

Because $b \in (0, 1]$, concavity gives $(1-\epsilon)^b \le 1 - b\epsilon$. It follows that

$$\frac{c}{d^b}(1-\epsilon)^b + C_p^2 \Delta^2 \epsilon^2 \le \frac{c}{d^b} - \frac{bc}{d^b}\epsilon + C_p^2 \Delta^2 \epsilon^2.$$

The quadratic on the right is minimized at $\epsilon_0 = bc/(2C_p^2\Delta^2 d^b)$. Accounting for the largest available window, its feasible optimizer is

$$\epsilon^* = \min\left\{\frac{bc}{2C_p^2\Delta^2 d^b}, 1 - \frac{d}{n}\right\}, \qquad s^* = \frac{d}{1-\epsilon^*}.$$

If $bc/(2C_p^2\Delta^2 d^b) \le 1 - d/n$, substitution into the upper bound gives

$$\mathbb{E}[L(X_{t+1}, \hat{h}_{w(s^*)})] - \mathbb{E}[L(X_{t+1}, h^*_{P_{t+1}})] \lesssim \frac{c}{d^b}\left(1 - \frac{b^2 c}{4C_p^2\Delta^2 d^b}\right).$$

Otherwise the optimizer of the upper bound is attained at the boundary $s^* = n$, where the generic bound equals

$$cn^{-b} + C_p^2\Delta^2\left(1 - \frac{d}{n}\right)^2.$$

Substituting the three pairs (c, b) gives the structural-break column of Table 1.

D Experimental setup

All three panels of Figure 1 use the scalar location model

$$Y_t = \mu_t + \eta_t, \qquad \hat{\mu}_w = \sum_{i=1}^n w_i Y_i,$$

where the mean path is deterministic and independent of the noise. The target is the next mean μ_{n+1}.

Panel specifications.

- *Panel (a): learning rates.* We use uniform weights on all n observations, set $\mu_t = 0$, and compare $\mathbb{E}[(n^{-1}\sum_{i=1}^n \eta_i)^2]$ under iid and dependent noise as n increases.
- *Panel (b): structural break.* We set $n = 2048$ and

  $$\mu_t = \begin{cases} -0.05, & t \le n - 256, \\ 0.05, & t > n - 256, \end{cases} \qquad \mu_{n+1} = 0.05.$$

  The estimator averages the most recent s observations, and the horizontal axis reports the window length s.
- *Panel (c): smooth drift.* We again set $n = 2048$, now with $\mu_t = 0.5t/(n+1)$ for $t = 1, \dots, n+1$. The normalized exponential weights satisfy

  $$w_i(\theta) \propto \exp[-\theta(n-i)],$$

  and the horizontal axis gives their decay horizon $1/\theta$.

Noise processes. For the dependent process, let the consecutive block lengths be independent copies of

$$L = 5\lfloor U^{-1/\gamma}\rfloor, \qquad U \sim \mathrm{Unif}(0,1), \qquad \gamma = 3.5.$$

Thus $\Pr(L \ge 5\ell) = \ell^{-\gamma}$. Each block receives an independent Rademacher sign, and η_t is held fixed at 0.35 times that sign throughout the block. We initialize the process in equilibrium: the block containing the first observation is drawn from the length-biased distribution

$$\Pr(L_0 = l) = \frac{l\Pr(L = l)}{\mathbb{E}[L]},$$

and its starting position is uniform within the block. For numerical sampling, the support of $L_0/5$ is truncated at 10^5; all subsequent blocks follow the ordinary length distribution. The resulting autocovariance is

$$\mathrm{Cov}(\eta_t, \eta_{t+k}) = 0.35^2\frac{\mathbb{E}[(L-k)_+]}{\mathbb{E}[L]},$$

which decays as $k^{-(\gamma-1)}$ and is summable. In the iid benchmark, the Rademacher sign is instead redrawn independently at every time point.

Risk curves. For either weight family, the expected excess risk separates exactly into

$$\underbrace{\sum_{i,j=1}^{n} w_i w_j \operatorname{Cov}(\eta_i, \eta_j)}_{\text{learning error}} + \underbrace{\left(\sum_{i=1}^{n} w_i \mu_i - \mu_{n+1}\right)^2}_{\text{drift error}}.$$

We compute both terms from the known mean path and autocovariance. We also validate the resulting excess-risk curves using 1,200 independent replications. The vertical dotted lines indicate the minimizers of the computed curves. Panels (b)–(c) therefore display directly how a longer averaging horizon reduces the learning error while increasing the drift error.

E WEIGHT FAMILIES

Lemma E.1. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and $S = \{x \in \mathbb{R}^d : \|x\| \le r\}$. Then*

$$N(\epsilon, S, \|\cdot\|) \le \left(1 + \frac{2r}{\epsilon}\right)^d.$$

Proof. The result follows from Vershynin (2018), Proposition 4.2.10. □

In the following, we consider n-dimensional weight vectors $w^{(t)} \in \mathbb{R}^n$, for $t \in \{1, \dots, n\}$, whose entries satisfy $w_i \in \mathbb{R}$ if $i \le t$ and $w_i = 0$ if $i > t$.

E.1 UNIFORM WINDOW WEIGHTS

Lemma E.2. *Let $s \in \{1, \dots, t\}$ and $t \in \{1, \dots, n\}$. Define the weight vector $w^{(t)}(s) \in \mathbb{R}^n$ by*

$$w_i^{(t)}(s) = \mathbb{1}\{i \le t\} \times \begin{cases} 1/s, & i \in \{t-s+1, \dots, t\}, \\ 0, & i \in \{1, \dots, t-s\}. \end{cases}$$

Let $\mathcal{W}_t = \{w^{(t)}(s) : s = 1, \dots, t\}$. Then

$$B_{\mathcal{W}} = 1, \qquad C_1 = 1, \qquad N_1(\epsilon, \mathcal{W}_t) \le t, \qquad N_1\left(\epsilon, \bigcup_{t=1}^{n} \mathcal{W}_t\right) \le \frac{n(n+1)}{2}.$$

Proof. Fix $t \in \{1, \dots, n\}$ and $s \in \{1, \dots, t\}$. For simplicity, we abbreviate $w = w^{(t)}(s) \in \mathcal{W}_t$ in the following. By construction, w has exactly s nonzero entries, each equal to $1/s$. Hence $\|w\|_\infty = 1/s$. Moreover,

$$\|w\|^2 = \sum_{i=1}^{t} w_i^2 = s \cdot \left(\frac{1}{s}\right)^2 = \frac{1}{s}.$$

The ratio is then $\|w\|_\infty / \|w\|^2 = 1$, which is independent of t and s. Therefore $B_{\mathcal{W}} = 1$. The second claim follows from $\|w\|_1 = s(1/s) = 1$, so $C_1 = 1$. For the covering number, $|\mathcal{W}_t| = t$, and thus we can cover the entire space with at most t balls for any $\epsilon > 0$, i.e., $N_1(\epsilon, \mathcal{W}_t) \le t$. The final statement follows from $|\bigcup_{t=1}^{n} \mathcal{W}_t| \le \sum_{t=1}^{n} |\mathcal{W}_t| = \sum_{t=1}^{n} t = n(n+1)/2$. □

E.2 EXPONENTIAL WEIGHTS

Lemma E.3. *Fix $t \in \{1, \dots, n\}$. For $\theta \in (0, \infty)$ define the weight vector $w^{(t)}(\theta) \in \mathbb{R}^n$, where for $i = 1, \dots, t$:*

$$w_i^{(t)}(\theta) = \mathbb{1}\{i \le t\} \frac{\exp(-\theta(t-i))}{\sum_{j=1}^{t} \exp(-\theta(t-j))}. \tag{9}$$

Let $\Theta = (0, R)$ with $R < \infty$ and define the class $\mathcal{W}_t(\Theta) = \{w^{(t)}(\theta) : \theta \in \Theta\}$. Then for every $\epsilon > 0$,

$$N_1(\epsilon, \mathcal{W}_t(\Theta)) \le \frac{3R(t-1)}{\epsilon} \quad (t \ge 2;\ t = 1 \text{ is trivial}), \qquad N_1\left(\epsilon, \bigcup_{t=1}^{n} \mathcal{W}_t(\Theta)\right) \le \frac{3Rn^2}{2\epsilon}.$$

Proof. Fix $t \in \{1, \dots, n\}$ and $\theta \in (0, \infty)$. We again abbreviate $w(\theta) = w^{(t)}(\theta)$ in what follows. Let $x_i = (t - i)$ for $i = 1, \dots, t$ and notice $x_i \in [0, (t-1)]$. Differentiating gives

$$\begin{aligned}
\frac{d}{d\theta} w_i(\theta) &= \frac{-x_i e^{-\theta x_i}}{\sum_{j=1}^t e^{-\theta x_j}} - \frac{e^{-\theta x_i}}{(\sum_{j=1}^t e^{-\theta x_j})^2} \left(- \sum_{j=1}^t x_j e^{-\theta x_j} \right) \\
&= \frac{-x_i e^{-\theta x_i}}{\sum_{j=1}^t e^{-\theta x_j}} - \frac{e^{-\theta x_i}}{(\sum_{j=1}^t e^{-\theta x_j})^2} \left(- \left(\sum_{j=1}^t e^{-\theta x_j} \right) \sum_{j=1}^t w_j(\theta) x_j \right) \\
&= -w_i(\theta) x_i + \frac{e^{-\theta x_i}}{\sum_{j=1}^t e^{-\theta x_j}} \sum_{j=1}^t w_j(\theta) x_j \\
&= -w_i(\theta) x_i + w_i(\theta) \sum_{j=1}^t w_j(\theta) x_j \\
&= w_i(\theta) \left(\sum_{j=1}^t w_j(\theta) x_j - x_i \right).
\end{aligned}$$

Now bound the $\|\cdot\|_1$-norm of the derivative:

$$\begin{aligned}
\left\| \frac{d}{d\theta} w(\theta) \right\|_1 = \sum_{i=1}^t \left| \frac{d}{d\theta} w_i(\theta) \right| &= \sum_{i=1}^t w_i(\theta) \left| \sum_{j=1}^t w_j(\theta) x_j - x_i \right| \\
&\le (t-1) \sum_{i=1}^t w_i(\theta) \\
&\le (t-1).
\end{aligned}$$

Fix $\theta, \theta' \in \Theta$. By the mean value theorem, we then have

$$\|w(\theta) - w(\theta')\|_1 \le (t-1)|\theta - \theta'|.$$

Set $\eta = \epsilon/(t-1)$ and let $\{\theta_1, \dots, \theta_N\}$ be an η-net of Θ with size equal to $N_1(\eta, \Theta)$, the covering number with respect to $|\cdot|$. For any $\theta \in \Theta$ choose θ_k with $|\theta - \theta_k| \le \eta$. This implies $\|w(\theta) - w(\theta_k)\|_1 \le (t-1)\eta = \epsilon$. Hence $\{w(\theta_1), \dots, w(\theta_N)\}$ is an ϵ-net of $\mathcal{W}_t(\Theta)$ in $\|\cdot\|_1$, and therefore $N_1(\epsilon, \mathcal{W}_t(\Theta)) \le N_1\left(\epsilon/(t-1), \Theta\right)$. Applying Lemma E.1 with $d = 1$ and radius R gives

$$N_1(\eta, \Theta) \le \frac{3R}{\eta} = \frac{3R(t-1)}{\epsilon},$$

and the final statement follows from

$$N_1(\epsilon, \bigcup_{t=1}^{n} \mathcal{W}_t(\Theta)) \le \sum_{t=1}^{n} \frac{3R(t-1)}{\epsilon} \le \frac{3Rn^2}{2\epsilon}.$$

□

Lemma E.4. *Let $\theta \in (0, \infty)$. For the exponential weights in (9),*

$$B_{\mathcal{W}(\Theta)} \le 2, \qquad C_1 = 1.$$

Proof. Define $\rho = \exp(-\theta)$. We can rewrite the exponential weights for $k = t - i \in \{0, \dots, t-1\}$ as

$$w_i^{(t)}(\theta) = \frac{\rho^{t-i}}{\sum_{j=1}^{t} \rho^{t-j}} = \frac{\rho^{t-i}}{\sum_{k=0}^{t-1} \rho^k} = \frac{(1-\rho)\rho^{t-i}}{1-\rho^t}.$$

Now compute

$$\|w^{(t)}(\theta)\|^2 = \sum_{i=1}^{t} (w_i^{(t)})^2(\theta) = \sum_{k=0}^{t-1} \left(\frac{(1-\rho)\rho^k}{1-\rho^t} \right)^2 = \frac{(1-\rho)^2}{(1-\rho^t)^2} \sum_{k=0}^{t-1} \rho^{2k} = \frac{(1-\rho)^2(1-\rho^{2t})}{(1-\rho^t)^2(1-\rho^2)}.$$

Because the maximum is attained at $i = t$, we also have

$$\|w^{(t)}(\theta)\|_\infty = \frac{1}{\sum_{k=0}^{t-1} \rho^k} = \frac{1-\rho}{1-\rho^t}.$$

The ratio is then

$$\frac{\|w^{(t)}(\theta)\|_\infty}{\|w^{(t)}(\theta)\|^2} = \frac{1-\rho}{1-\rho^t} \cdot \frac{(1-\rho^t)^2(1-\rho^2)}{(1-\rho)^2(1-\rho^{2t})} = \frac{(1-\rho^t)(1-\rho^2)}{(1-\rho)(1-\rho^{2t})} = \frac{1+\rho}{1+\rho^t} \le 1 + \rho \le 2,$$

which holds for any $\theta \in (0, \infty)$. The second claim follows by noting that $w^{(t)}(\theta)$ is a probability vector with nonnegative entries for any $t \in \{1, \dots, n\}$. □

E.3 Brown double exponential smoothing weights

Lemma E.5. *Fix $t \in \{1, \dots, n\}$ and let $\Theta = (0, 1)$. For $\theta \in \Theta$, define the weight vector $w^{(t)}(\theta) \in \mathbb{R}^n$ by*

$$w_i^{(t)}(\theta) = \mathbb{1}\{i \le t\} \frac{\theta[2 - \theta(t-i+1)](1-\theta)^{t-i}}{\sum_{j=1}^{t} \theta[2 - \theta(t-j+1)](1-\theta)^{t-j}}. \tag{10}$$

Let $\mathcal{W}_t(\Theta) = \{w^{(t)}(\theta) : \theta \in \Theta\}$. Then for every $\epsilon > 0$,

$$N_1(\epsilon, \mathcal{W}_t(\Theta)) \le \frac{60t}{\epsilon}, \qquad N_1\left(\epsilon, \bigcup_{t=1}^{n} \mathcal{W}_t(\Theta)\right) \le \frac{60n^2}{\epsilon}.$$

Proof. Fix $t \in \{1, \dots, n\}$, $\theta \in (0, 1)$ and write $k = t - i \in \{0, \dots, t-1\}$ and $r = 1 - \theta \in (0, 1)$. We again abbreviate $w(\theta) = w^{(t)}(\theta)$. Observe that

$$w_i(\theta) = w_{t-k}(\theta) = \frac{\theta[2 - \theta(k+1)]r^k}{\sum_{j=0}^{t-1} \theta[2 - \theta(j+1)]r^j} = \frac{[2 - \theta(k+1)]r^k}{\sum_{j=0}^{t-1} [2 - \theta(j+1)]r^j} = \frac{b_k(\theta)}{\sum_{j=0}^{t-1} b_j(\theta)}.$$

Differentiating $w_{t-k}(\theta)$ gives

$$\frac{d}{d\theta} w_{t-k}(\theta) = \frac{b_k'(\theta)}{\sum_{j=0}^{t-1} b_j(\theta)} - \frac{b_k(\theta)(\sum_{j=0}^{t-1} b_j'(\theta))}{(\sum_{j=0}^{t-1} b_j(\theta))^2}.$$

Therefore, by the triangle inequality,

$$\begin{aligned} \left\| \frac{d}{d\theta} w(\theta) \right\|_1 &\le \frac{\sum_{k=0}^{t-1} |b_k'(\theta)|}{\sum_{j=0}^{t-1} b_j(\theta)} + \frac{|\sum_{j=0}^{t-1} b_j'(\theta)|}{(\sum_{j=0}^{t-1} b_j(\theta))^2} \sum_{k=0}^{t-1} |b_k(\theta)| \\ &\le \frac{\sum_{k=0}^{t-1} |b_k'(\theta)|}{\sum_{j=0}^{t-1} b_j(\theta)} \left(1 + \frac{\sum_{k=0}^{t-1} |b_k(\theta)|}{\sum_{j=0}^{t-1} b_j(\theta)} \right). \end{aligned}$$

We start by bounding $\sum_{k=0}^{t-1} |b_k|$. Using geometric-sum identities, we obtain

$$\begin{aligned} \sum_{j=0}^{t-1} b_j(\theta) &= \sum_{j=0}^{t-1} \big(2 - \theta(j+1)\big) r^j = 2 \sum_{j=0}^{t-1} r^j - \theta \sum_{j=0}^{t-1} (j+1) r^j \\ &= \frac{2(1-r^t)}{\theta} - \frac{1 - (t+1)r^t + t r^{t+1}}{\theta} = \frac{1 + r^t(t\theta - 1)}{\theta}. \end{aligned}$$

Hence $\sum_{j=0}^{t-1} b_j(\theta) > 0$ for all $\theta \in (0,1)$ and, in particular,

$$\sum_{j=0}^{t-1} b_j(\theta) \geq \begin{cases} t, & \theta \leq 1/t, \\ 1/\theta, & \theta > 1/t. \end{cases}$$

Indeed, if $\theta \leq 1/t$ then $t\theta - 1 \leq 0$, $r^t(t\theta - 1) \geq (t\theta - 1)$ and thus $1 + r^t(t\theta - 1) \geq t\theta$, so $\sum_{j=0}^{t-1} b_j(\theta) \geq t$. If $\theta > 1/t$ then $t\theta - 1 > 0$ and $1 + r^t(t\theta - 1) \geq 1$, so $\sum_{j=0}^{t-1} b_j(\theta) \geq 1/\theta$. Using $|2 - \theta(k+1)| \leq 2 + \theta(k+1)$,

$$|b_k(\theta)| \leq (2 + \theta(k+1))r^k = 2r^k + \theta(k+1)r^k.$$

Summing and applying the same series arguments as before,

$$\sum_{k=0}^{t-1} |b_k(\theta)| \leq 2\sum_{k=0}^{t-1} r^k + \theta \sum_{k=0}^{t-1} (k+1)r^k \leq \frac{3}{\theta}.$$

We now bound $\sum_{k=0}^{t-1} |b_k'| / \sum_{j=0}^{t-1} b_j(\theta)$. First, compute

$$\frac{d}{d\theta}((2 - \theta(k+1))r^k) = -(k+1)r^k - k(2 - \theta(k+1))r^{k-1} = b_k'(\theta).$$

From the expression for $b_k'(\theta)$ and $|2 - \theta(k+1)| \leq 2 + \theta(k+1)$,

$$|b_k'(\theta)| \leq (k+1)r^k + k(2 + \theta(k+1))r^{k-1} \leq (k+1)r^k + 2kr^{k-1} + \theta k(k+1)r^{k-1}.$$

Hence

$$\sum_{k=0}^{t-1} |b_k'(\theta)| \leq \sum_{k=0}^{t-1} (k+1)r^k + 2\sum_{k=0}^{t-1} kr^{k-1} + \theta \sum_{k=0}^{t-1} k(k+1)r^{k-1}.$$

We bound the sums in the two regimes.

Case 1: $\theta \leq 1/t$. Use the finite-sum bounds $r^k \leq 1$ and $r^{k-1} \leq 1$. Since $\theta \leq 1/t$, we obtain

$$\begin{aligned} \sum_{k=0}^{t-1} |b_k'(\theta)| &\leq \sum_{k=0}^{t-1} (k+1) + 2\sum_{k=0}^{t-1} k + \theta \sum_{k=0}^{t-1} k(k+1) \\ &\leq \frac{t(t+1)}{2} + 2 \cdot \frac{t(t-1)}{2} + \theta t^3 \\ &\leq \frac{t(t+1)}{2} + t(t-1) + t^2 \\ &\leq 3t^2. \end{aligned}$$

Moreover, $\sum_{j=0}^{t-1} b_j(\theta) \geq t$, hence

$$\frac{\sum_{k=0}^{t-1} |b_k'(\theta)|}{\sum_{j=0}^{t-1} b_j(\theta)} \leq \frac{3t^2}{t} = 3t.$$

We also have $2 - \theta(k+1) \geq 1$ for all $k \leq t-1$, so $b_k(\theta) \geq 0$ and therefore $\sum_{k=0}^{t-1} |b_k(\theta)| = \sum_{k=0}^{t-1} b_k(\theta)$, i.e., $\|w^{(t)}(\theta)\|_1 = \sum_{k=0}^{t-1} |b_k| / \sum_{j=0}^{t-1} b_j(\theta) = 1$.

Case 2: $\theta > 1/t$. Extend the sums to infinity and use standard identities:

$$\begin{aligned} \sum_{k=0}^{t-1} |b_k'(\theta)| &\leq \sum_{k=0}^{\infty} (k+1)r^k + 2\sum_{k=0}^{\infty} kr^{k-1} + \theta \sum_{k=0}^{\infty} k(k+1)r^{k-1} \\ &= \frac{1}{(1-r)^2} + \frac{2}{(1-r)^2} + \frac{2\theta}{(1-r)^3} \\ &= \frac{1}{\theta^2} + \frac{2}{\theta^2} + \frac{2\theta}{\theta^3} \\ &= \frac{5}{\theta^2}. \end{aligned}$$

Since $\sum_{j=0}^{t-1} b_j(\theta) \geq 1/\theta$ in this regime,

$$\frac{\sum_{k=0}^{t-1} |b_k'(\theta)|}{\sum_{j=0}^{t-1} b_j(\theta)} \leq \frac{5/\theta^2}{1/\theta} = \frac{5}{\theta} \leq 5t,$$

and it further holds $\|w^{(t)}(\theta)\|_1 = \sum_{k=0}^{t-1} |b_k(\theta)| / \sum_{j=0}^{t-1} b_j(\theta) \leq 3$ because weights may be signed. The bound from the second case, where $\theta > 1/t$ and the weights may be negative, is larger than the bound from the first case, where the weights lie on a probability simplex. Therefore, for any $\theta \in (0, 1)$, we obtain

$$\left\| \frac{d}{d\theta} w^{(t)}(\theta) \right\|_1 \leq (5t)(1+3) = 20t.$$

Fix $\theta, \theta' \in \Theta$. By the mean value theorem, $\|w^{(t)}(\theta) - w^{(t)}(\theta')\|_1 \leq 20t\,|\theta - \theta'|$. Set $\eta = \epsilon/(20t)$ and let $\{\theta_1, \ldots, \theta_N\}$ be an η-net of Θ. Arguing as in Lemma E.3, we obtain $N_1(\epsilon, \mathcal{W}_t(\Theta)) \leq N_1(\eta, \Theta)$. Applying Lemma E.1 with $d = 1$ and radius $R = 1$ yields $N_1(\eta, \Theta) \leq 3/\eta = 60t/\epsilon$, and the last statement follows immediately from bounding the covering number of the union over $t = 1, \ldots, n$. □

Lemma E.6. *Let $\Theta = (0, 1)$ and $\mathcal{W}_t(\Theta) = \{w^{(t)}(\theta) : \theta \in \Theta\}$. For weights of the form (10),*

$$B_{\mathcal{W}} \leq 18e^2, \qquad C_1 \leq 3.$$

Proof. We will use the same notation as in the preceding proof. Therefore, let $r = 1 - \theta \in (0, 1)$ and $b_k(\theta) = (2 - \theta(k+1))r^k$ for $k = 0, \ldots, t-1$. Moreover, we again have to consider two regimes for θ.

Case 1: $\theta \leq 1/t$. For a fixed θ and all $k \leq t-1$, we have $2 - \theta(k+1) \geq 2 - \theta t \geq 1$, which implies $b_k(\theta) \geq r^k \geq r^{t-1}$. Moreover, since $\theta \leq 1/t$, we have $r \geq 1 - 1/t$ and thus $b_k(\theta) \geq r^{t-1} \geq \left(1 - \frac{1}{t}\right)^{t-1} \geq e^{-1}$ for all k. Therefore

$$\sum_{k=0}^{t-1} b_k(\theta)^2 \geq \sum_{k=0}^{t-1} e^{-2} = te^{-2}.$$

Furthermore, $b_0(\theta) = 2 - \theta \leq 2$, so $\max_{0 \leq k \leq t-1} |b_k(\theta)| \leq 2$, and trivially $\sum_{k=0}^{t-1} b_k(\theta) \leq \sum_{k=0}^{t-1} 2 = 2t$. Therefore

$$\frac{\|w^{(t)}(\theta)\|_\infty}{\|w^{(t)}(\theta)\|^2} = \frac{(\sum_{k=0}^{t-1} b_k(\theta)) \max_{0 \leq k \leq t-1} |b_k(\theta)|}{\sum_{k=0}^{t-1} b_k(\theta)^2} \leq \frac{(2t) \cdot 2}{te^{-2}} = 4e^2.$$

Case 2: $\theta > 1/t$. Fix $\theta > 1/t$ and set $m = \lfloor 1/(2\theta) \rfloor$. Then $m \leq t-1$, so the indices $k = 0, \ldots, m$ are available. For $k \leq m$ we have $\theta(k+1) \leq \theta(m+1) \leq 1 \Rightarrow 2 - \theta(k+1) \geq 1$, and also $r^k \geq r^m$. Since $\log(1-\theta) \geq -2\theta$ for $\theta \in (0, 1/2]$, we obtain

$$r^m = (1-\theta)^m \geq (1-\theta)^{1/(2\theta)} \geq e^{-1},$$

whenever $\theta \in (1/t, 1/2]$. Hence $b_k(\theta) \geq e^{-1}$ for all $k \leq m$ in this regime, and therefore

$$\sum_{k=0}^{t-1} b_k(\theta)^2 \geq \sum_{k=0}^{m} e^{-2} = (m+1)e^{-2}.$$

Since $m = \lfloor 1/(2\theta) \rfloor$, we have $\lfloor x \rfloor + 1 \geq x$ for all $x \geq 0$, and thus $m + 1 = \lfloor \frac{1}{2\theta} \rfloor + 1 \geq \frac{1}{2\theta}$. Consequently, for $\theta \in (1/t, 1/2]$,

$$\sum_{k=0}^{t-1} b_k(\theta)^2 \geq \frac{e^{-2}}{2\theta}.$$

For $\theta \in (1/2, 1)$ we instead use the trivial bound $\sum_{k=0}^{t-1} b_k(\theta)^2 \geq b_0(\theta)^2 = (2-\theta)^2 \geq 1$.

Moreover, the previous lemma shows that, for all $\theta \in (0, 1)$,

$$\sum_{k=0}^{t-1} b_k(\theta) \leq \sum_{k=0}^{t-1} |b_k(\theta)| \leq \frac{3}{\theta}.$$

Finally, for all $k \geq 0$, $|b_k(\theta)| \leq (2 + \theta(k+1))r^k \leq 2 + \theta(k+1)r^k$. Using $r^k = (1-\theta)^k \leq e^{-\theta k}$ and maximizing the function $x \mapsto \theta(x+1)e^{-\theta x}$ over $x \geq 0$ yields the maximizer $x^* = 1/\theta - 1$ and

$$\sup_{x \geq 0} \theta(x+1)e^{-\theta x} = \theta \left(\frac{1}{\theta}\right) e^{-\theta(1/\theta - 1)} = e^{-1+\theta} \leq 1.$$

Therefore, $\max_{0 \leq k \leq t-1} |b_k(\theta)| \leq 2 + 1 = 3$, and we obtain the uniform bound

$$\frac{\|w^{(t)}(\theta)\|_\infty}{\|w^{(t)}(\theta)\|^2} = \frac{|\sum_{k=0}^{t-1} b_k(\theta)| \max_{0 \leq k \leq t-1} |b_k(\theta)|}{\sum_{k=0}^{t-1} b_k(\theta)^2} \leq \begin{cases} \frac{\frac{3}{\theta} \cdot 3}{\frac{e^{-2}}{2\theta}} = 18e^2, & \theta \in (1/t, 1/2], \\ \frac{\frac{3}{\theta} \cdot 3}{1} \leq 18, & \theta \in (1/2, 1). \end{cases}$$

The proof of the preceding lemma gives $\sup_\theta \|w^{(t)}(\theta)\|_1 \leq 3$ for any t, yielding the second claim. $\square$