Sharpe 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 (\(\beta\)-mixing) data and square loss in a hypothesis class \(\mathcal{F} \subset L_{\Psi_p}\), where \(\Psi_p\) is the norm \(\|f\|_{\Psi_p} \equiv \sup_{m \geq 1} m^{-1/p} \|f\|_{L^m}\) 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 \(\Psi_p\) are comparable on our hypothesis class \(\mathcal{F}\)—that is, \(\mathcal{F}\) is a weakly sub-Gaussian class: \(\|f\|_{\Psi_p} \lesssim \|f\|_{L^2}\) 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 complexities 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 (\(\beta\)-mixing) setting through blocking (Bernstein, 1927; Yu, 1994). 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).

1University of Pennsylvania 2University 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. Copyright 2024 by the author(s).
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).\footnote{We say that a sequence $Z_{1:n}$ is $k$-wise independent if each of the blocks $Z_{(j+1)k:j+1}$ ($j = 0, 1, \ldots, n/k - 1$) are independent of each other.} By applying Bernstein’s inequality to the $bk$-bounded variables $V_{i/n/k}$, $V_i \triangleq \sum_{j=ik-k+1}^{ik} V_j$ we find that with probability at least $1 - \delta$:

$$
\frac{1}{n} \sum_{i=1}^{n} V_i \leq 2 \sqrt{\frac{k^{-1} \mathbb{E}(V_1^2) \ln(1/\delta)}{n}} + \frac{4bk \log(1/\delta)}{3n}. \quad (1)
$$

If the data instead were completely independent, then in the small and moderate deviations regime $\delta \geq \exp(-n\mathbb{E}V_1^2/b^2k)$, (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 $\mathcal{F}$. 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 $\{V_{1:n}(f) : f \in \mathcal{F}\}$. To obtain fast rates, this uniform control needs to be 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. Nâïvely 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 $\mathcal{F}$ 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.

1.1. Contribution

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

$$
\hat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \|f(X_i) - Y_i\|^2. \quad (2)
$$

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

$$
f_* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}\|f(X) - Y\|^2, \quad (X, Y) \sim P_{X,Y}. \quad (3)
$$

Let us also denote by $\mathcal{F}_*$ the star-hull of $\mathcal{F}$ around $f_*$. That is, $\mathcal{F}_* \triangleq \{\rho(f - f_*) : f \in \mathcal{F}, \rho \in [0, 1]\}$, which, for a convex class $\mathcal{F}$ coincides with the shifted class $\mathcal{F} - \{f_*\}$. We further equip $\mathcal{F}_*$ with the $L^2$-norm: $\|f\|_2^2 \triangleq \mathbb{E}\|f(X)\|^2$, $f \in \mathcal{F}_*, X \sim P_X$. Let us also define the “noise” $W_{1:n}$ by $W_i \triangleq Y_i - f_*(X_i), i \in [n]$. We focus on the case when $\mathcal{F}$ is either (1) convex or (2) realizable (i.e., $\mathbb{E}[W_i|X_i] = 0$ 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:

\[
r_\star \asymp \sup_{g \in \mathcal{F} \cap r_\star S_{L^2}} \mathbf{V} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_{1,n} (g(X_i)) \right) \right) \times \frac{\text{complexity}(\mathcal{F} \cap r_\star S_{L^2})}{r_\star \sqrt{n}},
\]

where for \( r \in \mathbb{R}, r > 0 \), \( S_{L^2} \) 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, \( \sup_{g \in \mathcal{F} \cap r_\star S_{L^2}} \mathbf{V} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_{1,n} (g(X_i)) \right) \right) \). 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 \( \mathcal{F} \) 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 \( \hat{f} \) over such a class \( \mathcal{F} \) 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 \( \mathcal{F} \). That is with probability \( 1 - \delta \):

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

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 \( \mathcal{F} \) satisfies a certain topological condition. Recall that for a random variable \( Z \) the \( \psi_p \)-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 \( f \in \mathcal{F} \) satisfies the inequality \( \| f \|_{\psi_p} \leq L \| f \|_{L^2} \). 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 \( \mathcal{F} \) 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:

\[
Q_n(f) \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \| f(X_i) - f_\star(X_i) \|^2 - \frac{1 + \varepsilon}{n} \sum_{i=1}^{n} \| f(X_i) - f_\star(X_i) \|^2,
\]

for some \( \varepsilon \in (0, 1) \). Under our assumptions, we will show that the process \( Q_n(f) \) is eventually nonpositive uniformly
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 \( \{Q_n(f) \leq 0, \forall f \in \mathcal{F}_* \cap rB_{L^2} \} \), using optimality of \( \hat{f} \) to (2) we also have the following deterministic (basic) inequality:

\[
\|\hat{f} - f_*\|_{L^2} \leq r + \frac{1 + \varepsilon}{r n} \sum_{i=1}^{n} 2\left(1 - \mathbb{E}'\right) \left[ \frac{r}{\|\hat{f} - f_*\|_{L^2}} \right]
\]

(7)

where \( \mathbb{E}' \) denotes expectation with respect to a fresh copy of randomness independent of \( \hat{f} \).

Hence, we also need to control the multiplier process:

\[
M_n(f) \triangleq \frac{1 + \varepsilon}{n} \sum_{i=1}^{n} 2\left(1 - \mathbb{E}'\right)(W_i, f(X_i)).
\]

(8)

It is the uniform control of \( M_n(f) \) over the class \( \mathcal{F}_* \) 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 2.1. 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 \( \mathcal{F}_* \) 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 \( \beta \)-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 \( \Psi_p \)-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 \( \mathcal{F}_* \) 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 \( \Psi_p \)-component.

- Our assumption that \( \mathcal{F}_* \) is a weakly sub-Gaussian class now comes into play by ensuring that, past a burn-in, the \( \Psi_p \)-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 \( \Psi_p \)-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 \( \|f\|_{\Psi_p} \leq L \|f\|_{L^2}^q \) 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 \( \mathbb{E} \) (resp. \( \mathbb{P} \)). For \( q \in [1, \infty) \) the \( 2q \)-variance of a random variable \( Z \) is defined as \( V_{2q}(Z) \triangleq (\mathbb{E}(Z - \mathbb{E}(Z))^{2q})^{1/q} \) with \( V_2 = \mathbb{V} \) being the standard variance. For \( p \in [1, \infty) \), we also introduce the \( \Psi_p \)-norm \( \|Z\|_{\Psi_p} \triangleq \sup_{m \geq 1} m^{-1/p} \|Z\|_{L^m} \) 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 \( \mathbb{P} \) and \( \mathbb{Q} \) defined on the same probability space, their total variation is denoted \( \|\mathbb{P} - \mathbb{Q}\|_{TV} \). Maxima (resp. minima) of two numbers \( a, b \in \mathbb{R} \) are denoted by \( a \lor b = \max(a, b) \) (resp. \( a \land b = \min(a, b) \)). For an integer \( n \in \mathbb{N} \), we also define the shorthand \( \{1, \ldots, n\} \). For a symmetric positive semidefinite matrix \( M \), \( \lambda_{\min}(M) \) denotes its smallest nonzero eigenvalue.

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

\[
\Delta_d(\mathcal{H}) \triangleq \sup_{h, h' \in \mathcal{H}} d(h, h').
\]

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

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

(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 $\gamma_\eta(\mathcal{H},d')$ for $d$ replaced with $d'$ in (9) (while being mindful of the fact that $d'$ is not a metric in general). Finally, since entropy integrals upper-bound $\gamma_\eta$-functionals, it will also be useful to introduce the covering number $\mathcal{N}_{\gamma_\eta}(\mathcal{H},s)$, which denotes the minimal number of $L^2$-balls of radius $s$ required to cover $\mathcal{H}$.

2. $\Psi_p$-N 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 $\Psi_p$-norm. The lemma comes from (Maurer & Pontil, 2021).

**Lemma 2.1 ($\Psi_p$-Bernstein MGF Bound).** Fix a random variable $Z$ and $p \in [1, \infty]$ such that $E[Z] \leq 0$ and $\|Z\|_{\Psi_p} < \infty$. Let $q$ and $q'$ be Hölder conjugates and suppose that $\lambda \in [0,1/(q'e)^{1/p}\|Z\|_{\Psi_p}^2]$. We have that:

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

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 $\lambda \in (0, (2(\xi')^{1/p}\|Z\|_{\Psi_p}^2)^{-1}]$ 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 $\Psi_p$-norm for any constant choice of $\lambda$ 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 $\mathcal{G}$ is $(L,\eta)$-$\Psi_p$ if for every $g \in \mathcal{G}$ we have that:

$$\|g\|_{\Psi_p} \leq L \|g\|_{L^2}. \quad (11)$$

If (11) holds for $\mathcal{G}$ with $\eta \in (0, 1)$ and some $L$ we will call $\mathcal{G}$ a weak $\Psi_p$-class. If (11) instead holds for $\eta = 1$ it is simply a $\Psi_p$-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 $\mathcal{F}$, the model $P_{(X,Y)}$, and the shifted target $W_{1:n} = (Y - f_n(X))_{1:n}$ at resolution $\mathcal{G}$ is given by

$$V_{2q}(\mathcal{F},\mathcal{G},P_{(X,Y)}) \triangleq \sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( W_i, \frac{g(X_i)}{\|g\|_{L^2}} \right). \quad (12)$$

We stress that, even though Definition 2.2 measures noise uniformly over a function class, it does not necessarily grow with the complexity of the class. For instance, under the additional hypotheses that $P_{(X,Y)}$ is drawn iid and that $W_i$ is independent of $X_i$ for $i \in [n]$, it is easy to see that $V_2(\mathcal{F},\mathcal{G},P_{(X,Y)}) = V_2(W)$ for every such well-specified class $\mathcal{G}$. Rather, Definition 2.2 is a measure of how well the targets $Y_{1:n}$ align with a given class $\mathcal{G}$.

2.1. The Multiplier Process

We will not directly control the multiplier process for $\beta$-mixing variables. Instead we first suppose that the model $P_{(X,Y)}$ is $k$-wise independent (where $k$ divides $n$). We then port these results to the $\beta$-mixing setting by blocking (see Appendix E.1). We use the following shorthand notation regarding $V_{2q}(\mathcal{F},\mathcal{G},P_{(X,Y)})$: we take the class $\mathcal{F}$ and the probability model $P_{(X,Y)}$ as fixed and thus omit the dependence on $\mathcal{F}$ (via $f_n$) and $P_{(X,Y)}$ and write $V_{2q}(\mathcal{G}) = V_{2q}(\mathcal{F},\mathcal{G},P_{(X,Y)})$. 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 $P_{(X,Y)}$ 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:

$$P \left( \sum_{i=1}^{n} (1 - E)(W_i, g(X_i) - g'(X_i)) \right)
\geq \frac{1}{4n} \|g - g'\|_{L^2}^2 V_{2q}(\{g\} - \{g'\}) u + 4(q'e)^{2/p}k \|W\|_{\Psi_p} \|g - g'\|_{\Psi_p} u \leq 2e^{-u}. \quad (13)$$

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 \( \mathcal{G} \) is \((L,\eta)\)-\( \Psi_p \), 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 \( \mathbb{V}_{2q}(\mathcal{F}_* \cap rS_{L^2}) \) 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 \( \mathbb{V}_{2q}(\mathcal{F}_* \cap rS_{L^2}) \). 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 \( \mathbb{V}_{2q}(\mathcal{F}_* \cap rS_{L^2}) = \mathbb{V}(\mathcal{F}_* \cap rS_{L^2}) + o(1) \). 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 \( P_{(X,Y)_{i=1}^n} \) (e.g. for the Markovian situation cf. Meyn & Tweedie, 1993, Theorem 17.3.6):

\[
\frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \left( \frac{g(X_i)}{\|g\|_{L^2}} \right) \sim \mathcal{N}(0, \mathbb{V}(\mathcal{F}_*, \{g\}, P_{(X,Y)_{i=1}^n})) ,
\]

where the variance term on the right is:

\[
\mathbb{V}(\mathcal{F}_*, \{g\}, P_{(X,Y)_{i=1}^n}) \overset{\text{Def}}{=} \lim_{n \to \infty} \mathbb{V}(\mathcal{F}_*, \{g\}, P_{(X,Y)_{i=1}^n}) .
\]

Now, since \( r = o(1) \) in all practical situations one expects \( \mathbb{V}(\mathcal{F}_*, \{g\}) \) as long as the map \( f \mapsto \mathbb{V}(f/\|f\|_{L^2}) \) is sufficiently regular near \( f_* \).

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 \( \mathcal{F} \). Suppose that \( \mathcal{F} \neq \mathcal{F}_* \) is \((L,\eta)\)-\( \Psi_p \). Suppose further that the model \( P_{(X,Y)_{i=1}^n} \) 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 \( 1 - \delta \):

\[
\sup_{f \in \mathcal{F}_* \cap rS_{L^2}} \frac{1}{rn} \sum_{i=1}^{n} (1 - \mathbb{E}) \langle W_i, f \rangle \\
\leq c_2 \sqrt{\mathbb{V}_{2q}(\mathcal{F}_* \cap rS_{L^2})} \left( \frac{1}{r\sqrt{n}} \right) \\
\times \gamma_2(\mathcal{F}_* \cap rS_{L^2}, d_{L^2}) + \frac{\log(1/\delta)}{n} \\
+ c_1(q')^{2/3}Lk\|W\|_{\mathcal{P}_p} \\
\times \left( \frac{1}{rn} \gamma_\eta(\mathcal{F}_* \cap rS_{L^2}, d_{L^2}) + \frac{r^{\eta-1}}{n} \log(1/\delta) \right) .
\]

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 \( \mathcal{F} \neq \mathcal{F}_* \) is \((L,\eta)\)-\( \Psi_p \). Suppose further that the model \( P_{(X,Y)_{i=1}^n} \) is stationary and \( k \)-wise independent where \( k \) divides \( n \). There exist a universal positive constant \( c \) such that uniformly for all \( f \in \mathcal{F}_* \setminus \{rB_{L^2}\} \) we have that with probability at least \( 1 - \delta \):

\[
\frac{1}{n} \sum_{i=1}^{n} \|f(X_i)\|^2 \geq r^2(1 - \varepsilon^2) \\
- c\left( n^{-1/2} \sqrt{kL^{1+3/4}r^q} \left( \frac{\log \left( \frac{4^{2/p}L^2}{\varepsilon r} \right) }{n} \right)^{1/p} \right. \\
\times \left( \gamma_{2+\eta/2}(\mathcal{F}_* \cap rS_{L^2}, d_{L^2}) + n^{-1}q^{1/3}L^{1+3/4}r^{q}-\eta \sqrt{\log(1/\delta)} \right) \\
+ n^{-1}(q')^{1/3}L^{1+3/4}r^q \left( \log \left( \frac{4^{2/p}L^2}{\varepsilon r} \right) \right)^{1/p}L^2 \\
\times \gamma_\eta(\mathcal{F}_* \cap rS_{L^2}, d_{L^2}) + n^{\eta/3} \log(1/\delta) \right) .
\]

### 2.3. \( \beta \)-Mixing Processes

We extend the empirical process results of the preceding two sections to \( \beta \)-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 \( \beta \)-mixing coefficients of \( Z \), denoted \( \beta_Z(i) \), for \( i \in [n] \):

\[
\beta_Z(i) \overset{\text{Def}}{=} \sup_{t \in [n]; t+i \leq n} \mathbb{E}\|P_{Z_{1:t}}(\cdot|Z_{1:i}) - P_{Z_{1:i}}(\cdot)\|_{TV} .
\]

### 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 \( n_{\text{quad}}, n_{\text{mult}}, k_{\text{mix}} \) which together dictate the minimal sample size necessary for our result to be sharp. The first of these, \( n_{\text{quad}} \), 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, $n_{\text{mult}}$, is required for the multiplier process, $M_n(f)$, to have a dominant variance term (informally—when the CLT-like rate becomes accurate). Finally, $k_{\text{mix}}$ is the minimal block-size it takes for the $\beta$-mixing model $P_{(X,Y),n}$ to be well-approximated by a corresponding $k$-wise independent model. These are given as follows below:

$$n_{\text{quad}}(r) = \inf \left\{ n \in \mathbb{N} \left| n^{-1/2} \sqrt{KL} + 3/4 \eta \right| \times \left( \log \left( \frac{4^2/p + 1/2L}{r} \right) \right)^{1/p} \times \left( \gamma_{2\beta_{k\text{mix}}} (\mathcal{F}, r S_{L^2}, d_{L^2}) + r \log(1/\delta) \right) + n^{-2} \left( \frac{4^2/p + 1/2L}{r} \right)^{1/p} \times \left( \gamma_{\eta} (\mathcal{F}, r S_{L^2}, d_{L^2}) + r \log(1/\delta) \right) \leq r^2 \right\},$$

and

$$n_{\text{mult}}(r) = \inf \left\{ n \in \mathbb{N} \left| \left( q' \right)^2/p L k \|W\|_{\Psi_p} \times \left( \frac{1}{\gamma_{\eta}} \gamma_{\eta} (\mathcal{F}, r S_{L^2}, d_{L^2}) + r \log(1/\delta) \right) \leq r \right\},$$

$$k_{\text{mix}} = \inf \{ k \in [n] | k \beta_{X,Y}^r (k) \geq n \delta^{-1} \}. \quad (18)$$

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, $\delta$, dominates the mixing term, $\frac{n}{k} \beta_{X,Y}^r (k)$, 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—$eta_{X,Y}^r (k) \leq \exp(-k/r_{\text{mix}})$ for some $r_{\text{mix}} \in \mathbb{R}_+$—this requirement is satisfied by $k \leq \frac{1}{\gamma_{\text{mix}}} \log(n/\delta)$. 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 $\mathcal{F}$ that is either (1) convex or (2) realizable. Suppose that $\mathcal{F} - \mathcal{F}$ is $(L, \eta)$-$\Psi_p$. Suppose further that the model $P_{(X,Y),n}$ 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 $r_s$ solves

$$r \geq c_1 \sqrt{\mathbb{V}_{2q} (\mathcal{F}, r S_{L^2})} \times \frac{1}{r^{2/3} \eta} \gamma_{2\beta_{X,Y}^r (k)} (\mathcal{F}, r S_{L^2}, d_{L^2}), \quad (19)$$

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

$$\|\hat{f} - f_s\|_2^2 \leq c_2 \left( r_s^2 + \mathbb{V}_{2q} (\mathcal{F}, r S_{L^2}) \frac{\log(1/\delta)}{n} \right), \quad (20)$$

as long as $n \geq c_3 \max \{ n_{\text{quad}}(r_s), n_{\text{mult}}(r_s) \}$ and $k \geq k_{\text{mix}}$ (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 $r_s$ given in (19). This radius depends on local complexity of the class $\mathcal{F}$ measured in $L^2$ distance as per the $\gamma_2$-functional and through the weak variance $\mathbb{V}_{2q} (\mathcal{F}, r S_{L^2})$.

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 $r_s$ in (19) becomes a variance-proxy scaled dimensional factor and that the burn-in requirement (18) amounts to a polynomial in problem data and $\log(1/\delta)$.

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

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

$$\log N_{L^2} (\mathcal{F}, s) \leq d_{\mathcal{F}} \log \left( \frac{1}{s} \right), \quad (21)$$

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

$$\|\hat{f} - f_s\|_2^2 \leq c \mathbb{V}_{2q} \left( \mathcal{F}, \frac{d_{\mathcal{F}} k \|W\|_{\Psi_p}^2}{n} S_{L^2} \right) \times \left( \frac{d_{\mathcal{F}} + \log(1/\delta)}{n} \right), \quad (22)$$

as long as $k \beta_{X,Y}^r (k) \geq n \delta^{-1}$ and

$$n \geq \phi_\eta \left( d_{\mathcal{F}}, k, \|W\|_{\Psi_p}, L, q, q', \right. \mathbb{V}^{-1} \left( \mathcal{F}, \frac{d_{\mathcal{F}} k \|W\|_{\Psi_p}^2}{n} S_{L^2} \right), \log(1/\delta) \right). \quad (23)$$
Consequently, after a polynomial burn-in and up to a universal positive constant, we are able to recover the optimal parametric rate $n^{-1}(d + \log(1/\delta))$ 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 $X_i$-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 $X = \mathbb{R}^d$ and $Y = \mathbb{R}$. Suppose that $k$ divides $n/2$ and that the model $P_{(X,Y)_{1:n}}$ is stationary and satisfies $Y_i = (\beta_x, X_i) + W_i$ for $i \in [n]$. Suppose further that:

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

There exist universal positive constants $c_1$ and $c_2$ such that if

$$n/k \geq c_1 \left( B_X / \sqrt{\lambda_{\min}(EXX^T)} \right)^{3+1/2} \left( k B_W^2 / V(W) \right) \times (d + \log(1/\delta)) \quad \text{and} \quad k \beta^{-1}(k) \geq nd^{-1}$$

(24)

we have that:

$$\| \hat{f} - f_* \|^2_{L^2} \leq c_2 V(W) \left( d + \log(1/\delta) \right) / n.$$ (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 Tsiarnak 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 $k B_W^2 / V(W)$ in (24) arises from the multiplier process: it is the cost of having $V(W)$ appear in (25) instead of $k B_W^2$. 

...continues...
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 $\mathcal{F}$ be a set of uniformly bounded functions $f : X \to \mathbb{R}$. Fix an integer $s \in \mathbb{N}$ and suppose that there exists a constant $C_{\mathcal{F}}$ such that $\sum_{|\alpha| \leq s} \|D^{\alpha} f\|_L \leq C_{\mathcal{F}}$. Suppose further that the distribution of the covariates $\mathcal{P}_X$ has density $\mu_X$ with respect to the Lebesgue measure and that there exists $\mu, \pi \in \mathbb{R}_+$ such that $\mu \leq \mu_X \leq \pi$. Under the above hypotheses there exists a positive constant $c$ only depending on $X$, $d$ and $s$ such that $\mathcal{F}$ is $(L, \eta) \cdot \Psi_{\infty}$ with $L = c \mu^{-\frac{2s+2}{2s+1}} \pi^\frac{1}{2s+1}$ and $\eta = \frac{2s}{2s+1}$.

**Proof.** The result for $\mathcal{P}_X$ equal to the (normalized) Lebesgue measure is immediate by the main result of (Nirenberg, 1959) instantiating to the correct smoothness class. The general case follows by our hypothesis that $\mathcal{P}_X$ is equivalent to the Lebesgue measure.

We stress that the quantities $L$ and $\eta$ 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 sufficiently 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 $M \subset \mathbb{R}^{d_M}$ equipped with the Euclidean norm $\| \cdot \|$. Consider a function $\phi : X \times M \to \mathbb{R}$ that generates a parametric class of functions $\mathcal{F} = \{ \phi(\cdot; \theta) \mid \theta \in M \}$. Define $M_* \triangleq \arg\min_{\theta \in M} \mathbb{E}(\phi(X, \theta) - Y)^2$ to be the set of population risk minimizers. Suppose that:

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

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

Then $\mathcal{F} - \{ f_* \}$ is $(Cba^{-1}, 2b) \cdot \Psi_\infty$.

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 $x \in X$:

$$|\phi(x; \theta) - \phi(x; \theta_*)| = |(\nabla_{\theta} \phi(x, \tilde{\theta}), \theta - \theta_*)| \leq \|\nabla_{\theta} \phi(x, \tilde{\theta})\| \|\theta - \theta_*\| \leq C \|\theta - \theta_*\|$$

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

$$\sup_{x \in X} \|\phi(x, \theta) - \phi(x, \theta_*)\| \leq C b a^{-1} (\mathbb{E}(\phi(X, \theta) - \phi(X, \theta_*))^2)^\frac{1}{2}$$

Equivalently, $\|f\|_{L_\infty} \leq C b a^{-1} \|f\|_{L_2}^\frac{1}{2}$ for every $f \in \mathcal{F} - \{ f_* \}$ 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 $\mathcal{F}$ be the class of linear functionals on $\mathbb{R}^d$. Suppose that $\lambda_{\min}(\mathbb{E}XX^T) > 0$. Then $\mathcal{F}$ is $(L, 1) \cdot \Psi_2$ with $L = \sup_{v \in \mathbb{R}^d : \|v\| = 1} \|\mathbb{E}X^T v\|_2$.

**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 $\mathcal{F}$ be a finite subset of $L_{\Psi_2}$. Then $\mathcal{F}$ is $(L, 1) \cdot \Psi_2$ with $L = \max_{f \in \mathcal{F}} \|f\|_{\Psi_2} / \|f\|_{L_2}$.

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

**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.
Sharp Rates in Dependent Learning Theory

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


# Contents

1. **Introduction**
   - 1.1 Contribution .................................................. 2
   - 1.2 Proof Outline .................................................. 3
   - 1.3 Further Preliminaries ........................................ 4

2. **$\Psi_p$-Norms, Bernstein's Inequality and Empirical Processes**
   - 2.1 The Multiplier Process ...................................... 5
   - 2.2 The Quadratic Process ........................................ 6
   - 2.3 $\beta$-Mixing Processes .................................... 6

3. **The Main Result**
   - 3.1 Further Comparison to Related Work ...................... 8

4. **Examples of Weakly sub-Gaussian Classes** ............. 9

References .................................................................... 10

A. **Summary** .......................................................... 13

B. **Properties of $\Psi_p$- and $L^p$-Norms** ................. 13
   - B.1 Proof of Lemma 2.1 ........................................... 13

C. **Controlling the Multiplier Process** ...................... 14

D. **Controlling the Quadratic Process** ....................... 17

E. **Results for Mixing Empirical Processes** ................. 20
   - E.1 Blocking .......................................................... 20
   - E.2 Controlling Empirical Processes for $\beta$-Mixing Data 20

F. **Finishing the Proof of Theorem 3.1** ....................... 21

G. **Proof of the Corollaries to Theorem 3.1** ................. 22
   - G.1 Proof of Corollary 3.1 ....................................... 22
   - G.2 Proof of Corollary 3.2 ....................................... 23
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, \( \beta \)-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 \( 2^{\eta} \) of the process at hand (where typically \( \eta = 1 + o(1) \)). 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 \( \beta \)-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 \( \Psi_p \)- 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:

\[
\| \langle Z, Z' \rangle \|_{\Psi_p/2} \leq 2^{2/p} \| Z \|_{\Psi_p} \| Z' \|_{\Psi_p}.
\]

**Proof.** We compute:

\[
\| \langle Z, Z' \rangle \|_{\Psi_p/2} = \sup_{m \geq 1} \frac{\| \langle Z, Z' \rangle \|_{L^m}}{m^{2/p}} = \sup_{m \geq 1} \frac{(E|\langle Z, Z' \rangle|)^{1/m}}{m^{2/p}} \leq \sup_{m \geq 1} \frac{(E\| Z \|^m \| Z' \|^m)^{1/m}}{m^{2/p}} \leq 2^{2/p} \sup_{m \geq 1} \frac{(E\| Z \|^{2m} \| Z' \|^{2m})^{1/2m}}{2m^{1/p}} \leq 2^{2/p} \| Z \|_{\Psi_p} \| Z' \|_{\Psi_p},
\]

as was required. \( \blacksquare \)

B.1. Proof of Lemma 2.1

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

\[
E \exp (\lambda Z) \leq \exp \left( \frac{\lambda^2}{2} \frac{(E(Z)^{2q})^{1/q}}{1 - \lambda(q' \epsilon)^{1/p} \| Z \|_{\Psi_p}} \right).
\]

**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 \( L^\infty \) norm is replaced by a \( \Psi_p \)-norm. We begin by expanding the exponential function:
\[ \mathbb{E} \exp(\lambda Z) = \mathbb{E} \left[ \sum_{m=0}^{\infty} \frac{(\lambda Z)^m}{m!} \right] \]
\[ \leq 1 + \sum_{m=0}^{\infty} \mathbb{E} \left[ (\lambda Z)^2 \right]^m \frac{(\lambda Z)^{m+2}}{(m+2)!} \]  
\[ (\mathbb{E} Z \leq 0) \]  
\[ \leq 1 + \lambda^2 \left( \mathbb{E} (Z)^2 q \right)^{1/q} \sum_{m=0}^{\infty} \left( \mathbb{E} \left[ |\lambda Z|^{mq'} \right] \right)^{1/q'} \frac{(\lambda Z)^{m+2}}{(m+2)!} \]  
\[ \text{(Hölder’s Ineq.)} \]

We next have:
\[ \left( \mathbb{E} \left[ |\lambda Z|^{mq'} \right] \right)^{1/q'} = \|Z\|_{L^{mq'}}^{m} \]
\[ \leq (mq')^m/p \|Z\|_{\Psi_p}^m \]  
\[ \text{(df. of } \Psi_p) \]
\[ \leq (m!)^{1/p} \langle q', e \rangle^{m/p} \|Z\|_{\Psi_p}^m \]  
\[ \text{(Stirling’s Approximation)} \]

Upon combining (30) with (31) we arrive at
\[ \mathbb{E} \exp(\lambda Z) \]
\[ \leq 1 + \lambda^2 \left( \mathbb{E} (Z)^2 q \right)^{1/q} \sum_{m=0}^{\infty} \frac{(m!)^{1/p} \lambda^m (q')^m/p \|Z\|_{\Psi_p}^m}{(m+2)!} \]
\[ \leq 1 + \frac{\lambda^2}{2} \left( \mathbb{E} (Z)^2 q \right)^{1/q} \sum_{m=0}^{\infty} \left( \lambda (q')^1/p \|Z\|_{\Psi_p} \right)^m \]
\[ (p \in [1, \infty], m \in \mathbb{N} \Rightarrow \frac{(m!)^{1/p}}{(m+2)!} \leq \frac{1}{2}) \]
\[ = 1 + \frac{\lambda^2}{2} \left( \mathbb{E} (Z)^2 q \right)^{1/q} \frac{1-x}{1-\lambda (q')^1/p \|Z\|_{\Psi_p}} \]
\[ \leq \exp \left( \frac{\lambda^2}{2} \left( \mathbb{E} (Z)^2 q \right)^{1/q} \frac{1-x}{1-\lambda (q')^1/p \|Z\|_{\Psi_p}} \right) \]
\[ (x \in \mathbb{R} \Rightarrow 1+x \leq e^x) \]

as per requirement.

\[ \text{C. Controlling the Multiplier Process} \]

**Lemma 2.2 (Pointwise Control).** Fix two Hölder conjugates \( q \) and \( q' \). Suppose that the model \( P_{(X,Y)_1} \) 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:

\[ P \left( \sum_{i=1}^{n} (1-E)(W_i, g(X_i) - g'(X_i)) \right) \]
\[ \geq \sqrt{4n\|g-g'\|_{L^2}^2} V_{2q} (\{g\} - \{g'\}) u \]
\[ + 4(q'^2/p k \|W\|_{\Psi_p} \|g-g'\|_{\Psi_p} u) \leq 2e^{-u}. \]  
\[ (13) \]

**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:

\[ \mathbb{E} \exp \left( \sum_{i=1}^{k} (1-E)(W_i, g(X_i) - g'(X_i)) \right) \]
\[ \leq \exp \left( \frac{\lambda^2 k \|g-g'\|_{L^2}^2 V_{2q} \left( \sum_{i=1}^{k} (1-E)(W_i, g(X_i) - g'(X_i)) \right) \|g-g'\|_{\Psi_p / 2}^2}{2 \left( 1-\lambda (q')^2/p \| \sum_{i=1}^{k} (1-E)(W_i, g(X_i) - g'(X_i)) \|_{\Psi_p / 2} \right)} \]  
\[ (33) \]
Taking the above exponential inequality as a starting point, for a fixed \( \lambda > 0 \), \( (q' e)^{2/p} \| \sum_{i=1}^{k} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle \|_{\psi_{p/2}} \) \(^{-1} \). Now, by triangle inequality and Lemma B.1,

\[
\lambda(q' e)^{2/p} \left\| \sum_{i=1}^{k} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle \right\|_{\psi_{p/2}} \leq \lambda(2q' e)^{2/p} k \| W \|_{\psi_p} \| g(X_i) - g'(X_i) \|_{\psi_p}.
\]

Consequently:

\[
E \exp \left( \lambda \sum_{i=1}^{n} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle \right) \leq \exp \left( \frac{\lambda^2 k \| g - g' \|_{L_2}^2 \mathcal{V}_{2q} (\{ g \} - \{ g' \})}{2 (1 - \lambda(2q' e)^{2/p} k \| W \|_{\psi_p} \| g - g' \|_{\psi_p})} \right).
\]

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

\[
E \exp \left( \lambda \sum_{i=1}^{n} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle \right) \leq \exp \left( \frac{\lambda^2 n \| g - g' \|_{L_2}^2 \mathcal{V}_{2q} (\{ g \} - \{ g' \})}{2 (1 - \lambda(2q' e)^{2/p} k \| W \|_{\psi_p} \| g - g' \|_{\psi_p})} \right).
\]

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:

\[
E \exp \left( \lambda \sum_{i=1}^{n} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle \right) \leq \exp \left( \lambda^2 n \| g - g' \|_{L_2}^2 \mathcal{V}_{2q} (\{ g \} - \{ g' \}) \right).
\]

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

\[
P \left( \sum_{i=1}^{n} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle > u \right) \leq \inf_{\lambda > 0} E \exp \left( -u \lambda + \lambda \sum_{i=1}^{n} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle \right)
\]

\[
\leq \inf_{\lambda \in \Lambda} \left( -u \lambda + \frac{\lambda^2 n \| g - g' \|_{L_2}^2 \mathcal{V}_{2q} (\{ g \} - \{ g' \})}{2 (1 - \lambda(2q' e)^{2/p} k \| W \|_{\psi_p} \| g - g' \|_{\psi_p})} \right)
\]

\[
\leq \exp \left( \frac{-u^2}{4n \| g - g' \|_{L_2}^2 \mathcal{V}_{2q} (\{ g \} - \{ g' \})} \right) \quad \text{otherwise.}
\]

Rescaling and summing the failure probabilities in either case yields:

\[
P \left( \sum_{i=1}^{n} (1 - E) \langle W_i, g(X_i) - g'(X_i) \rangle > \sqrt{4n \| g - g' \|_{L_2}^2 \mathcal{V}_{2q} (\{ g \} - \{ g' \})} u \right)
\]

\[
+ 4(2q' e)^{2/p} k \| W \|_{\psi_p} \| g - g' \|_{\psi_p} u \leq 2e^{-u},
\]

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 a failure probability \( \delta \in (0,1) \), a positive scalar \( r \in (0,\infty) \), two Hölder conjugates \( q \) and \( q' \), and a class \( \mathcal{F} \). Suppose that \( \mathcal{F} - \mathcal{F} \) is \((L,\eta)\)-\( \Psi_p \). Suppose further that the model \( P_{(X,Y),\nu,\eta} \) 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 \( 1 - \delta \):

\[
\sup_{f \in \mathcal{F} \cap \mathcal{S}_{L^2}} \frac{1}{r n} \sum_{i=1}^{n} (1 - \mathbb{E}(W_i,f)) \leq c_1 \left( \gamma_2(\mathcal{F}_*, \mathcal{S}_{L^2}, d_2) + \gamma_1(\mathcal{F}_*, \mathcal{S}_{L^2}, d_1) \right) \\
+ c_2 \left( \Delta_2(\mathcal{F}_*, \mathcal{S}_{L^2}) \sqrt{\log(1/\delta)} + \Delta_1(\mathcal{F}_*, \mathcal{S}_{L^2}) \log(1/\delta) \right).
\]

**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 \((L,\eta)\)-\( \Psi_p \). We begin with \( d_2 \), which is just a dilated \( L^2 \)-metric:

\[
\gamma_2(\mathcal{F}_*, \mathcal{S}_{L^2}, d_2) = \inf_{\{F_m\}} \sup_{f \in \mathcal{F}_* \cap \mathcal{S}_{L^2}} \sum_{m=0}^{\infty} 2^{m/2} d_2(f, F_m) \\
= \sqrt{4n (V_{2q}(\mathcal{F}_* \cap \mathcal{S}_{L^2})) \gamma_2(\mathcal{F}_*, \mathcal{S}_{L^2}, d_{L^2})},
\]

and

\[
\Delta_2(\mathcal{F}_*, \mathcal{S}_{L^2}) \leq r \sqrt{4n (V_{2q}(\mathcal{F}_* \cap \mathcal{S}_{L^2}))}.
\]

Turning to the \( d_1 \), we have:

\[
\gamma_1(\mathcal{F}_*, \mathcal{S}_{L^2}, d_1) = \inf_{\{F_m\}} \sup_{f \in \mathcal{F}_* \cap \mathcal{S}_{L^2}} \sum_{m=0}^{\infty} 2^m d_1(f, F_m) \\
= \inf_{\{F_m\}} \sup_{f \in \mathcal{F}_* \cap \mathcal{S}_{L^2}} \sum_{m=0}^{\infty} 2^m (q' e)^{2/p} k \|W\|_{\Psi_p} d_{\Psi_p}(f, F_m) \\
\leq 2 (q' e)^{2/p} k \|W\|_{\Psi_p} \gamma_1(\mathcal{F}_*, \mathcal{S}_{L^2}, d_{L^2}) \\
\leq 2 (q' e)^{2/p} k \|W\|_{\Psi_p} \gamma_1(\mathcal{F}_*, \mathcal{S}_{L^2}, d_{L^2}),
\]
where the first inequality uses that $\mathcal{G}$ is $(L, \eta)$-$\Psi_p$ and the last inequality uses that $\gamma_1(\mathcal{F}_* \cap rS_{L^2}, d_{L^2}^n) \leq \gamma_0(\mathcal{F}_* \cap rS_{L^2}, d_{L^2})$ as long as $r \leq 1$. Similarly:

$$\Delta_1(\mathcal{F}_* \cap rS_{L^2}) \leq 2(q/c)^{2/p}Lk\|W\|_{\Psi_p}r^n. \tag{44}$$

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 $\mathcal{G}$ be $(L, \eta)$-$\Psi_p$, and let $g \in \mathcal{G}$ be such that $\|g\|_{L^2} = r$. For $\tau \in \mathbb{R}_+$, define $g_\tau \triangleq g1_{\|g\| \leq \tau}$. There exists a truncation level $\tau$ and a universal positive constant $c > 0$ such that:

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

and

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

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

$$\|g\|_{L^2}^2 - \|g_\tau\|_{L^2}^2 \leq \mathbf{E}[\|g\|^21_{\|g\| > \tau}] \leq \mathbf{E}[\|g\|^41_{\|g\| > \tau}] \leq \sqrt{\mathbf{E}[\|g\|^4]} \exp \left( -cr^p/\|g\|_{\Psi_p}^2 \right). \tag{47}$$

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

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

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

$$\|g\|_{\Psi_p} \leq L\|g\|_{L^2} = Lr^n,$$

and

$$\mathbf{E}[\|g\|^4 \leq 4^{4/p}\|g\|_{\Psi_p}^4 \leq 4^{4/p}Lr^{4\eta}. \tag{49}$$

Hence our choice of $\tau$ satisfies:

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

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 $\mathcal{F}_* - \mathcal{F}_* = (L, \eta)$-$\Psi_p$. Suppose further that the model $\mathcal{P}_{(X, Y), 1, n}$ is stationary and $k$-wise independent where $k$ divides $n$. There exist a universal positive constant $c$ such that uniformly for all $f \in \mathcal{F}_* \setminus \{rB_{L^2}\}$ we have that with probability at least $1 - \delta$:

$$\frac{1}{n} \sum_{i=1}^n \|f(X_i)\|^2 \geq r^2(1 - \varepsilon^2)$$

$$- c \left\{ \frac{n^{-1/2}}{k^L1^{3/4p}r^n} \left( \log \left( \frac{A^{2/p}L}{\varepsilon r} \right) \right)^{1/p} \times \left( \gamma_2(q') \mathcal{G}_* \cap rS_{L^2}, d_{L^2} \right) + r^{1+3n} \sqrt{\log(1/\delta)} \right\}$$

$$+ n^{-1}(q')^{1/p}kr^n \left( \log \left( \frac{A^{2/p}L}{\varepsilon r} \right) \right)^{1/p} L^2$$

$$\times \left( \gamma_0(\mathcal{F}_* \cap rS_{L^2}, d_{L^2}) + r^n \log(1/\delta) \right) . \tag{16}$$
Proof. By star-shapedness, we may assume without loss of generality that $f \in \mathcal{F}_s \cap rS_{L^2}$. Fix $\tau$ such that for all such $f$

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

(51)

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

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

(52)

and, for well-chosen $\varepsilon, r$, 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 $\lambda$:

$$\mathbb{E}\exp \left( \lambda \left[ \sum_{i=1}^{n} \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2 \right] \right) \leq \exp \left( \frac{\lambda^2n\mathbb{V}_4 \left( \frac{1}{\sqrt{k}}\sum_{i=1}^{k} \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2 \right)}{2 \left( 1 - \lambda(q'\varepsilon)^1/pk \right) \left[ \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2 \right]^{p/2}} \right),$$  

(53)

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

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

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

(54)

Consequently by Lemma B.1 and $\tau$-truncation:

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

(55)

A centering argument thus gives:

$$\left\| \left[ \sum_{i=1}^{n} \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2 \right]^{p/2} \right\|_{\Psi_{p/2}} \leq 2^{2+2/p}\|f - g\|_{\Psi_p}.$$  

(56)

wherefore we set

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

(57)

Let us next address the variance term:

$$\mathbb{V}_4 \left( \frac{1}{\sqrt{k}}\sum_{i=1}^{k} \|f_\tau(X_i)\|^2 - \|f_\tau(X_i)\|^2 \right)$$

$$= \mathbb{V}_4 \left( \frac{1}{\sqrt{k}}\sum_{i=1}^{k} (f_\tau(X_i) + g_\tau(X_i), f_\tau(X_i) - g_\tau(X_i)) \right)$$  

(58)

(use (54))

$$\leq \frac{1}{k} \mathbb{E} \left( \frac{1}{\sqrt{k}}\sum_{i=1}^{k} \langle f_\tau(X_i) + g_\tau(X_i), f_\tau(X_i) - g_\tau(X_i) \rangle \right)^4$$

$$\leq k\mathbb{E} \left( \langle f_\tau(X) + g_\tau(X), f_\tau(X) - g_\tau(X) \rangle \right)^4$$

(Cauchy-Schwarz)

$$\leq 4k\tau^2\|f - g\|_{L^2}^2 \triangleq d_2^2(f, g).$$

($\tau$-boundedness and Cauchy-Schwarz)

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

$$\mathbb{E}\exp \left( \lambda \left[ \sum_{i=1}^{n} \|f_\tau(X_i)\|^2 - \|f_\tau\|_{L^2}^2 - \|g_\tau(X_i)\|^2 + \|g_\tau\|_{L^2}^2 \right] \right) \leq \exp \left( \frac{\lambda^2\mathbb{V}_4^2(f, g)}{2 (1 - \lambda d_1(f, g))} \right).$$  

(59)
We thus obtain the probability estimate \((u > 0)\):

\[
P\left( \frac{1}{n} \sum_{i=1}^{n} \| f_i(X_i) \|^2 - \| f_i \|_{L^2}^2 > c' \sqrt{u/n} d_2(f, g) + c(u/n) d_1(f, g) \right) \leq 2e^{-u}
\]

(60)

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

\[
\tilde{d}_1(f, g) \equiv c_1 n^{-1} d_1(f, g) = c_1 n^{-1} (q' e)^{1/2} \sqrt{k n^{2+2/p} \tau} \| f - g \|_{\Psi_p},
\]

(61)

\[
\tilde{d}_2(f, g) \equiv c' n^{-1/2} d_2(f, g) = c_2 n^{-1/2} (k n^{2} \| f - g \|_4^2),
\]

(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:

\[
\sup_{f \in \mathcal{F}, rB_L} \left\{ \frac{1}{n} \sum_{i=1}^{n} \| f_i(X_i) \|^2 - \| f_i \|_{L^2}^2 \right\} \leq c_3 (\gamma_2(\mathcal{F}_s \cap rB_{L^2}, \tilde{d}_1, \tilde{d}_2) + \sqrt{u} \Delta_{\tilde{d}_1}(\mathcal{F}_s \cap rB_{L^2})) + c_4 (\gamma_1(\mathcal{F}_s \cap rB_{L^2}, \tilde{d}_1) + u \Delta_{\tilde{d}_1}(\mathcal{F}_s \cap rB_{L^2}))
\]

(63)

for two universal positive constants \(c_3, c_4\).

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

\[
\Delta_{\tilde{d}_1}(\mathcal{F}_s \cap rB_{L^2}) = c_1 n^{-1} (q' e)^{1/2} \sqrt{k n^{2+2/p} \tau} \Delta_{\Psi_p}(\mathcal{F}_s \cap rB_{L^2}) \leq c_1 n^{-1} (q' e)^{1/2} k n^{2+2/p} \tau L \tau^n
\]

(64)

and also (recalling (62)):

\[
\Delta_{\tilde{d}_2}(\mathcal{F}_s \cap rB_{L^2}) \leq c' n^{-1/2} \sqrt{k n^{4}} \Delta_{\Psi_p}(\mathcal{F}_s \cap rB_{L^2}) \leq c' n^{-1/2} \sqrt{k n^{4}} L^{3/4} \gamma^n,
\]

(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 \(\gamma\)-functionals, we have:

\[
\gamma_1(\mathcal{F}_s \cap rS_{L^2}, \tilde{d}_1) \leq c_1 n^{-1} (q' e)^{1/2} k n^{2+2/p} \tau \gamma_n(\mathcal{F}_s \cap rS_{L^2}, d_{L^2})
\]

(66)

and

\[
\gamma_2(\mathcal{F}_s \cap rS_{L^2}, \tilde{d}_2) \leq c' n^{-1/2} \sqrt{k n^{4}} \gamma_{2+\tau \gamma}^n(\mathcal{F}_s \cap rS_{L^2}, d_{L^2}).
\]

(67)

Putting everything together we thus obtain that:

\[
\sup_{f \in \mathcal{F}, rB_L} \left\{ \frac{1}{n} \sum_{i=1}^{n} \| f_i(X_i) \|^2 - \| f_i \|_{L^2}^2 \right\} \leq C L^{3/4} n^{-1/2} \sqrt{k} \left( \gamma_{2+\tau \gamma}^n(\mathcal{F}_s \cap rS_{L^2}, d_{L^2}) + n^{4+3\tau} \sqrt{\log(1/\delta)} \right)
\]

\[
+ C' n^{4/p} n^{-1} (q')^{1/2} k n^{2+2/p} \tau L \gamma_n(\mathcal{F}_s \cap rS_{L^2}, d_{L^2}) + n^{4 \tau} \gamma_n(\mathcal{F}_s \cap rS_{L^2}, d_{L^2})
\]

\[
= C n^{-1/2} \sqrt{k n^{4+3\tau}} \left( c^{-1} \log \left( \frac{4^2 p L}{\varepsilon \tau} \right) \right)^{1/p} \left( \gamma_{2+\tau \gamma}^n(\mathcal{F}_s \cap rS_{L^2}, d_{L^2}) + r^{1+3\tau} \sqrt{\log(1/\delta)} \right)
\]

\[
+ C' n^{4/p} n^{-1} (q')^{1/2} k n^{2+2/p} \tau n \left( c^{-1} \log \left( \frac{4^2 p L}{\varepsilon \tau} \right) \right)^{1/p} L^2 \left( \gamma_n(\mathcal{F}_s \cap rS_{L^2}, d_{L^2}) + r^{4 \tau} \log(1/\delta) \right),
\]

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:

$$Z_{1:O}^o \triangleq (Z_{a_1}, \ldots, Z_{a_{2m}-1}), \quad Z_{1:E}^e \triangleq (Z_{a_2}, \ldots, Z_{a_{2m}}),$$

(68)

Let $\tilde{Z}_{1:O}^o$ and $\tilde{Z}_{1:E}^e$ be blockwise decoupled versions of (68). That is we posit that $\tilde{Z}_{1:O}^o \sim P_{\tilde{Z}_{1:O}^o}$ and $\tilde{Z}_{1:E}^e \sim P_{\tilde{Z}_{1:E}^e}$, where:

$$P_{\tilde{Z}_{1:O}^o} \triangleq P_{Z_{a_1}} \otimes P_{Z_{a_3}} \otimes \cdots \otimes P_{Z_{a_{2m}-1}} \quad \text{and} \quad P_{\tilde{Z}_{1:E}^e} \triangleq P_{Z_{a_2}} \otimes P_{Z_{a_4}} \otimes \cdots \otimes P_{Z_{a_{2m}}},$$

(69)

The process $\tilde{Z}_{1:n}$ with the same marginals as $\tilde{Z}_{1:O}^o$ and $\tilde{Z}_{1:E}^e$ is said to be the decoupled version of $Z_{1:n}$. To be clear: $P_{\tilde{Z}_{1:n}} \triangleq P_{Z_{a_1}} \otimes P_{Z_{a_2}} \otimes \cdots \otimes P_{Z_{a_{2m}}}$, so that $\tilde{Z}_{1:O}^o$ and $\tilde{Z}_{1:E}^e$ 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 $\beta$-mixing process $Z_{1:n}$ and let $\tilde{Z}_{1:n}$ be its decoupled version. For any measurable function $f$ of $Z_{1:O}^o$ (resp. $g$ of $Z_{1:E}^e$) with joint range $[0, 1]$ we have that:

$$|E(f(Z_{1:O}^o)) - E(f(\tilde{Z}_{1:O}^o))| \leq \sum_{i \in E \setminus \{2m\}} \beta_Z(a_i),$$

$$|E(g(Z_{1:E}^e)) - E(g(\tilde{Z}_{1:E}^e))| \leq \sum_{i \in O \setminus \{1\}} \beta_Z(a_i).$$

(70)

E.2. Controlling Empirical Processes for $\beta$-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 $\beta$-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 $\mathcal{F}$. Suppose that $\mathcal{F}_* - \mathcal{F}_*$ is $(L, \eta)\Psi_p$. Suppose further that the model $P_{(X,Y)}^{\mathcal{F}_*}$ is stationary and $\beta$-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 $1 - \delta - \frac{4}{r^2} \beta(k)$:

$$\sup_{f \in \mathcal{F}_* \cap rS_{L^2}} \frac{1}{n} \sum_{i=1}^{n} (1 - E)(W_i, f) \leq c_2 \sqrt{V_2 q(\mathcal{F}_* \cap rS_{L^2}, \mathcal{L}_2)} \left( \frac{1}{r \sqrt{n}} \gamma_2(\mathcal{F}_* \cap rS_{L^2}, \mathcal{L}_2) + \sqrt{\frac{\log(1/\delta)}{n}} \right) + c_1 (q'e)^{2/p} L k ||W||_{\Psi_p} \left( \frac{1}{r \sqrt{n}} \gamma_q(\mathcal{F}_* \cap rS_{L^2}, \mathcal{L}_2) + \frac{r^{q-1}}{n} \log(1/\delta) \right)$$

(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 $\mathcal{F}_* - \mathcal{F}_*$ is $(L, \eta)\Psi_p$. Suppose further that the model $P_{(X,Y)}^{\mathcal{F}_*}$ is stationary and $\beta$-mixing and suppose further that $k \in \mathbb{N}$ divides $n/2$. There exists a universal positive constant $c$ such that uniformly for all $f \in \mathcal{F}_* \cap rS_{L^2}$ we have that with probability at least $1 - \delta - \frac{4}{r^2} \beta(k)$:
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) $\mathcal{F}$ is convex or (2) $\mathcal{F}$ is realizable. For every $r > 0$ we have that:

$$\|\hat{f} - f_*\|^2 \leq r^2 + \frac{1}{r^2} \left( \sup_{g \in \mathcal{F}, r S_{L_2}} M_n(g) \right)^2 + \sup_{g \in \mathcal{F}} Q_n(g).$$

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

$$\frac{1}{n} \sum_{i=1}^n \|\hat{f}(X_i) - f_*(X_i)\|^2 \leq 2 \frac{1}{n} \sum_{i=1}^n (W_i, (\hat{f} - f_*)(X_i)).$$

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

$$\frac{1}{n} \sum_{i=1}^n \|\hat{f}(X_i) - f_*(X_i)\|^2 \leq 2 \frac{1}{n} \sum_{i=1}^n (1 - E')(W_i, (\hat{f} - f_*)(X_i))$$

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

Consequently we also have that:

$$\|\hat{f} - f_*\|^2 = \left(1 + \varepsilon\right) \frac{1}{n} \sum_{i=1}^n \|\hat{f}(X_i) - f_*(X_i)\|^2 + \|\hat{f} - f_*\|^2 - \left(1 + \varepsilon\right) \frac{1}{n} \sum_{i=1}^n \|\hat{f}(X_i) - f_*(X_i)\|^2$$

$$\leq \frac{2(1 + \varepsilon)}{n} \sum_{i=1}^n (1 - E')(W_i, (\hat{f} - f_*)(X_i)) + \|\hat{f} - f_*\|^2 - \left(1 + \varepsilon\right) \frac{1}{n} \sum_{i=1}^n \|\hat{f}(X_i) - f_*(X_i)\|^2$$

(76)

Fix now a radius $r$ and set $g = \frac{r}{\|\hat{f} - f_*\|_{L^2}} (\hat{f} - f_*)$. If $\|\hat{f} - f_*\|_{L^2} \geq r$, dividing both sides above by $\|\hat{f} - f_*\|_{L^2}$ yields for the first term above in (76):\n
$$\frac{2(1 + \varepsilon)}{n\|\hat{f} - f_*\|_{L^2}} \sum_{i=1}^n (1 - E')(W_i, (\hat{f} - f_*)(X_i))$$

$$= \frac{1}{n} \sum_{i=1}^n \left\{2(1 + \varepsilon)(1 - E')(W_i, r^{-1} g(X_i))\right\} \quad \text{(df. of } g \text{ and divide)}$$

$$\leq r^{-1} \sup_{g \in \mathcal{F}, r S_{L_2}} M_n(g).$$

(77)
Either the above inequality holds or $\|\hat{f} - f^*\|_{L^2} \leq r$. For every $r > 0$ it is thus true that:

$$\|\hat{f} - f^*\|_{L^2}^2 \leq r^2 + \left( r^{-1} \sup_{g \in \mathcal{F} \cap rS_{L^2}} M_n(g) \right)^2 + \sup_{g \in \mathcal{F}} Q_n(g) \quad (78)$$

This proves the claim. ■

**Finishing the proof of Theorem 3.1.** We apply Lemma F.1 with $r = r^*$, $\varepsilon = 1/2$ and note that $n \geq c_3 \max \{ n_{\text{quad}}(r^*), n_{\text{mult}}(r^*) \}$ implies: (1) in combination with Proposition E.3 that $\sup_{g \in \mathcal{F}} Q_n(g) \lesssim r_{\ast}^2$; and (2) in combination with Proposition E.2 that $\left( \sup_{g \in \mathcal{F} \cap rS_{L^2}} M_n(g) \right)^2$ 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 $k \geq k_{\text{mix}}$. ■

**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 $\mathcal{F}$ that is either (1) convex or (2) realizable. Suppose that $\mathcal{F} - \mathcal{F}^*$ is $(L, \eta)\cdot \Psi_p$. Suppose further that the model $P_{(X,Y)_{1:n}}$ is stationary and that $k$ divides $n/2$.

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

$$\log N_{L^2}(\mathcal{F}, s) \leq d_{\mathcal{F}} \log \left( \frac{1}{s} \right) \quad (21)$$

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

$$\|\hat{f} - f^*\|_{L^2}^2 \leq c V_{2q} \left( \mathcal{F} \cap \sqrt{\frac{d_{\mathcal{F}} k \|W\|_{L^2}^2}{n}} S_{L^2} \right) \times \left( \frac{d_{\mathcal{F}} + \log(1/\delta)}{n} \right) \quad (22)$$

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

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

**Proof.** Let us begin by observing that for some constant $c_\eta$ only depending on $\eta$ we have that:

$$\gamma_\eta(\mathcal{F} \cap rS_{L^2}, d_{L^2}) \leq c_\eta \int_0^r \left( d_{\mathcal{F}} \log \left( \frac{1}{s} \right) \right)^{1/\eta} ds$$

$$= c_\eta d_{\mathcal{F}}^{1/\eta} r \int_0^1 \left( \log \left( \frac{r}{s} \right) \right)^{1/\eta} ds$$

$$\leq c_\eta d_{\mathcal{F}}^{1/\eta} r \Gamma(1/\eta + 1). \quad (79)$$

Hence for some universal positive constant $c$:

$$\sqrt{V(\mathcal{F} \cap rS_{L^2}) \times \frac{1}{r\sqrt{n}}} \gamma_2(\mathcal{F} \cap rS_{L^2}, d_{L^2}) \leq c \sqrt{V(\mathcal{F} \cap rS_{L^2}) \times \frac{1}{\sqrt{n}} d_{\mathcal{F}}^{1/2}}. \quad (80)$$

22
A few applications of the Cauchy-Schwarz inequality now yields for any $r$:

$$
\mathbf{V}(\mathcal{F} \cap r S_L^2) \leq k\|W\|_{L^2}^2.
$$

(81)

A candidate choice is therefore $r_* = \left( \frac{d \mathbf{V}(\mathcal{F} \cap r S_L^2)}{\mathbf{V}(\mathcal{F} \cap r S_L^2)} \right)^{1/n}$. A straightforward but tedious calculation now reveals that the inequality $n \geq \max \{ n_{\text{quad}}(r_*), n_{\text{mult}}(r_*) \}$ has a solution depending polynomially on problem data as long as $n > 1/4$.

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 $X = \mathbb{R}^d$ and $Y = \mathbb{R}$. Suppose that $k$ divides $n/2$ and that the model $\mathcal{P}_{(X,Y)_{1:n}}$ is stationary and satisfies $Y_i = (\beta^*, X_i) + W_i$ 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

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

(24)

we have that:

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

(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

$$
\gamma_1(\mathcal{F}_* \cap r S_L^2, d L^2) \leq c dr, \quad \text{and} \quad \gamma_2(\mathcal{F}_* \cap r S_L^2, d L^2) \leq c \sqrt{d r}.
$$

(82)

Moreover, since $W_{1:n}$ is a martingale difference sequence we have $\mathbf{V}(\mathcal{F}_* \cap r S_L^2) = \frac{1}{n} \sum_{i=1}^n \mathbf{V}(W_i)$. 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 $r_* \propto \sqrt{\mathbf{V}(W) \times \frac{d}{n}}$.

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

$$
n_{\text{quad}}(r_*) = \inf \left\{ n \in \mathbb{N} \left| \begin{array}{l}
\left[ n^{-1/2} \sqrt{k L} L^{1+3/4} r_* \times \left( r_* \sqrt{d} + r_* \log(1/\delta) \right) \\
+ n^{-1} L^2 k r_*(r_* \sqrt{d} + r_* \log(1/\delta)) \right] \leq r_*^2
\end{array} \right. \right\}
$$

$$
\leq \inf \left\{ n \in \mathbb{N} \left| \begin{array}{l}
\left[ n^{-1/2} \sqrt{k L} L^{1+3/4} \times \left( \sqrt{d} + \sqrt{\log(1/\delta)} \right) \right] \leq 1
\end{array} \right. \right\}
$$

$$
+ \inf \left\{ n \in \mathbb{N} \left| \begin{array}{l}
\left[ n^{-1} L^2 k (d + \log(1/\delta)) \right] \leq 1
\end{array} \right. \right\}
$$

$$
\leq 2k(L \vee 1)^{3+1/2} (d + \log(1/\delta)).
$$
Next, we turn to $n_{\text{mult}}$:

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

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

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

as we sought to prove.