Sharp Rates in Dependent Learning Theory: Avoiding Sample Size Deflation for the Square Loss

Ingvar Ziemann [1] **Stephen Tu** [2] **George J. Pappas** [1] **Nikolai Matni** [1]

Abstract

In this work, we study statistical learning with dependent (β-mixing) data and square loss in a hypothesis class $\mathscr{F} \subset L_{\Psi_p}$ where Ψ_p is the norm $\|f\|_{\Psi_p} \triangleq \sup_{m\geq 1} m^{-1/p}\|f\|_{L^m}$ for some $p \in [2,\infty]$. Our inquiry is motivated by the search for a sharp noise interaction term, or variance proxy, in learning with dependent data. Absent any realizability assumption, typical non-asymptotic results exhibit variance proxies that are deflated *multiplicatively* by the mixing time of the underlying covariates process. We show that whenever the topologies of L^2 and Ψ_p are comparable on our hypothesis class $\mathscr{F}$—that is, $\mathscr{F}$ is a weakly sub-Gaussian class: $\|f\|_{\Psi_p} \lesssim \|f\|_{L^2}^{\eta}$ for some $\eta \in (0,1]$—the empirical risk minimizer achieves a rate that only depends on the complexity of the class and second order statistics in its leading term. Our result holds whether the problem is realizable or not and we refer to this as a *near mixing-free rate*, since direct dependence on mixing is relegated to an additive higher order term. We arrive at our result by combining the above notion of a weakly sub-Gaussian class with mixed tail generic chaining. This combination allows us to compute sharp, instance-optimal rates for a wide range of problems. Examples that satisfy our framework include sub-Gaussian linear regression, more general smoothly parameterized function classes, finite hypothesis classes, and bounded smoothness classes.

1. Introduction

While a significant portion the data used in modern learning algorithms exhibits temporal dependencies, we still lack a sharp theory of supervised learning from dependent data. Examples exhibiting such dependencies are far ranging and abundant, and include forecasting applications and data from controls/robotics systems. Over the last several decades, an order-wise rather sharp theory of learning with *independent* data has emerged. An entirely incomplete list of these advances includes the introduction of local Rademacher compleixities by Bartlett et al. (2005), sharp rates in misspecified linear regression by Hsu et al. (2012), and culminates in the learning without concentration framework by Mendelson (2014), which enables an instance-optimal understanding of many standard learning problems through a *critical radius* that is sensitive to both the noise scale and the (local) geometry of the hypothesis class.

In principle, one expects these results to be carried over to the dependent (β-mixing) setting through *blocking* (Bernstein, 1927; Yu, 1994).[1] At a high level, the blocking technique involves splitting the original data (of length $n \in \mathbb{N}$) into consecutive blocks, each of length $k \in \mathbb{N}$, with the length chosen such that the starting points of each block are approximately independent. Indeed, several prior works pursue this route (Mohri & Rostamizadeh, 2008; Kuznetsov & Mohri, 2017; Roy et al., 2021). However, the drawback with this approach is that it typically deflates the original sample size by the block length factor k. If such a deflation were to appear in the final rate of convergence, this would clearly constitute worst-case behavior; it corresponds to every data point being revealed repeatedly, k times and with perfect dependence, within a sequence of n observations.

In the context of the square loss function, the typical approach to sidestep this sample size deflation relies on the "noise" (residual term) forming a martingale difference sequence. This approach has been carried out for parametric inference in (generalized) linear dynamical systems by Simchowitz et al. (2018) and Kowshik et al. (2021) and also for more general hypothesis classes and supervised learning with square loss by Ziemann & Tu (2022). For the square loss function the martingale approach requires that the problem is strongly realizable: the best predictor in the hypothesis class should coincide with the regression function (conditional expectation of targets given past inputs).

[1] See Appendix E.1 for a description of this technique.

[1]University of Pennsylvania [2]University of Southern California. Correspondence to: Ingvar Ziemann <ingvarz@seas.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Put differently, one requires that the hypothesis class is rich enough such that conditional expectation function (of the targets and given past inputs) can be realized by it.

In this paper, we instead show how the blocking approach can be salvaged for a wide range of hypotheses classes and the square loss function. In contrast to the just-mentioned references, our analysis does not require a realizability assumption. Instead, we show how to extend the analysis of Ziemann et al. (2023b) for linear regression to more general hypothesis classes. At a high level, this analysis involves combining the above-mentioned blocking technique with Bernstein's inequality. To motivate this approach, let us consider what happens in Bernstein's inequality when we are given $V_{1:n}$ b-bounded random variables that are k-wise independent, where k divides n, and with identical marginals (for simplicity alone).[2] By applying Bernstein's inequality to the bk-bounded variables $\bar{V}_{i:n/k}, \bar{V}_i \triangleq \sum_{j=ik-k+1}^{ik} V_j$ we find that with probability at least $1-\delta$:

$$\frac{1}{n}\sum_{i=1}^{n} V_i \le 2\sqrt{\frac{k^{-1}\mathbf{E}(\bar{V}_1)^2 \ln(1/\delta)}{n}} + \frac{4bk\log(1/\delta)}{3n}. \quad (1)$$

If the data instead were completely independent, then in the small and moderate deviations regime $\delta \gtrsim \exp(-n\mathbf{E}V_1^2/b^2k)$, (1) is just as sharp as directly applying Bernstein's inequality to the independent sum. In this regime for this problem, nothing is lost by blocking, even if the data happens to be iid and we use the blocked version of Bernstein's inequality. By contrast, if one were to carry out the same computation using Hoeffding's inequality (for bounded random variables) instead of Bernstein's, we would incur an irreducible factor k in the leading term in all regimes—even if the dependent bound is instantiated for independent variables. This suggests that the variance interacts much more gracefully with blocking arguments than higher order moments.

The difficulty in combining blocking with Bernstein's inequality lies in making Bernstein's inequality uniform across the correct portion of the hypothesis class $\mathscr{F}$. Namely, in statistical learning it typically does not suffice to control sums of a single sequence of random variables $V_{1:n}$ but rather we need to uniformly control sums of an indexed family $\{V_{1:n}(f) : f \in \mathscr{F}\}$. To obtain fast rates, this uniform control needs to combined with a localization argument, so that one does "pay" for hypotheses too far away from the ground truth but only those within a certain critical radius. Naïvely union-bounding (or chaining) over such a family unfortunately again reintroduces a sample-size deflation by the block-length factor k. This happens because the variance term in (1) starts to balance the boundedness term at the above-mentioned critical radius without further assumption. Ziemann et al. (2023b) show how to overcome the issue of uniformity when $\mathscr{F}$ is a linear class via the Fuk-Nagaev inequality (Einmahl & Li, 2008). Unfortunately, this inequality cannot be applied beyond the linear setting. Here, we introduce machinery based on a refinement of sub-Gaussian classes (Lecué & Mendelson, 2013), and a refinement of Bernstein's inequality (due to Maurer & Pontil (2021)), that we combine with mixed-tail generic chaining (as introduced by Dirksen (2015)). Our approach allows us to overcome this issue with blocking and Bernstein's inequality for a surprisingly wide range of function classes, thereby relegating any dependence on mixing to additive higher order terms, instead of the typical multiplicative deflation term.

[2] We say that a sequence $Z_{1:n}$ is k-wise independent if each of the blocks $Z_{jk+1:(j+1)k}$ $(j = 0, 1, \dots, n/k-1)$ are independent of each other.

1.1. Contribution

Let us now make our contribution more precise. We are given stationary β-mixing data $(X, Y)_{1:n}$ where the X_i (resp. Y_i) assume values in a subset of a normed space denoted $(\mathsf{X}, \|\cdot\|_{\mathsf{X}})$ (resp. a Hilbert space $(\mathsf{Y}, \langle\cdot,\cdot\rangle, \|\cdot\|)$). We assume that $(X, Y)_{1:n}$ is stationary and denote for any $i \in [n]$ the joint distribution of (X_i, Y_i) by $\mathsf{P}_{X,Y}$, and the corresponding marginals are denoted P_X and P_Y. We study empirical risk minimization over a hypothesis class $\mathscr{F}$, containing functions $f : \mathsf{X} \to \mathsf{Y}$, and with the square loss function. In this scenario, we study the performance of the (any) empirical risk minimizer

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \|f(X_i) - Y_i\|^2. \quad (2)$$

Our main contribution is to characterize the rate of convergence of (2) to the best possible predictor $f_\star$ in the class $\mathscr{F}$ defined as:

$$f_\star \in \operatorname*{argmin}_{f\in\mathscr{F}} \mathbf{E}\|f(X) - Y\|^2, \qquad (X, Y) \sim \mathsf{P}_{X,Y}. \quad (3)$$

Let us also denote by $\mathscr{F}_\star$ the star-hull of $\mathscr{F}$ around $f_\star$. That is, $\mathscr{F}_\star \triangleq \{\rho(f - f_\star) : f \in \mathscr{F}, \rho \in [0,1]\}$, which, for a convex class $\mathscr{F}$ coincides with the shifted class $\mathscr{F} - \{f_\star\}$. We further equip $\mathscr{F}_\star$ with the L^2-norm: $\|f\|_{L^2}^2 \triangleq \mathbf{E}\|f(X)\|^2, f \in \mathscr{F}_\star, X \sim \mathsf{P}_X$. Let us also define the "noise" $W_{1:n}$ by $W_i \triangleq Y_i - f_\star(X_i), i \in [n]$. We focus on the case when $\mathscr{F}$ is either (1) convex or (2) realizable (i.e., $\mathbf{E}[W_i|X_i] = 0$ for $i \in [n]$). Note that this restriction is due to a known shortcoming of ERM which holds even in iid settings, and can be removed by modifying the estimator itself; we will discuss this issue in more detail shortly.

As is typical in the learning theory literature, we characterize the rate of convergence of (2) through a fixed point, or critical radius. This critical radius takes the form as a

solution to:

$$r_\star \asymp \sup_{g \in \mathscr{F}_\star \cap r_\star S_{L^2}} \mathbf{V}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\langle W_i, \frac{g(X_i)}{\|g\|_{L^2}} \right\rangle\right) \times \frac{\text{complexity}(\mathscr{F}_\star \cap r_\star S_{L^2})}{r_\star \sqrt{n}}, \quad (4)$$

where for $r \in \mathbb{R}, r > 0$, rS_{L^2} is the unit sphere of radius r in L^2 (the space of square integrable functions, and with the corresponding unit ball denoted rB_{L^2}) and $\mathbf{V}(\cdot)$ denotes the variance operator. This critical radius is akin to the one in Bartlett et al. (2005), but also resembles the noise interaction term of Mendelson (2014, introduced following Equation 2.2) in that our radius depends on the *weak variance*, $\sup_{g \in \mathscr{F}_\star \cap r_\star S_{L^2}} \mathbf{V}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\langle W_i, \frac{g(X_i)}{\|g\|_{L^2}} \right\rangle\right)$.[3] To aid in the interpretation of $r_\star$, we will instantiate our main result, Theorem 3.1, for parametric classes and show that this radius exhibits the desired "dimension counting" scaled with noise-to-signal behavior, see Corollary 3.1 and Corollary 3.2. Moreover, the weak variance term takes into account how targets $Y_{1:n}$ interact with the function class $\mathscr{F}$ through $W_{1:n}$, locally at radius $r_\star$ near the minimizer $f_\star$, via a second-order statistic. In particular, this variance term is always sharper than the corresponding iid variance term deflated by a factor of the mixing-time (or block-length).

With these preliminaries in place we are ready to state an informal version of our main result.

Informal version of Theorem 3.1. *Given data that mixes sufficiently fast, for a wide range of (1) convex or (2) realizable hypothesis classes, any empirical risk minimizer $\widehat{f}$ over such a class $\mathscr{F}$ converges at least as fast a rate characterized by the critical radius $r_\star$ given by the solution to (4) depending on the variance of the noise-class interaction and local scale of the class $\mathscr{F}$. That is with probability $1-\delta$:*

$$\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim r_\star^2 + \frac{(\text{weak variance}) \times \log(1/\delta)}{n} + \text{terms of higher order}(r_\star, n^{-1}, \text{mixing}, \log(1/\delta)). \quad (5)$$

Moreover, for d-dimensional parametric classes the leading term is $(\text{weak variance}) \times \frac{d+\log(1/\delta)}{n}$.

The crux of this result is that past a burn-in, the ERM excess risk does not directly depend on mixing times, but only on the relevant second order statistics. Put differently, the effect of slow mixing has been relegated to a small *additive* term with higher order dependence on $1/n$. Indeed, both $r_\star$ and the variance term in (5) do not directly depend on slow mixing (i.e., are not deflated by the block-length k) but only on relevant second order statistics. Slow mixing only affects higher order additive terms that can be pushed into the burn-in.

The qualifier "wide range" above refers to the requirement that the class $\mathscr{F}$ satisfies a certain topological condition. Recall that for a random variable Z the Ψ_p-norm is the norm $\|Z\|_{\Psi_p} = \sup_{m \geq 1} m^{-1/p} \|Z\|_{L^m}$. We will ask that for some $\eta \in (0, 1]$ and $L > 0$, every $f \in \mathscr{F}_\star$ satisfies the inequality $\|f\|_{\Psi_p} \leq L \|f\|_{L^2}^\eta$. We will say that such classes are *weakly sub-Gaussian* and will verify that such an inequality indeed holds true for a range of examples in Section 4:

- bounded smoothness classes, see Proposition 4.1;
- parametric classes that are Lipschitz in their parameterization, see Proposition 4.2;
- sub-Gaussian linear regression, see Proposition 4.3;
- finite hypothesis classes, see Proposition 4.4.

Finally, the requirement that $\mathscr{F}$ be either (1) convex or (2) realizable can easily be removed with a few modifications if one replaces the empirical risk minimizer by the star estimator of Audibert (2007). In this case (but with the L^2-error replaced with the no-longer directly comparable excess risk functional) the geometric inequality by Liang et al. (2015, Lemma 1) takes a similar role to the basic inequality we use below. The necessity of imposing (1) or (2) is due to a known shortcoming of empirical risk minimization outside of convex (or realizable) classes, and not an issue directly related to dependent data (see e.g. the discussion in Mendelson, 2019).

1.2. Proof Outline

From a more technical standpoint, our contribution is a novel analysis of two empirical processes that arise in (but are not restricted to) empirical risk minimization, and which are sharp even for dependent data. Following the language of Mendelson (2014), we refer to these as the quadratic and multiplier empirical processes. The first of these, the quadratic process, controls a one-sided discrepancy between the empirical and population L^2-norms:

$$Q_n(f) \triangleq \frac{1}{n} \sum_{i=1}^n \mathbf{E}\|f(X_i) - f_\star(X_i)\|^2 - \frac{(1+\varepsilon)}{n} \sum_{i=1}^n \|f(X_i) - f_\star(X_i)\|^2, \quad (6)$$

for some $\varepsilon \in (0, 1)$. Under our assumptions, we will show that the process $Q_n(f)$ is eventually nonpositive uniformly

[3] The terminology weak variance comes from the empirical processes literature in that the supremum in Definition 2.2 is on the outside of the expectation.

for all sufficiently large f, implying that the empirical L^2-norm dominates the population L^2-norm.

Now for $r \in \mathbb{R}_+$ and conditionally on the event $\{Q_n(f) \le 0,\ \forall f \in \mathscr{F}_\star \setminus rB_{L^2}\}$, using optimality of $\widehat{f}$ to (2) we also have the following deterministic (basic) inequality:

$$\|\widehat{f} - f_\star\|_{L^2} \le r + \frac{1+\varepsilon}{rn}\sum_{i=1}^n 2(1-\mathbf{E}')\left[\left\langle W_i, \frac{r\left[\widehat{f}(X_i) - f_\star(X_i)\right]}{\|\widehat{f} - f_\star\|_{L^2}}\right\rangle\right], \tag{7}$$

where $\mathbf{E}'$ denotes expectation with respect to a fresh copy of randomness independent of $\widehat{f}$.

Hence, we also need to control the multiplier process:

$$M_n(f) \triangleq \frac{1+\varepsilon}{n}\sum_{i=1}^n 2(1-\mathbf{E}')\langle W_i, f(X_i)\rangle. \tag{8}$$

It is the uniform control of $M_n(f)$ over the class $\mathscr{F}_\star$ intersected with the radius r ball rS_{L^2} balanced with the first term of (7) that gives rise to the critical radius (4). This argument is formalized in Lemma F.1. Just as in Mendelson (2014), it is the multiplier process (8) that yields the dominant contribution to the error (5) (after a burn-in). This is important as it allows us to use blocking to control (6) without affecting the leading term of the final rate.

We reiterate that our analysis of the above two empirical processes ((6) and (8)) rests crucially on the assumption that $\mathscr{F}_\star$ is a weakly sub-Gaussian class. Let us also point out that we first make a simplifying assumption, namely that our model is k-wise independent. We later port all results to the β-mixing setting by blocking, cf. Section 2.3. A sketch of the analysis of $M_n(f)$—found in Section 2.1 with proofs relegated to Appendix C—now goes as follows:

- We invoke a refinement of Bernstein's inequality (Lemma 2.1) to gain pointwise control of $M_n(f)$. The benefit of this over the standard version is that we do not require boundedness, but rather finite Ψ_p-norm suffices. Unless $p = \infty$ (boundedness), the price we pay for this is that the variance proxy is degraded to a moment of order $2q, q > 1$ instead of order 2.
- We make this refinement of Bernstein's inequality uniform over the class $\mathscr{F}_\star$ intersected with the radius r ball rS_{L^2} by invoking mixed-tail generic chaining (Dirksen, 2015). This splits the tail into an L^{2q}-component and a Ψ_p-component.
- Our assumption that $\mathscr{F}_\star$ is a weakly sub-Gaussian class now comes into play by ensuring that, past a burn-in, the Ψ_p-component of the mixed tail is of lesser magnitude than the L^{2q}-part of the tail. Just as in our introductory example with Bernstein's inequality (1), any dependence on mixing is relegated to this smaller Ψ_p-component (which now assumes the role of boundedness).
- Combining these steps with (7) yields control of the multiplier process and is summarized in Theorem 2.1.

The analysis of $Q_n(f)$ is relatively standard and amounts to showing that the norm-bound $\|f\|_{\Psi_p} \le L\|f\|_{L^2}^\eta$ is sufficient to modify a standard truncation argument (see e.g. Wainwright, 2019, Theorem 14.12). We then proceed to control the remainder of said truncation argument completely analogously to our above approach for $M_n(f)$. We detail these arguments in Section 2.2 and prove them in Appendix D. Finally, we combine our control of the multiplier and quadratic processes (Theorem 2.1 and Theorem 2.2) with blocking to arrive at our main result, Theorem 3.1.

1.3. Further Preliminaries

Notation. Expectation (resp. probability) with respect to all the randomness of the underlying probability space is denoted by $\mathbf{E}$ (resp. $\mathbf{P}$). For $q \in [1,\infty)$ the $2q$-variance of a random variable Z is defined as $\mathbf{V}_{2q}(Z) \triangleq (\mathbf{E}(Z - \mathbf{E}Z)^{2q})^{1/q}$ with $\mathbf{V}_2 = \mathbf{V}$ being the standard variance. For $p \in [1,\infty)$, we also introduce the Ψ_p-norm $\|Z\|_{\Psi_p} \triangleq \sup_{m\ge 1} m^{-1/p}\|Z\|_{L^m}$ and also set $\|Z\|_{\Psi_\infty} \triangleq \|Z\|_{L^\infty}$. Two extended real numbers $q, q' \in [1,\infty]$ are said to be Hölder conjugates if $1/q + 1/q' = 1$, where, as we do throughout, $1/\infty$ is interpreted as 0. For two probability measures P and Q defined on the same probability space, their total variation is denoted $\|\mathsf{P} - \mathsf{Q}\|_{\mathsf{TV}}$. Maxima (resp. minima) of two numbers $a, b \in \mathbb{R}$ are denoted by $a \vee b = \max(a,b)$ (resp. $a \wedge b = \min(a,b)$). For an integer $n \in \mathbb{N}$, we also define the shorthand $[n] \triangleq \{1,\dots,n\}$. For a symmetric positive semidefinite matrix M, $\lambda_{\min}(M)$ denotes its smallest nonzero eigenvalue.

Talagrand's functionals. The complexity$(\mathscr{F}_\star \cap r_\star S_{L^2})$ term in (4) is made precise through Talagrand's γ_α-functional (with $\alpha = 2$ being the dominant term in our result). Let be $(\mathscr{H}, d)$ a metric space. We denote the diameter of $\mathscr{H}$ with respect to d by

$$\Delta_d(\mathscr{H}) \triangleq \sup_{h,h'\in\mathscr{H}} d(h,h').$$

A sequence $\mathcal{H} = (H_m)_{m\in\mathbb{Z}_+}$ of subsets of $\mathscr{H}$ is called admissible if $|H_0| = 1$ and $|H_m| \le 2^{2^m}$ for all $m \ge 1$. For $\alpha \in (0,\infty)$, the γ_α-functional of $(\mathscr{H}, d)$ is defined by

$$\gamma_\alpha(\mathscr{H}, d) \triangleq \inf_{\mathcal{H}} \sup_{h\in\mathscr{H}} \sum_{m=0}^\infty 2^{m/\alpha} d(h, H_n), \tag{9}$$

where the infimum is taken over all admissible sequences (we write $d(h, H) = \inf_{s\in H} d(h, s)$ whenever H is a set). For $\eta \in (0, 1)$, we slightly abuse notation and write $\gamma_\alpha(\mathscr{H}, d^\eta)$ for d replaced with d^η in (9) (while being mindful of that fact that d^η is not a metric in general). Finally, since entropy integrals upper-bound γ_α-functionals, it will also be useful to introduce the covering number $\mathcal{N}_{L^2}(\mathscr{H}, s)$, which denotes the minimal number of L^2-balls of radius s required to cover $\mathscr{H}$.

2. Ψ_p-Norms, Bernstein's Inequality and Empirical Processes

In this section we establish a few preliminary technical lemmas that will be useful for controlling the multiplier and quadratic processes ((8) and (6)). We begin with a version of Bernstein's inequality that controls the Laplace transform of Z in terms of its L^{2q}-norm ($q \geq 1$) and some Ψ_p-norm. The lemma comes from (Maurer & Pontil, 2021).

Lemma 2.1 (Ψ_p-Bernstein MGF Bound)**.** *Fix a random variable Z and $p \in [1, \infty]$ such that $\mathbf{E}Z \leq 0$ and $\|Z\|_{\Psi_p} < \infty$. Let q and q' be Hölder conjugates and suppose that $\lambda \in [0, 1/(q'e)^{1/p}\|Z\|_{\Psi_p}]$. We have that:*

$$\mathbf{E}\exp(\lambda Z) \leq \exp\left(\frac{\frac{\lambda^2}{2}\left(\mathbf{E}(Z)^{2q}\right)^{1/q}}{1-\lambda(q'e)^{1/p}\|Z\|_{\Psi_p}}\right). \tag{10}$$

Our intention is to use Lemma 2.1 to afford us—pointwise in g—control of the multiplier process introduced in (8). Indeed, notice that in the regime $\lambda \in (0, (2(q'e)^{1/p}\|Z\|_{\Psi_p})^{-1}]$ the dominant term in (10) is $2q$-variance of Z. Since (7) can be localized to a ball of radius r in L^2 it suffices that the L^2-norm provides some weak control of the Ψ_p-norm for any constant choice of λ to be admissible once the localization radius r is chosen small enough. This in turn motivates the following definition.

Definition 2.1 (Weakly sub-Gaussian Class)**.** *Fix $\eta \in (0, 1]$ and $L \in [1, \infty)$. We say that a class $\mathscr{G}$ is (L, η)-Ψ_p if for every $g \in \mathscr{G}$ we have that:*

$$\|g\|_{\Psi_p} \leq L\|g\|_{L^2}^\eta. \tag{11}$$

If (11) holds for $\mathscr{G}$ with $\eta \in (0, 1)$ and some L we will call $\mathscr{G}$ a weak Ψ_p-class. If (11) instead holds for $\eta = 1$ it is simply a Ψ_p-class. This generalizes the notion of a sub-Gaussian class from (Lecué & Mendelson, 2013), which corresponds to $\eta = 1$ and $p = 2$. Let us further point out that by homogeneity, if $\eta \in (0, 1)$ in (11), then one should expect L to depend polynomially on some other norm (or homogenous functional) of g. Indeed, by the Gagliardo-Nirenberg interpolation inequality, the above relaxation ($\eta < 1$) covers smoothness classes (Proposition 4.1), whereas the strict sub-Gaussian class assumption ($\eta = 1$) of (Lecué & Mendelson, 2013) is difficult to verify beyond linear functionals.

As we have pointed out above, our intention is to apply Lemma 2.1 pointwise to the multiplier process (8). However, this yields a different variance term for each index point of the empirical process. The solution to this is simply to define a uniform variance term, as is done below.

Definition 2.2 (Noise Level)**.** *The $2q$-noise-class-interaction between $\mathscr{F}$, the model $\mathsf{P}_{(X,Y)_{1:n}}$, and the shifted target $W_{1:n} = (Y - f_\star(X))_{1:n}$ at resolution $\mathscr{G}$ is given by*

$$\mathbf{V}_{2q}\left(\mathscr{F}, \mathscr{G}, \mathsf{P}_{(X,Y)_{1:n}}\right) \triangleq \sup_{g\in\mathscr{G}} \mathbf{V}_{2q}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\langle W_i, \frac{g(X_i)}{\|g\|_{L^2}}\right\rangle\right). \tag{12}$$

We stress that, even though Definition 2.2 measures noise uniformly over a function class, it does not generally grow with the complexity of the class. For instance, under the additional hypotheses that $\mathsf{P}_{(X,Y)_{1:n}}$ is drawn iid and that W_i is independent of X_i for $i \in [n]$, it is easy to see that $\mathbf{V}_2\left(\mathscr{F}, \mathscr{G}, \mathsf{P}_{(X,Y)_{1:n}}\right) = \mathbf{V}_2(W)$ for every such well-specified class $\mathscr{G}$. Rather, Definition 2.2 is a measure of how well the targets $Y_{1:n}$ align with a given class $\mathscr{G}$.

2.1. The Multiplier Process

We will not directly control the multiplier process for β-mixing variables. Instead we first suppose that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is k-wise independent (where k divides n). We then port these results to the β-mixing setting by blocking (see Appendix E.1). We use the following shorthand notation regarding $\mathbf{V}_{2q}\left(\mathscr{F}, \mathscr{G}, \mathsf{P}_{(X,Y)_{1:k}}\right)$: we take the class $\mathscr{F}$ and the probability model $\mathsf{P}_{(X,Y)_{1:k}}$ as fixed and thus omit the dependence on $\mathscr{F}$ (via $f_\star$) and $\mathsf{P}_{(X,Y)_{1:k}}$ and write $\mathbf{V}_{2q}(\mathscr{G}) = \mathbf{V}_{2q}\left(\mathscr{F}, \mathscr{G}, \mathsf{P}_{(X,Y)_{1:k}}\right)$. With these remarks in place, we now turn to establishing pointwise control of (8) using Lemma 2.1.

Lemma 2.2 (Pointwise Control)**.** *Fix two Hölder conjugates q and q'. Suppose that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and k-wise independent where k divides n. For every $g, g' \in L_{\Psi_p}$ and $u \in (0, \infty)$ we have that:*

$$\mathbf{P}\Bigg(\sum_{i=1}^n (1-\mathbf{E})\langle W_i, g(X_i) - g'(X_i)\rangle > \sqrt{4n\|g-g'\|_{L^2}^2 \mathbf{V}_{2q}\left(\{g\} - \{g'\}\right)u} + 4(q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}u\Bigg) \leq 2e^{-u}. \tag{13}$$

In the main development we will instantiate Lemma 2.2 with $r = \|g - g'\|_{L^2}^2$ decaying to 0 (which should be thought of as a fixed point upper-bounding the rate of convergence

of ERM) at a polynomial rate in n. If furthermore $\mathscr{G}$ is (L,η)-Ψ_p, then the second term (linear in u) of (13) can be rendered negligible at every scale r, which allows us to invoke mixed-tail generic chaining (Dirksen, 2015) to show that the weak variance $\mathbf{V}_{2q}(\mathscr{F}_\star \cap rS_{L^2})$ dominates the noise level in the small-to-moderate deviations regime.

Put differently, at the scale of localization considered here, the noise level of the empirical process is almost entirely dictated by the weak variance $\mathbf{V}_{2q}(\mathscr{F}_\star \cap rS_{L^2})$. Now, since q' is the Hölder conjugate of q this further implies that we may choose $q = 1 + o(1)$ so that we might expect $\mathbf{V}_{2q}(\mathscr{F}_\star \cap rS_{L^2}) = \mathbf{V}(\mathscr{F}_\star \cap rS_{L^2}) + o(1)$. Moreover if $p = \infty$ this always the case and we may choose $q = 1$. In principle no better variance proxy is possible, since already for a single function g as $n \to \infty$, by the central limit theorem under mild ergodicity assumptions on $\mathsf{P}_{(X,Y)_{1:\infty}}$ (e.g. for the Markovian situation cf. Meyn & Tweedie, 1993, Theorem 17.3.6):

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\mathbf{E})\left\langle W_i, \frac{g(X_i)}{\|g\|_{L^2}}\right\rangle \rightsquigarrow N(0, \mathbf{V}(\mathscr{F}_\star, \{g\}, \mathsf{P}_{(X,Y)_{1:\infty}})), \quad (14)$$

where the variance term on the right is:
$\mathbf{V}(\mathscr{F}_\star, \{g\}, \mathsf{P}_{(X,Y)_{1:\infty}}) \triangleq \lim_{n\to\infty} \mathbf{V}(\mathscr{F}_\star, \{g\}, \mathsf{P}_{(X,Y)_{1:n}})$.

Now, since $r = o(1)$ in all practical situations one expects $\mathbf{V}(\mathscr{F}_\star \cap rS_{L^2}) \approx \mathbf{V}(\{f_\star\})$ as long as the map $f \mapsto \mathbf{V}(f/\|f\|_{L^2})$ is sufficiently regular near $f_\star$.

We arrive at our main result for the multiplier process by making uniform the pointwise control afforded to use by Lemma 2.2 via an instantiation of mixed-tail generic chaining (Dirksen, 2015) (for ease of reference, we restate a corollary of his result as Lemma C.1 in the appendix). This yields the following result.

Theorem 2.1. *Fix a failure probability $\delta \in (0,1)$, a positive scalar $r \in (0,\infty)$, two Hölder conjugates q and q', and a class $\mathscr{F}$. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and k-wise independent where k divides n. There exist universal positive constants c_1, c_2 such that for any $r \in (0,1]$ we have that with probability at least $1-\delta$:*

$$\begin{aligned}\sup_{f\in\mathscr{F}_\star\cap rS_{L^2}} &\frac{1}{rn}\sum_{i=1}^{n}(1-\mathbf{E})\langle W_i, f\rangle \\ &\le c_2\sqrt{\mathbf{V}_{2q}(\mathscr{F}_\star\cap rS_{L^2})}\left(\frac{1}{r\sqrt{n}}\right. \\ &\quad\left.\times\gamma_2(\mathscr{F}_\star\cap rS_{L^2}, d_{L^2}) + \sqrt{\frac{\log(1/\delta)}{n}}\right) \\ &\quad + c_1(q'e)^{2/p}Lk\|W\|_{\Psi_p} \\ &\quad\times\left(\frac{1}{rn}\gamma_\eta(\mathscr{F}_\star\cap rS_{L^2}, d_{L^2}) + \frac{r^{\eta-1}}{n}\log(1/\delta)\right). \quad (15)\end{aligned}$$

In the sequel, we will see that the first term on the right of (15) is typically dominant.

2.2. The Quadratic Process

A slight modification of the argument leading to Theorem 2.2 combined with a truncation argument detailed in Lemma D.1 also yields control of the quadratic process.

Theorem 2.2 (Lower Uniform Law)**.** *Fix a failure probability $\delta \in (0,1)$, a tolerance $\varepsilon > 0$, a localization radius $r \in (0,1]$, and two Hölder conjugates q and q'. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and k-wise independent where k divides n. There exist a universal positive constant c such that uniformly for all $f \in \mathscr{F}_\star \setminus \{rB_{L^2}\}$ we have that with probability at least $1-\delta$:*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|f(X_i)\|^2 &\ge r^2(1-\varepsilon^2) \\ &- c\left\{n^{-1/2}\sqrt{k}L^{1+3/4}r^\eta\left(\log\left(\frac{4^{2/p}L}{\varepsilon r}\right)\right)^{1/p}\right. \\ &\times\left(\gamma_{\frac{2+6\eta}{4}}(\mathscr{F}_\star\cap rS_{L^2}, d_{L^2}) + r^{\frac{1+3\eta}{4}}\sqrt{\log(1/\delta)}\right) \\ &+ n^{-1}(q')^{1/p}kr^\eta\left(\log\left(\frac{4^{2/p}L}{\varepsilon r}\right)\right)^{1/p}L^2 \\ &\left.\times(\gamma_\eta(\mathscr{F}_\star\cap rS_{L^2}, d_{L^2}) + r^\eta\log(1/\delta))\right\}. \quad (16)\end{aligned}$$

2.3. β-Mixing Processes

We extend the empirical process results of the preceding two sections to β-mixing processes in Appendix E.2. We do so by a simple blocking argument that we review in Appendix E.1, and for which we have already set the stage by establishing our results for k-wise independent processes. Here, we state the definition of dependence we rely on in the sequel.

Definition 2.3. *Let $Z_{1:n}$ be a stochastic process. The β-mixing coefficients of Z, denoted $\beta_Z(i)$, are for $i \in [n]$:*

$$\beta_Z(i) \triangleq \sup_{t\in[n]:t+i\le n}\mathbf{E}\|\mathsf{P}_{Z_{i+t}}(\cdot \mid Z_{1:t}) - \mathsf{P}_{Z_{i+t}}(\cdot)\|_{\mathsf{TV}}. \quad (17)$$

3. The Main Result

Before we state our main result, we will need to establish one more preliminary matter. Let us define the burn-in times $n_{\mathsf{quad}}, n_{\mathsf{mult}}, k_{\mathsf{mix}}$ which together dictate the minimial sample size necessary for our result to be sharp. The first of these, n_{quad}, is required for the population L^2 error to

be dominated by the empirical L^2 error: i.e., the quadratic process $Q_n(f)$ is nonpositive on our class of interest. The second of these, n_{mult}, is required for the multiplier process, $M_n(f)$, to have a dominant variance term (informally—when the CLT-like rate becomes accurate). Finally, k_{mix} is the minimal block-size it takes for the β-mixing model $\mathsf{P}_{(X,Y)_{1:n}}$ to be well-approximated by a corresponding k-wise independent model. These are given as follows below:

$$
\begin{aligned}
n_{\mathsf{quad}}(r) &= \inf\Bigg\{ n \in \mathbb{N} \Bigg| \Bigg[n^{-1/2}\sqrt{k} L^{1+3/4} r^{\eta} \\
&\quad \times \left(\log\left(\frac{4^{2/p+1/2}L}{r}\right)\right)^{1/p} \\
&\quad \times \left(\gamma_{\frac{2+6\eta}{4}}(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + r^{\frac{1+3\eta}{4}}\sqrt{\log(1/\delta)}\right) \\
&\quad + n^{-1} L^2 (q')^{1/p} k r^{\eta} \left(\log\left(\frac{4^{2/p+1/2}L}{r}\right)\right)^{1/p} \\
&\quad \times \left(\gamma_\eta(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + r^\eta \log(1/\delta)\right) \Bigg] \le r^2 \Bigg\}, \\
n_{\mathsf{mult}}(r) &= \inf\left\{ n \in \mathbb{N} \middle| (q'e)^{2/p} L k \|W\|_{\Psi_p} \right. \\
&\quad \left. \times \left(\frac{1}{rn}\gamma_\eta(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + \frac{r^{\eta-1}}{n}\log(1/\delta)\right) \le r \right\} \\
k_{\mathsf{mix}} &= \inf\{k \in [n] | k\beta^{-1}_{X,Y}(k) \ge n\delta^{-1}\}.
\end{aligned}
\tag{18}
$$

The first two of these are calculated by requiring the remainder terms in Proposition E.2 and Proposition E.3 to be of negligible order. The last term is obtained by requiring that failure term, δ, dominates the mixing term, $\frac{n}{k}\beta_{X,Y}(k)$, in the failure probability of these propositions. At this point, as a practical example, it is worth to point out that if the process $(X,Y)_{1:n}$ is geometrically ergodic—$\beta_{X,Y}(k) \lesssim \exp(-k/\tau_{\mathsf{mix}})$ for some $\tau_{\mathsf{mix}} \in \mathbb{R}_+$—this requirement is satisfied by $k \lesssim (1 \vee \tau_{\mathsf{mix}})\log(n/\delta)$. With these burn-in times in place, we are now ready to state the main result of our paper.

Theorem 3.1. *Fix a failure probability $\delta \in (0,1)$, two Hölder conjugates q and q', and a class $\mathscr{F}$ that is either (1) convex or (2) realizable. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and that k divides $n/2$. There exist universal positive constants c_1, c_2, c_3 such that the following holds. If $r_\star$ solves*

$$
r \ge c_1 \sqrt{\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap rS_{L^2}\right)} \times \frac{1}{r\sqrt{n}}\gamma_2(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}),
\tag{19}
$$

we have that with probability $1-4\delta$ that:

$$
\|\widehat{f} - f_\star\|^2_{L^2} \le c_2\left(r_\star^2 + \mathbf{V}_{2q}\left(\mathscr{F}_\star \cap r_\star S_{L^2}\right)\frac{\log(1/\delta)}{n}\right)
\tag{20}
$$

as long as $n \ge c_3 \max\{n_{\mathsf{quad}}(r_\star), n_{\mathsf{mult}}(r_\star)\}$ and $k \ge k_{\mathsf{mix}}$ (given in (18)).

Theorem 3.1 informs us that past a burn-in, the rate of convergence of empirical risk minimization is dictated by the critical radius $r_\star$ given in (19). This radius depends on local complexity of the class $\mathscr{F}$ measured in L^2 distance as per the γ_2-functional and through the weak variance $\mathbf{V}_{2q}(\mathscr{F}_\star \cap r_\star S_{L^2})$. We point out that we may choose $q=1$ if $p=\infty$, so that the variance term in (22) is the actual variance $\mathbf{V}_2$. As indicated at the discussion following (14), this variance term cannot be improved in general. Otherwise we can typically let q approach 1 as the sample size becomes sufficiently large. Moreover, unless the class exhibits large nonparametric behavior, the dependency on the complexity is also the best possible even in the iid case (Lecué & Mendelson, 2013).

We now turn to parsing Theorem 3.1 by specializing it to parametric classes. First, in Corollary 3.1 we show that for parametric classes the complexity term dictated by the critical radius $r_\star$ in (19) becomes a variance(-proxy)-scaled dimensional factor and that the burn-in requirement (18) amounts to a polynomial in problem data and $\log(1/\delta)$. Second, we simplify matters further and study bounded and realizable linear regression in Corollary 3.2. In this case, we will see that the variance term $\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap r_\star S_{L^2}\right)$ in (22) simply reduces to 2-variance of the noise variable W. Moreover, the first two burn-in requirements n_{quad} and n_{mult} are in this case satisfied as soon as $n/k \gtrsim d + \log(1/\delta)$.

Corollary 3.1 (Parametric Classes)**.** *Fix a failure probability $\delta \in (0,1)$, two Hölder conjugates q, q', and a class $\mathscr{F}$ that is either (1) convex or (2) realizable. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and that k divides $n/2$.*

There exists a universal positive constant c and a polynomial function ϕ_η such that the following holds true. Suppose that there exists $d_{\mathscr{F}} \in \mathbb{R}_+$ such that for $s > 0$:

$$
\log \mathcal{N}_{L^2}(\mathscr{F}_\star, s) \le d_{\mathscr{F}} \log\left(\frac{1}{s}\right)
\tag{21}
$$

We have with probability $1-4\delta$ that:

$$
\begin{aligned}
\|\widehat{f} - f_\star\|^2_{L^2} \le c\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap \sqrt{\frac{d_{\mathscr{F}} k \|W\|^2_{L^2}}{n}} S_{L^2}\right) \\
\times \left(\frac{d_{\mathscr{F}} + \log(1/\delta)}{n}\right)
\end{aligned}
\tag{22}
$$

as long as $k\beta^{-1}(k) \ge n\delta^{-1}$ and

$$
\begin{aligned}
n \ge \phi_\eta\Bigg(& d_{\mathscr{F}}, k, \|W\|_{\Psi_p}, L, q, q', \\
& \mathbf{V}^{-1}\left(\mathscr{F}_\star \cap \sqrt{\frac{d_{\mathscr{F}} k \|W\|^2_{L^2}}{n}} S_{L^2}\right), \log(1/\delta)\Bigg).
\end{aligned}
\tag{23}
$$

Consequently, after a polynomial burn-in and up to a universal positive constant, we are able to recover the optimal parametric rate $n^{-1}\big(d_{\mathscr{F}}+\log(1/\delta)\big)$ scaled by the appropriate noise term. Stated in its most general form, the burn-in term (18) can be somewhat hard to parse. The next corollary shows that in the case $\eta = 1$ our burn-in coincides with the familiar requirement that the (effective) sample size exceeds the number of degrees of freedom. To simplify matters further we now specialize our result to realizable bounded linear regression. Here, one can think of this burn-in as requiring the empirical covariance matrix of the X-process to be invertible with high probability.[4]

Corollary 3.2 (Realizable Linear Regression)**.** *Fix a failure probability $\delta \in (0,1)$, a covariate bound $B_X \in (0,\infty)$ and a noise bound $B_W \in (0,\infty)$ and let $\mathsf{X} = \mathbb{R}^d$ and $\mathsf{Y} = \mathbb{R}$. Suppose that k divides $n/2$ and that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and satisfies $Y_i = \langle \beta_\star, X_i\rangle + W_i$ for $i \in [n]$. Suppose further that:*

1. *$X_{1:n}$ is bounded $|\langle v, X_i\rangle| \leq B_X, \forall i \in [n]$ and $v \in \mathbb{R}^d$ with $\|v\| = 1$; and*
2. *$W_{1:n}$ is a bounded martingale difference sequence—$\mathbf{E}[W_i|X_{1:i}] = 0$ and $|W_i| \leq B_W$, $\forall i \in [n]$.*

There exist universal positive constants c_1 and c_2 such that if

$$\frac{n}{k} \geq c_1 \left(B_X/\sqrt{\lambda_{\min}(\mathbf{E}XX^\mathsf{T})}\right)^{3+1/2} \left(\frac{kB_W^2}{\mathbf{V}(W)}\right) \times (d+\log(1/\delta)) \quad \text{and} \quad k\beta^{-1}(k) \geq n\delta^{-1} \tag{24}$$

we have that:

$$\|\widehat{f} - f_\star\|_{L^2}^2 \leq c_2 \mathbf{V}(W) \left(\frac{d+\log(1/\delta)}{n}\right). \tag{25}$$

3.1. Further Comparison to Related Work

In terms of technical development, this work is most closely related to the work on iid learning in sub-Gaussian classes by Lecué & Mendelson (2013) and the result for misspecified (agnostic) dependent linear regression by Ziemann et al. (2023b)—which we generalize to more general function classes at the cost of more stringent moment assumptions. Returning to Lecué & Mendelson (2013), and beside the fact that they work with independent data, the biggest difference is in how we deal with the multiplier process. We employ chaining with a mixed tail (Dirksen, 2015), instead of a single tail. On a practical level, the advantage of the mixed tail result is that it allows us to push the dependence on, mixing, L (the norm equivalence parameter in Theorem 3.1) and any higher order norms into the burn-in. Crucially, we make the observation that chaining with a mixed tail allows us to work with weaker norm relations ($\eta < 1$ in Definition 2.1). We do not require equivalence of norms but rather a weaker notion of topological equivalence. Such equivalences hold in significantly wider generality than the sub-Gaussian class assumption as we show in Section 4 below. In particular we are able to handle smoothness classes in Proposition 4.1, which cannot be covered in the baseline sub-Gaussian class framework. Another advantage of this approach is that it allows to relegate the parameter L to a higher order term, which appears multiplicatively instead of additively in the bound by Lecué & Mendelson (2013). This is important in order to achieve the correct scaling with temporal dependency as there are typically no obvious bounds on this parameter other than in terms of the block-length k. Hence, if our dependence on L were multiplicative instead of additive it would thereby re-introduce the sample-size deflation we sought to sidestep. Again, it is the invocation of the mixed-tail chaining result of Dirksen (2015) that allows for this.

Another closely related line of work studies parameter identification in auto-regressive models (for an overview, see Tsiamis et al., 2023; Ziemann et al., 2023a). When the noise model is strictly realizable—the variables $W_{1:n}$ form a martingale difference sequence with respect to the filtration generated by $X_{1:n}$—parameter identification is possible at the iid rate even in the absence of mixing (Simchowitz et al., 2018; Faradonbeh et al., 2018; Sarkar & Rakhlin, 2019; Kowshik et al., 2021). Our results do not cover the mixing-free regime as we consider the agnostic setting in which self-normalized martingale arguments (Peña et al., 2009; Abbasi-Yadkori et al., 2011) are not available. We consider providing a unified analysis of the martingale and mixing situations an interesting future direction.

More generally, several authors have considered learning under various weak dependency notions. Kuznetsov & Mohri (2017) give generalization bounds in a more general setting using the same blocking technique—due to Yu (1994)—used here. Statements similar in spirit can also be found in e.g., Steinwart & Christmann (2009), Duchi et al. (2012) and most recently Roy et al. (2021). However, they all suffer the dependency deflation discussed above and in our introduction (Section 1). We also note that Ziemann & Tu (2022) and Maurer (2023) obtain rates—similar to ours here—that relegate mixing times into additive burn-in factors. On the one hand, the work of Ziemann & Tu (2022) operates at a similar level of generality when it comes to hypothesis classes and also relies on the square loss function but requires a stringent realizability assumption to be applicable. Moreover, both our noise term and our complexity parameter are sharper than theirs. On the other hand, the work of

[4] A small caveat to this remark is that the factor $\frac{kB_W^2}{\mathbf{V}(W)}$ in (24) arises from the multiplier process: it is the cost of having $\mathbf{V}(W)$ appear in (25) instead of kB_W^2.

Maurer (2023) operates at a higher level of generality than us, but does not seem to be able to reproduce sharp rates when specialized to our situation.

4. Examples of Weakly sub-Gaussian Classes

We conclude by collecting a few examples of weakly sub-Gaussian classes (Definition 2.1). Arguably the most compelling example identified in the present manuscript are smoothness classes, which are not covered even in the iid setting by Lecué & Mendelson (2013).

Proposition 4.1 (Smoothness Classes)**.** *Let X be a measurable, open, connected and bounded subset of $\mathbb{R}^d$ with Lipschitz boundary and let $\mathscr{F}$ be a set of uniformly bounded functions $f : \mathsf{X} \to \mathbb{R}$. Fix an integer $s \in \mathbb{N}$ and suppose that there exists a constant $C_{\mathscr{F}}$ such that $\sum_{|\alpha| \leq s} \|D^\alpha f\|_{L^\infty} \leq C_{\mathscr{F}}$.[5] Suppose further that the distribution of the covariates P_X has density μ_{X} with respect to the Lebesque measure and that there exists $\underline{\mu}, \overline{\mu} \in \mathbb{R}_+$ such that $\underline{\mu} \leq \mu_{\mathsf{X}} \leq \overline{\mu}$. Under the above hypotheses there exists a positive constant c only depending on X, d and s such that $\mathscr{F}$ is (L, η)-Ψ_∞ with $L = c\overline{\mu}\underline{\mu}^{-\frac{2s}{2s+d}} C_{\mathscr{F}}^{\frac{d}{2s+d}}$ and $\eta = \frac{2s}{2s+d}$.*

Proof. The result for P_X equal to the (normalized) Lebesque measure is immediate by the main result of (Nirenberg, 1959) instantiated to the correct smoothness class. The general case follows by our hypothesis that P_X is equivalent to the Lebesque measure. ■

We stress that the quantities L and η only appear in the burn-in of Theorem 3.1. In other words, Theorem 3.1 provides sharp rates almost universally, or at least as long as the hypothesis class is sufficiently smooth and bounded (although the latter can be relaxed). However, one important caveat is that this burn-in can be exponentially large (curse of dimensionality) unless the class is sufficently smooth: s is proportional to d above.

Our next example relies on smoothness in parameter space instead of smoothness in terms of inputs.

Proposition 4.2. *Fix an open parameter set $\mathsf{M} \subset \mathbb{R}^{d_{\mathscr{F}}}$ equipped with the Euclidean norm $\|\cdot\|$. Consider a function $\phi : \mathsf{X} \times \mathsf{M} \to \mathbb{R}$ that generates a parametric class of functions $\mathscr{F} = \{\phi(\cdot; \theta) \mid \theta \in \mathsf{M}\}$. Define $\mathsf{M}_\star \triangleq \operatorname{argmin}_{\theta \in \mathsf{M}} \mathbf{E}(\phi(X, \theta) - Y)^2$ to be the set of population risk minimizers. Suppose that:*

(i) for $a, b \in \mathbb{R}_+$, the estimation error functional of the model $\mathscr{F}$ is (a, b)-sharp, that is: $\forall \theta \in \mathsf{M}$ there exists $\theta_\star \in \mathsf{M}_\star$ such that $ab^{-1}\|\theta - \theta_\star\| \leq \left(\mathbf{E}(\phi(X, \theta) - \phi(X, \theta_\star))^2\right)^b$;

(ii) the partial gradient $\nabla_\theta \phi(x, \theta)$ exists and is uniformly norm-bounded by $C > 0$ for all $(x, \theta) \in \mathsf{X} \times \mathsf{M}$.

Then $\mathscr{F} - \{f_\star\}$ is $(Cba^{-1}, 2b)$-Ψ_∞.

The sharpness condition (i) in Proposition 4.2 is standard in optimization (see e.g Roulet & d'Aspremont, 2017). This condition holds somewhat generically (Łojasiewicz, 1993), but the exact constants a and b are typically difficult to obtain. Fortunately, downstream use of Proposition 4.2 only relies on these constants in the burn-in.

Proof. By the mean value form of Taylor's Theorem and Cauchy-Schwarz we write for fixed $x \in \mathsf{X}$:

$$|\phi(x; \theta) - \phi(x; \theta_\star)| = |\langle \nabla_\theta \phi(x, \tilde{\theta}), \theta - \theta_\star \rangle| \leq \|\nabla_\theta \phi(x, \tilde{\theta})\| \|\theta - \theta_\star\| \leq C\|\theta - \theta_\star\| \quad (26)$$

for some $\tilde{\theta} \in [\theta, \theta_\star]$. Consequently by our sharpness hypothesis and by optimizing over the left hand side of (26) we have that for some $\theta_\star \in \mathsf{M}_\star$ and every $\theta \in \mathsf{M}$:

$$\sup_{x \in \mathsf{X}} \|\phi(x, \theta) - \phi(x, \theta_\star)\| \leq \frac{Cb}{a} \left(\mathbf{E}(\phi(X, \theta) - \phi(X, \theta_\star))^2\right)^b \quad (27)$$

Equivalently, $\|f\|_{L^\infty} \leq Cba^{-1}\|f\|_{L^2}^{2b}$ for every $f \in \mathscr{F} - \{f_\star\}$ as per requirement. ■

There is also a more direct argument that easily covers linear functionals on $\mathbb{R}^d$.

Proposition 4.3. *Let X be a sub-Gaussian random variable taking values in $\mathbb{R}^d$ and let $\mathscr{F}$ be the class of linear functionals on $\mathbb{R}^d$. Suppose that $\lambda_{\min}\left(\mathbf{E}XX^\mathsf{T}\right) > 0$. Then $\mathscr{F}$ is $(L, 1)$-Ψ_2 with $L = \sup_{v \in \mathbb{R}^d : \|v\| = 1} \frac{\|\langle v, X \rangle\|_{\Psi_2}}{\|\langle v, X \rangle\|_{L^2}}$.*

Proof. The only observation we need to make is that it suffices to prove the result for $\|v\| = 1$ by homogeneity. The result is then immediate by construction. ■

Analogously, finite hypothesis classes are also covered.

Proposition 4.4. *Let $\mathscr{F}$ be a finite subset of L_{Ψ_2}. Then $\mathscr{F}$ is $(L, 1)$-Ψ_2 with $L = \max_{f \in \mathscr{F}} \frac{\|f\|_{\Psi_2}}{\|f\|_{L^2}}$.*

Proof. The result is immediate since the maximum in the quantity L above is achieved since $|\mathscr{F}| < \infty$. ■

Acknowledgements

Ingvar Ziemann is supported by a Swedish Research Council international postdoc grant. Nikolai Matni is supported in part by NSF award CPS-2038873, NSF award SLES-2331880 and NSF CAREER award ECCS-2045834. George J. Pappas acknowledges support from NSF award EnCORE-2217062.

[5] Summation over D^α uses multi-index notation—the sum is over all partial derivative operators of order less than or equal to s.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Audibert, J.-Y. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20, 2007.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Bernstein, S. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97:1–59, 1927.

Dirksen, S. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20:1 – 29, 2015. doi: 10.1214/EJP.v20-3760. URL https://doi.org/10.1214/EJP.v20-3760.

Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. I. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Einmahl, U. and Li, D. Characterization of lil behavior in banach space. *Transactions of the American Mathematical Society*, 360(12):6677–6693, 2008.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

Kowshik, S., Nagaraj, D., Jain, P., and Netrapalli, P. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8518–8531, 2021.

Kuznetsov, V. and Mohri, M. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106 (1):93–117, 2017.

Lecué, G. and Mendelson, S. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

Liang, T., Rakhlin, A., and Sridharan, K. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pp. 1260–1285. PMLR, 2015.

Łojasiewicz, S. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pp. 1575–1595, 1993.

Maurer, A. Generalization for slowly mixing processes. *arXiv preprint arXiv:2305.00977*, 2023.

Maurer, A. and Pontil, M. Concentration inequalities under sub-gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34: 7588–7597, 2021.

Mendelson, S. Learning without concentration. In *Conference on Learning Theory*, pp. 25–39. PMLR, 2014.

Mendelson, S. An unrestricted learning procedure. *J. ACM*, 66(6), nov 2019. ISSN 0004-5411. doi: 10.1145/3361699. URL https://doi.org/10.1145/3361699.

Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

Mohri, M. and Rostamizadeh, A. Rademacher complexity bounds for non-i.i.d. processes. *Advances in Neural Information Processing Systems*, 21, 2008.

Nirenberg, L. On elliptic partial differential equations. *Annali della Scuola Normale Superiore di Pisa-Scienze Fisiche e Matematiche*, 13(2):115–162, 1959.

Oliveira, R. I. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016.

Peña, V. H., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

Roulet, V. and d'Aspremont, A. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.

Roy, A., Balasubramanian, K., and Erdogdu, M. A. On empirical risk minimization with dependent and heavy-tailed data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Sarkar, T. and Rakhlin, A. Near Optimal Finite Time Identification of Arbitrary Linear Dynamical Systems. In *International Conference on Machine Learning*, pp. 5610–5618, 2019.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pp. 439–473. PMLR, 2018.

Steinwart, I. and Christmann, A. Fast learning from non-i.i.d. observations. *Advances in Neural Information Processing Systems*, 22, 2009.

Tsiamis, A., Ziemann, I., Matni, N., and Pappas, G. J. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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

Ziemann, I. and Tu, S. Learning with little mixing. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4626–4637, 2022.

Ziemann, I., Tsiamis, A., Lee, B., Jedra, Y., Matni, N., and Pappas, G. J. A tutorial on the non-asymptotic theory of system identification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 8921–8939. IEEE, 2023a.

Ziemann, I., Tu, S., Pappas, G. J., and Matni, N. The noise level in linear regression with dependent data. *to appear in the proceedings of Advances in Neural Information Processing Systems and arXiv:2305.11165*, 2023b.

Contents

A. Summary

In this work, we obtain instance-optimal convergence rates for learning with the square loss function and dependent data. We overcome the typical deflation, by the mixing time, of the sample size. The main technical step to arrive at this result is a refined analysis of the multiplier process (8) via mixed tail generic chaining that is suitable for dependent, β-mixing, random variables. Indeed, the leading order term of our main result, Theorem 3.1, does not directly depend on any mixing-time type quantities. It mimics the correct asymptotic rate and scales solely in terms of the statistics of order $2q$ of the process at hand (where typically $q = 1 + o(1)$). Finally, our result also allows us to evaluate said multiplier process for a wider range of hypothesis classes. Typically, sharp closed form expressions for this process are only available for linear functionals, covered in the iid setting by Lecué & Mendelson (2013) and Oliveira (2016), and extended to the β-mixing setting by Ziemann et al. (2023b). By contrast, since our result relies on a weaker notion of topological equivalence, it is applicable to more general classes, such as smoothness classes (Proposition 4.1) and parametric classes with sufficiently regular parameterization (Proposition 4.2).

B. Properties of Ψ_p- and L^p-Norms

We begin with an elementary property.

Lemma B.1. *For every two random variables $Z, Z' \in L_{\Psi_p}$ we have that:*

$$\|\langle Z, Z'\rangle\|_{\Psi_{p/2}} \leq 2^{2/p}\|Z\|_{\Psi_p}\|Z'\|_{\Psi_p}. \tag{28}$$

Proof. We compute:

$$\begin{aligned}
\|\langle Z, Z'\rangle\|_{\Psi_{p/2}} &= \sup_{m\geq 1}\frac{\|\langle Z, Z'\rangle\|_{L^m}}{m^{2/p}} \\
&= \sup_{m\geq 1}\frac{(\mathbf{E}|\langle Z, Z'\rangle|^m)^{1/m}}{m^{2/p}} \\
&\leq \sup_{m\geq 1}\frac{(\mathbf{E}\|Z\|^m\|Z'\|^m)^{1/m}}{m^{2/p}} && (\langle\cdot,\cdot\rangle\text{-Cauchy-Schwarz}) \\
&\leq \sup_{m\geq 1}\frac{(\mathbf{E}\|Z\|^{2m}\mathbf{E}\|Z'\|^{2m})^{1/2m}}{m^{2/p}} && (L^2\text{-Cauchy-Schwarz}) \\
&\leq 2^{2/p}\sup_{m\geq 1}\frac{(\mathbf{E}\|Z\|^{2m})^{1/2m}}{(2m)^{1/p}}\sup_{m\geq 1}\frac{(\mathbf{E}\|Z'\|^{2m})^{1/2m}}{(2m)^{1/p}} \\
&\leq 2^{2/p}\|Z\|_{\Psi_p}\|Z'\|_{\Psi_p}, && (\{m\geq 1\}\subset\{2m\geq 1\})
\end{aligned} \tag{29}$$

as was required. ■

B.1. Proof of Lemma 2.1

Lemma 2.1 (Ψ_p-Bernstein MGF Bound)**.** *Fix a random variable Z and $p \in [1, \infty]$ such that $\mathbf{E}Z \leq 0$ and $\|Z\|_{\Psi_p} < \infty$. Let q and q' be Hölder conjugates and suppose that $\lambda \in [0, 1/(q'e)^{1/p}\|Z\|_{\Psi_p}]$. We have that:*

$$\mathbf{E}\exp(\lambda Z) \leq \exp\left(\frac{\frac{\lambda^2}{2}\left(\mathbf{E}(Z)^{2q}\right)^{1/q}}{1-\lambda(q'e)^{1/p}\|Z\|_{\Psi_p}}\right). \tag{10}$$

Proof. The idea of the proof is very much the same as that of the standard Bernstein MGF bound but with the modification made in Maurer & Pontil (2021) by which the L^∞ norm is replaced by a Ψ_p-norm. We begin by expanding the exponential function:

$$\begin{aligned}\mathbf{E}\exp(\lambda Z) &= \mathbf{E}\left[\sum_{m=0}^{\infty}\frac{(\lambda Z)^m}{m!}\right]\\ &\leq 1+\sum_{m=0}^{\infty}\frac{\mathbf{E}\left[(\lambda Z)^2(\lambda Z)^m\right]}{(m+2)!} && (\mathbf{E}Z\leq 0)\\ &\leq 1+\lambda^2\left(\mathbf{E}(Z)^{2q}\right)^{1/q}\sum_{m=0}^{\infty}\frac{\left(\mathbf{E}\left[|\lambda Z|^{mq'}\right]\right)^{1/q'}}{(m+2)!}. && (\text{Hölder's Ineq.})\end{aligned} \tag{30}$$

We next have:

$$\begin{aligned}\left(\mathbf{E}\left[|Z|^{mq'}\right]\right)^{1/q'} &= \|Z\|_{L^{mq'}}^m\\ &\leq (mq')^{m/p}\|Z\|_{\Psi_p}^m && (\text{df. of } \Psi_p)\\ &\leq (m!)^{1/p}(q'e)^{m/p}\|Z\|_{\Psi_p}^m. && (\text{Stirling's Approximation})\end{aligned} \tag{31}$$

Upon combining (30) with (31) we arrive at

$$\begin{aligned}&\mathbf{E}\exp(\lambda Z)\\ &\leq 1+\lambda^2\left(\mathbf{E}(Z)^{2q}\right)^{1/q}\sum_{m=0}^{\infty}\frac{(m!)^{1/p}\lambda^m(q'e)^{m/p}\|Z\|_{\Psi_p}^m}{(m+2)!}\\ &\leq 1+\frac{\lambda^2\left(\mathbf{E}(Z)^{2q}\right)^{1/q}}{2}\sum_{m=0}^{\infty}\left(\lambda(q'e)^{1/p}\|Z\|_{\Psi_p}\right)^m && \left(p\in[1,\infty], m\in\mathbb{N}\Rightarrow\frac{(m!)^{1/p}}{(m+2)!}\leq\frac{1}{2}\right)\\ &= 1+\frac{\frac{\lambda^2}{2}\left(\mathbf{E}(Z)^{2q}\right)^{1/q}}{1-\lambda(q'e)^{1/p}\|Z\|_{\Psi_p}} && \left(x\in[0,1)\Rightarrow\sum_{m=0}^{\infty}x^m=\frac{1}{1-x}\right)\\ &\leq \exp\left(\frac{\frac{\lambda^2}{2}\left(\mathbf{E}(Z)^{2q}\right)^{1/q}}{1-\lambda(q'e)^{1/p}\|Z\|_{\Psi_p}}\right), && (x\in\mathbb{R}\Rightarrow 1+x\leq e^x)\end{aligned} \tag{32}$$

as per requirement. ■

C. Controlling the Multiplier Process

Lemma 2.2 (Pointwise Control)**.** *Fix two Hölder conjugates q and q'. Suppose that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and k-wise independent where k divides n. For every $g, g'\in L_{\Psi_p}$ and $u\in(0,\infty)$ we have that:*

$$\begin{aligned}\mathbf{P}\Bigg(\sum_{i=1}^{n}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle &\\ > \sqrt{4n\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\{g\}-\{g'\}\right)u} &\\ +4(q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}u\Bigg)\leq 2e^{-u}.&\end{aligned} \tag{13}$$

Proof. First, note that we may assume throughout the proof that $\|g-g'\|_{L^2}>0$, for otherwise the result is trivial. We now begin by applying Lemma 2.1:

$$\begin{aligned}&\mathbf{E}\exp\left(\lambda\sum_{i=1}^{k}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\right)\\ &\leq\exp\left(\frac{\lambda^2k\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\frac{1}{\sqrt{k}\|g-g'\|_{L^2}}\sum_{i=1}^{k}\langle W_i, g(X_i)-g'(X_i)\rangle\right)}{2\left(1-\lambda(q'e)^{2/p}\|\sum_{i=1}^{k}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\|_{\Psi_{p/2}}\right)}\right)\end{aligned} \tag{33}$$

as long as $\lambda < \left((q'e)^{2/p}\| \sum_{i=1}^{k}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\|_{\Psi_{p/2}}\right)^{-1}$. Now, by triangle inequality and Lemma B.1,

$$\lambda(q'e)^{2/p}\left\|\sum_{i=1}^{k}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\right\|_{\Psi_{p/2}} \leq \lambda(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g(X_i)-g'(X_i)\|_{\Psi_p}. \tag{34}$$

Consequently:

$$\begin{aligned}&\mathbf{E}\exp\left(\lambda\sum_{i=1}^{k}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\right)\\ &\leq \exp\left(\frac{\lambda^2 k\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\{g\}-\{g'\}\right)}{2\left(1-\lambda(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}\right)}\right).\end{aligned} \tag{35}$$

Since the process is k-wise independent and mean zero we thus have that:

$$\begin{aligned}&\mathbf{E}\exp\left(\lambda\sum_{i=1}^{n}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\right)\\ &\leq \exp\left(\frac{\lambda^2 n\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\{g\}-\{g'\}\right)}{2\left(1-\lambda(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}\right)}\right).\end{aligned} \tag{36}$$

Hence for every $\lambda \in \left[0, \left(2(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}\right)^{-1}\right) \triangleq \Lambda$ we have:

$$\begin{aligned}&\mathbf{E}\exp\left(\lambda\sum_{i=1}^{n}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\right)\\ &\leq \exp\left(\lambda^2 n\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\{g\}-\{g'\}\right)\right).\end{aligned} \tag{37}$$

Taking the above exponential inequality as a starting point, for a fixed $u \in (0,\infty)$, a Chernoff argument now yields:

$$\begin{aligned}&\mathbf{P}\left(\sum_{i=1}^{n}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle > u\right)\\ &\leq \inf_{\lambda>0}\mathbf{E}\exp\left(-u\lambda+\lambda\sum_{i=1}^{n}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle\right)\\ &\leq \inf_{\lambda\in\Lambda}\left(-\lambda u+\frac{\lambda^2 n\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\{g\}-\{g'\}\right)}{2\left(1-\lambda(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}\right)}\right)\\ &\leq \begin{cases}\exp\left(\frac{-u^2}{4n\left(\|g-g'\|_{L^2}^2\mathbf{V}_{2q}(\{g\}-\{g'\})\right)}\right) & u \leq \frac{\left(n\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\{g\}-\{g'\}\right)\right)}{2(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}},\\ \exp\left(\frac{-u}{4(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}}\right) & \text{otherwise.}\end{cases}\end{aligned} \tag{38}$$

Rescaling and summing the failure probabilities in either case yields:

$$\begin{aligned}\mathbf{P}\Bigg(\sum_{i=1}^{n}(1-\mathbf{E})\langle W_i, g(X_i)-g'(X_i)\rangle &\\ > \sqrt{4n\|g-g'\|_{L^2}^2\mathbf{V}_{2q}\left(\{g\}-\{g'\}\right)u} &\\ + 4(2q'e)^{2/p}k\|W\|_{\Psi_p}\|g-g'\|_{\Psi_p}u\Bigg) &\leq 2e^{-u},\end{aligned} \tag{39}$$

as was required. ■

Let us turn to making Lemma 2.2 uniform. By instantiating Theorem 3.5 of (Dirksen, 2015) combined with the pointwise control of Lemma 2.2, we immediately have the following result.

Lemma C.1 (Corollary of Theorem 3.5 in (Dirksen, 2015))**.** *Fix $\delta \in (0,1), r > 0$ and consider the space $\mathscr{F}_\star \cap rS_{L^2}$ endowed with the metrics*

$$\begin{aligned} d_1(g,g') &= 4(q'e)^{2/p} k \|W\|_{\Psi_p} \|g-g'\|_{\Psi_p}, \\ d_2(g,g') &= \sqrt{4n\left(\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap rS_{L^2}\right)\right)} \|g-g'\|_{L^2}, \end{aligned} \tag{40}$$

and denote the corresponding diameters by $\Delta_i(\mathscr{F}_\star \cap rS_{L^2}) \triangleq \sup_{g,g' \in \mathscr{F}_\star \cap rS_{L^2}} d_i(g,g'), i \in [2]$. There exist universal positive constants c_1 and c_2 such that with probability at least $1-\delta$:

$$\begin{aligned} \sup_{f \in \mathscr{F}_\star \cap rS_{L^2}} \sum_{i=1}^n (1-\mathbf{E})\langle W_i, f\rangle &\le c_1 \left(\gamma_2(\mathscr{F}_\star \cap rS_{L^2}, d_2) + \gamma_1(\mathscr{F}_\star \cap rS_{L^2}, d_1)\right) \\ &\quad + c_2 \left(\Delta_2(\mathscr{F}_\star \cap rS_{L^2})\sqrt{\log(1/\delta)} + \Delta_1(\mathscr{F}_\star \cap rS_{L^2}) \log(1/\delta)\right). \end{aligned} \tag{41}$$

We now restate and prove the result for the multiplier process.

Theorem 2.1. *Fix a failure probability $\delta \in (0,1)$, a positive scalar $r \in (0,\infty)$, two Hölder conjugates q and q', and a class $\mathscr{F}$. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and k-wise independent where k divides n. There exist universal positive constants c_1, c_2 such that for any $r \in (0,1]$ we have that with probability at least $1-\delta$:*

$$\begin{aligned} &\sup_{f \in \mathscr{F}_\star \cap rS_{L^2}} \frac{1}{rn} \sum_{i=1}^n (1-\mathbf{E})\langle W_i, f\rangle \\ &\qquad \le c_2 \sqrt{\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap rS_{L^2}\right)} \left(\frac{1}{r\sqrt{n}} \right. \\ &\qquad \left. \times \gamma_2(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + \sqrt{\frac{\log(1/\delta)}{n}}\right) \\ &\qquad + c_1 (q'e)^{2/p} Lk \|W\|_{\Psi_p} \\ &\qquad \times \left(\frac{1}{rn}\gamma_\eta(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + \frac{r^{\eta-1}}{n}\log(1/\delta)\right). \end{aligned} \tag{15}$$

Proof. We need to translate the metrics (appearing in Lemma C.1) d_1, d_2 and their diameters into the standard L^2-metric using that the class is (L,η)-Ψ_p. We begin with d_2, which is just a dilated L^2-metric:

$$\begin{aligned} \gamma_2(\mathscr{F}_\star \cap rS_{L^2}, d_2) &= \inf_{\{F_m\}} \sup_{f \in \mathscr{F}_\star \cap rS_{L^2}} \sum_{m=0}^\infty 2^{m/2} d_2(f, F_m) \\ &= \sqrt{4n\left(\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap rS_{L^2}\right)\right)} \gamma_2(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}), \end{aligned} \tag{42}$$

and

$$\Delta_2(\mathscr{F}_\star \cap rS_{L^2}) \le r\sqrt{4n\left(\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap rS_{L^2}\right)\right)}. \tag{43}$$

Turning to the d_1, we have:

$$\begin{aligned} \gamma_1(\mathscr{F}_\star \cap rS_{L^2}, d_1) &= \inf_{\{F_m\}} \sup_{f \in \mathscr{F}_\star \cap rS_{L^2}} \sum_{m=0}^\infty 2^m d_1(f, F_m) \\ &= \inf_{\{F_m\}} \sup_{f \in \mathscr{F}_\star \cap rS_{L^2}} \sum_{m=0}^\infty 2^m 2(q'e)^{2/p} k \|W\|_{\Psi_p} d_{\Psi_p}(f, F_m) \\ &\le 2(q'e)^{2/p} Lk \|W\|_{\Psi_p} \gamma_1(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}^\eta) \\ &\le 2(q'e)^{2/p} Lk \|W\|_{\Psi_p} \gamma_\eta(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}), \end{aligned}$$

where the first inequality uses that $\mathscr{G}$ is (L,η)-Ψ_p and the last inequality uses that $\gamma_1(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}^\eta) \le \gamma_\eta(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2})$ as long as $r \le 1$. Similarly:

$$\Delta_1(\mathscr{F}_\star \cap rS_{L^2}) \le 2(q'e)^{2/p} Lk \|W\|_{\Psi_p} r^\eta. \tag{44}$$

The result follows by substituting the above expressions into the result of (Dirksen, 2015) captured as Lemma C.1. ■

D. Controlling the Quadratic Process

Lemma D.1 (Truncation Accuracy). *Fix $\varepsilon, r > 0$, let $\mathscr{G}$ be (L,η)-Ψ_p, and let $g \in \mathscr{G}$ be such that $\|g\|_{L^2} = r$. For $\tau \in \mathbb{R}_+$, define $g_\tau \triangleq g\mathbf{1}_{\|g\| \le \tau}$. There exists a truncation level τ and a universal positive constant $c > 0$ such that:*

$$\|g\|_{L^2}^2 - \|g_\tau\|_{L^2}^2 \le r^2 \varepsilon \tag{45}$$

and

$$\tau \le Lr^\eta \left(c^{-1} \log\left(\frac{4^{2/p} L^{1/2} r^{2\eta}}{\varepsilon r^4} \right) \right)^{1/p}. \tag{46}$$

Proof. Fix a level $\tau > 0$ to be determined later. For any such level we have that:

$$\|g\|_{L^2}^2 - \|g_\tau\|_{L^2}^2 \le \mathbf{E}[\|g\|^2 \mathbf{1}_{\|g\| > \tau}] \le \sqrt{\mathbf{E}\|g\|^4 \mathbf{P}(\|g\| > \tau)} \le \sqrt{\mathbf{E}\|g\|^4} \exp\left(-c\tau^p / \|g\|_{\Psi_p}^p \right). \tag{47}$$

Hence if we choose $\tau^p = c^{-1}\|g\|_{\Psi_p}^p \log\left(\frac{\sqrt{E\|g\|^4}}{\varepsilon r^2 \mathbf{E}\|g\|^2} \right)$ we have:

$$\|g\|_{L^2}^2 - \|g_\tau\|_{L^2}^2 \le \varepsilon^2. \tag{48}$$

It remains to derive an upper bound on τ. Since $\mathscr{G}$ is (L,η)-Ψ_p and $\|g\|_{L^2} = r$ we have that

$$\begin{aligned} &\|g\|_{\Psi_p} \le L\|g\|_2^\eta = Lr^\eta, \text{ and} \\ &\mathbf{E}\|g\|^4 \le 4^{4/p}\|g\|_{\Psi_p}^4 \le 4^{4/p} L r^{4\eta}. \end{aligned} \tag{49}$$

Hence our choice of τ satisfies:

$$\tau \le Lr^\eta \left(c^{-1} \log\left(\frac{4^{2/p} L^{1/2} r^{2\eta}}{\varepsilon r^4} \right) \right)^{1/p} \tag{50}$$

and so the result has been established. ■

Theorem 2.2 (Lower Uniform Law). *Fix a failure probability $\delta \in (0,1)$, a tolerance $\varepsilon > 0$, a localization radius $r \in (0,1]$, and two Hölder conjugates q and q'. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and k-wise independent where k divides n. There exist a universal positive constant c such that uniformly for all $f \in \mathscr{F}_\star \setminus \{rB_{L^2}\}$ we have that with probability at least $1-\delta$:*

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \|f(X_i)\|^2 \ge r^2(1-\varepsilon^2) \\ - c\Bigg\{ n^{-1/2}\sqrt{k} L^{1+3/4} r^\eta \left(\log\left(\frac{4^{2/p} L}{\varepsilon r} \right) \right)^{1/p} \\ \times \left(\gamma_{\frac{2+6\eta}{4}}(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + r^{\frac{1+3\eta}{4}} \sqrt{\log(1/\delta)} \right) \\ + n^{-1}(q')^{1/p} k r^\eta \left(\log\left(\frac{4^{2/p} L}{\varepsilon r} \right) \right)^{1/p} L^2 \\ \times \left(\gamma_\eta(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + r^\eta \log(1/\delta) \right) \Bigg\}. \end{aligned} \tag{16}$$

Proof. By star-shapedness, we may assume without loss of generality that $f \in \mathscr{F}_\star \cap rS_{L^2}$. Fix τ such that for all such f

$$\|f\|_{L^2}^2 - \|f_\tau\|_{L^2}^2 \leq \varepsilon^2, \tag{51}$$

and note that the existence of such a level is guaranteed by Lemma D.1.

It is then clear that:

$$\frac{1}{n}\sum_{i=1}^n \|f(X_i)\|^2 \geq r^2(1-\varepsilon^2) - \sup_{f\in\mathscr{F}_\star\cap rB_{L^2}} \left\{ \frac{1}{n}\sum_{i=1}^n \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 \right\} \tag{52}$$

and, for well-chosen ε, r, it therefore suffices to control the supremum of empirical process to the right of (52) and we will use Dirksen's theorem again to do so (Dirksen, 2015, Theorem 3.5). A preliminary estimate using k-wise independence, stationarity and Lemma 2.1 gives that for every f, g and admissible λ:

$$\begin{aligned}
&\mathbf{E}\exp\left(\lambda\left[\sum_{i=1}^n \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2\right]\right) \\
&\qquad\leq \exp\left(\frac{\lambda^2 n \mathbf{V}_4\left(\frac{1}{\sqrt{k}}\sum_{i=1}^k \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2\right)}{2\left(1-\lambda(q'e)^{1/p}k\left\|\|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2\right\|_{\Psi_{p/2}}\right)}\right).
\end{aligned} \tag{53}$$

This is almost the exponential inequality we need, but we will want increment conditions for the above empirical process in terms Ψ_p and L^2.

The increment condition in Ψ_p is simple. We observe that for any two f, g and any x in their domain:

$$\|f_\tau(x)\|^2 - \|g_\tau(x)\|^2 = \langle (f_\tau + g_\tau)(x), (f_\tau + g_\tau)(x)\rangle. \tag{54}$$

Consequently by Lemma B.1 and τ-truncation:

$$\|\|f_\tau(X)\|^2 - \|g_\tau(X)\|^2\|_{\Psi_{p/2}} \leq 2^{2/p}\|f_\tau + g_\tau\|_{\Psi_p}\|f_\tau - g_\tau\|_{\Psi_p} \leq 2^{1+2/p}\tau\|f-g\|_{\Psi_p}. \tag{55}$$

A centering argument thus gives:

$$\|\|f_\tau(X)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X)\|^2 + \|g_\tau\|_{L^2}^2\|_{\Psi_{p/2}} \leq 2^{2+2/p}\tau\|f-g\|_{\Psi_p} \tag{56}$$

wherefore we set

$$d_1(f,g) \triangleq (q'e)^{1/p}k2^{2+2/p}\tau\|f-g\|_{\Psi_p}. \tag{57}$$

Let us next address the variance term:

$$\begin{aligned}
&\mathbf{V}_4\left(\frac{1}{\sqrt{k}}\sum_{i=1}^k \|f_\tau(X_i)\|^2 - \|g_\tau(X_i)\|^2\right) \\
&= \mathbf{V}_4\left(\frac{1}{\sqrt{k}}\sum_{i=1}^k \langle f_\tau(X_i) + g_\tau(X_i), f_\tau(X_i) - g_\tau(X_i)\rangle\right) && \text{(use (54))} \\
&\leq \sqrt{\mathbf{E}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^k \langle f_\tau(X_i) + g_\tau(X_i), f_\tau(X_i) - g_\tau(X_i)\rangle\right)^4} && (\mathbf{V}_4[\cdot] \leq \sqrt{\mathbf{E}[|\cdot|^4]}) \\
&\leq k\sqrt{\mathbf{E}\left(\langle f_\tau(X) + g_\tau(X), f_\tau(X) - g_\tau(X)\rangle\right)^4} && \text{(Cauchy-Schwarz)} \\
&\leq 4k\tau^2\|f-g\|_{L^4}^2 \triangleq d_2^2(f,g). && (\tau\text{-boundedness and Cauchy-Schwarz})
\end{aligned} \tag{58}$$

With d_1, d_2 as in (57) and (58), we can now estimate (53) as:

$$\mathbf{E}\exp\left(\lambda\left[\sum_{i=1}^n \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2\right]\right) \leq \exp\left(\frac{\lambda^2 n d_2^2(f,g)}{2\left(1-\lambda d_1(f,g)\right)}\right). \tag{59}$$

We thus obtain the probability estimate ($u > 0$):

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2 > c'\sqrt{u/n}d_2(f,g) + c(u/n)d_1(f,g)\right) \leq 2e^{-u} \tag{60}$$

for two universal positive constants c, c'. After defining (normalizing) for some universal positive constants c_1, c_2:

$$\tilde{d}_1(f,g) \triangleq cn^{-1}d_1(f,g) = c_1 n^{-1}(q'e)^{1/p}k2^{2+2/p}\tau\|f-g\|_{\Psi_p}, \tag{61}$$

$$\tilde{d}_2(f,g) \triangleq c'n^{-1/2}d_2(f,g) = c_2 n^{-1/2}\left(k\tau^2\|f-g\|_{L^4}^2\right), \tag{62}$$

we notice that (60) is consistent with the mixed tail generic chaining condition in Dirksen (2015, Equation 12) for metrics $\tilde{d}_1, \tilde{d}_2$. Consequently, by Theorem 3.5 in (Dirksen, 2015) we have that:

$$\sup_{f\in\mathscr{F}_\star\cap rB_{L^2}}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2\right\} \leq c_3(\gamma_2(\mathscr{F}_\star\cap rB_{L^2},\tilde{d}_2) + \sqrt{u}\Delta_{\tilde{d}_2}(\mathscr{F}_\star\cap rB_{L^2})) + c_4(\gamma_1(\mathscr{F}_\star\cap rB_{L^2},\tilde{d}_1) + u\Delta_{\tilde{d}_1}(\mathscr{F}_\star\cap rB_{L^2})) \tag{63}$$

for two universal positive constants c_3, c_4.

To finish the proof, we turn to relating the quantities γ and Δ in terms of problem data. We have (recalling (61)):

$$\Delta_{\tilde{d}_1}(\mathscr{F}_\star\cap rB_{L^2}) = cn^{-1}(q'e)^{1/p}k2^{2+2/p}\tau\Delta_{\Psi_p}(\mathscr{F}_\star\cap rB_{L^2}) \leq cn^{-1}(q'e)^{1/p}k2^{2+2/p}\tau Lr^\eta \tag{64}$$

and also (recalling (62)):

$$\begin{aligned}\Delta_{\tilde{d}_2}(\mathscr{F}_\star\cap rB_{L^2}) &\leq c'n^{-1/2}2\sqrt{k}\tau\Delta_{d_{L^2}}(\mathscr{F}_\star\cap rB_{L^4}) \\ &\leq c'n^{-1/2}2\sqrt{k}\tau L^{3/4}r^{\frac{1+3\eta}{4}},\end{aligned} \tag{65}$$

where we used Cauchy-Schwarz and the class assumption in the last step to control the L^4 norm by the L^2 norm.

As for γ-functionals, we have:

$$\gamma_1(\mathscr{F}_\star\cap rS_{L^2},\tilde{d}_1) \leq cn^{-1}(q'e)^{1/p}k2^{2+2/p}\tau L\gamma_\eta(\mathscr{F}_\star\cap rS_{L^2},d_{L^2}) \tag{66}$$

and

$$\gamma_2(\mathscr{F}_\star\cap rS_{L^2},\tilde{d}_2) \leq c'n^{-1/2}2\sqrt{k}\tau L^{3/4}\gamma_{\frac{2+6\eta}{4}}(\mathscr{F}_\star\cap rS_{L^2},d_{L^2}). \tag{67}$$

Putting everything together we thus obtain that:

$$\begin{aligned}&\sup_{f\in\mathscr{F}_\star\cap rB_{L^2}}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2\right\} \\ &\leq CL^{3/4}n^{-1/2}\sqrt{k}\tau\left(\gamma_{\frac{2+6\eta}{4}}(\mathscr{F}_\star\cap rS_{L^2},d_{L^2}) + r^{\frac{1+3\eta}{4}}\sqrt{\log(1/\delta)}\right) \\ &\quad + C'2^{C''/p}n^{-1}(q')^{1/p}k\tau L\left(\gamma_\eta(\mathscr{F}_\star\cap rS_{L^2},d_{L^2}) + r^\eta\log(1/\delta)\right) \\ &= Cn^{-1/2}\sqrt{k}L^{1+3/4}r^\eta\left(c^{-1}\log\left(\frac{4^{2/p}L}{\varepsilon r}\right)\right)^{1/p}\left(\gamma_{\frac{2+6\eta}{4}}(\mathscr{F}_\star\cap rS_{L^2},d_{L^2}) + r^{\frac{1+3\eta}{4}}\sqrt{\log(1/\delta)}\right) \\ &\quad + C'2^{C''/p}n^{-1}(q')^{1/p}kr^\eta\left(c^{-1}\log\left(\frac{4^{2/p}L}{\varepsilon r}\right)\right)^{1/p}L^2\left(\gamma_\eta(\mathscr{F}_\star\cap rS_{L^2},d_{L^2}) + r^\eta\log(1/\delta)\right),\end{aligned}$$

for universal positive constants C, C', C''. Since $p \geq 1$ we may replace all the terms containing upper-case universal constants by a single universal constant as in the theorem statement. ■

E. Results for Mixing Empirical Processes

E.1. Blocking

Recall that we partition $[n]$ into $2m$ consecutive intervals, denoted a_j for $j \in [2m]$, so that $\sum_{j=1}^{2m} |a_j| = n$. Denote further by O (resp. by E) the union of the oddly (resp. evenly) indexed subsets of $[n]$. We further abuse notation by writing $\beta_Z(a_i) = \beta_Z(|a_i|)$ in the sequel.

We split the process $Z_{1:n}$ as:

$$Z^o_{1:|O|} \triangleq (Z_{a_1}, \dots, Z_{a_{2m-1}}), \quad Z^e_{1:|E|} \triangleq (Z_{a_2}, \dots, Z_{a_{2m}}). \tag{68}$$

Let $\tilde{Z}^o_{1:|O|}$ and $\tilde{Z}^e_{1:|E|}$ be blockwise decoupled versions of (68). That is we posit that $\tilde{Z}^o_{1:|O|} \sim \mathsf{P}_{\tilde{Z}^o_{1:|O|}}$ and $\tilde{Z}^e_{1:|E|} \sim \mathsf{P}_{\tilde{Z}^e_{1:|E|}}$, where:

$$\mathsf{P}_{\tilde{Z}^o_{1:|O|}} \triangleq \mathsf{P}_{Z_{a_1}} \otimes \mathsf{P}_{Z_{a_3}} \otimes \cdots \otimes \mathsf{P}_{Z_{a_{2m-1}}} \quad \text{and} \quad \mathsf{P}_{\tilde{Z}^e_{1:|E|}} \triangleq \mathsf{P}_{Z_{a_2}} \otimes \mathsf{P}_{Z_{a_4}} \otimes \cdots \otimes \mathsf{P}_{Z_{a_{2m}}}. \tag{69}$$

The process $\tilde{Z}_{1:n}$ with the same marginals as $\tilde{Z}^o_{1:|O|}$ and $\tilde{Z}^e_{1:|E|}$ is said to be the decoupled version of $Z_{1:n}$. To be clear: $\mathsf{P}_{\tilde{Z}_{1:n}} \triangleq \mathsf{P}_{Z_{a_1}} \otimes \mathsf{P}_{Z_{a_2}} \otimes \cdots \otimes \mathsf{P}_{Z_{a_{2m}}}$, so that $\tilde{Z}^o_{1:|O|}$ and $\tilde{Z}^e_{1:|E|}$ are alternatingly embedded in $\tilde{Z}_{1:n}$. The following result is key—by skipping every other block, $\tilde{Z}_{1:n}$ may be used in place of $Z_{1:n}$ for evaluating bounded scalar functionals, such as probabilities of measurable events, at the cost of an additive mixing-related term.

Proposition E.1 (Lemma 2.6 in (Yu, 1994); Proposition 1 in (Kuznetsov & Mohri, 2017))**.** *Fix a β-mixing process $Z_{1:n}$ and let $\tilde{Z}_{1:n}$ be its decoupled version. For any measurable function f of $Z^o_{1:|O|}$ (resp. g of $Z^e_{1:|E|}$) with joint range $[0,1]$ we have that:*

$$\begin{aligned} |\mathbf{E}(f(Z^o_{1:|O|})) - \mathbf{E}(f(\tilde{Z}^o_{1:|O|}))| &\le \sum_{i \in E \setminus \{2m\}} \beta_Z(a_i), \\ |\mathbf{E}(g(Z^e_{1:|E|})) - \mathbf{E}(g(\tilde{Z}^e_{1:|E|}))| &\le \sum_{i \in O \setminus \{1\}} \beta_Z(a_i). \end{aligned} \tag{70}$$

E.2. Controlling Empirical Processes for β-Mixing Data

Applying Proposition E.1 to Theorem 2.1 and Theorem 2.2 yields the desired control of the multiplier and quadratic processes also for β-mixing data.

Proposition E.2. *Fix a failure probability $\delta \in (0,1)$, a positive scalar $r \in (0,\infty)$, two Hölder conjugates q and q', and a class $\mathscr{F}$. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and β-mixing and suppose further that $k \in \mathbb{N}$ divides $n/2$. There exist universal positive constants c_1, c_2 such that for any $r \in (0,1]$ we have that with probability at least $1 - \delta - \frac{n}{k}\beta(k)$:*

$$\begin{aligned} &\sup_{f \in \mathscr{F}_\star \cap r S_{L^2}} \frac{1}{rn} \sum_{i=1}^{n} (1 - \mathbf{E}) \langle W_i, f \rangle \\ &\qquad \le c_2 \sqrt{\mathbf{V}_{2q}\left(\mathscr{F}_\star \cap r S_{L^2}\right)} \left(\frac{1}{r\sqrt{n}} \gamma_2(\mathscr{F}_\star \cap r S_{L^2}, d_{L^2}) + \sqrt{\frac{\log(1/\delta)}{n}} \right) \\ &\qquad\qquad + c_1 (q'e)^{2/p} L k \|W\|_{\Psi_p} \left(\frac{1}{rn} \gamma_\eta(\mathscr{F}_\star \cap r S_{L^2}, d_{L^2}) + \frac{r^{\eta-1}}{n} \log(1/\delta) \right) \end{aligned} \tag{71}$$

Proposition E.3. *Fix a failure probability $\delta \in (0,1)$, a tolerance $\varepsilon > 0$, a localization radius $r \in (0,1]$, and two Hölder conjugates q and q'. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and β-mixing and suppose further that $k \in \mathbb{N}$ divides $n/2$. There exists a universal positive constant c such that uniformly for all $f \in \mathscr{F}_\star \cap r S_{L^2}$ we have that with probability at least $1 - \delta - \frac{n}{k}\beta(k)$:*

$$\frac{1}{n}\sum_{i=1}^{n}\|f(X_i)\|^2 \geq r^2(1-\varepsilon^2)$$
$$- c\Bigg\{ n^{-1/2}\sqrt{k}L^{1+3/4}r^{\eta}\left(\log\left(\frac{4^{2/p}L}{\varepsilon r}\right)\right)^{1/p}\left(\gamma_{\frac{2+6\eta}{4}}(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + r^{\frac{1+3\eta}{4}}\sqrt{\log(1/\delta)}\right)$$
$$+ n^{-1}(q')^{1/p}kr^{\eta}\left(\log\left(\frac{4^{2/p}L}{\varepsilon r}\right)\right)^{1/p} L^2\left(\gamma_{\eta}(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) + r^{\eta}\log(1/\delta)\right)\Bigg\}. \quad (72)$$

F. Finishing the Proof of Theorem 3.1

Before we finish the proof of the main result, let us first make formal the justification for the introduction of quadratic and multiplier processes in Section 1.1. The following lemma bounds the excess risk of empirical risk minimizer in terms of these.

Lemma F.1 (Localized Basic Inequality)**.** *Suppose that either (1) $\mathscr{F}$ is convex or (2) $\mathscr{F}$ is realizable. For every $r > 0$ we have that:*

$$\|\widehat{f} - f_\star\|_{L^2}^2 \leq r^2 + \frac{1}{r^2}\left(\sup_{g\in\mathscr{F}_\star\cap rS_{L^2}} M_n(g)\right)^2 + \sup_{g\in\mathscr{F}_\star} Q_n(g). \quad (73)$$

Proof. We begin by observing that the optimality of $\widehat{f}$ to (2) yields the basic inequality:

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{f}(X_i) - f_\star(X_i)\|^2 \leq \frac{2}{n}\sum_{i=1}^{n}\langle W_i, (\widehat{f} - f_\star)(X_i)\rangle. \quad (74)$$

If $\mathscr{F}$ is convex, we have that $\mathbf{E}\langle W_i, (f - f_\star)(X_i)\rangle \leq 0$ for every f (by optimality of $f_\star$ to the population objective). If instead $\mathscr{F}$ is realizable the same holds true but with equality. Hence, in either case:

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{f}(X_i) - f_\star(X_i)\|^2 \leq \frac{2}{n}\sum_{i=1}^{n}(1-\mathbf{E}')\langle W_i, (\widehat{f} - f_\star)(X_i)\rangle \quad (75)$$

where $\mathbf{E}'$ denotes expectation with respect to a fresh copy of randomness (independent of the data used to construct $\widehat{f}$).

Consequently we also have that:

$$\|\widehat{f} - f_\star\|_{L^2}^2 = \frac{(1+\varepsilon)}{n}\sum_{i=1}^{n}\|\widehat{f}(X_i) - f_\star(X_i)\|^2 + \|\widehat{f} - f_\star\|_{L^2}^2 - \frac{(1+\varepsilon))}{n}\sum_{i=1}^{n}\|\widehat{f}(X_i) - f_\star(X_i)\|^2$$
$$\leq \frac{2(1+\varepsilon)}{n}\sum_{i=1}^{n}(1-\mathbf{E}')\langle W_i, (\widehat{f} - f_\star)(X_i)\rangle + \|\widehat{f} - f_\star\|_{L^2}^2 - \frac{(1+\varepsilon))}{n}\sum_{i=1}^{n}\|\widehat{f}(X_i) - f_\star(X_i)\|^2 \quad (76)$$

Fix now a radius r and set $g = \frac{r}{\|\widehat{f}-f_\star\|_{L^2}}(\widehat{f} - f_\star)$. If $\|\widehat{f} - f_\star\|_{L^2} \geq r$, dividing both sides above by $\|\widehat{f} - f_\star\|_{L^2}$ yields for the first term above in (76):

$$\frac{2(1+\varepsilon)}{n\|\widehat{f} - f_\star\|_{L^2}}\sum_{i=1}^{n}(1-\mathbf{E}')\langle W_i, (\widehat{f} - f_\star)(X_i)\rangle$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{2(1+\varepsilon)(1-\mathbf{E}')\langle W_i, r^{-1}g(X_i)\rangle\right\} \quad \text{(df. of } g \text{ and divide)} \quad (77)$$
$$\leq r^{-1}\sup_{g\in\mathscr{F}_\star\cap rS_{L^2}} M_n(g).$$

Either the above inequality holds or $\|\widehat{f} - f_\star\|_{L^2} \leq r$. For every $r > 0$ it is thus true that:

$$\|\widehat{f} - f_\star\|_{L^2}^2 \leq r^2 + \left(r^{-1} \sup_{g \in \mathscr{F}_\star \cap r S_{L^2}} M_n(g) \right)^2 + \sup_{g \in \mathscr{F}_\star} Q_n(g) \tag{78}$$

This proves the claim. ■

Finishing the proof of Theorem 3.1. We apply Lemma F.1 with $r = r_\star$, $\varepsilon = 1/2$ and note that $n \geq c_3 \max\{n_{\mathsf{quad}}(r_\star), n_{\mathsf{mult}}(r_\star)\}$ implies: (1) in combination with Proposition E.3 that $\sup_{g \in \mathscr{F}_\star} Q_n(g) \lesssim r_\star^2$; and (2) in combination with Proposition E.2 that $\left(\sup_{g \in \mathscr{F}_\star \cap r_\star S_{L^2}} M_n(g)\right)^2$ scales at most like the RHS of (22). The result follows by a union bound over the failure events of Proposition E.2 and Proposition E.3, all the while taking into account the fact that we posit $k \geq k_{\mathsf{mix}}$. ■

G. Proof of the Corollaries to Theorem 3.1

G.1. Proof of Corollary 3.1

Corollary 3.1 (Parametric Classes)**.** *Fix a failure probability $\delta \in (0,1)$, two Hölder conjugates q, q', and a class $\mathscr{F}$ that is either (1) convex or (2) realizable. Suppose that $\mathscr{F}_\star - \mathscr{F}_\star$ is (L,η)-Ψ_p. Suppose further that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and that k divides $n/2$.*

There exists a universal positive constant c and a polynomial function ϕ_η such that the following holds true. Suppose that there exists $d_{\mathscr{F}} \in \mathbb{R}_+$ such that for $s > 0$:

$$\log \mathcal{N}_{L^2}(\mathscr{F}_\star, s) \leq d_{\mathscr{F}} \log\left(\frac{1}{s}\right) \tag{21}$$

We have with probability $1 - 4\delta$ that:

$$\|\widehat{f} - f_\star\|_{L^2}^2 \leq c \mathbf{V}_{2q}\left(\mathscr{F}_\star \cap \sqrt{\frac{d_{\mathscr{F}} k \|W\|_{L^2}^2}{n}} S_{L^2} \right) \times \left(\frac{d_{\mathscr{F}} + \log(1/\delta)}{n} \right) \tag{22}$$

as long as $k\beta^{-1}(k) \geq n\delta^{-1}$ and

$$n \geq \phi_\eta\Bigg(d_{\mathscr{F}}, k, \|W\|_{\Psi_p}, L, q, q', \mathbf{V}^{-1}\left(\mathscr{F}_\star \cap \sqrt{\frac{d_{\mathscr{F}} k \|W\|_{L^2}^2}{n}} S_{L^2} \right), \log(1/\delta) \Bigg). \tag{23}$$

Proof. Let us begin by observing that for some constant c_η only depending on η we have that:

$$\begin{aligned} \gamma_\eta(\mathscr{F}_\star \cap r S_{L^2}, d_{L^2}) &\leq c_\eta \int_0^r \left(d_{\mathscr{F}} \log\left(\frac{1}{s}\right) \right)^{1/\eta} ds \\ &= c_\eta d_{\mathscr{F}}^{1/\eta} r \int_0^1 \left(\log\left(\frac{r}{s}\right) \right)^{1/\eta} ds \\ &\leq c_\eta d_{\mathscr{F}}^{1/\eta} r \Gamma(1/\eta + 1). \end{aligned} \tag{79}$$

Hence for some universal positive constant c:

$$\sqrt{\mathbf{V}\left(\mathscr{F}_\star \cap r S_{L^2}\right)} \times \frac{1}{r\sqrt{n}} \gamma_2(\mathscr{F}_\star \cap r S_{L^2}, d_{L^2}) \leq c \sqrt{\mathbf{V}\left(\mathscr{F}_\star \cap r S_{L^2}\right)} \times \frac{1}{\sqrt{n}} d_{\mathscr{F}}^{1/2}. \tag{80}$$

A few applications of the Cauchy-Schwarz inequality now yields for any r:

$$\mathbf{V}\left(\mathscr{F}_\star \cap rS_{L^2}\right) \leq k\|W\|_{L^2}^2. \tag{81}$$

A candidate choice is therefore $r_\star = \sqrt{\frac{d_{\mathscr{F}}\mathbf{V}\left(\mathscr{F}_\star \cap \sqrt{\frac{d_{\mathscr{F}}k\|W\|_{L^2}^2}{n}} S_{L^2}\right)}{n}}$. A straighforward but tedious calculation now reveals that the inequality $n \geq \max\left\{n_{\mathsf{quad}}(r_\star), n_{\mathsf{mult}}(r_\star)\right\}$ has a solution depending polynomially on problem data as long as $\eta > 1/4$. ■

G.2. Proof of Corollary 3.2

Corollary 3.2 (Realizable Linear Regression)**.** *Fix a failure probability $\delta \in (0,1)$, a covariate bound $B_X \in (0,\infty)$ and a noise bound $B_W \in (0,\infty)$ and let $\mathsf{X} = \mathbb{R}^d$ and $\mathsf{Y} = \mathbb{R}$. Suppose that k divides $n/2$ and that the model $\mathsf{P}_{(X,Y)_{1:n}}$ is stationary and satisfies $Y_i = \langle \beta_\star, X_i\rangle + W_i$ for $i \in [n]$. Suppose further that:*

1. *$X_{1:n}$ is bounded $|\langle v, X_i\rangle| \leq B_X$, $\forall i \in [n]$ and $v \in \mathbb{R}^d$ with $\|v\| = 1$; and*
2. *$W_{1:n}$ is a bounded martingale difference sequence—$\mathbf{E}[W_i|X_{1:i}] = 0$ and $|W_i| \leq B_W$, $\forall i \in [n]$.*

There exist universal positive constants c_1 and c_2 such that if

$$\frac{n}{k} \geq c_1 \left(B_X/\sqrt{\lambda_{\min}(\mathbf{E}XX^{\mathsf{T}})}\right)^{3+1/2} \left(\frac{kB_W^2}{\mathbf{V}(W)}\right) \times \left(d + \log(1/\delta)\right) \quad \text{and} \quad k\beta^{-1}(k) \geq n\delta^{-1} \tag{24}$$

we have that:

$$\|\widehat{f} - f_\star\|_{L^2}^2 \leq c_2 \mathbf{V}(W)\left(\frac{d+\log(1/\delta)}{n}\right). \tag{25}$$

Proof. We apply Theorem 3.1 with $p = \infty$, $q = 1$ and $\eta = 1$. As in the proof of the preceding corollary (see (79)), notice that

$$\begin{aligned} \gamma_1(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) &\leq cdr, \qquad \text{and} \\ \gamma_2(\mathscr{F}_\star \cap rS_{L^2}, d_{L^2}) &\leq c\sqrt{d}r. \end{aligned} \tag{82}$$

Moreover, since $W_{1:n}$ is a martingale difference sequence we have $\mathbf{V}(\mathscr{F}_\star \cap rS_{L^2}) = \frac{1}{n}\sum_{i=1}^n \mathbf{V}(W_i)$. Consequently, the critical radius inequality (19) becomes

$$r \geq c \times \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbf{V}(W_i) \times \frac{d}{n}}$$

so that (using stationarity) we may choose $r_\star \propto \sqrt{\mathbf{V}(W) \times \frac{d}{n}}$.

Let us now turn to evaluating (18) for this model. n_{quad} reads:

$$\begin{aligned} n_{\mathsf{quad}}(r_\star) &= \inf\left\{n \in \mathbb{N} \middle| \left[n^{-1/2}\sqrt{k}L^{1+3/4}r_\star \times \left(r_\star\sqrt{d} + r_\star\sqrt{\log(1/\delta)}\right)\right.\right. \\ &\quad \left.\left. + n^{-1}L^2 k r_\star\left(r_\star d + r_\star\log(1/\delta)\right)\right] \leq r_\star^2\right\} \\ &\leq \inf\left\{n \in \mathbb{N} \middle| \left[n^{-1/2}\sqrt{k}L^{1+3/4} \times \left(\sqrt{d} + \sqrt{\log(1/\delta)}\right)\right] \leq 1\right\} \\ &\quad + \inf\left\{n \in \mathbb{N} \middle| \left[n^{-1}L^2 k\left(d + \log(1/\delta)\right)\right] \leq 1\right\} \\ &\leq 2k(L \vee 1)^{3+1/2}\left(d + \log(1/\delta)\right). \end{aligned}$$

Next, we turn to n_{mult}:

$$n_{\mathsf{mult}}(r_\star) = \inf\left\{n \in \mathbb{N} \middle| LkB_W\left(\frac{d+\log(1/\delta)}{n}\right) \leq \sqrt{\mathbf{V}(W)d/n}\right\} \quad \leq L^2k^2\frac{B_W^2}{\mathbf{V}(W)}\left(d+\log(1/\delta)\right).$$

Moreover, it is easy to see that may choose $L = B_X/\sqrt{\lambda_{\min}(\mathbf{E}XX^{\mathsf{T}})} \geq 1$. Hence the desired result follows under the burn-in requirement that:

$$\frac{n}{k} \geq c\left(B_X/\sqrt{\lambda_{\min}(\mathbf{E}XX^{\mathsf{T}})}\right)^{3+1/2}\left(\frac{kB_W^2}{\mathbf{V}(W)}\right)\left(d+\log(1/\delta)\right)$$

as we sought to prove. ■