A Finite-Sample Generalization Bound for Semiparametric Regression: Partially Linear Models

Ruitong Huang **Csaba Szepesvári**
Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8

Abstract

In this paper we provide generalization bounds for semiparametric regression with the so-called partially linear models where the regression function is written as the sum of a linear parametric and a nonlinear, nonparametric function, the latter taken from a some set $\mathcal{H}$ with finite entropy-integral. The problem is technically challenging because the parametric part is unconstrained and the model is underdetermined, while the response is allowed to be unbounded with subgaussian tails. Under natural regularity conditions, we bound the generalization error as a function of the metric entropy of $\mathcal{H}$ and the dimension of the linear model. Our main tool is a ratio-type concentration inequality for increments of empirical processes, based on which we are able to give an exponential tail bound on the size of the parametric component. We also provide a comparison to alternatives of this technique and discuss why and when the unconstrained parametric part in the model may cause a problem in terms of the expected risk. We also explain by means of a specific example why this problem cannot be detected using the results of classical asymptotic analysis often seen in the statistics literature.

1 INTRODUCTION

In this paper we consider finite-time risk bounds for empirical risk-minimization algorithms for *partially linear stochastic models* of the form

$$Y_i = \phi(X_i)^\top \theta + h(X_i) + \varepsilon_i, \quad 1 \le i \le n, \qquad (1)$$

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

where X_i is an input, Y_i is an observed, potentially unbounded response, ε_i is noise, ϕ is the known basis function, θ is an unknown, finite dimensional parameter vector and h is a nonparametric function component. The most well-known example of this type of model in machine learning is the case of Support Vector Machines (SVMs) with offset (in this case $\phi(x) \equiv 1$). The general partially linear stochastic model, which perhaps originates from the econometrics literature [e.g., Engle et al., 1986, Robinson, 1988, Stock, 1989], is a classic example of semiparametric models that combine parametric (in this case $\phi(\cdot)^\top\theta$) and nonparametric components (here h) into a single model. The appeal of semiparametric models has been widely discussed in statistics, machine learning, control theory or other branches of applied sciences [e.g., Bickel et al., 1998, Smola et al., 1998, Härdle et al., 2004, Gao, 2007, Kosorok, 2008, Greblicki and Pawlak, 2008, Horowitz, 2009]. In a nutshell, whereas a purely parametric model gives rise to the best accuracy if correct, it runs the risk of being misspecified. On the other hand, a purely nonparametric model avoids the risk of model misspecification, therefore achieving greater applicability and robustness, though at the price of the estimates perhaps converging at a slower rate. Semiparametric models, by combining parametric and nonparametric components into a single model, aim at achieving the best of both worlds. Another way of looking at them is that they allow to add prior "structural" knowledge to a nonparametric model, thus potentially significantly boosting the convergence rate when the prior is correct. For a convincing demonstration of the potential advantages of semiparametric models, see, e.g., the paper by Smola et al. [1998].

Despite all the interest in semiparametric modeling, to our surprise we were unable to find any work that would have been concerned with the finite-time *predictive performance* (i.e., risk) of semiparametric methods. Rather, existing theoretical works in semiparametrics are concerned with discovering conditions and algorithms for constructing statistically efficient estimators of the unknown parameters of the parametric part. This problem has been more or less settled in the book by Bickel et al. [1998], where sufficient and necessary conditions are described along with recipes for constructing statistically efficient procedures. Although statistical efficiency (which roughly means achieving the Cramer-Rao lower bound as the sample size increases indefinitely) is of major interest, statistical efficiency does not give rise to finite-time bounds on the excess risk, the primary quantity of interest in machine learning. In this paper, we make the first initial steps to provide these missing bounds.

The closest to our work are the papers of Chen et al. [2004] and Steinwart [2005], who both considered the risk of SVMs with offset (a special case of our model). Here, as noted by both authors, the main difficulty is bounding the offset. While Chen et al. [2004] bounded the offset based on a property of the optimal solution for the hinge loss and derived finite-sample risk bounds, Steinwart [2005] considered consistency for a larger class of "convex regular losses". Specific properties of the loss functions were used to show high probability bounds on the offset. For our more general model, similarly to these works the bulk of the work will be to prove that with high probability the parametric model will stay bounded (we assume $\sup_x \|\phi(x)\|_2 < +\infty$). The difficulty is that the model is underdetermined and in the training procedures only the nonparametric component is penalized. This suggests that perhaps one could modify the training procedure to penalize the parametric component, as well. However, it appears that the semiparametric literature largely rejects this approach. The main argument is that a penalty would complicate the tuning of the method (because the strength of the penalty needs to be tuned, too), and that the parametric part is added based on a strong prior belief that the features added will have a significant role and thus rather than penalizing them, the goal is to encourage their inclusion in the model. Furthermore, the number of features in the parametric part are typically small, thus penalizing them is largely unnecessary. However, we will return to discussing this issue at the end of the article.

Finally, let us make some comments on the computational complexity of training partially linear models. When the nonparametric component belongs to an RKHS, an appropriate version of the representer theorem can be used to derive a finite-dimensional optimization problem [Smola et al., 1998], leading to quadratic optimization problem subject to linear constrains. Recent work by Kienzle and Schölkopf [2005] and Lee and Wright [2009] concern specialized solvers to find an approximate optimizer of the arising problem. In particular, in their recent work Lee and Wright [2009] proposed a decomposition algorithm that is capable to deal with large-scale semiparametric SVMs.

The main tool in the paper is a ratio-type concentration inequality due to van de Geer [2000]. With this, the boundedness of the parameter vector is derived from the properties of the loss function: The main idea is to use the level sets of the empirical loss to derive the required bounds. Although our main focus is the case of the quadratic loss, we study the problem more generally. In particular, we require the loss function to be smooth, Lipschitz, "non-flat" and convex, of which the quadratic loss is one example.

The paper is organized as follows. We first define the notation we use and give the details of the problem setting. In the next section, we state our assumptions and the results, together with a comparison to alternative approaches. All the proofs are in the Appendix due to space limits.

2 PROBLEM SETTING AND NOTATION

Throughout the paper, the input space $\mathcal{X}$ will be a separable, complete metric space, and $\mathcal{Y}$, the label space, will be a subset of the reals $\mathbb{R}$. In this paper, we allow $Y \in \mathcal{Y}$ to be unbounded. Given the independent, identically distributed sample $Z_{1:n} = (Z_1, ..., Z_n)$, $Z_i = (X_i, Y_i)$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, the partially constrained empirical risk minimization problem with the partially linear stochastic model (1) is to find a minimizer of

$$\min_{\theta \in \mathbb{R}^d, h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(Y_i, \phi(X_i)^\top \theta + h(X_i)\right) ,$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is a loss function, $\phi : \mathcal{X} \to \mathbb{R}^d$ is a basis function and $\mathcal{H}$ is a set of real-valued functions over $\mathcal{X}$, holding the "nonparametric" component h. Our main interest is when the loss function is quadratic, i.e., $\ell(y, y') = \frac{1}{2}(y - y')^2$, but for the sake of exploring how much we exploit the structure of this loss, we will present the results in an abstract form.

Introducing $\mathcal{G} = \left\{\phi(\cdot)^\top \theta : \theta \in \mathbb{R}^d\right\}$, the above problem can be written in the form

$$\min_{g \in \mathcal{G}, h \in \mathcal{H}} L_n(g + h), \tag{2}$$

where $L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$. Typically, $\mathcal{H}$ arises as the set $\{h : \mathcal{X} \to \mathbb{R} : J(h) \leq K\}$ with some $K > 0$ and some functional J that takes larger values for "rougher" functions.[1]

[1] The penalized empirical risk-minimization problem, $\min_{g \in \mathcal{G}, h} L_n(h + g) + J(h)$ is closely related to (2) as suggested by the identity $\min_{g \in \mathcal{G}, h} L_n(g + h) + \lambda J(h) = \min_{K \geq 0} \lambda K + \min_{g \in \mathcal{G}, h: J(h) \leq K} L_n(g + h)$ explored in a specific context by Blanchard et al. [2008].

The goal of learning is to find a predictor with a small expected loss. Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, the expected loss, or *risk*, of f is defined to be $L(f) = \mathbb{E}\left[\ell(Y, f(X))\right]$, where $Z = (X, Y)$ is an independent copy of $Z_i = (X_i, Y_i)$ $(i = 1, \ldots, n)$. Let (g_n, h_n) be a minimizer[2] of (2) and let $f_n = g_n + h_n$.

When analyzing a learning procedure returning a function f_n, we compare the risk $L(f_n)$ to the best risk possible over the considered set of functions, i.e., to $L^* = \inf_{g\in\mathcal{G},h\in\mathcal{H}} L(g + h)$. A bound on the *excess risk* $L(f_n) - L^*$ is called a generalization (error) bound. In this paper, we seek bounds in terms of the entropy-integral of $\mathcal{H}$. Our main result, Theorem 3.2, provides such a bound, essentially generalizing the analogue result of Bartlett and Mendelson [2002]. In particular, our result shows that, in line with existing empirical evidence, the price of including the parametric component in terms of the increase of the generalization bound is modest, which, in favourable situations, can be far outweighed by the decrease of L^* that can be attributed to including the parametric part. However, in terms of the expected excess risk, the unconstrained parametric part may cause a problem in some case.

By the standard reasoning, the excess risk is decomposed as follows:

$$L(f_n) - L(f^*) = (L(f_n) - L_n(f_n)) + \underbrace{(L_n(f_n) - L_n(f^*))}_{\le 0} + (L_n(f^*) - L(f^*))\,, \tag{3}$$

where $f^* = \arg\min_{f\in\mathcal{G}+\mathcal{H}} L(f)$. Here, the third term can be upper bounded as long as f^* is "reasonable" (e.g., bounded). On the other hand, the first term is more problematic, at least for unbounded loss functions and when Y is unbounded. Indeed, in this case f_n can take on large values and correspondingly $L(f_n)$ could also be rather large. Note that this is due to the fact that the parametric component is unconstrained.

The classical approach to deal with this problem is to introduce, clipping, or truncation of the predictions (cf. Theorem 11.5 of Györfi et al. [2002]). However, clipping requires additional knowledge such as that Y is bounded with a known bound. Furthermore, the clipping level appears in the bounds, making the bounds weak when the level is conservatively estimated In fact, one suspects that clipping is unnecessary in our setting where we will make strong enough assumptions on the tails of Y (though much weaker than assuming that Y is bounded). In fact, in practice, it is quite rare to see clipping implemented. Hence, in what follows we will keep to our original goal and analyze the procedure with no clipping. Further comparison to results with clipping will be given after our main results are presented.

To analyze the excess risk we will proceed by showing that with large probability, $\|g_n\|_\infty$ is controlled. This is, in fact, where the bulk of the work will lie.

[2]For simplicity, we assume that this minimizer and in fact all the others that we will need later exist. This is done for the sake of simplifying the presentation: The analysis is simple to extend to the general case. Further, if there are multiple minimizers, we choose one.

3 ASSUMPTIONS AND RESULTS

In this section we state our assumptions, which will be followed by stating our main result. We also discuss a potential problem caused by including the unconstrained parametric part, and explain why standard asymptotic analysis can not detect this problem. Due to the space limit, all the proofs are postponed to the Appendix. Before stating our assumptions and results, we introduce some more notation. We will denote the Minkowski-sum of $\mathcal{G} + \mathcal{H}$ of $\mathcal{G}$ and $\mathcal{H}$ by $\mathcal{F}$: $\mathcal{F} = \mathcal{G} + \mathcal{H} \doteq \{g + h : g \in \mathcal{G}, h \in \mathcal{H}\}$. The L^2-norm of a function is defined as $\|f\|_2^2 \doteq \mathbb{E}\left[f^2(X)\right]$, while given the random sample $X_{1:n} = (X_1, \ldots, X_n)$, the n-norm of a function is defined as the (scaled) ℓ^2-norm of the restriction of the function to $X_{1:n}$: $\|f\|_n^2 = \frac{1}{n}\sum_i f(X_i)^2$. The vector $(f(X_1), \ldots, f(X_n))^\top$ is denoted by $f(X_{1:n})$. The matrix $(\phi(X_1), \ldots, \phi(X_n))^\top \in \mathbb{R}^{n\times d}$ is denoted by Φ (or $\Phi(X_{1:n})$ if we need to indicate its dependence on $X_{1:n}$). We let $\hat{G} = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ be the empirical Grammian matrix and $G = \mathbb{E}[\phi(X)\phi(X)^\top]$ be the population Grammian matrix underlying ϕ. Denote the minimal positive eigenvalue of G by $\lambda_{\min}$, while let $\hat{\lambda}_{\min}$ be the same for $\hat{G}$. The rank of G is denoted by $\rho = \mathrm{rank}(G)$. Lastly, let $L_{h,n}(g) = L_n(h+g)$, $\overline{L}_n(f) = \mathbb{E}\left[L_n(f) \mid X_{1:n}\right]$ and $\overline{L}_{h,n}(g) = \mathbb{E}\left[L_n(h + g) \mid X_{1:n}\right]$.

3.1 Assumptions

In what follows we will assume that the functions in $\mathcal{H}$ are bounded by $r > 0$. If $\mathcal{K}$ is an RKHS space with a continuous reproducing kernel κ and $\mathcal{X}$ is compact (a common assumption in the literature, e.g., Cucker and Zhou 2007, Steinwart and Christmann 2008), this assumption will be satisfied if $J(h) = \|h\|_{\mathcal{K}}$ and $\mathcal{H} = \{h \in \mathcal{K} : J(h) \le r\}$, where, without loss of generality (WLOG), we assume that the maximum of κ is below one.

We will also assume that $R = \sup_{x\in\mathcal{X}} \|\phi(x)\|_2$ is finite. If ϕ is continuous and $\mathcal{X}$ is compact, this assumption will be satisfied, too. In fact, by rescaling the basis functions if needed, we will assume WLOG that $R = 1$.

Definition 1. *Let β, Γ be positive numbers. A (non-centered) random variable X is subgaussian with pa-*

rameters (β, Γ) *if*

$$\mathbb{E}\left[\exp\left(|\beta X|^2\right)\right] \leq \Gamma < \infty.$$

Let us start with our assumptions that partly concern the loss function, ℓ, partly the joint distribution of (X, Y).

Assumption 3.1 (Loss function)**.**

(i) Convexity: The loss function ℓ is convex with respect to its second argument, i.e., $\ell(y, \cdot)$ is a convex function for all $y \in \mathcal{Y}$.

(ii) There exists a bounded measurable function $\hat{h}$ and a constant $Q < \infty$ such that

$$\mathbb{E}\left[\ell\left(Y, \hat{h}(X)\right)|X\right] \leq Q \quad \textit{almost surely.}$$

(iii) Subgaussian Lipschitzness: There exists a function $K_\ell : \mathcal{Y} \times (0, \infty) \to \mathbb{R}$ such that for any constant $c > 0$ and $c_1, c_2 \in [-c, c]$,

$$|\ell(y, c_1) - \ell(y, c_2)| \leq K_\ell(y, c)\,|c_1 - c_2|\,,$$

and such that $\mathbb{E}\left[\exp(|\beta K_\ell(Y, c)|^2)|\,X\right] \leq \Gamma_c < \infty$ for some constant Γ_c depending only on c almost surely. WLOG, we assume that $K_\ell(y, \cdot)$ is a monotonically increasing function for any $y \in \mathcal{Y}$.

(iv) Level-Set: For any $X_{1:n} \subset \mathcal{X}$, and any $c \geq 0$, $R_c = \sup_{f \in \mathcal{F}: \mathbb{E}[L_n(f)|X_{1:n}] \leq c} \|f\|_n$ is finite and independent of n.

The convexity assumption is standard.

Remark 3.1. Assumption 3.1(ii) basically requires Y, even if it is unbounded, still can be approximated by a function in $\mathcal{H}$ at every X with constant expected loss.

Remark 3.2. The subgaussian Lipschitzness assumption is a general form of Lipschitzness property which allows the Lipschitzness coefficient to depend on y.

Remark 3.3. If the loss function is the quadratic loss, the subgaussian Lipschitzness assumption is an immediate corollary of the subgaussian property of Y conditioning on X. In particular, $|(Y - c_1)^2 - (Y - c_1)^2| = |2Y - c_1 - c_2||c_1 - c_2|$. Thus we can pick $K_\ell(Y, c) = 2|Y| + 2c$ and $\beta = \frac{1}{2\sqrt{2}}$, then $\mathbb{E}\left[\exp(|\beta K_\ell(Y, c)|^2)\right] = \mathbb{E}\left[\exp(\frac{1}{2}(|Y| + c)^2)\right] \leq \mathbb{E}\left[\exp(|Y|^2)\right] + \exp(c^2)$.

Remark 3.4. Unlike the first three assumptions, Assumption 3.1(iv), which requires that the sublevel sets of $\mathbb{E}\left[L_n(\cdot)\,|\,X_{1:n}\right]$ are bounded in $\|\cdot\|_n$, is nonstandard. This assumption will be crucial for showing the boundedness of the parametric component of the model. We argue that in some sense this assumption, given the method considered, is necessary. The idea is that since f_n minimizes the empirical loss it should also have a small value of $\mathbb{E}\left[L_n(\cdot)\,|\,X_{1:n}\right]$ (in fact, this is not that simple to show given that it is not known whether f_n is bounded). As such, it will be in some sublevel set of $\mathbb{E}\left[L_n(\cdot)\,|\,X_{1:n}\right]$. Otherwise, nothing prevents the algorithm from choosing a minimizer (even when minimizing $\mathbb{E}\left[L_n(\cdot)\,|\,X_{1:n}\right]$ instead of $L_n(\cdot)$) with an unbounded $\|\cdot\|_n$ norm.

Remark 3.5. One way of weakening Assumption 3.1(iv) is to assume that there exist a minimizer of $\mathbb{E}\left[L_n(\cdot)\,|\,X_{1:n}\right]$ over $\mathcal{F}$ that has a bounded norm and then modify the procedure to pick the one with the smallest $\|\cdot\|_n$ norm.

Example 3.1 (Quadratic Loss). In the case of quadratic loss, i.e., when $\ell(y, y') = \frac{1}{2}(y - y')^2$, $R_c^2 \leq 4c + 8Q + 4s^2$ where $s = \|\hat{h}\|_\infty$. Indeed, this follows from

$$\begin{aligned} \|f\|_n^2 &\leq \frac{2}{n}\sum_i \mathbb{E}\left[(f(X_i) - Y_i)^2\,|X_{1:n}\right] + \mathbb{E}\left[Y_i^2\,|X_{1:n}\right] \\ &\leq 4\mathbb{E}\left[L_n(f)\,|X_{1:n}\right] + \frac{2}{n}\sum_i \mathbb{E}\left[Y_i^2\,|X_i\right]. \end{aligned}$$

Then $\mathbb{E}\left[Y_i^2\,|X_i\right] \leq 2\mathbb{E}\left[(Y_i - \hat{h}(X_i))^2\,|X_i\right] + 2\hat{h}^2(X_i) \leq 4Q + 2s^2$. Here, the last inequality is by Assumption 3.1(ii) and the boundedness of $\hat{h}$.

Example 3.2 (Exponential Loss). In the case of exponential loss, i.e., when $\ell(y, y') = \exp(-yy')$ and if $\mathcal{Y} = \{+1, -1\}$ the situation is slightly more complex. R_c will be finite as long as the posterior probability of seeing either of the labels is uniformly bounded away from one, as assumed e.g., by Blanchard et al. [2008]. Specifically, if $\eta(x) \doteq \mathbb{P}(Y = 1|X = x) \in [\varepsilon, 1 - \varepsilon]$ for some $\varepsilon > 0$ then a simple calculation shows that $R_c^2 \leq c/\varepsilon$.

It will be convenient to introduce the alternate notation $\ell((x, y), f)$ for $\ell(y, f(x))$ (i.e., $\ell((x, y), f) \doteq \ell(y, f(x))$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, $f : \mathcal{X} \to \mathbb{R}$. Given $h \in \mathcal{H}$, let $g_{h,n} = \arg\min_{g \in \mathcal{G}} L_n(h + g) = \arg\min_{g \in \mathcal{G}} L_{h,n}(g)$ and $\overline{g}_{h,n} = \arg\min_{g \in \mathcal{G}} \overline{L}_{h,n}(g)$ ($\overline{L}_{h,n}$ and $L_{h,n}$ are defined at the end of Section 2). The next assumption states that the loss function is locally "not flat":

Assumption 3.2 (Non-flat Loss)**.** *Assume that there exists $\varepsilon > 0$ such that for any $h \in \mathcal{H}$ and vector $a \in [-\varepsilon, \varepsilon]^n \cap \mathrm{Im}(\Phi)$,*

$$\begin{aligned} \frac{\varepsilon}{n}\|a\|_2^2 \leq \mathbb{E}&\left[\frac{1}{n}\sum_i \ell(Z_i, h + \overline{g}_{h,n} + a_i)\,\middle|\, X_{1:n}\right] \\ &- \mathbb{E}\left[\frac{1}{n}\sum_i \ell(Z_i, h + \overline{g}_{h,n})\,\middle|\, X_{1:n}\right] \end{aligned}$$

holds a.s., where recall that $Z_i = (X_i, Y_i)$.

Note that it is key that the "perturbation" a is in the image space of Φ and that it is applied at $h + \overline{g}_{h,n}$ and not at an arbitrary function h, as shown by the next example:

Example 3.3 (Quadratic loss). In the case of the quadratic loss, note that $g(X_{1:n}) = \Phi(X_{1:n})\theta$. Let $\overline{\theta}_{h,n}$ be a minimizer of $\overline{L}_{h,n}(\cdot)$ satisfying $\overline{\theta}_{h,n} = \left(\Phi^\top\Phi\right)^{+}\Phi^\top(\mathbb{E}\left[Y_{1:n}|X_{1:n}\right] - h(X_{1:n}))$. Therefore,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_i \ell((X_i,Y_i), h+\overline{g}_{h,n}+a_i)\,|\,X_{1:n}\right]\\&\qquad - \mathbb{E}\left[\frac{1}{n}\sum_i \ell((X_i,Y_i), h+\overline{g}_{h,n})\,|\,X_{1:n}\right]\\&= \frac{1}{n}\sum_i \mathbb{E}\left[a_i\left\{2(\overline{g}_{h,n}(X_i)+h(X_i)-Y_i)+a_i\right\}\,|\,X_{1:n}\right],\end{aligned}$$

which is equal to $\frac{1}{n}\|a\|_2^2 + \frac{2}{n}a^\top\left\{\Phi\left(\Phi^\top\Phi\right)^{+}\Phi^\top - I\right\}\left\{\mathbb{E}\left[Y_{1:n}|X_{1:n}\right] - h(X_{1:n})\right\} = \frac{1}{n}\|a\|_2^2$, where the last equality follows since $a \in \mathrm{Im}(\Phi)$.

We will need an assumption that the entropy of $\mathcal{H}$ satisfies an integrability condition. For this, recall the definition of entropy numbers:

Definition 2. *For $\varepsilon > 0$, the ε-covering number $N(\varepsilon,\mathcal{H},d)$ of a set $\mathcal{H}$ equipped with a pseudo-metric d is the number of balls with radius ε measured with respect to d necessary to cover $\mathcal{H}$. The ε-entropy of $\mathcal{H}$ is $H(\varepsilon,\mathcal{H},d) = \log N(\varepsilon,\mathcal{H},d)$.*

We will allow d to be replaced by a pseudo-norm, meaning the covering/entropy-numbers defined by the pseudo-distance generated by the chosen pseudo-norm. Note that if $d' \le d$ then the ε-balls w.r.t. d' are bigger than the ε-balls w.r.t. d. Hence, any ε-cover w.r.t. d is also gives an ε-cover w.r.t. d'. Therefore, $N(\varepsilon,H,d') \le N(\varepsilon,H,d)$ and also $H(\varepsilon,H,d') \le H(\varepsilon,H,d)$.

Let $\|\cdot\|_{\infty,n}$ be the infinity empirical norm: For $f : \mathcal{X} \to \mathbb{R}$, $\|f\|_{\infty,n} = \max_{1\le k\le n}|f(X_k)|$. Note that trivially $\|f\|_n \le \|f\|_{\infty,n} \le \|f\|_\infty$. We use $\|\cdot\|_{\infty,n}$ in our next assumption:

Assumption 3.3 (Integrable Entropy Numbers of $\mathcal{H}$)**.** *There exists a (non-random) constant C_H such that, $\int_0^1 H^{1/2}(v,\mathcal{H},\|\cdot\|_{\infty,n})\,dv \le C_H$ holds a.s.*

Remark 3.6. Assumption 3.3 is well-known in the literature of empirical processes to guarantee the uniform laws of large numbers [Dudley, 1984, Giné and Zinn, 1984, Tewari and Bartlett, 2013]. The assumption essentially requires that the entropy numbers of $\mathcal{H}$ should not grow very fast as the scale approaches to zero. For example, this assumption holds if for any $0 < u \le 1$, $H(u,\mathcal{H},\|\cdot\|_{\infty,n}) \le cu^{-(2-\varepsilon)}$ for some $c > 0$, $\varepsilon > 0$. Based on our previous discussion, $H(u,\mathcal{H},\|\cdot\|_{\infty,n}) \le H(u,\mathcal{H},\|\cdot\|_\infty)$; the latter entropy numbers are well-studied for a wide range of function spaces (and enjoy the condition required here). For examples see, e.g., [Dudley, 1984, Giné and Zinn, 1984, Tewari and Bartlett, 2013].

For the next assumption let $G_{\lambda_{\min}}$ be the event when $\hat{\lambda}_{\min} \ge \lambda_{\min}/2$.

Assumption 3.4 (Lipschitzness of the Parametric Solution Path)**.** *Let P_X denote the distribution of X. There exists a constant K_h such that on $G_{\lambda_{\min}}$ for $[P_X]$ almost all $x \in \mathcal{X}$, $h \mapsto \overline{g}_{h,n}(x)$ is K_h-Lipschitz w.r.t. $\|\cdot\|_{\infty,n}$ over $\mathcal{H}$.*

Remark 3.7. When $\overline{g}_{h,n}$ is uniquely defined, Assumption 3.4 will be satisfied whenever ℓ is sufficiently smooth w.r.t. its first argument, as follows, e.g., from the Implicit Function Theorem.

Example 3.4 (Quadratic loss). In the case of the quadratic loss, by Example 3.3,

$$\begin{aligned}\overline{g}_{h,n}(x) &= \langle\phi(x), \left(\Phi^\top\Phi\right)^{+}\Phi^\top(\mathbb{E}\left[Y_{1:n}|X_{1:n}\right] - h(X_{1:n}))\rangle\\&= \frac{1}{n}\sum_i\langle\phi(x), \hat{G}^{+}\phi(X_i)\left(\mathbb{E}\left[Y_i|X_{1:n}\right] - h(X_i)\right)\rangle\end{aligned}$$

Thus, for $h, h' \in \mathcal{H}$, on $G_{\lambda_{\min}}$,

$$\begin{aligned}&|\overline{g}_{h,n}(x) - \overline{g}_{h',n}(x)|\\&\quad = \left|\langle\phi(x), \left(\Phi^\top\Phi\right)^{+}\Phi^\top\left(h'(X_{1:n}) - h(X_{1:n})\right\rangle\right|\\&\quad \le \frac{2\,\|\phi(x)\|_2}{\lambda_{\min}}\,\frac{1}{n}\sum_i|h'(X_i) - h(X_i)|\;\|\phi(X_i)\|_2\\&\quad \le \frac{2}{\lambda_{\min}}\,\|h' - h\|_{\infty,n}\end{aligned}$$

where we used $\|\phi(x)\|_2 \le 1$ multiple times which holds $[P_X]$ a.e. on $\mathcal{X}$.

3.2 Results

Our first main result implies that g_n is bounded with high probability:

Theorem 3.1. *Let Assumptions 3.1 to 3.4 hold. Then, there exist positive constants c_1, c_2, U such that for any $0 < \delta < 1$ and n such that $n \ge c_1 + c_2\frac{\log\left(\frac{2\rho}{\delta}\right)}{\lambda_{\min}}$, it holds that*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\|g_{h,n}\|_\infty \ge U\right) \le \delta\,. \qquad (4)$$

The result essentially states that for some specific value of U, the probability that the event $\sup_{h\in\mathcal{H}}\|g_{h,n}\|_\infty > U$ happens is exponentially small as a function of the sample size n. The constant U is inversely proportional to $\lambda_{\min}$ and depends on both R_c from the level-set assumption and r. Here c depends on Q, $\|\hat{h}\|_\infty$ from Assumption 3.1(ii) and the subgaussian parameters. The actual value of U can be read out from the proof.

The main challenges in the proof of this result are that the bound has to hold uniformly over $\mathcal{H}$ (this allows us to bound $\|g_n\|_\infty$), and also that the response Y is unbounded, as are the functions in $\mathcal{G}$. The main tool is a ratio type tail inequality for empirical processes, allowing us to deal both with the unbounded responses and functions, which is then combined with our assumptions on the loss function, in particular, with the level-set assumption.

Given Theorem 3.1, various high-probability risk bounds can be derived using more or less standard techniques, although when the response is not bounded and clipping is not available, we were not able to identify any result in the literature that would achieve this. In our proof, we use the technique of van de Geer [1990], which allows us to work with unbounded responses without clipping the predictions, to derive our high-probability risk bound. Since this technique was developed for the fixed design case, we combine it with a method, which uses Rademacher complexities, upper bounded in terms of the entropy integral, We use the technique of van de Geer [1990], which allows us to work with unbounded responses without clipping the predictions. Since this technique was developed for the fixed design case, we combine it with a method, which uses Rademacher complexities, upper bounded in terms of the entropy integral, so as to get an out-of-sample generalization bound.[3] The bound in our result is of the order $1/\sqrt{n}$, which is expected given our constraints on the nonparametric class $\mathcal{H}$. However, we note in passing, that under stronger conditions, such as $L(f^*) = 0$ [Pollard, 1995, Haussler, 1992], or the convexity of $\mathcal{F}$ (which does not hold in our case unless we take the convex hull of $\mathcal{F} = \mathcal{G} + \mathcal{H}$), that the true regression function belongs to $\mathcal{F}$, the loss is the quadratic loss (or some other loss which is strongly convex), a faster rate of $O(1/n)$ can also be proved [Lee et al., 1998, Györfi et al., 2002, Bartlett et al., 2005, Koltchinskii, 2006, 2011], though the existing works seem to make various assumptions about Y which we would like to avoid. Hence, we leave the proof of such faster rates for future work.

Let $(x)_+ = \max(x, 0)$ denote the positive part of $x \in \mathbb{R}$.

Theorem 3.2. *Let Assumptions 3.1 to 3.4 hold and let $f^* = g^* + h^*$ be a minimizer of L over $\mathcal{G} + \mathcal{H}$ (i.e., $g^* \in \mathcal{G}$, $h^* \in \mathcal{H}$). There exist positive constants $c, c_1, c_2, c_3, c_4, \alpha$ and $U \geq \|g^*\|_\infty$ such that for any $0 < \delta < 1$ satisfying $\log \frac{1}{\delta} \geq c$ and $n \geq c_1 + c_2 \log\left(\frac{4\rho}{\delta}\right)/\lambda_{\min}$, with probability at least $1 - 3\delta$,*

$$L(f_n) - L(f^*) \leq c_3 \frac{C_H + \rho^{1/2}(\log(U))_+}{\sqrt{n}} + 2(r+U)\sqrt{\frac{\log\frac{2}{\delta}}{\alpha n}} + c_4\sqrt{\frac{\log\frac{1}{\delta}}{2n}} \tag{5}$$

where $f_n = h_n + g_n$ is a minimizer of $L_n(\cdot)$ over $\mathcal{H} + \mathcal{G}$.

Remark 3.8. The constants ρ and $\lambda_{\min}$ appear both in U and in the lower bound constraint of n. Defining $\overline{\ell}(x, p) = \mathbb{E}[\ell(Y, p)|X = x]$, Constant c_3 depends on the (essential) Lipschitz coefficient of $\overline{\ell}(X, p)$ when $p \in [-r - U, r + U]$ and constant c_4 depends on the (essential) range of $\ell(X, p)$. Both of them can be shown to be finite based on Assumption 3.1. The bound has a standard form: The first and the last of the three terms comes from bounding the out-of-sample generalization error, while the term in the middle (containing α) bounds the in-sample generalization error. We use a measure-disintegration technique to transfer the results of van de Geer [1990] which are developed for the fixed design setting (i.e., when the covariates $X_{1:n}$ are deterministic) to the random design setting that we consider in this paper.

Notice that the above high probability result holds only if n is large compared to $\log(1/\delta)$, or, equivalently when δ is not too small compared to n, a condition that is inherited from Theorem 3.1. Was this constraint absent, the tail of $L(f_n) - L(f^*)$ would be of a subgaussian type, which we could integrate to get an expected risk bound. However, because of the constraint, this does not work. With no better idea, one can introduce clipping, to limit the magnitude of the prediction errors on an event of probability (say) $1/n$. This still result in an expected risk bound of the order (i.e., $O(1/\sqrt{n})$), as expected, although with an extra logarithmic factor. However, if one needs to introduce clipping, this could be done earlier, reducing the problem to studying the metric entropy of the clipped version of $\mathcal{F}$ (which is almost what is done in Lemma A.2 given in the supplementary material). For this, assuming Y is bounded, one can use Theorem 11.5 of Györfi et al. [2002]. Note, however, that in this result, for example, the clipping level, which one would probably select conservatively in practice, appears raised to the 4th power. We do not know whether this is a proof artifact (this might worth to be investigated). In comparison, with our technique, the clipping level could actually be made appear only through its logarithm in our bound if we choose $\delta = 1/(Ln)$. On the other hand, our bound scales with $\lambda_{\min}^{-1}$ through U. This is alarming unless the eigenvalues of the Grammian are well-controlled, in which case $\lambda_{\min}^{-1} = O(\sqrt{\rho})$.

Given the imbroglio that the constraint connecting n

[3] "In-sample" generalization bounds concern the deviation $\overline{L}_n(f_n) - \overline{L}_n(f^*)$, while "out-of-sample bounds" concern $L(f_n) - L(f^*)$.

and δ causes, the question arises whether this condition could be removed from Theorem 3.2. The following example, based on Problem 10.3 of Györfi et al. [2002], shows that already in the purely parametric case, there exist perfectly innocent looking problems which make ordinary least squares fail:

Example 3.5 (Failure of Ordinary Least Squares). Let $\mathcal{X} = [0,1]$, $\mathcal{Y} = \mathbb{R}$, $\ell(y,p) = (y-p)^2$, $\phi : \mathcal{X} \to \mathbb{R}^3$, $\phi_1(x) = \mathbb{I}_{[0,1/2]}(x)$, $\phi_2(x) = x \cdot \mathbb{I}_{[0,1/2]}(x)$, $\phi_3(x) = \mathbb{I}_{(1/2,1]}(x)$, where $\mathbb{I}_A$ denotes the indicator of set $A \subset \mathcal{X}$. Let $f_\theta(x) = \phi(x)^\top \theta$, $\theta \in \mathbb{R}^3$. As to the data, let $(X,Y) \in \mathcal{X} \times \{-1,+1\}$ be such that X and Y are independent of each other, X is uniform on $\mathcal{X}$ and $\mathbb{P}(Y=+1) = \mathbb{P}(Y=-1) = 1/2$. Note that $\mathbb{E}[Y|X] = 0$, hence the model is well-specified (the true regression function lies in the span of basis functions). Further, $f^*(x) = 0$. Now, let $(X_1,Y_1),\dots,(X_n,Y_n)$ be n independent copies of (X,Y) and let $\hat{\theta}_n = \arg\min_{\theta\in\mathbb{R}^3} L_n(\phi^\top\theta)$. Denote the empirical Grammian on the data by $\hat{G}_n = \frac{1}{n}\sum_k \phi(X_k)\phi(X_k)^\top$, $\hat{\lambda}_{\min}(n) = \lambda_{\min}(\hat{G}_n)$, $\lambda_{\min} = \lambda_{\min}(\mathbb{E}\left[\phi(X)\phi(X)^\top\right])$. The following hold:

(a) $\mathbb{E}\left[L_n(f_{\hat{\theta}_n})\right] = \infty$ (infinite risk!);

(b) $\mathbb{E}\left[\overline{L}_n(f_{\hat{\theta}_n}) - L(f^*)\right] \to 0$ as $n \to \infty$ (well-behaved in-sample generalization);

(c) For some event B_n with $\mathbb{P}(B_n) \sim e^{-n}$, $c(\sqrt{t}-2t) \le \mathbb{P}\left(\hat{\lambda}_{\min}(n) \le t\lambda_{\min}|B_n\right) \le c'(\sqrt{t}-2t)$ for some $0 < c < c'$;

(d) $\mathbb{E}\left[\hat{\lambda}^{-1}_{\min}(n)\right] = +\infty$.

To understand what happens in this example, consider the event A_n. On this event, which has a probability proportional to e^{-n}, $\hat{\theta}_{n,1} = (Y_1 + Y_2)/2$ and $\hat{\theta}_{n,2} = \frac{Y_1-Y_2}{X_1-X_2}$, so that $f_{\hat{\theta}_n}(X_i) = Y_i$, $i = 1,2$. Then, the out-of-sample risk can be lower bounded using $\mathbb{E}\left[(f_{\hat{\theta}_n}(X)-Y)^2\right] = \mathbb{E}\left[f_{\hat{\theta}_n}(X)^2\right] + 1 \ge (\mathbb{E}\left[|f_{\hat{\theta}_n}(X)|\,|A_n\right] P(A_n))^2 + 1$. Now, $\mathbb{E}\left[|f_{\hat{\theta}_n}(X) - Y|\,|A_n\right] = 2\mathbb{E}\left[X/|X_1-X_2|\,|A_n\right] = \mathbb{E}\left[1/|X_1-X_2|\,|A_n\right] = +\infty$. A similar calculation shows the rest of the claims.

This example leads to multiple conclusions: *(i)* Ordinary least squares is guaranteed to have finite expected risk if and only $\mathbb{E}\left[\lambda_{\min}(G_n)^{-1}\right] < +\infty$, a condition which is independent to previous conditions such as "good statistical leverage" [Hsu et al., 2012]. *(ii)* The constraint connecting δ and n cannot be removed from Theorem 3.2 without imposing additional conditions. *(iii)* Not all high probability bounds are equal. In particular, the type of in Theorem 3.2 constraining n to be larger than $\log(1/\delta)$ does not guarantee small expected risk. *(iv)* Under the additional condition that the inverse moment of $\lambda_{\min}(G_n)$ is finite, Theorem 3.2 gives rise to an expected risk bound. *(v)* Good in-sample generalization, or in-probability parameter convergence, or that the estimated parameter satisfies the central limit theorem (which all hold in the above example) does not lead to good expected risk for ordinary least-squares; demonstrating a practical example where out-of-sample generalization error is not implied by any of these "classical" results that are extensively studied in statistics (e.g., [Bickel et al., 1998]). *(vi)* Although the "Eigenvalue Chernoff Bound" (Theorem 4.1) of Gittens and Tropp [2011] captures the probability of the smallest positive eigenvalue being significantly underestimated correctly as a function of the sample size, it fails to capture the actual behavior of the left-tail, and this behavior can be significantly different for different distributions. Understanding this phenomenon remains an important problem to study.

Based on this example, we see that another option to get an expected risk bound without clipping the predictions or imposing an additional restriction on the basis functions and the data generating distribution, is to clip the eigenvalues of the data Grammian before inversion at a level of $O(1/n)$ or to add this amount to all the eigenvalues. One way of implementing the increase of eigenvalues is to employ ridge regression by introducing penalty of form $\|\theta\|_2^2$ in the empirical loss minimization criterion. Then, by slightly modifying our derivations and setting $\delta = O(1/n^2)$, an expected risk bound can be derived from Theorem 3.2, e.g., for the squared loss, since then outside of an event with probability $O(1/n^2)$, the risk is controlled by the high probability bound of Theorem 3.2, while on the remaining "bad event", the prediction error will stay bounded by n^2. Although numerical algebra packages implement pseudo-inverses by cutting the minimum eigenvalue, this may be insufficient since they usually cut at the machine precision level, which translates into sample size which may not be available in practice.

4 CONCLUSIONS AND FUTURE WORK

In this paper we set out to investigate the question whether current practice in semiparametric regression of not penalizing the parametric component is a wise choice from the point of view of finite-time performance. We found that for any error probability level, for sample sizes $n = \Omega(\log(1/\delta))$, the risk of such a procedure can indeed be bounded with high probability, proving the first finite-sample generalization bound for partially linear stochastic models. The main difficulty of the proof is to guarantee the parametric component is bounded in the supremum norm. However, we have also found that an additional restriction connecting the data generating distribution and the parametric part is necessary to derive an expected risk bound. This second observation is based on an example where the model is purely parametric. Thus, unless this additional knowledge is available, we think that it is too risky to follow current practice and recommend introducing some form of regularization for the parametric part and/or clipping the predictions when suitable bounds are available on the range of the Bayes predictor. We have also identified that existing bounds in the literature do not capture the behavior of the distribution of the minimum positive eigenvalue of empirical Grammian matrices, which would be critical to understand for improving our understanding of the basic question of how the expected risk of ordinary least-squares behaves.

Acknowledgements

This work was supported by the Alberta Innovates Technology Futures and NSERC. Part of this work was done while Csaba Szepesvári was visiting Microsoft Research. We are indebted to the excellent anonymous reviewer, who pointed out that clipping of predictions is a viable option and whose constructive comments pushed us to dig deeper, which eventually led to a much improved paper. All the remaining mistakes and omissions are ours.

References

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36:489–531, 2008.

J.T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Nederlandica*, 51(3):287–317, 1997.

D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou. Support vector machine soft margin classifiers: Error analysis. *Journal of Machine Learning Research*, 5:1143–1175, December 2004.

F. Cucker and D.-X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

R.M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

R.M. Dudley. *A course on empirical processes*. Ecole d'Eté de Probabilités de St. Flour, 1982, Lecture Notes in Mathematics. Springer, 1984.

R.F. Engle, C.W.J. Granger, J. Rice, and A. Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81:310–320, 1986.

J. Gao. *Nonlinear Time Series: Semiparametric and Nonparametric Methods*, volume 108 of *Monographs on Statistics and Applied Probability*. Taylor & Francis, 2007.

E. Giné and J. Zinn. On the central limit theorem for empirical processes. *Annals of Probability*, 12: 929–989, 1984.

A. Gittens and J.A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. *arXiv preprint arXiv:1104.4513*, 2011.

W. Greblicki and M. Pawlak. *Nonparametric system identification*. Cambridge University Press, 2008.

L. Györfi, M. Kohler, A. Kryżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, 2004.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

J.L. Horowitz. *Semiparametric and nonparametric methods in econometrics*. Springer, 2009.

D. Hsu, S.M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *COLT*, pages 9.1–9.24, 2012.

W. Kienzle and B. Schölkopf. Training support vector machines with multiple equality constraints. In *Proceedings of 16th European Conference on Machine Learning*, pages 182–193, 2005.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

S. Lee and S.J. Wright. Decomposition algorithms for training large-scale semiparametric support vector machines. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD '09, pages 1–14, Berlin, Heidelberg, 2009. Springer-Verlag.

W.S. Lee, P.L. Bartlett, and R.C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44 (5):1974–1980, 1998.

D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3): 271–278, 1995.

A. Rakhlin and K. Sridharan. Stat928: Statistical learning theory and sequential prediction. *Lecture Notes in University of Pennsyvania*, 2014.

P.M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

A.J. Smola, T.T. Frieß, and B. Schölkopf. Semiparametric support vector and linear programming machines, 1998.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer-Verlag New York, 2008.

J.H. Stock. Nonparametric policy analysis. *Journal of the American Statistical Association*, 84(406):567–575, 1989.

A. Tewari and P.L. Bartlett. Learning theory. In R. Chellappa and S. Theodoridis, editors, *Academic Press Library in Signal Processing*, volume 1, chapter 14. Elsevier, 1st edition, 2013. to appear.

S. van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990.

S. van de Geer. *Empirical processes in M-estimation*, volume 45. Cambridge University Press, 2000.

A Appendix

The appendix is devoted to the proof of Theorem 3.1 and Theorem 3.2. Although Theorem 3.2 uses Theorem 3.1, we still present the proof of Theorem 3.2 first, as it is a more standard proof. After this proof, we will present that of Theorem 3.1. However, first we present some results that we will need later multiple times.

Lemma A.1 (Elementary Properties of Subgaussian Random Variables)**.** *Let U be a subgaussian random variable with parameters (β, Γ). Then,*

(i) $\mathbb{E}[|U|] \le \frac{1}{\beta}\sqrt{\Gamma - 1}$*;*

(ii) for any constant $c \ge 0$, $U + c$ is subgaussian.

Proof. (i) follows from

$$\Gamma \ge \mathbb{E}\left[\exp\left(|\beta U|^2\right)\right] \ge \exp\left(\mathbb{E}\left[|\beta U|^2\right]\right) \ge \exp\left(\mathbb{E}\left[|\beta U|\right]^2\right) \ge 1 + \mathbb{E}\left[|\beta U|\right]^2 .$$

(ii) follows from

$$\mathbb{E}\left[\exp\left(\tfrac{1}{2}\beta^2(U+c)^2\right)\right] \le \mathbb{E}\left[\exp\left(\beta^2 U^2 + \beta^2 c^2\right)\right] \le e^{\beta^2 c^2}\mathbb{E}\left[\exp\left(|\beta U|^2\right)\right] \le e^{\beta^2 c^2}\Gamma .$$

□

We will also need the following result:

Lemma A.2. *Let $U > 0$, $\mathcal{C} = \mathcal{H} + \mathcal{G}(U)$. Then, a.s.*

$$\int_0^1 \sqrt{H(u, \mathcal{C}, \|\cdot\|_n)}\, du \le 2C_H + 2C_G(U) ,$$

where $C_G(U) = \rho^{1/2}\int_0^1 \log^{1/2}\left(\frac{4U+u}{u}\right) du \, (= O(\sqrt{\rho(\log(U))_+})$.

Proof of Lemma A.2. Since $\mathcal{C} = \mathcal{H} + \mathcal{G}(U)$, a standard argument shows that

$$H(u;\sigma) \le H(u/2; \mathcal{H}; \|\cdot\|_n) + H(u/2, \mathcal{G}(U), \|\cdot\|_n) . \tag{6}$$

Now, note that $\|\cdot\|_n \le \|\cdot\|_{\infty,n}$. Thus,

$$\begin{aligned}\int_0^1 H^{1/2}(u/2, \mathcal{H}, \|\cdot\|_n)\, du &= 2\int_0^{1/2} H^{1/2}(u, \mathcal{H}, \|\cdot\|_n)\, du \le 2\int_0^1 H^{1/2}(u, \mathcal{H}, \|\cdot\|_n)\, du \\ &\le 2\int_0^1 H^{1/2}(u, \mathcal{H}, \|\cdot\|_{\infty,n})\, du \le 2C_H,\end{aligned}$$

where the last inequality is by Assumption 3.3. Moreover, since $\|g\|_n \le \|g\|_\infty$, $\mathcal{G}(U)$ is a subset of the ball $B_{\mathcal{G},\|\cdot\|_n}(0, U)$. Thus,

$$\begin{aligned}\int_0^1 H^{1/2}(u/2, \mathcal{G}(U), \|\cdot\|_n)\, du &\le 2\int_0^1 H^{1/2}(u, \mathcal{G}(U), \|\cdot\|_n)\, du \le 2\int_0^1 H^{1/2}(u, B_{\mathcal{G},\|\cdot\|_n}(0,U), \|\cdot\|_n)\, du \\ &\le 2\rho^{1/2}\int_0^1 \log^{1/2}\left(\frac{4U+u}{u}\right) du = 2C_G(U) ,\end{aligned}$$

where the second inequality is by Corollary 2.6 of [van de Geer, 2000], which states that $H(\varepsilon, B_{\mathcal{G},\|\cdot\|_n}(0,\sigma)) \le \rho\log(\frac{4\sigma+\varepsilon}{\varepsilon})$. Using (6) and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ which holds for $a, b \ge 0$, we conclude that

$$\int_0^1 \sqrt{H(u;\sigma)}\, du \le 2C_H + 2C_G(U),$$

finishing the proof of the claim. □

B The Proof of Theorem 3.2

In this section we prove Theorem 3.2 assuming that Theorem 3.1 holds.

Let U be as in Theorem 3.1 and let E denote the event when

$$\sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty \le U\,.$$

For any $z \ge 0$,

$$\begin{aligned}\mathbb{P}\left(L(f_n) - L(f^*) > z\right) &= \mathbb{P}\left(L(f_n) - L(f^*) > z, E^c\right) + \mathbb{P}\left(L(f_n) - L(f^*) > z, E\right) \\ &\le \mathbb{P}\left(E^c\right) + \mathbb{P}\left(L(f_n) - L(f^*) > z, E\right).\end{aligned} \tag{7}$$

Thus, to study the tail probabilities of $L(f_n) - L(f^*)$, it suffices to study $L(f_n) - L(f^*)$ *on the event* E.

Define $\mathcal{G}(U) = \{g \in \mathcal{G} : \|g\|_\infty \le U\}$ and $\mathcal{C} = \mathcal{H} + \mathcal{G}(U)$. On E, we claim that $f_n \in \mathcal{C}$. We have $f_n = h_n + g_n$ and since $h_n \in \mathcal{H}$ by definition, it remains to show that $g_n \in \mathcal{G}(U)$. By appropriately selecting $g_{h,n}$, we can arrange for $g_n = g_{h_n,n}$. Hence, $\|g_n\|_\infty \le \sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty \le U$, showing that $f_n \in \mathcal{C}$ indeed holds.

Now, by increasing U if necessary, we can always arrange for $f^* = h^* + g^* \in \mathcal{C}$ (for this we may need to increase U so that $\|g^*\|_\infty \le U$). Hence, in what follows, we will assume this.[4] By (3), on E it holds almost surely that

$$\begin{aligned}L(f_n) - L(f^*) &\le L_n(f^*) - L(f^*) - (L_n(f_n) - L(f_n)) \\ &= (\widetilde{\Delta}_n(f_n) - \widetilde{\Delta}_n(f^*)) + (\overline{\Delta}_n(f_n) - \overline{\Delta}_n(f^*)) \\ &\le \underbrace{\sup_{f\in\mathcal{C}} \widetilde{\Delta}_n(f) - \widetilde{\Delta}_n(f^*)}_{\widetilde{\Delta}^*_n(\mathcal{C})} + \underbrace{\sup_{f\in\mathcal{C}} |\overline{\Delta}_n(f) - \overline{\Delta}_n(f^*)|}_{\overline{\Delta}^*_n(\mathcal{C})},\end{aligned} \tag{8}$$

where we introduced $\overline{\Delta}_n(f) = L_n(f) - \overline{L}_n(f)$ and $\widetilde{\Delta}_n(f) = L(f) - \overline{L}_n(f)$ with $\overline{L}_n(f) = \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\ell(Y_k, f(X_k))|X_k\right]$. Note that the first term does not depend on the (unbounded) responses $Y_1, \ldots, Y_n$. Furthermore, by our assumptions, $\widetilde{\Delta}_n(f)$ is bounded for f bounded. Hence, we can analyze these terms using tools developed for bounded random variables and empirical processes. Now, while the last term involves $Y_1, \ldots, Y_n$, $\overline{\Delta}_n$ compares average losses *over the sample* $X_1, \ldots, X_n$, this last term concerns *in-sample* generalization. Hence, as we will show it below, it can be analyzed using tools developed for the so-called "fixed design" setting. In fact, the following result gives tail bounds for this part:

Lemma B.1. *Let Assumptions 3.1 to 3.4 hold and WLOG assume that* $U \ge \max(1, \|g^*\|_\infty)$. *Then, there exist constants* $c, \alpha > 0$ *such that for any* $0 < \delta < 1$ *satisfying* $\log\frac{1}{\delta} \ge c$ *with probability at least* $1 - \delta$,

$$\overline{\Delta}^*_n(\mathcal{C}) \le 2(r + U)\sqrt{\frac{\log\frac{2}{\delta}}{\alpha n}}\,. \tag{9}$$

The proof is based on Theorem 3.3 of van de Geer [1990], which we quote below for completeness. Let (Λ, d) be a pseudo-metric space and for $u > 0$ let $B_{\Lambda,d}(\lambda, u)$ be the d-ball in Λ that has radius u and is centered at λ. We will allow d to be replaced with a pseudo-norm meaning the ball where the pseudo-metric is defined by the chosen pseudo-norm. The theorem of van de Geer bounds the tails of the suprema of centered, Lipschitz empirical processes of Λ over balls of Λ:

Theorem B.2 (Theorem 3.3 of van de Geer [1990])**.** *Let* (Λ, d) *be a pseudo-metric space with* $d^2 = (1/n)\sum_{k=1}^n d_k^2$ *where* $d_1, \ldots, d_n$ *pseudo-metrics on* Λ. *Let* $U_1, \ldots, U_n$ *be real-valued, independent, centered process on* Λ *such that for* $Z_n = \frac{1}{\sqrt{n}}\sum U_k$, $Z_n(\lambda_0) = 0$ *for some* $\lambda_0 \in \Lambda$. *For* $u > 0$, *denote by* $H(u;\sigma) = H(u, B_{\Lambda,d}(\lambda_0,\sigma), d)$, *the* u*-entropy of the ball* $B_{\Lambda,d}(\lambda_0,\sigma)$. *Assume further that* $|U_k(\lambda) - U_k(\lambda')| \le M_k d_k(\lambda, \lambda')$ *with* $M_k \ge 0$ *random such that* $\mathbb{E}[\exp(|\beta M_k|^2)] \le \Gamma < \infty$ *for some positive constants* β *and* Γ. *Then, there exist* $\alpha, \eta, C_1, C_2 > 0$ *depending only on* β *and* Γ *such that*

$$\mathbb{P}\left(\sup_{\lambda\in B_{\Lambda,d}(\lambda_0,\sigma)} |Z_n(\lambda)| \ge t\right) \le 2\exp\left(-\frac{\alpha t^2}{\sigma^2}\right)$$

[4]Note that this is assumed only to simplify the presentation.

holds for any $t>0$ and $\sigma>0$ that satisfies $t/\sigma > C_1$ and $t > C_2 \int_0^{t_0} \sqrt{H(u;\sigma)}\,du$ where $t_0 \ge \inf\{u : H(u;\sigma) \le \eta t^2/\sigma^2\}$.

Let us now turn to the proof of Lemma B.1.

Proof of Lemma B.1. Let $(W, \mathcal{W}, \mathbb{P})$ be the probability space that holds $(X_1, Y_1), \ldots, (X_n, Y_n)$. Note that with no loss of generality, we can assume that $(W, \mathcal{W})$ is a Borel-space (this is because all our random variables leave in complete, separable metric spaces). For $x_1, \ldots, x_n \in \mathcal{X}$, let $x_{1:n} = (x_1, \ldots, x_n)$. Similarly, let $X_{1:n} = (X_1, \ldots, X_n)$. Define $(\mathbb{P}_{x_{1:n}})_{x_{1:n}\in\mathcal{X}^n}$ to be the disintegration of the probability measure $\mathbb{P}$ with respect to $X_{1:n}$, also known as the regular conditional probability measure obtained from $\mathbb{P}$ by conditioning on $X_{1:n}$.[5] The expectation operator corresponding to $\mathbb{P}_{x_{1:n}}$ will be denoted by $\mathbb{E}_{x_{1:n}}$. Note that, by the definition of $\mathbb{P}_{x_{1:n}}$, for any random variable Z on $(W, \mathcal{W}, \mathbb{P})$, $\mathbb{E}_{X_{1:n}}[Z] = \mathbb{E}[Z|X_{1:n}]$ holds everywhere and in particular, for a measurable function $s : \mathcal{Y} \to \mathbb{R}$, $\mathbb{E}_{X_{1:n}}[s(Y_k)] = \mathbb{E}[s(Y_k)|X_{1:n}] = \mathbb{E}[s(Y_k)|X_k]$, where the last equality holds since by assumption (X_t, Y_t) is i.i.d.

Let

$$U_{k,x_{1:n}}(f) = \Delta_{k,x_{1:n}}(f) - \Delta_{k,x_{1:n}}(f^*)$$

where

$$\Delta_{k,x_{1:n}}(f) = \ell(Y_k, f(x_k)) - \mathbb{E}_{x_{1:n}}[\ell(Y_k, f(x_k))]\,, \qquad f \in \Lambda\,.$$

Note that $U_{k,x_{1:n}}$'s are independent, centered processes over $W_{x_{1:n}}$. Let $Z_{n,x_{1:n}} = \frac{1}{\sqrt{n}}\sum_{k=1}^n U_{k,x_{1:n}}$. By construction, $Z_{n,x_{1:n}}(f^*) = 0$. We now show that it is enough to study the deviations of the suprema of $Z_{n,x_{1:n}}(f)$ over the probability spaces $W_{x_{1:n}}$.

We have

$$\overline{\Delta}_n(f) = L_n(f) - \overline{L}_n(f) = \frac{1}{n}\sum_{k=1}^n \Delta_{k,X_{1:n}}(f)$$

and so

$$\sqrt{n}(\overline{\Delta}_n(f) - \overline{\Delta}_n(f^*)) = Z_{n,X_{1:n}}(f)\,.$$

By the construction of $\mathbb{P}_{x_{1:n}}$, for $z \ge 0$,

$$\mathbb{P}(\overline{\Delta}_n^*(\mathcal{C}) \ge z) = \int \mathbb{P}_{x_{1:n}}(\overline{\Delta}_n^*(\mathcal{C}) \ge z) P_{X_{1:n}}(dx_{1:n})\,, \tag{10}$$

and

$$\mathbb{P}_{x_{1:n}}(\overline{\Delta}_n^*(\mathcal{C}) \ge z/\sqrt{n}) = \mathbb{P}_{x_{1:n}}(\sqrt{n}\sup_{f\in\mathcal{C}}|\overline{\Delta}_n(f) - \overline{\Delta}_n(f^*)| \ge z) = \mathbb{P}_{x_{1:n}}(\sup_{f\in\mathcal{C}} Z_{n,x_{1:n}}(f) \ge z)\,. \tag{11}$$

Let $\Lambda = \mathcal{C}$, $d_{k,x_{1:n}}(f, f') = |f(x_k) - f'(x_k)|$ and $d^2_{x_{1:n}}(f, f') = \frac{1}{n}\sum_{k=1}^n d^2_{k,x_{1:n}}(f, f')$. By construction, $d_{x_{1:n}}(f, f') = \|f - f'\|_n$. Since for any $f = h + g \in \mathcal{C}$,

$$\|f - f^*\|_n = \|h - h^*\|_n + \|g - g^*\|_n \le \|h - h^*\|_\infty + \|g - g^*\|_\infty \le 2(r + U) =: \sigma\,,$$

thus, $\mathcal{C} \subset B_{\Lambda, d_{x_{1:n}}}(f^*, \sigma) \subset \Lambda = \mathcal{C}$ and

$$\mathbb{P}_{x_{1:n}}(\overline{\Delta}_n^*(\mathcal{C}) > z/\sqrt{n}) = \mathbb{P}_{x_{1:n}}\left(\sup_{f\in B_{\Lambda,d_{x_{1:n}}}(f^*,\sigma)} Z_{n,x_{1:n}}(f) > z\right)\,. \tag{12}$$

Thus, it remains to bound this latter probability. Fix $x_{1:n} \in \mathcal{X}^n$ such that

$$\mathbb{E}_{x_{1:n}}[\exp((\beta K_\ell(Y, c)^2)] \le \Gamma_c, \text{ for all } c > 0\,. \tag{13}$$

[5] The defining properties of $(\mathbb{P}_{x_{1:n}})$ are that for each $x_{1:n} \in \mathcal{X}^n$, $\mathbb{P}_{x_{1:n}}$ is a probability measure on $(W, \mathcal{W})$ concentrated on $\{X_{1:n} = x_{1:n}\}$, $x_{1:n} \mapsto \mathbb{P}_{x_{1:n}}$ is measurable and for any $f : (W, \mathcal{W}) \to [0, \infty)$ measurable function $\int f(w)\mathbb{P}(dw) = \int(\int f(w)\mathbb{P}_{x_{1:n}}(dw))P_{X_{1:n}}(dx_{1:n})$. The existence of $(\mathbb{P}_{x_{1:n}})$, which is also called a regular conditional probability distribution is ensured thanks to the assumption that $(W, \mathcal{W})$ is Borel. Moreover, $(\mathbb{P}_{x_{1:n}})$ is unique up to an almost sure equivalence in the sense that if $(\hat{\mathbb{P}}_{x_{1:n}})$ is another disintegration of $\mathbb{P}$ w.r.t. $X_{1:n}$ then $P_X(\{x_{1:n} : \mathbb{P}_{ux} \ne \hat{\mathbb{P}}_{x_{1:n}}\}) = 0$. For background on disintegration and conditioning, the reader is referred to Chang and Pollard [1997].

Let us now apply Theorem B.2 to $W_{x_{1:n}} = (W, \mathcal{W}, \mathbb{P}_{x_{1:n}})$ with Λ, $(d_{k,x_{1:n}})$ and $(U_{k,x_{1:n}})$ $(k = 1, \ldots, n)$, as defined above. To verify the uniform subgaussian property of the Lipschitz coefficient of $U_{k,x_{1:n}}$, note that for $f, f' \in \mathcal{C}$, by Assumption 3.1(iii),

$$\begin{aligned}|U_{k,x_{1:n}}(f) - U_{k,x_{1:n}}(f')| &= |\Delta_{k,x_{1:n}}(f) - \Delta_{k,x_{1:n}}(f')| \\ &\le |\ell(Y_k, f(x_k)) - \ell(Y_k, f'(x_k))| + |\mathbb{E}_{x_{1:n}}[\ell(Y_k, f(x_k)) - \ell(Y_k, f'(x_k))]| \\ &\le K_l(Y_k, r+U)|f(x_k) - f'(x_k)| + \mathbb{E}_{x_{1:n}}[K_l(Y_k, r+U)]|f(x_k) - f'(x_k)|\,.\end{aligned}$$

By Lemma A.1(i), $\mathbb{E}_{x_{1:n}}[K_l(Y_k, r+U)] \le \frac{1}{\beta}\sqrt{\Gamma_{r+U} - 1}$ and so by part (ii) of the same lemma, $K_l(Y_k, r+U) + \mathbb{E}_{x_{1:n}}[K_l(Y_k, r+U)]$ is subgaussian, with parameters β' and Γ' only depending on $r+U$.

Therefore, from Theorem B.2 we conclude that there exists $C_1, C_2, \eta > 0$ such that for any $t > 0$ satisfying $\eta t^2/\sigma^2 \ge H(1;\sigma)$, $t > C_1\sigma$ and $t > C_2 \int_0^1 \sqrt{H(u;\sigma)}\,du$, it holds that

$$\mathbb{P}_{x_{1:n}}\left(\sup_{f\in\mathcal{C}} |Z_{n,x_{1:n}}(f)| \ge t\right) = \mathbb{P}_{x_{1:n}}\left(\sup_{f\in B_{\Lambda,d_{x_{1:n}}}(f^*,\sigma)} |Z_{n,x_{1:n}}(f)| \ge t\right) \le 2\exp\left(-\frac{\alpha t^2}{\sigma^2}\right)\,. \tag{14}$$

It still remains to check that $H(1,\sigma)$ and $\int_0^1 \sqrt{H(u;\sigma)}\,du$ are finite (otherwise the result is vacuous). By definition, $H(u;\sigma) = H(u, B_{\Lambda,d_{x_{1:n}}}(f^*,\sigma), d_{x_{1:n}}) = H(u, \mathcal{C}, d_{x_{1:n}}) = H(u, \mathcal{C}, \|\cdot\|_n)$. Hence, by Lemma A.2, $\int_0^1 H^{1/2}(u;\sigma)\,du \le 2C_H + 2C_G(U)$. Noting that $H(u;\sigma)$ is monotonically decreasing in u, we calculate $H^{1/2}(1;\sigma) \le \int_0^1 H^{1/2}(u;\sigma)\,du \le 2C_H + 2C_G(U)$ and so $H(1;\sigma) \le (2C_H + 2C_G(U))^2 < \infty$. We conclude that (14) holds for any $t \ge t_{\min} := \max\{C_1\sigma, C_2(2C_H + 2C_G(U)), (2C_H + 2C_G(U))\sigma\eta^{-1/2}\}$.

Since by Assumption 3.1(iii), (13) holds $[P_X]$-almost surely, combining (10), (12) and (14), we get

$$\mathbb{P}\left(\overline{\Delta}_n^*(\mathcal{C}) \ge t/\sqrt{n}\right) \le 2\exp\left(-\frac{\alpha t^2}{\sigma^2}\right)\,. \tag{15}$$

Inverting this inequality, we see that for any $0 < \delta < 1$ such that $\log(2/\delta) \ge t_{\min}^2\alpha/\sigma^2$, with probability at least $1-\delta$,

$$\overline{\Delta}_n^*(\mathcal{C}) \le \sigma\sqrt{\frac{\log\frac{2}{\delta}}{\alpha n}}\,,$$

finishing the proof. ☐

It remains to bound $\widetilde{\Delta}_n^*(\mathcal{C}) = \sup_{f\in\mathcal{C}} \widetilde{\Delta}_n(f) - \widetilde{\Delta}_n(f^*)$. For this, define

$$\overline{\ell}(x,p) = \mathbb{E}\left[\ell(Y,p)\,|\,X = x\right]\,.$$

With a slight abuse of notation, we also introduce $\overline{\ell}(x,f) = \overline{\ell}(x, f(x))$. Let

$$B(\overline{\ell}, U) = \left\|\sup\nolimits_{p\in[-r-U, r+U]} \overline{\ell}(X,p)\right\|_{L^\infty}\,,$$

where $\|\cdot\|_{L^\infty}$ denotes the essential supremum of its argument. We also let $\overline{L}$ be the Lipschitz constant of $\overline{\ell}$ when $p \in [-r-U, r+U]$:

$$\mathrm{Lip}(\overline{\ell}, U) = \left\|\sup\nolimits_{p,p'\in[-r-U,r+U],p\neq p'} \tfrac{\overline{\ell}(X,p) - \overline{\ell}(X,p')}{|p-p'|}\right\|_{L^\infty}\,.$$

The next lemma shows that both quantities are finite:

Lemma B.3. *Let $r' = \max(r+U, \|\hat{h}\|_\infty)$. Then, $B(\overline{\ell}, U) \le Q + \frac{2r'}{\beta}\sqrt{\Gamma_{r'} - 1} < +\infty$ and $\mathrm{Lip}(\overline{\ell}, U) < \frac{\sqrt{\Gamma_{r+U} - 1}}{\beta} < +\infty$.*

Proof. For the second statement, for any $t, s \in [-b, b]$ we have

$$\overline{\ell}(X,t) - \overline{\ell}(X,s) \le \mathbb{E}\left[|\ell(Y,t) - \ell(Y,s)|\,\middle|\,X\right] \le \mathbb{E}\left[K_\ell(Y,b)|t-s|\,\middle|\,X\right] \le \tfrac{\sqrt{\Gamma_b - 1}}{\beta}|t-s|\,,$$

where we used Assumption 3.1(iii) and Lemma A.1(i). Thus, $\mathrm{Lip}(\overline{\ell}, U) \le \frac{\sqrt{\Gamma_{r+U-1}}}{\beta} < +\infty$.

For the first statement take some $|p| \le r + U$ and write

$$\begin{aligned}\overline{\ell}(X,p) &\le \overline{\ell}(X,\hat{h}(X)) + |\overline{\ell}(X,p) - \overline{\ell}(X,\hat{h}(X))| \le Q + \mathrm{Lip}(\overline{\ell}, r')|p - \hat{h}(X)| \le Q + \mathrm{Lip}(\overline{\ell}, r')(|r+U| + \|\hat{h}\|_\infty|) \\ &\le Q + \tfrac{\Gamma_{r'-1}}{\beta}(2r'),\end{aligned}$$

where in the second inequality we used Assumption 3.1(ii), while in the last one we used the bound on the Lipschitz coefficient. ∎

As it is well known, the Rademacher complexity of $\mathcal{C}$, defined next, captures *exactly* the behavior of $\mathbb{E}\left[\widetilde{\Delta}_n^*(\mathcal{C})\right]$ (e.g., Tewari and Bartlett [2013]).

Definition 3 (Rademacher Complexity of Subsets of $\mathbb{R}^n$)**.** *Let $A \subset \mathbb{R}^n$, $(\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n$ be independent Rademacher random variables (i.e., $\mathbb{P}(\sigma_k = 1) = 1/2$). The Rademacher complexity of A, $\mathfrak{R}(A)$ is*

$$\mathfrak{R}(A) = \frac{1}{n}\,\mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^n \sigma_i a_i\right].$$

Definition 4 (Rademacher Complexity of Function Sets)**.** *Let $\mathcal{F} \subset \{f : f : \mathcal{X} \to \mathbb{R}\}$ and P be a measure on $\mathcal{X}$. Then, the nth Rademacher number of $\mathcal{F}$ induced by P is*

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\left[\mathfrak{R}(\mathcal{F}(X_{1:n}))\right],$$

where $\mathcal{F}(X_{1:n}) = \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$ is the projection of $\mathcal{F}$ to an i.i.d. sample $X_{1:n} = (X_1, \ldots, X_n)$ from P. When n and P are uniquely identified from the context, we also call $\mathfrak{R}_n(\mathcal{F})$ the Rademacher-complexity of $\mathcal{F}$.

The Rademacher complexity enjoys a number of useful properties, amongst which we need the following contraction property:

Theorem B.4. *For $A \subset \mathbb{R}^n$ and $\phi = (\phi_1, \ldots, \phi_n) : \mathbb{R}^n \to \mathbb{R}^n$, define $\phi \circ A = \{(\phi_1(a_1), \ldots, \phi_n(a_n)) : a \in A\}$. Assume that all the component functions ϕ_i are L-Lipschitz over A. Then, $\mathfrak{R}(\phi \circ A) \le L\mathfrak{R}(A)$.*

Note that this theorem is usually stated for the case when $\phi_1 = \ldots = \phi_n$. The simpler form is sufficient for "margin based losses" (used in classification) that have the form $\ell(y,p) = g(yp)$ with some g. As we will see, here we need this more general form as our losses are less constrained. However, the proof of this more general result still follows the standard reasoning.

Proof. We follow the proof of Theorem 11.9 in [Rakhlin and Sridharan, 2014] and write

$$\begin{aligned}n\mathfrak{R}(\phi \circ A) &= \mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^n \sigma_i \phi_i(a_i)\right] \\ &= \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i) + \phi_n(a_n) \,|\, \sigma_n = 1\right] + \mathbb{E}\left[\sup_{b\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(b_i) - \phi_n(b_n) \,|\, \sigma_n = -1\right]\right\} \\ &= \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a,b\in A}\sum_{i=1}^{n-1}\sigma_i(\phi_i(a_i) + \phi_i(b_i)) + (\phi_n(a_n) - \phi_n(b_n))\right]\right\} \\ &\le \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a,b\in A}\sum_{i=1}^{n-1}\sigma_i(\phi_i(a_i) + \phi_i(b_i)) + L|a_n - b_n|\right]\right\}.\end{aligned}$$

Now assume that some (a^*, b^*) achieves the supremum (the proof when the supremum is not achieved is easy once we know how to prove the statement for the case when the supremums involved are all achieved). If $a_n^* \ge b_n^*$, the absolute value can be removed. Otherwise, (b^*, a^*) will achieve the same supremum, and again the absolute value can be removed. Thus, the last expression is bounded by

$$
\begin{aligned}
&\frac{1}{2}\left\{\mathbb{E}\left[\sup_{a,b\in A}\sum_{i=1}^{n-1}\sigma_i(\phi_i(a_i)+\phi_i(b_i))+L(a_n-b_n)\right]\right\} \\
&\qquad = \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i)+La_n \middle|\, \sigma_n=1\right]+\mathbb{E}\left[\sup_{b\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(b_i)-Lb_n \middle|\, \sigma_n=-1\right]\right\} \\
&\qquad = \mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i)+L\sigma_n a_n\right].
\end{aligned}
$$

Continuing this way,

$$
\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i)+L\sigma_n a_n\right] \le \mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-2}\sigma_i\phi_i(a_i)+L(\sigma_{n-1}a_{n-1}+\sigma_n a_n)\right] \le L\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right],
$$

thus finishing the proof. □

Let $\mathcal{L} = \left\{s_f : \mathcal{X}\to\mathbb{R} \,:\, s_f(x) = \bar{\ell}(x,f)-\bar{\ell}(x,f^*), f\in\mathcal{C}, x\in\mathcal{X}\right\}$. Note that $\widetilde{\Delta}_n(f)-\widetilde{\Delta}_n(f^*) = (L(f)-\overline{L}_n(f)) - (L(f^*)-\overline{L}_n(f^*)) = \mathbb{E}\left[\bar{\ell}(X,f)-\bar{\ell}(X,f^*)\right] - \frac{1}{n}\sum_{k=1}^n(\bar{\ell}(X_k,f)-\bar{\ell}(X_k,f^*)) = \mathbb{E}\left[s_f(X)\right]-\frac{1}{n}\sum_{k=1}^n s_f(X_k)$. Following the standard argument, since the range of functions in $\mathcal{L}$ is bounded by $B(\bar{\ell},U)$, by McDiamid's inequality, for any $0<\delta<1$, with probability at least $1-\delta$,

$$
\widetilde{\Delta}_n^*(\mathcal{C}) = \sup_{s\in\mathcal{L}}\mathbb{E}\left[s(X)\right]-\tfrac{1}{n}\textstyle\sum_{k=1}^n s(X_k) \le \mathbb{E}\left[\sup_{s\in\mathcal{L}}\mathbb{E}\left[s(X)\right]-\tfrac{1}{n}\sum_{k=1}^n s(X_k)\right]+B(\bar{\ell},U)\sqrt{\frac{2\log\frac{1}{\delta}}{n}}.
$$

Following the calculation before Theorem 7 in Section 3.2 of Tewari and Bartlett [2013],

$$
\mathbb{E}\left[\sup_{s\in\mathcal{L}}\mathbb{E}\left[s(X)\right]-\tfrac{1}{n}\textstyle\sum_{k=1}^n s(X_k)\right] \le 2\mathfrak{R}_n(\mathcal{L}).
$$

Let us now bound $\mathfrak{R}_n(\mathcal{L}) = \mathbb{E}\left[\mathfrak{R}(\mathcal{L}(X_{1:n}))\right]$. We can write

$$
\mathcal{L}(X_{1:n}) = \{s_f(X_{1:n}) \,:\, f\in\mathcal{C}\} = \phi\circ\mathcal{C}(X_{1:n}),
$$

where $\phi = (\phi_1,\dots,\phi_n) : \mathbb{R}^n\to\mathbb{R}^n$ is defined by $\phi_k(t) = \bar{\ell}(X_k,t)-\bar{\ell}(X_k,f^*(X_k))$ (note that ϕ is random). By definition, each component of ϕ is almost surely Lipschitz over any bounded interval $[-b,b]$ with the same Lipschitz constant (depending on b). Indeed, for any $t,s\in[-b,b]$,

$$
\begin{aligned}
\||\phi_k(t)-\phi_k(s)|\|_{L^\infty} &= \||\bar{\ell}(X_k,t)-\bar{\ell}(X_k,s)|\|_{L^\infty} \\
&= \inf\left\{a\in\mathbb{R} \,:\, \mathbb{P}\left(|\bar{\ell}(X_k,t)-\bar{\ell}(X_k,s)|>a\right)\right\} \\
&= \inf\left\{a\in\mathbb{R} \,:\, \mathbb{P}\left(|\bar{\ell}(X,t)-\bar{\ell}(X,s)|>a\right)\right\} \\
&= \||\bar{\ell}(X,t)-\bar{\ell}(X,s)|\|_{L^\infty} \\
&\le \mathrm{Lip}(\bar{\ell},b)|t-s|,
\end{aligned}
$$

where the second and fourth equalities used the definition of $\|\cdot\|_{L^\infty}$ and the third used that X_k and X are identically distributed. Now, since $\mathcal{C}$ contains functions bounded by $r+U$, by Theorem B.4,

$$
\mathfrak{R}(\phi\circ\mathcal{C}(X_{1:n})) \le \mathrm{Lip}(\bar{\ell},U)\,\mathfrak{R}(\mathcal{C}(X_{1:n})) \quad \text{a.s.}
$$

and hence

$$
\mathfrak{R}_n(\mathcal{L}) = \mathbb{E}\,\mathfrak{R}(\mathcal{L}(X_{1:n})) = \mathbb{E}\,\mathfrak{R}(\phi\circ\mathcal{C}(X_{1:n})) \le \mathrm{Lip}(\bar{\ell},U)\,\mathbb{E}\,\mathfrak{R}(\mathcal{C}(X_{1:n})) = \mathrm{Lip}(\bar{\ell},U)\,\mathfrak{R}_n(\mathcal{C}).
$$

Our next goal is to bound $\mathfrak{R}_n(\mathcal{C})$. By Dudley's entropy integral bound [Dudley, 1967] (e.g., Theorem 10 of Tewari and Bartlett [2013], for a statement with a proof see Theorem 11.4 of Rakhlin and Sridharan [2014]),

$$\mathfrak{R}_n(\mathcal{C}) \le \frac{12}{\sqrt{n}} \mathbb{E} \int_0^1 H^{1/2}(u, \mathcal{C}, \|\cdot\|_n)\, du \le \frac{12}{\sqrt{n}}(2C_H + 2C_G(U)),$$

where the second inequality holds thanks to Lemma A.2 and we also used that Dudley's bound holds regardless the scale of the range of functions in $\mathcal{C}$, which is not hard to check by inspecting the proof of the bound. Combining all the inequalities we get that with probability at least $1-\delta$,

$$\widetilde{\Delta}_n^*(\mathcal{C}) \le \frac{48(C_H + C_G(U))\ \mathrm{Lip}(\bar{\ell}, U)}{\sqrt{n}} + B(\bar{\ell}, U)\sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \tag{16}$$

Combining Equations (7) and (8), we have for any $z \ge 0$,

$$\mathbb{P}\left(L(f_n) - L(f^*) > z\right) \le \mathbb{P}\left(E^c\right) + \mathbb{P}\left(\widetilde{\Delta}_n^*(\mathcal{C}) + \overline{\Delta}_n^*(\mathcal{C}) > z\right). \tag{17}$$

Now, by Lemma B.1 and (16), for any $0 < \delta < 1$ such that $\log(1/\delta) \ge c$, with probability at least $1 - 2\delta$,

$$\widetilde{\Delta}_n^*(\mathcal{C}) + \overline{\Delta}_n^*(\mathcal{C}) \le \frac{48(C_H + C_G(U))\ \mathrm{Lip}(\bar{\ell}, U)}{\sqrt{n}} + 2(r+U)\sqrt{\frac{\log \frac{2}{\delta}}{\alpha n}} + B(\bar{\ell}, U)\sqrt{\frac{\log \frac{1}{\delta}}{2n}} =: \pi(\delta)\,.$$

Together with (17) and Theorem 3.1, we thus get that with probability $1 - 3\delta$, provided that $\log(1/\delta) \ge c$ and $n \ge c_1 + c_2 \frac{\log\left(\frac{2\rho}{\delta}\right)}{\lambda_{\min}}$,

$$L(f_n) - L(f^*) \le \pi(\delta)\,,$$

thus finishing the proof.

C The Proof of Theorem 3.1

In this section we present the proof of Theorem 3.1, which calls for a bound of

$$\sup_{h \in \mathcal{H}} \|g_{h,n}\|_\infty$$

that holds with high probability. Fix $h \in \mathcal{H}$. Then, $g_{h,n}(x) = \langle \theta, \phi(x) \rangle \le \|\theta_{h,n}\|_2 \|\phi(x)\|_2$, where $\theta_{h,n}$ is the parameter vector of $g_{h,n}$. Since $\|\phi(x)\|_2 \le 1$, it suffices to bound $\|\theta_{h,n}\|_2$. On $G_{\lambda_{\min}}$, which is defined as the event $\{\hat{\lambda}_{\min} \ge \lambda_{\min}/2\}$, we have

$$g_{h,n}^2(x) \le \|\theta_{h,n}\|_2^2 \le \frac{\theta_{h,n}^\top \hat{G}\, \theta_{h,n}}{\hat{\lambda}_{\min}} = \frac{2\,\|g_{h,n}\|_n}{\lambda_{\min}}\,. \tag{18}$$

Hence, the problem is reduced to proving a uniform (h-independent) upper bound on the empirical norm of $g_{h,n}$ and showing that $G_{\lambda_{\min}}$ happens with "large probability".

For the latter, we use a result of Gittens and Tropp [2011]. This is summarized in the lemma which also includes some observations that will prove to be useful later:

Lemma C.1. *The following hold:*

(i) With probability one, for any $\theta \in \mathbb{R}^d$, $\theta^\top \hat{G}\theta \le \frac{\theta^\top G\theta}{\lambda_{\min}}$.

(ii) Assuming that $n \in \mathbb{N}$ and $\delta \in (0,1)$ are such that

$$n \ge \frac{2}{\lambda_{\min} \log\left(\frac{e}{2}\right)} \log\left(\frac{\rho}{\delta}\right), \tag{19}$$

where ρ and $\lambda_{\min}$ are respectively the rank and the smallest positive eigenvalue of G, with probability at least $1-\delta$, it holds that $\hat{\lambda}_{\min} \ge \frac{\lambda_{\min}}{2} > 0$.

(iii) For any n, δ satisfying (19), with probability $1-\delta$ it holds that for any $\theta \in \mathbb{R}^d$ and $[P_X]$ almost every $x \in \mathcal{X}$, $|\langle \theta, \phi(x) \rangle| \le \sqrt{\frac{2\theta^\top \hat{G}\theta}{\lambda_{\min}}}$.

The (easy) proof of the lemma is deferred to Appendix C.2.

To get an upper bound on the empirical norm of $g_{h,n}$, we will use

$$\|g_{h,n}\|_n \le \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n + \left\|\overline{g}_{h,n}\right\|_n \tag{20}$$

and develop uniform bound on the two terms on the r.h.s..

Lemma C.2. *It holds almost surely that*

$$\sup_{h\in\mathcal{H}} \left\|\overline{g}_{h,n}\right\|_n \le \bar{R},$$

where $\bar{R} = R_{C_0} + r$, $C_0 = \frac{2\hat{r}}{\beta}\sqrt{\Gamma_{\hat{r}} - 1} + Q$, $\hat{r} = \max(r, \|\hat{h}\|_\infty)$ and $\hat{h}$ is the function from Assumption 3.1(ii).

The constant R_{C_0} that appears in the statement is defined in our "level-set assumption" (cf. Assumption 3.1(iv)).

Proof. Fix some $h \in \mathcal{H}$. We have $\left\|\overline{g}_{h,n}\right\|_n = \left\|h + \overline{g}_{h,n} + (-h)\right\|_n \le \left\|h + \overline{g}_{h,n}\right\|_n + \|-h\|_n \le \left\|h + \overline{g}_{h,n}\right\|_n + r$ thanks to $\|h\|_\infty \le r$. Hence, it remains to bound $\left\|h + \overline{g}_{h,n}\right\|_n$.

By Assumption 3.1(iv), for this it suffices if we show a bound on $\overline{L}_n(h + \overline{g}_{h,n})$ since by this assumption if $\overline{L}_n(h + \overline{g}_{h,n}) \le c$ then $\left\|h + \overline{g}_{h,n}\right\|_n \le R_c$. By the optimizing property of $\overline{g}_{h,n}$, we have $\overline{L}_n(h + \overline{g}_{h,n}) = \overline{L}_{n,h}(\overline{g}_{h,n}) \le \overline{L}_{n,h}(0) = \overline{L}_n(h)$. Now, by definition

$$\overline{L}_n(h) = \mathbb{E}\left[\frac{1}{n}\sum_i \ell(Y_i, h(X_i)) \Big| X_{1:n}\right],$$

hence, it suffices to bound $\mathbb{E}\left[\ell(Y_i, h(X_i)) | X_i\right]$. For this, we have

$$\mathbb{E}\left[\ell(Y_i, h(X_i))|X_i\right] \le \mathbb{E}\left[|\ell(Y_i, h(X_i)) - \ell(Y_i, \hat{h}(X_i))| \,|X_i\right] + \mathbb{E}\left[\ell(Y_i, \hat{h}(X_i)) \,|X_i\right],$$

where we used that by assumption the loss is nonnegative. By Assumption 3.1(ii),

$$\mathbb{E}\left[\ell(Y_i, \hat{h}(X_i)) \,|X_i\right] \le Q.$$

Therefore it is sufficient to bound

$$\mathbb{E}\left[|\ell(Y_i, h(X_i)) - \ell(Y_i, \hat{h}(X_i))| \,|X_i\right].$$

Note that by Assumption 3.1(iii), almost surely $\mathbb{E}\left[\exp\left(|\beta K_\ell(Y, r)|^2\right) | X\right] \le \Gamma_r$. So, by Lemma A.1 (i), $\mathbb{E}\left[K_\ell(Y, r) | X\right] \le \frac{1}{\beta}\sqrt{\Gamma_r - 1}$ a.s.. Thus, with $\hat{r} = \max(r, \|\hat{h}\|_\infty)$,

$$\mathbb{E}\left[|\ell(Y_i, h(X_i)) - \ell(Y_i, \hat{h}(X_i))| \,|X_i\right] \le \mathbb{E}\left[2\hat{r}K_\ell(Y_i, \hat{r}) | X_i\right] \le \frac{2\hat{r}}{\beta}\sqrt{\Gamma_{\hat{r}} - 1}.$$

Putting together the inequalities, we obtain that $\overline{L}_n(h + \overline{g}_{h,n}) \le \frac{2\hat{r}}{\beta}\sqrt{\Gamma_{\hat{r}} - 1} + Q =: C_0$ and thus $\left\|h + \overline{g}_{h,n}\right\|_n \le R_{C_0}$. □

Let us now consider bounding $\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$. In fact, we will only bound this on the event $G_{\lambda_{\min}}$ when $\hat{\lambda}_{\min} \ge \lambda_{\min}/2$. Since we use this event to upper bound $1/\hat{\lambda}_{\min}$ by $2/\lambda_{\min}$, there is no loss in bounding $\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$ on this event only. Note that by Lemma C.1 (ii), $G_{\lambda_{\min}}$ holds with probability at least $1-\delta$.

Lemma C.3. *There exist problem-dependent positive constants C_0 and $L_0 \ge 1$ such that for any $n \ge 16L_0^4$, it holds that*

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \ge 1\,, G_{\lambda_{\min}}\Big) \le \exp\left(-\frac{C_0 n}{4}\right). \tag{21}$$

The proof of this lemma follows the proofs in the paper of van de Geer [1990], who studied the deviations $\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$ for $h = 0$ (see also van de Geer 2000). It turns out the techniques of the mentioned paper are just strong enough to bound the uniform deviation $\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$. As the proof is lengthy and technical, it is developed in a separate section.

Now, combining (18), (20) and Lemma C.2 we get that on $G_{\lambda_{\min}}$,

$$G_{n,\infty} \doteq \sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty \leq \frac{2}{\lambda_{\min}} \sup_{h\in\mathcal{H}} \|g_{h,n}\|_n \leq \frac{2}{\lambda_{\min}} \left(\overline{R} + \sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \right) . \tag{22}$$

Since for any $A > 0$,

$$\mathbb{P}\left(G_{n,\infty} > A\right) \leq \mathbb{P}\left(G^c_{\lambda_{\min}}\right) + \mathbb{P}\left(G_{n,\infty} > A, G_{\lambda_{\min}}\right)$$

and by (22),

$$\mathbb{P}\left(G_{n,\infty} > A, G_{\lambda_{\min}}\right) \leq \mathbb{P}\left(\frac{2}{\lambda_{\min}} \left(\overline{R} + \sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \right) > A, G_{\lambda_{\min}} \right),$$

choosing $A = \frac{2}{\lambda_{\min}}\left(\overline{R} + 1\right)$, we see that

$$\mathbb{P}\left(G_{n,\infty} > \frac{2}{\lambda_{\min}} \left(\overline{R} + 1\right) \right) \leq \mathbb{P}\left(G^c_{\lambda_{\min}}\right) + \mathbb{P}\left(\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \geq 1, G_{\lambda_{\min}} \right) .$$

By Eq. (19) and Lemma C.3, provided that $n \geq \frac{2}{\lambda_{\min} \log\left(\frac{e}{2}\right)} \log\left(\frac{2\rho}{\delta}\right)$, $n \geq 16 L_0^4$ and $n \geq \frac{4\log\left(\frac{2}{\delta}\right)}{C_0}$ we get that

$$\mathbb{P}\left(G_{n,\infty} > \frac{2}{\lambda_{\min}} \left(\overline{R} + 1\right) \right) \leq \delta\,,$$

which is the desired statement. In particular, we can choose $U = \frac{2}{\lambda_{\min}}\left(\overline{R} + 1\right)$.

C.1 The Proof of Lemma C.3

The proof follows the ideas from the paper of van de Geer [1990]. Lemma C.3 calls for a uniform (in $h \in \mathcal{H}$) bound for $\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$. Fix $h \in \mathcal{H}$. We consider a self-normalized "version" of the differences $g_{h,n} - \overline{g}_{h,n}$, which are easier to deal with. This is done as follows: For $g \in \mathcal{G}$, define

$$\omega_{g,h} = \frac{g - \overline{g}_{h,n}}{1 + K\left\|g - \overline{g}_{h,n}\right\|_n} \quad \text{and} \quad \Omega_{h,n} = \left\{\omega_{g,h} \,:\, g \in \mathcal{G}\right\},$$

where $K > 0$ is to be chosen later. Then, for any $\omega \in \Omega_{h,n}$, $\|\omega\|_n < \frac{1}{K}$ and

$$\begin{aligned} \left\|g - \overline{g}_{h,n}\right\|_n &= \frac{\left\|g - \overline{g}_{h,n}\right\|_n}{1 + K\left\|g - \overline{g}_{h,n}\right\|_n} \left(1 + K\left\|g - \overline{g}_{h,n}\right\|_n\right) = \|\omega_{g,h}\|_n \left(1 + K\left\|g - \overline{g}_{h,n}\right\|_n\right) \\ &= \frac{\|\omega_{g,h}\|_n}{1 - K\|\omega_{g,h}\|_n} . \end{aligned} \tag{23}$$

Thus, we see that is enough to control the empirical norm of

$$\hat{\omega}_{h,n} = \omega_{g_{h,n},h} = \frac{g_{h,n} - \overline{g}_{h,n}}{1 + K\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n}.$$

The first step is to bound this norm in terms of the increments of the empirical process

$$\Delta_{h,n}(g) := L_{h,n}(g) - \overline{L}_{h,n}(g)\,.$$

Lemma C.4 ("Basic Inequality"). *Let Assumption 3.2 hold. There exists a constant η, such that on the event $G_{\lambda_{\min}}$, for any $h \in \mathcal{H}$,*

$$\eta \left\| \hat{\omega}_{h,n} \right\|_n^2 \leq \Delta_{h,n}(\overline{g}_{h,n}) - \Delta_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n})\,.$$

The proof, which is stated in Appendix C.3, follows standard arguments. Based on this, we can reduce the study of the supremum of the empirical norm of $\hat{\omega}_{h,n}$ to that of the supremum of the increments $\mathcal{V}_{h,n}(\omega) = \sqrt{n}\big(\Delta_{h,n}(\overline{g}_{h,n}) - \Delta_{h,n}(\overline{g}_{h,n} + \omega)\big)$ normalized by ω. In particular, it follows from Lemma C.4 that for $L, \sigma > 0$,

$$\begin{aligned} &\mathbb{P}\big(\sup_{h\in\mathcal{H}} \left\| \hat{\omega}_{h,n} \right\|_n \geq L\sigma\,, G_{\lambda_{\min}} \big) \\ &\quad = \mathbb{P}\Big(\exists h \in \mathcal{H} \,:\, \left\| \hat{\omega}_{h,n} \right\|_n \geq L\sigma,\, \frac{\mathcal{V}_{h,n}(\hat{\omega}_{h,n})}{\left\| \hat{\omega}_{h,n} \right\|_n^2} \geq \eta\sqrt{n}\,, G_{\lambda_{\min}} \Big) \\ &\quad \leq \mathbb{P}\left(\sup_{(g,h)\in\mathcal{G}\times\mathcal{H}:\left\|\omega_{g,h}\right\|_n \geq L\sigma} \frac{\mathcal{V}_{h,n}(\omega_{g,h})}{\left\| \omega_{g,h} \right\|_n^2} \geq \eta\sqrt{n}\,, G_{\lambda_{\min}} \right). \end{aligned} \tag{24}$$

The supremum of normalized increments similar to the one appearing above was studied by van de Geer [1990]. In fact, we will adapt Lemma 3.4 of this paper to our purposes. The lemma requires minimal modifications: In our case, the empirical process is indexed with elements of $\{\omega_{g,h} \,:\, g \in \mathcal{G}, h \in \mathcal{H}\}$, the product set $\mathcal{G} \times \mathcal{H}$, whereas van de Geer [1990] considers a similar result for $h = 0$. As a result, whereas van de Geer [1990] reduces the study of this probability to bounding the "size" of balls in the the index space, we will reduce it to bounding the size of "tubes".

To state the generalization of Lemma 3.4 of van de Geer [1990], we introduce the following abstract setting: Let $(V, d_{V,k})$, $(\Lambda, d_{\Lambda,k})$ be pseudo-metric spaces ($k = 1, \dots, n$), d_k^2 be the pseudo-metric on $V \times \Lambda$, which for $\gamma = (\nu, \lambda)$, $\tilde{\gamma} = (\tilde{\nu}, \tilde{\lambda})$ in $V \times \Lambda$ is defined by $d_k^2(\gamma, \tilde{\gamma}) = d_{V,k}^2(\nu, \tilde{\nu}) + d_{\Lambda,k}^2(\lambda, \tilde{\lambda})$. Further, let d^2 be the pseudo-metric on $V \times \Lambda$ defined by $d^2 = \frac{1}{n}\sum_{k=1}^n d_k^2$. Consider the real-valued processes $U_1, U_2, \dots, U_n$ on $V \times \Lambda$ and the process

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n U_k\,.$$

For $\sigma > 0$, denote by $H(\varepsilon, \sigma) \doteq H(\varepsilon, T(\sigma), d)$, the metric entropy of the σ-"tube"

$$T(\sigma) = \cup_{\nu\in V} \{\nu\} \times \{\lambda \in \Lambda_\nu \,:\, d_\Lambda(\lambda_\nu, \lambda) \leq \sigma\} \subset V \times \Lambda\,,$$

where for $\nu \in V$, $\Lambda_\nu \subset \Lambda$ and d_Λ (defining the "tube") is the a pseudo-metric on Λ defined by $d_\Lambda^2(\lambda, \tilde{\lambda}) = \frac{1}{n}\sum_k d_{\Lambda,k}^2(\lambda, \tilde{\lambda})$. For $L > 0$, define

$$\alpha_n(L, \sigma) = \frac{\int_0^1 \sqrt{H(uL\sigma, L\sigma)}du}{\sqrt{n}L\sigma}.$$

With this, we are ready to state our generalization of Lemma 3.4 of van de Geer [1990]:

Lemma C.5. *Assume that the following conditions hold:*

(i) $U_1, U_2, \cdots, U_n$ *are independent, centered; for all* $\nu \in V$, $Z_n(\nu, \lambda_\nu) = 0$ *for some* $\lambda_\nu \in \Lambda$, *and*

$$|U_k(\gamma) - U_k(\tilde{\gamma})| \leq M_k d_k(\gamma, \tilde{\gamma})\,, \quad \gamma, \tilde{\gamma} \in V \times \Lambda,$$

where $M_1, M_2, \cdots, M_n$ *are uniformly subgaussian, i.e., for some positive* β *and* Γ,

$$\mathbb{E}[\exp(|\beta M_k|^2)] \leq \Gamma < \infty, k = 1, 2, \dots, n.$$

(ii) Assume that $\sigma > 0$ *is such that* $\sqrt{n}\sigma \geq 1$ *and suppose*

$$\lim_{L\to\infty} \alpha_n(L, \sigma) = 0\,.$$

Then, there exist constants $L_0 \ge 1$ and C_0, depending only on (β, Γ) and the map $L \mapsto \alpha_n(L, \sigma)$, such that for all $L \ge L_0$,

$$\mathbb{P}\Big(\sup_{\nu \in V} \sup_{\substack{\lambda \in \Lambda_\nu: \\ d_\Lambda(\lambda_\nu, \lambda) > L\sigma}} \frac{|Z_n(\nu, \lambda)|}{d_\Lambda^2(\lambda_\nu, \lambda)} \ge \sqrt{n} \Big) \le \exp(-C_0 L^2 \sigma^2 n).$$

Remark C.1. The proof is obtained by modifying the proof of van de Geer [1990]'s Lemma 3.4 in a straightforward manner and hence it is omitted. A careful investigation of the original proof will find that the result also holds if we find L_0 and C_0 depending on an *upper bound* $\tilde{\alpha}_n(L, \sigma)$ for $\alpha_n(L, \sigma)$ provided that $\lim_{L \to \infty} \tilde{\alpha}_n(L, \sigma) = 0$ still holds. Moreover, if the upper bound is selected such that it does not depend on n and σ but only on L and the "size" of the spaces V, $(\Lambda_\nu)_{\nu \in V}$, then L_0 and C_0 will depend only on (β, Γ) and the mentioned "size".

To apply Lemma C.5 to our problem, we choose the spaces to be $V = \mathcal{H}$, $\Lambda = \cup_{h \in \mathcal{H}} \Lambda_h$, where $\Lambda_h = \Omega_{h,n}$. Further, we choose the pseudo-metrics to be $d_{V,k}^2(h, \tilde{h}) = |h(X_k) - \tilde{h}(X_k)|^2 + \|h - \tilde{h}\|_{\infty,n}^2$ $(h, \tilde{h} \in V)$, and $d_{\Lambda,k}(\omega, \tilde{\omega}) = |\omega(X_k) - \tilde{\omega}(X_k)|$ $(\omega, \tilde{\omega} \in \Lambda)$. We also choose $\Lambda_h = \Omega_{h,n} \subset \Lambda$. Since these pseudo-metrics are random (they depend on $X_{1:n}$), for a proper use of Lemma C.5 we again need to "condition" on $X_{1:n}$ when using this lemma. Making this argument formal has been discussed in Appendix B.

For $f \in L^1(\mathcal{X}, P_X)$, $\omega \in \Lambda$, $h \in \mathcal{H}$ set

$$\Delta_k(f) = \frac{1}{\eta} \left(\ell(Z_k, f) - \mathbb{E}_{x_{1:n}}[\ell(Z_k, f)] \right),$$
$$U_k(h, \omega) = \Delta_k(h + \overline{g}_{h,n}) - \Delta_k(h + \overline{g}_{h,n} + \omega).$$

(We remind the reader that, although not shown to minimize clutter, Δ_k and U_k do depend on $x_{1:n}$.)

Now, for $h \in \mathcal{H}$, we set $\lambda_h = 0$. Thus, $U_k(h, \lambda_h) = U_k(h, 0) = 0$. Furthermore, for $Z_n(h, \omega) = \frac{1}{\sqrt{n}} \sum_{k=1}^n U_k(h, \omega)$ we have $Z_n(h, \omega) = \frac{1}{\eta} \mathcal{V}_{h,n}(\omega)$ and therefore (using that $\lambda_h = 0$ and $d_\Lambda(\omega, \tilde{\omega}) = \|\omega - \tilde{\omega}\|_n$)

$$\sup_{h \in \mathcal{H}} \sup_{\substack{\omega \in \Lambda_h: \\ d_\Lambda(\lambda_h, \omega) > L\sigma}} \frac{Z_n(h, \omega)}{d_\Lambda^2(\lambda_h, \omega)} = \sup_{h \in \mathcal{H}} \sup_{\substack{\omega \in \Omega_{h,n}: \\ \|\omega\|_n > L\sigma}} \frac{\mathcal{V}_{h,n}(\omega)}{\eta \|\omega\|_n^2} =: Q_n(L\sigma), \tag{25}$$

showing that the conclusion of the lemma suffices to bound the quantity of interest appearing in (24).

We claim that the condition of Lemma C.5 are satisfied for $[P_X]$ almost every $x_{1:n} \in \mathcal{X}^n$ such that $\lambda_{\min}(x_{1:n}) \doteq \lambda_{\min}(\Phi(x_{1:n})^\top \Phi(x_{1:n})) \ge \lambda_{\min}/2$. Let $\mathcal{N} \subset \mathcal{X}^n$ be the $[P_X]$ null-set where the claim is not required to hold (we will construct $\mathcal{N}$ in the proof). That U_k are centered and $Z_n(h, \lambda_h) = 0$ for any $h \in \mathcal{H}$ holds by construction. As far as the remaining conditions are concerned, we have:

Condition (i)*, the independence of* (U_k): This follows from the definition of $\mathbb{P}_{x_{1:n}}$ and the independence of (X_k, Y_k).

Condition (i)*, the Lipschitzness of* U_k: Our goal is to show (for later use) that the Lipschitz coefficients M_k can be chosen independently of n and $x_{1:n}$ as long $\lambda_{\min}(x_{1:n}) \ge \lambda_{\min}/2$. For this, we will assume that

$$K \ge 1. \tag{26}$$

Since U_k is defined as a function of Δ_k, we consider the Lipschitzness of Δ_k first. Using the definition of Δ_k and the Lipschitzness of ℓ (cf. Assumption 3.1(iii)), for any $f, f' \in L^1(\mathcal{X}, P_X)$ we have

$$|\Delta_k(f) - \Delta_k(f')| \le \frac{1}{\eta} \left(\frac{|\ell(Z_k, f)) - \ell(Z_k, f')|}{|f(X_k) - f'(X_k)|} + \frac{\mathbb{E}\left[|\ell(Z_k, f) - \ell(Z_k, f')|\, |X_k\right]}{|f(X_k) - f'(X_k)|} \right) |f(X_k) - f'(X_k)|.$$

Denote $\frac{|\ell(Z_k, f)) - \ell(Z_k, f')|}{|f(X_k) - f'(X_k)|} + \frac{\mathbb{E}\left[|\ell(Z_k, f) - \ell(Z_k, f')|\, |X_k\right]}{|f(X_k) - f'(X_k)|}$ by $N_k(f, f')$. Thus, for $h, \tilde{h} \in \mathcal{H}$, $\omega, \tilde{\omega} \in \Lambda$, letting $f = h + \overline{g}_{h,n}$, $\tilde{f} = \tilde{h} + \overline{g}_{\tilde{h},n}$,

$$\begin{aligned}
&|U_k(h, \omega) - U_k(\tilde{h}, \tilde{\omega})| \\
&\quad \le \frac{1}{\eta} N_k(f, \tilde{f}) \left| f(X_k) - \tilde{f}(X_k) \right| \\
&\qquad + \frac{1}{\eta} N_k(f + \omega, \tilde{f} + \tilde{\omega}) \left\{ \left| f(X_k) - \tilde{f}(X_k) \right| + |\omega(X_k) - \tilde{\omega}(X_k)| \right\}
\end{aligned}$$

Now, by assumption $|h(x_k)|, |\tilde{h}(x_k)| \le r$. From $\lambda_{\min}(x_{1:n}) \ge \lambda_{\min}/2$, (18) and Lemma C.2 it follows that $|\overline{g}_{h,n}(x_k)|, |\overline{g}_{\tilde{h},n}(x_k)| \le \frac{2\overline{R}}{\lambda_{\min}}$. Also, by the same argument as in Lemma C.7, again thanks to $\lambda_{\min}(x_{1:n}) \ge \lambda_{\min}/2$, $|\omega(x_k)|, |\tilde{\omega}(x_k)| \le \frac{1}{K(\lambda_{\min}/2)^{1/2}} \le \frac{1}{(\lambda_{\min}/2)^{1/2}}$, where we used (26). Hence,

$$N_k(f, \tilde{f}) \le K_\ell\left(Y_k, r + \frac{2\overline{R}}{\lambda_{\min}}\right) + \mathbb{E}\left[K_\ell\left(Y_k, r + \frac{2\overline{R}}{\lambda_{\min}}\right) \mid X_k\right]$$

and similarly,

$$N_k(f+\omega, \tilde{f}+\tilde{\omega}) \le K_\ell\left(Y_k, r + \frac{2\overline{R}}{\lambda_{\min}} + \frac{1}{(\lambda_{\min}/2)^{1/2}}\right) + \mathbb{E}\left[K_\ell\left(Y_k, r + \frac{2\overline{R}}{\lambda_{\min}} + \frac{1}{(\lambda_{\min}/2)^{1/2}}\right) \mid X_k\right]$$

Now,

$$\begin{aligned}|f(x_k) - \tilde{f}(x_k)| &\le |h(x_k) - \tilde{h}(x_k)| + |\overline{g}_{h,n}(x_k) - \overline{g}_{\tilde{h},n}(x_k)| \\ &\le |h(x_k) - \tilde{h}(x_k)| + K_h\|h - \tilde{h}\|_{\infty,n}\,,\end{aligned}$$

where the second inequality follows since by Assumption 3.4, $h \mapsto \overline{g}_{h,n}(x_k)$ is K_h-Lipschitz. Therefore, by the choice of $d_{V,k}$ and $d_{\Lambda,k}$,

$$|U_k(h,\omega) - U_k(\tilde{h},\tilde{\omega})| \le \frac{2M_k}{\eta}\left(d_{V,k}(h,\tilde{h}) + d_{\Lambda,k}(\omega,\tilde{\omega})\right) \le M_k' d_k\left((h,\omega),(\tilde{h},\tilde{\omega})\right)$$

where $M_k = 2K_\ell\left(Y_k, r + \frac{2\overline{R}}{\lambda_{\min}} + \frac{1}{(\lambda_{\min}/2)^{1/2}}\right) + 2\mathbb{E}\left[K_\ell\left(Y_k, r + \frac{2\overline{R}}{\lambda_{\min}} + \frac{1}{(\lambda_{\min}/2)^{1/2}}\right) \mid X_k\right]$. Note that by Lemma A.1(i), it holds almost surely that $\mathbb{E}\left[K_\ell\left(Y_k, r + \frac{2\overline{R}}{\lambda_{\min}} + \frac{1}{(\lambda_{\min}/2)^{1/2}}\right) \mid X_k\right] \le \frac{1}{\beta}\sqrt{\Gamma_{r+2\overline{R}/\lambda_{\min}+1/(\lambda_{\min}/2)^{1/2}} - 1}$. Then by Lemma A.1(ii), M_k is uniformly subgaussian, so is M_k'.

Condition (ii): We want to verify that $\alpha_n(L,\sigma) \to 0$ as $L \to \infty$ and show that in fact an upper bound $\tilde{\alpha}(L)$ on $\alpha_n(L,\sigma)$ which is independent of $x_{1:n}$, n, K and σ exists such that $\tilde{\alpha}(L) \to 0$ still holds. Since $\alpha_n(L,\sigma)$ depends on the entropy numbers $H(\varepsilon, T(\sigma), d)$ of the tube w.r.t. $d^2 = \frac{1}{n}\sum_k d_k^2$, first we need to estimate these entropy numbers. For $\gamma = (h,\omega)$, $\tilde{\gamma} = (\tilde{h},\tilde{\omega})$, we have

$$\begin{aligned}d^2(\gamma,\tilde{\gamma}) &= \frac{1}{n}\sum_k d_{V,k}^2(h,\tilde{h}) + \frac{1}{n}\sum_k d_{\Lambda,k}^2(\omega,\tilde{\omega}) \\ &= \|h-\tilde{h}\|_n^2 + \|h-\tilde{h}\|_{\infty,n}^2 + \|\omega-\tilde{\omega}\|_n^2 \le 2\left(\|h-\tilde{h}\|_{\infty,n}^2 + \|\omega-\tilde{\omega}\|_n^2\right)\,.\end{aligned}$$

Further, $d_\Lambda^2(\omega,\tilde{\omega}) = \|\omega-\tilde{\omega}\|_n^2$ and therefore by the choice $\Lambda_h = \Omega_{h,n}$ and $\lambda_h = 0$,

$$T(\sigma) = \{(h,\omega) \,:\, h \in \mathcal{H}, \omega \in \Omega_{h,n} \text{ s.t. } \|\omega\|_n \le \sigma\}\,.$$

Therefore, it suffices to estimate the metric entropy of $T(\sigma)$ at different scales w.r.t. the pseudo-norm $\|\cdot\|_T$ defined by $\|(h,\omega_{g,h})\|_T = \|h\|_{\infty,n} + \|\omega_{g,h}\|_n$. This is done in the following proposition, which also shows that the integrability assumption is satisfied (the proof is presented in the appendix):

Proposition C.6. *Let Assumptions 3.1 to 3.4 hold. Take $n \ge 1$, $K > 0$, $\varepsilon > 0$, $1 \ge \sigma \ge \varepsilon$ such that $K\sigma \le 1/2$. Then on $G_{\lambda_{\min}}$,*

$$H(\varepsilon, T(\sigma), \|\cdot\|_T) \le \rho\log(\sigma/\varepsilon) + \rho\log(241) + AH(\tfrac{\varepsilon}{A}, \mathcal{H}, \|\cdot\|_{\infty,n})$$

holds a.s. for some positive (non-random) constant A that depends only on K_h.

Furthermore, on $G_{\lambda_{\min}}$,

$$\int_0^1 H^{1/2}(u\sigma, T(\sigma), \|\cdot\|_T)\,du \le A'\sqrt{\rho} + \frac{A''}{\sigma}\,,$$

holds a.s. for some universal constant $A' > 0$ and some non-random constant A'' that depends on C_H and K_h only.

Now, $H(\varepsilon,\sigma) = H(\varepsilon, T(\sigma), d) \leq CH(\varepsilon, T(\sigma), \|\cdot\|_T)$ with some universal constant C, hence $H(uL\sigma, L\sigma) \leq CH(uL\sigma, T(L\sigma), \|\cdot\|_T)$ and by the previous result,

$$\int_0^1 H^{1/2}(uL\sigma, L\sigma)du \leq C^{1/2}\int_0^1 H^{1/2}(uL\sigma, T(L\sigma), \|\cdot\|_T)\, du \leq C'\left(1+\frac{1}{\sigma}\right) \leq \frac{2C'}{\sigma}$$

where C' is a constant that is independent of L, n, K, σ and we assumed that $\sigma \leq 1$. Hence,

$$\alpha_n(L,\sigma) \leq \frac{2C'}{\sqrt{n}L\sigma^2} \leq \frac{2C'}{L}$$

provided that $\sqrt{n}\sigma^2 \geq 1$. Thus, under this condition, $\alpha_n(L,\sigma) \to 0$ as $L \to \infty$, as required. Furthermore, the upper bound on $\alpha_n(L,\sigma)$ is independent of $x_{1:n}$, K, n and σ. Therefore, L_0 and C_0 can be selected independently of $x_{1:n}$, K, n and σ, finishing the verification of the conditions of Lemma C.5.

Therefore, using (25) we conclude that for any $L \geq L_0$, K, n, σ such that $\sqrt{n}\sigma^2 \geq 1$ and $K\sigma \leq 1/2$ and $K \geq 1$, for $[P_X]$ almost all $x_{1:n}$ such that $\lambda_{\min}(x_{1:n}) \geq \lambda_{\min}/2$,

$$\mathbb{P}_{x_{1:n}}\Big(Q_n(L\sigma) \geq \sqrt{n}\Big) \leq \exp(-C_0L^2\sigma^2 n).$$

Now, by the definition of $\mathbb{P}_{x_{1:n}}$,

$$\begin{aligned}
\mathbb{P}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) &= \int \mathbb{P}_{x_{1:n}}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) P_X(dx_{1:n}) \\
&= \int_{\lambda_{\min}(x_{1:n})\geq\lambda_{\min}/2} \mathbb{P}_{x_{1:n}}\left(Q_n(L\sigma) \geq \sqrt{n}\right) P_X(dx_{1:n}) \\
&\leq \int_{\lambda_{\min}(x_{1:n})\geq\lambda_{\min}/2} \exp(-C_0L^2\sigma^2 n) P_X(dx_{1:n}) \leq \exp(-C_0L^2\sigma^2 n)\,,
\end{aligned}$$

where the second equality follows since $G_{\lambda_{\min}}$ is $X_{1:n}$-measurable.

Hence, by combining (23) and (24), using the definition of $Q_n(L\sigma)$ in (25) and choosing $L = L_0$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \geq \frac{L_0\sigma}{1-KL_0\sigma}, G_{\lambda_{\min}}\right) \leq \mathbb{P}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) \leq \exp(-C_0L_0^2\sigma^2 n)\,.$$

Choosing $\sigma = 1/(2L_0)$ and $K = 1$, noting that $n \geq \sigma^{-4}$ then translates into $n \geq 16L_0^4$ gives that

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \geq 1, G_{\lambda_{\min}}\right) \leq \exp(-C_0 n/4),$$

which is the desired result (we also used that $L_0 \geq 1$ by assumption and hence $\sigma \leq 1$ which gives that $\sqrt{n}\sigma \geq \sqrt{n}\sigma^2 \geq 1$).

C.2 Eigenvalue Bound

Lemma C.1. *The following hold:*

(i) With probability one, for any $\theta \in \mathbb{R}^d$, $\theta^\top \hat{G}\theta \leq \frac{\theta^\top G\theta}{\lambda_{\min}}$.

(ii) Assuming that $n \in \mathbb{N}$ and $\delta \in (0,1)$ are such that

$$n \geq \frac{2}{\lambda_{\min}\log\left(\frac{e}{2}\right)}\log\left(\frac{\rho}{\delta}\right), \tag{19}$$

where ρ and $\lambda_{\min}$ are respectively the rank and the smallest positive eigenvalue of G, with probability at least $1-\delta$, it holds that $\hat{\lambda}_{\min} \geq \frac{\lambda_{\min}}{2} > 0$.

(iii) For any n, δ satisfying (19), with probability $1-\delta$ it holds that for any $\theta \in \mathbb{R}^d$ and $[P_X]$ almost every $x \in \mathcal{X}$, $|\langle\theta, \phi(x)\rangle| \leq \sqrt{\frac{2\theta^\top\hat{G}\theta}{\lambda_{\min}}}$.

Proof. Part (i): We first show that $\mathrm{Ker}(G) \subseteq \mathrm{Ker}(\hat{G})$ holds almost surely: In particular, this can be seen by proving that $G\theta = 0$ for some $\theta \in \mathbb{R}^d$ then with probability one, $\hat{G}\theta = 0$ also holds. Indeed, if the latter did not hold with probability one, then for some $\varepsilon > 0$, $\mathbb{P}(\theta^\top \hat{G}\theta \geq \varepsilon) > 0$ would hold. Then, $\theta^\top G\theta = \mathbb{E}[\theta^\top \hat{G}\theta] \geq \varepsilon \mathbb{P}(\theta^\top \hat{G}\theta \geq \varepsilon) > 0$, which means that $\theta \notin \mathrm{Ker}(G)$. Now, if we take a set of vectors $\{\theta_1, \ldots, \theta_m\}$ spanning $\mathrm{Ker}(G)$, then on some event E with $\mathbb{P}(E) = 1$, $\hat{G}\theta_i = 0$ holds for *all* $1 \leq i \leq m$. Now, on E, $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$. Indeed, take an arbitrary $\theta \in \mathrm{Ker}(G)$ and expand it using $\{\theta_i\}$: $\theta = \sum_{i=1}^m \lambda_i \theta_i$. Then, $\hat{G}t = \sum_i \lambda \hat{G}\theta_i$ and since $\hat{G}\theta_i = 0$ simultaneously for all i, the statement follows.

Now, for proving Part (i), consider the event E where $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$. We prove the result on E: Pick any $\theta \in \mathbb{R}^d$ and decompose it into $\theta = \theta_\perp + \theta_{||}$ such that $\theta_\perp \perp \mathrm{Im}(G)$ and $\theta_{||} \in \mathrm{Im}(G)$. Hence, $\theta^\top G\theta = \theta_{||}^\top G\theta_{||}$. Since $\theta_\perp \in \mathrm{Ker}(G)$ and $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$, we have $\hat{G}\theta_\perp = 0$. Hence, $\theta^\top \hat{G}\theta = \theta_{||}^\top \hat{G}\theta_{||}$. Now, since $\|\phi(x)\|_2 \leq 1$ it holds that $\hat{\lambda}_{\max} \leq 1$, where $\hat{\lambda}_{\max}$ denotes the largest eigenvalue of $\hat{G}$. Therefore, on E,

$$\theta^\top \hat{G}\theta = \theta_{||}^\top \hat{G}\theta_{||} \leq \|\theta_{||}\|_2^2 \leq \frac{\theta_{||}^\top G\theta_{||}}{\lambda_{\min}} = \frac{\theta^\top G\theta}{\lambda_{\min}}.$$

Since $\mathbb{P}(E) = 1$, the result follows.

Part (ii): By the "Eigenvalue Chernoff Bound" (Theorem 4.1) of Gittens and Tropp [2011], with probability at least $1 - \rho \exp\Big(-n\lambda_{\min}\big(\varepsilon + (1-\varepsilon)\log(1-\varepsilon)\big)\Big)$, $\hat{\lambda}_{\min} \geq (1-\varepsilon)\lambda_{\min}$. Choosing $\varepsilon = 1/2$ gives the result.

Part (iii): Fix n, δ as required. Let E be the event where $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$ and let F_δ be the event where the inequality of Part (ii) holds. Take the set S of those $x \in \mathrm{supp}(P_X)$ where $\mathrm{Ker}(G) \subset \mathrm{Ker}(\phi(x)\phi(x)^\top)$ holds. It follows from the argument presented in Part (i) that $P_X(\mathcal{X} \setminus S) = 0$.

Since $\mathbb{P}(E \cap F_\delta) \geq 1 - \delta$, it suffices to prove the statement on $E \cap F_\delta$. Hence, in what follows all statements are meant to hold on this event. Pick any $\theta \in \mathbb{R}^d$, $x \in S$ and decompose θ as before. Then, thanks to $x \in S$ it holds that $\theta_\perp \in \mathrm{Ker}(\phi(x)\phi(x)^\top)$. Hence, $\langle \theta, \phi(x)\rangle^2 = \theta^\top \phi(x)\phi(x)^\top \theta = \theta_{||}\phi(x)\phi(x)^\top \theta_{||} = \langle \theta_{||}, \phi(x)\rangle^2$. Now, owning to $\|\phi(x)\|_2 \leq 1$,

$$\langle \theta_{||}, \phi(x)\rangle^2 \leq \left\|\theta_{||}\right\|_2^2 \leq \frac{\theta_{||}^\top \hat{G}\theta_{||}}{\hat{\lambda}_{\min}} \leq \frac{2\,\theta_{||}^\top \hat{G}\theta_{||}}{\lambda_{\min}} = \frac{2\,\theta^\top \hat{G}\theta}{\lambda_{\min}},$$

where the last inequality follows from Part (ii). □

C.3 Proof of the "Basic Inequality" (Lemma C.4)

We start with a uniform bound for the infinity norm of elements in $\Omega_{h,n}$. Let

$$K_\infty = \frac{1}{K(\lambda_{\min}/2)^{1/2}}.$$

Recall that $G_{\lambda_{\min}}$ is the event when $\hat{\lambda}_{\min} \geq \lambda_{\min}/2$.

Lemma C.7. *On the event $G_{\lambda_{\min}}$,*

$$\sup_{\omega \in \Omega_{h,n}} \|\omega\|_\infty < K_\infty.$$

Proof. Introduce $\|x\|_M^2 = x^\top M x$ for M positive definite. Let $\overline{\theta}_{h,n}$ be the parameter of $\overline{g}_{h,n}$. Thus,

$$|\omega(x)| = \frac{|\langle \phi(x), \theta - \overline{\theta}_{h,n}\rangle|}{1 + K\sqrt{\left\|\theta - \overline{\theta}_{h,n}\right\|_{\hat{G}}^2}} \leq \frac{\|\theta - \overline{\theta}_{h,n}\|_2}{1 + K\hat{\lambda}_{\min}^{1/2}\|\theta - \overline{\theta}_{h,n}\|_2} < \frac{1}{K\hat{\lambda}_{\min}^{1/2}} \leq K_\infty.$$

□

With this, we can state the proof of Lemma C.4:

Lemma C.4 ("Basic Inequality")**.** *Let Assumption 3.2 hold. There exists a constant η, such that on the event $G_{\lambda_{\min}}$, for any $h \in \mathcal{H}$,*

$$\eta \left\|\hat{\omega}_{h,n}\right\|_n^2 \leq \Delta_{h,n}(\overline{g}_{h,n}) - \Delta_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}).$$

Proof. The proof follows the ideas underlying the proof of Lemma 12.2 of the book of van de Geer [2000].

First, we will prove that $L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) \geq 0$. Note that by the definition of $g_{h,n}$, $L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(g_{h,n}) \geq 0$. Thus,

$$0 \leq L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(g_{h,n} - \overline{g}_{h,n} + \overline{g}_{h,n}) \leq \frac{1}{\alpha}(L_{h,n}(\overline{g}_{h,n}) - L_{h,n}((1-\alpha)\overline{g}_{h,n} + \alpha g_{h,n}))$$

for any $0 < \alpha \leq 1$, because of the convexity of $L_{h,n}$. Taking $\alpha = \frac{1}{1+K\left\|g_{h,n}-\overline{g}_{h,n}\right\|_n}$, the previous inequality gives

$$\frac{1}{\alpha}(L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n})) \geq 0. \tag{27}$$

Now take $\varepsilon > 0$ small enough so that it satisfies Assumption 3.2 and also $\frac{\varepsilon}{K_\infty} \leq 1$. Then we have $\overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) \geq \frac{\varepsilon}{K_\infty}(\overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}))$ because $\overline{g}_{h,n}$ is a minimizer of $\overline{L}_{h,n}$ (thus $\overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) > 0$) and thus

$$\overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) \geq \overline{L}_{h,n}(\overline{g}_{h,n} + \frac{\varepsilon}{K_\infty}\hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) \geq \frac{\varepsilon^3}{K_\infty^2}\|\hat{\omega}_{h,n}\|_n^2. \tag{28}$$

Here, the first inequality holds by the convexity of $\overline{L}_{h,n}$. The second inequality follows from Assumption 3.2 used with $a = \hat{\omega}_{h,n}\Big|_{X_{1:n}}$, once we verify that its conditions. That $a \in [-\varepsilon, \varepsilon]^n$ follows from Lemma C.7, while $a \in \mathrm{Im}(\Phi)$ follows since both $g_{h,n}|_{X_{1:n}}$ and $\overline{g}_{h,n}|_{X_{1:n}}$ satisfy this, by construction. Combining (27) and (28) gives the desired result. □

C.4 Proof of Proposition C.6

The result we want to prove is as follows:

Proposition C.6. *Let Assumptions 3.1 to 3.4 hold. Take $n \geq 1$, $K > 0$, $\varepsilon > 0$, $1 \geq \sigma \geq \varepsilon$ such that $K\sigma \leq 1/2$. Then on $G_{\lambda_{\min}}$,*

$$H(\varepsilon, T(\sigma), \|\cdot\|_T) \leq \rho\log(\sigma/\varepsilon) + \rho\log(241) + AH(\tfrac{\varepsilon}{A}, \mathcal{H}, \|\cdot\|_{\infty,n})$$

holds a.s. for some positive (non-random) constant A that depends only on K_h.

Furthermore, on $G_{\lambda_{\min}}$,

$$\int_0^1 H^{1/2}(u\sigma, T(\sigma), \|\cdot\|_T)\, du \leq A'\sqrt{\rho} + \frac{A''}{\sigma}\,,$$

holds a.s. for some universal constant $A' > 0$ and some non-random constant A'' that depends on C_H and K_h only.

We start by showing that the mapping $g, h \mapsto (h, \omega_{g,h})$ is Lipschitz w.r.t $\|\cdot\|_T$ ($g \in \mathcal{G}$, $h \in \mathcal{H}$) as this will allow us to bound the entropy of $T(\sigma)$ in terms of the entropy of $\mathcal{H}$ and the entropy of the union of balls in $\cup_{h\in\mathcal{H}}\Omega_{h,n}$, in particular $\cup_{h\in\mathcal{H}}\Omega_{h,n}(\sigma)$.

Proposition C.8. *Let Assumption 3.4 hold Then, for any $K, \sigma > 0$ satisfying $K\sigma \leq 1/2$ and any $(g_1, h_1), (g_2, h_2) \in \mathcal{G} \times \mathcal{H}$ s.t. $\|\omega_{g_i,h_i}\|_n \leq \sigma$,*

$$\|\omega_{g_1,h_1} - \omega_{g_2,h_2}\|_n \leq K_g\|g_1 - g_2\|_n + K_gK_h\|h_1 - h_2\|_{\infty,n} \tag{29}$$

holds a.s. on the event $G_{\lambda_{\min}}$, where $K_g = 4\sqrt{2}$.

The constant K_h appearing in the bound is the Lipschitz constant defined in Assumption 3.4.

Proof. Take any $(g_1, h_1), (g_2, h_2) \in \mathcal{G} \times \mathcal{H}$ with the required property. By the triangle inequality, we have

$$\|\omega_{g_1,h_1} - \omega_{g_2,h_2}\|_T \leq \|\omega_{g_1,h_1} - \omega_{g_2,h_1}\|_T + \|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_T\ .$$

Let us consider bounding $\|\omega_{g_1,h_1} - \omega_{g_2,h_1}\|_T$ as the first step. To minimize clutter, introduce $h = h_1$, $\omega_i = \omega_{g_i,h}$, $i = 1, 2$. With this, our goal is to bound $\|\omega_1 - \omega_2\|_T$.

We have

$$
\begin{aligned}
|\omega_1(x) - \omega_2(x)| &= \left| \frac{(g_1 - \overline{g}_{h,n})(x)}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{(g_2 - \overline{g}_{h,n})(x)}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right| \\
&= \left| g_1(x) \left(\frac{1}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right) \right. \\
&\qquad + \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} (g_1 - g_2)(x) \\
&\qquad \left. - \overline{g}_{h,n}(x) \left(\frac{1}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right) \right| .
\end{aligned}
$$

By the triangle inequality,

$$
\left| \frac{1}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right| \le K \|g_1 - g_2\|_n .
$$

Thus,

$$
|\omega_1(x) - \omega_2(x)| \le K |g_1(x) - \overline{g}_{h,n}(x)| \; \|g_1 - g_2\|_n + |(g_1 - g_2)(x)|
$$

and therefore,

$$
\begin{aligned}
n\|\omega_1 - \omega_2\|_n^2 &\le \sum_{i=1}^n \left\{ K|g_1(X_i) - \overline{g}_{h,n}(X_i)| \; \|g_1 - g_2\|_n + |(g_1 - g_2)(X_i)| \right\}^2 \\
&\le 2 \sum_{i=1}^n \left\{ K^2 |g_1(X_i) - \overline{g}_{h,n}(X_i)|^2 \; \|g_1 - g_2\|_n^2 + |(g_1 - g_2)(X_i)|^2 \right\} \\
&\le 2n(K^2 \left\|g_1 - \overline{g}_{h,n}\right\|_n^2 + 1) \; \|g_1 - g_2\|_n^2 .
\end{aligned}
$$

By Equation (23),

$$
\left\|g_1 - \overline{g}_{h,n}\right\|_n = \frac{\|\omega_1\|_n}{1 - K\|\omega_1\|_n} .
$$

Since $\omega_1 \in \Omega_{h,n}(\sigma)$, $\|\omega_1\|_n \le \sigma$ and $K\sigma < 1$ by assumption, $\left\|g_1 - \overline{g}_{h,n}\right\|_n \le \frac{\sigma}{1-K\sigma}$. Combining this with the bound on $n\|\omega_1 - \omega_2\|_n^2$, after simplification we get

$$
\begin{aligned}
\|\omega_1 - \omega_2\|_n &\le \sqrt{2 + 2\left(\frac{K\sigma}{1 - K\sigma}\right)^2} \; \|g_1 - g_2\|_n \le \sqrt{2}\left(1 + \frac{K\sigma}{1 - K\sigma}\right) \|g_1 - g_2\|_n \\
&= \frac{2\sqrt{2}}{1 - K\sigma} \|g_1 - g_2\|_n \le K_g \; \|g_1 - g_2\|_n ,
\end{aligned} \tag{30}
$$

where $K_g = 4\sqrt{2}$ in the last two steps we used that by assumption $K\sigma \le 1/2$.

Let us now consider bounding

$$
\|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_n .
$$

Noticing that apart from a sign, $\overline{g}_{h,n}$ and g play a symmetric role in the definition of $\omega_{g,h}$, following the derivation in the first part we get that, similarly to (30),

$$
\|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_n \le \frac{2\sqrt{2}}{1 - K\sigma} \|\bar{g}_{h_1,n} - \bar{g}_{h_2,n}\|_n \le K_g \|\bar{g}_{h_1,n} - \bar{g}_{h_2,n}\|_n .
$$

Since by Assumption 3.4, $\|\bar{g}_{h_1,n} - \bar{g}_{h_2,n}\|_n \le K_h \|h_1 - h_2\|_{\infty,n}$ holds a.s. on $G_{\lambda_{\min}}$, we get

$$
\|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_n \le K_g K_h \|h_1 - h_2\|_{\infty,n} .
$$

Putting together the bounds obtained, we get that on the event $G_{\lambda_{\min}}$,

$$\|\omega_{g_1,h_1} - \omega_{g_2,h_2}\|_n \le K_g \|g_1 - g_2\|_n + K_g K_h \|h_1 - h_2\|_{\infty,n}$$

as required. □

With this, we can state the proof of Proposition C.6.

Proof of Proposition C.6. We can write

$$T(\sigma) = \cup_{h\in\mathcal{H}}\{h\} \times \Omega_{h,n}(\sigma)\,,$$

where

$$\Omega_{h,n}(\sigma) = \{\omega \in \Omega_{h,n} \,:\, \|\omega\|_n \le \sigma\}\ .$$

We first show that

$$H(\varepsilon, T(\sigma), \|\cdot\|_T) \le H(\tfrac{\varepsilon}{2}, \mathcal{H}, \|\cdot\|_{\infty,n}) + H(\tfrac{\varepsilon}{2}, \Omega_n(\sigma), \|\cdot\|_n)\,, \tag{31}$$

where $\Omega_n(\sigma) = \cup_{h\in\mathcal{H}}\Omega_{h,n}(\sigma)$. In short, this follows since $T(\sigma) \subset \mathcal{H} \times \Omega_n(\sigma)$ and since, by definition, $\|\cdot\|_T$ is obtained by "summing" $\|\cdot\|_{\infty,n}$ and $\|\cdot\|_n$.

In details, we have: Let C be an integer s.t. $C \ge \exp\left(H(\varepsilon/2, \mathcal{H}, \|\cdot\|_{\infty,n})\right)$. Then, there exists $\{h_1, \ldots, h_C\} \subset \mathcal{H}$ such that for any $h \in \mathcal{H}$, $\|h - h_i\|_{\infty,n} \le \varepsilon/2$ for some $i \in \{1, \ldots, C\}$. Similarly, let D be an integer s.t.

$$D \ge \exp(H(\tfrac{\varepsilon}{2}, \Omega_n(\sigma), \|\cdot\|_n)) \ge \max_{1\le i\le C} \exp(H(\tfrac{\varepsilon}{2}, \Omega_{h_i,n}(\sigma), \|\cdot\|_n))$$

and $\{\omega_1, \ldots, \omega_D\} \subset \Omega_n(\sigma)$ be an $\varepsilon/2$-net of $\Omega_n(\sigma)$ w.r.t. $\|\cdot\|_n$. Then,

$$\{(h_i, \omega_j \,:\, 1 \le i \le C, 1 \le j \le D\}$$

is an ε-net of $T(\sigma)$: To show this pick any $(h, \omega) \in T(\sigma)$. Then, take the index i such that $\|h - h_i\|_{\infty,n} \le \varepsilon/2$ and take the index j such that $\|\omega - \omega_j\|_n \le \varepsilon/2$. Then, $\|(h, \omega) - (h_i, \omega_j)\|_T = \|h - h_i\|_{\infty,n} + \|\omega - \omega_j\|_n \le \varepsilon$ as required. This shows that (31) indeed holds.

Next, we bound $H(\varepsilon, \Omega_n(\sigma), \|\cdot\|_n)$. We have

$$\begin{aligned}\Omega_n(\sigma) &= \left\{\omega_{g,h} \,:\, h \in \mathcal{H}, g \in \mathcal{G}, \|\omega_{g,h}\|_n \le \sigma\right\} \\ &\subset \left\{\omega_{g,h} \,:\, h \in \mathcal{H}, g \in \mathcal{G}, \left\|g - \overline{g}_{h,n}\right\|_n \le \tfrac{\sigma}{1-K\sigma}\right\}\,,\end{aligned} \tag{32}$$

where the containment follows since by Equation (23), $\left\|g - \overline{g}_{h,n}\right\|_n = \frac{\|\omega_{g,h}\|_n}{1-K\|\omega_{g,h}\|_n}$. For $s \ge 0$ define

$$\mathcal{G}_h(s) = \left\{g \in \mathcal{G} \,:\, \left\|g - \overline{g}_{h,n}\right\|_n \le s\right\}\ .$$

Pick $\hat{\mathcal{H}} \subset \mathcal{H}$ and an arbitrary "discretization" map $N : \mathcal{H} \to \hat{\mathcal{H}}$. We claim that on $G_{\lambda_{\min}}$,

$$\Omega_n(\sigma) \subset \left\{\omega_{g,h} \,:\, h \in \mathcal{H}, g \in \mathcal{G}_{N(h)}\left(\tfrac{\sigma}{1-K\sigma} + K_h \|N(h) - h\|_{\infty,n}\right)\right\} \text{ a.s.} \tag{33}$$

By (32) it suffices to show that for any $h \in \mathcal{H}$ and $g \in \mathcal{G}_h\left(\frac{\sigma}{1-K\sigma}\right)$,

$$g \in \mathcal{G}_{N(h)}\left(\tfrac{\sigma}{1-K\sigma} + K_h \|N(h) - h\|_{\infty,n}\right) \tag{34}$$

also holds true. For brevity introduce $h' = N(h)$. Thanks to the choice g and Assumption 3.4,

$$\left\|g - \overline{g}_{h',n}\right\|_n \le \left\|g - \overline{g}_{h,n}\right\|_n + \left\|\overline{g}_{h,n} - \overline{g}_{h',n}\right\|_n \le \frac{\sigma}{1-K\sigma} + K_h \|h - h'\|_{\infty,n}$$

holds a.s. on $G_{\lambda_{\min}}$, which shows that (34) indeed holds.

The following statements holds a.s. on $G_{\lambda_{\min}}$ – hence we will not mention this condition to minimize clutter. If $\hat{\mathcal{H}}$ is an $\varepsilon/(2K_gK_h)$-net of $\mathcal{H}$ w.r.t. $\|\cdot\|_{\infty,n}$ and $N(h) = \arg\min_{h'\in\mathcal{H}} \|h-h'\|_{\infty,n}$ then $K_h \|N(h)-h\|_{\infty,n} \le \varepsilon/(2K_g) \le \varepsilon/2$ and therefore for any $h' \in \hat{\mathcal{H}}$,

$$\mathcal{G}_{h'}\left(\tfrac{\sigma}{1-K\sigma} + K_h \|N(h)-h\|_{\infty,n}\right) \subset \mathcal{G}_{h'}\left(2\sigma + \frac{\varepsilon}{2}\right).$$

For each $h \in \hat{\mathcal{H}}$, let $\hat{\mathcal{G}}_{h'}$ be an $\varepsilon/2K_g$-net of $\mathcal{G}_{h'}\left(2\sigma + \frac{\varepsilon}{2}\right)$. We claim that

$$S = \left\{\omega_{g',h'} \,:\, h' \in \hat{\mathcal{H}},\ g' \in \hat{\mathcal{G}}_{h'}\right\}$$

is an ε-net of $\Omega_n(\sigma)$ w.r.t. $\|\cdot\|_n$. Indeed, let $\omega = \omega_{g,h} \in \Omega_n(\sigma)$ arbitrary. Let h' be the nearest neighbor of h in $\hat{\mathcal{H}}$ w.r.t. $\|\cdot\|_{\infty,n}$ and let g' be the nearest neighbor of g in $\hat{\mathcal{G}}_{h'}$ w.r.t. $\|\cdot\|_n$. Note that $g \in \mathcal{G}_{h'}(2\sigma+\varepsilon/2)$. Then, by Proposition C.8,

$$\|\omega_{g,h} - \omega_{g',h'}\|_n \le K_g \|g-g'\|_n + K_gK_h \|h-h'\|_{\infty,n}.$$

Now, because $g \in \mathcal{G}_{h'}(2\sigma+\varepsilon/2)$ and $\hat{\mathcal{G}}_{h'}$ is an $\varepsilon/(2K_k)$-net of this set, $K_g\|g-g'\|_n \le \varepsilon/2$. Similarly, by the choice of $\mathcal{H}$, $\|h-h'\|_{\infty,n} \le \varepsilon/2$, showing that S is indeed an ε-net of $\Omega_n(\sigma)$. Note that the cardinality of S can be bounded by

$$|S| \le |\hat{\mathcal{H}}| \max_{h'\in\hat{\mathcal{H}}} |\hat{\mathcal{G}}_{h'}|.$$

Hence,

$$H(\varepsilon, \Omega_n(\sigma), \|\cdot\|_n) \le H(\tfrac{\varepsilon}{2K_g}, \mathcal{G}_{h_0}(2\sigma+\varepsilon/2), \|\cdot\|_n) + H(\tfrac{\varepsilon}{2K_gK_h}, \mathcal{H}, \|\cdot\|_{\infty,n}),$$

for an arbitrary $h_0 \in \mathcal{H}$, where we used that $\|\cdot\|_n$ is translation invariant.

Combining this with (31), we get

$$\begin{aligned} H(\varepsilon, T(\sigma), \|\cdot\|_T) &\le H(\tfrac{\varepsilon}{2}, \Omega_n(\sigma), \|\cdot\|_n) + H(\tfrac{\varepsilon}{2}, \mathcal{H}, \|\cdot\|_{\infty,n}) \\ &\le H(\tfrac{\varepsilon}{4K_g}, \mathcal{G}_{h_0}(2\sigma+\varepsilon/2), \|\cdot\|_n) + H(\tfrac{\varepsilon}{4K_gK_h}, \mathcal{H}, \|\cdot\|_{\infty,n}) + H(\tfrac{\varepsilon}{2}, \mathcal{H}, \|\cdot\|_{\infty,n}) \\ &\le H(\tfrac{\varepsilon}{4K_g}, \mathcal{G}_{h_0}(2\sigma+\varepsilon/2), \|\cdot\|_n) + AH(\tfrac{\varepsilon}{A}, \mathcal{H}, \|\cdot\|_{\infty,n}) \end{aligned} \tag{35}$$

for A large enough.

By Corollary 2.6 in the book of van de Geer [2000], $H(\varepsilon, \mathcal{G}_{h_0}(\sigma), \|\cdot\|_n) \le \rho\log(\frac{4\sigma+\varepsilon}{\varepsilon})$ a.s.. Hence,

$$H(\tfrac{\varepsilon}{4K_g}, \mathcal{G}_{h_0}(2\sigma+\varepsilon/2), \|\cdot\|_n) \le \rho\log\left(\frac{32\sigma K_g + 8K_g\varepsilon + \varepsilon}{\varepsilon}\right) \le \rho\log(241) + \rho\log(\sigma/\varepsilon), \ a.s..$$

Here the second inequality follows from bounding the ε in numerator by σ (since $\sigma \ge \varepsilon$), and $K_g < 6$. Combining this with (35) finishes the proof of the first statement.

To prove the second part, note that $\int_0^1 (-\log(x))^{1/2}\,dx = \sqrt{\pi}/2$. Thus, with $\mathcal{I}(\sigma,\mathcal{H}) = \int_0^1 H^{1/2}(u\sigma, \mathcal{H}, \|\cdot\|_{\infty,n})\,du$,

$$\begin{aligned} \int_0^1 H^{1/2}(u\sigma, T(\sigma), \|\cdot\|_T)\,du &\le \rho^{\frac{1}{2}} \int_0^1 \log^{\frac{1}{2}}(\tfrac{1}{u})\,du + \rho^{\frac{1}{2}}\log^{\frac{1}{2}}(241) + A^{\frac{1}{2}}\mathcal{I}(\tfrac{\sigma}{A}, \mathcal{H}) \\ &\le \rho^{\frac{1}{2}}\sqrt{\pi}/2 + \rho^{\frac{1}{2}}\log^{\frac{1}{2}}(241) + A^{\frac{1}{2}}\mathcal{I}(\tfrac{\sigma}{A}, \mathcal{H}). \end{aligned}$$

Now, note that

$$\int_0^1 H^{1/2}(u\sigma, \mathcal{H}, \|\cdot\|_{\infty,n})\,du = \frac{1}{\sigma}\int_0^\sigma H^{1/2}(v, \mathcal{H}, \|\cdot\|_{\infty,n})\,dv \le \frac{1}{\sigma}\int_0^1 H^{1/2}(v, \mathcal{H}, \|\cdot\|_{\infty,n})\,dv \le C_H/\sigma,$$

where the first inequality follows since $\sigma \le 1$, while the last inequality follows by Assumption 3.3. The desired result follows by choosing $A' = \sqrt{\pi}/2 + \log^{\frac{1}{2}}(241)$ and $A'' = A^{3/2}C_H$. □