Generalization Bounds for Partially Linear Models

Ruitong Huang and **Csaba Szepesvári**

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. Under natural regularity conditions, we bound the generalization error as a function of the Rademacher complexity of $\mathcal{H}$ and that 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.

Introduction

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

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

where X_i is an input, Y_i is an observed 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.[1]

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.

Problem Setting and Notation

Throughout the paper, the input space $\mathcal{X}$ will be a metric space, and $\mathcal{Y}$, the label space, will be a subset of the reals $\mathbb{R}$. In particular, we will assume that $\mathcal{Y} \subset [-\Lambda, \Lambda]$. 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 penalized empirical risk minimization problem with the partially linear stochastic model (1) is to find a minimizer of

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

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. We assume that $0 \in \mathcal{H}$. 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

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

($L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$). Typically, $\mathcal{H}$ arises as the set $\{h : \mathcal{X} \to \mathbb{R} : J(h) \le K\}$ with some $K > 0$ and some penalty functional J that penalizes of the "roughness" of the functions, hence we call this problem the constrained empirical risk-minimization problem over $\mathcal{G} + \mathcal{H} \doteq \{g + h : g \in \mathcal{G}, h \in \mathcal{H}\}$.[2]

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[3] of (2) and let $f_n = g_n + h_n$.

When analyzing a learning procedure, we compare the risk $L(f_n)$ of the predictor f_n it returns to the best risk possible, i.e., to $L^* = \min_{g \in \mathcal{G}, h \in \mathcal{H}} L(g + h)$. 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 Rademacher complexity of $\mathcal{H}$ and an appropriate subset of $\mathcal{G}$. Our main result, Theorem 1, 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 increased estimation error is modest, which, in favourable situations, can be far outweighed by the decrease of L^* that can be attributed to including the parametric part.

As usual, we start with the decomposition of the excess risk

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

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. Indeed, for such losses if f_n can take on large values then $L(f_n)$ could be rather large. If the problem was purely nonparametric, $f \in \mathcal{H}$ and an assumption that essentially requires the uniform law of large

[1] This suggests that to avoid the difficulty, 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.

[2] 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 $\min_{g \in \mathcal{G}, h} L_n(g + h) + \lambda J(h) = \min_{K \ge 0} \lambda K + \min_{g \in \mathcal{G}, h : J(h) \le K} L_n(g+h)$. The reader interested in this relationship is advised to check out the paper of Blanchard et al. [2008], who explores this relationship in a specific context.

[3] For simplicity, we assume that this minimizer and in fact all the others that we will need later exist.

numbers holds over $\mathcal{H}$ would imply that this will not happen. However, in our case $f_n = g_n + h_n$ and while $h_n \in \mathcal{H}$ is well controlled, no uniform law holds over $\mathcal{G}$, the set that g_n belongs to. Hence, the bulk of the work will consist of showing that g_n is well-controlled.

Before introducing our assumptions, let us introduce some more notation. We will denote the Minkowski-sum of $\mathcal{G}$ and $\mathcal{H}$ by $\mathcal{F}$: $\mathcal{F} = \mathcal{G} + \mathcal{H}$. 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 $X_{1:n} = (X_1, \dots, X_n)$, the n-norm of a function is defined as the 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 $(f(X_1), \dots, f(X_n))^\top$ is denoted by $f(X_{1:n})$. The matrix $(\phi(X_1), \dots, \phi(X_n))^\top \in \mathbb{R}^{n\times d}$ 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 $\rho = \operatorname{rank}(G)$. Lastly, let $L_{h,n}(g) = L_n(h+g)$, $\overline{L}_n(f) = \mathbb{E}\left[L_n(f)\,|\,X_{1:n}\right]$ and $\overline{L}_{h,n}(g) = \mathbb{E}\left[L_n(h+g)\,|\,X_{1:n}\right]$.

Assumptions and Result

In this section we state our assumptions, which will be followed by stating our main result. We will also sketch the steps of the proof, leaving the details for the next section.

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. Again, if ϕ is continuous and $\mathcal{X}$ is compact, this assumption will also be automatically satisfied. In fact, by rescaling the basis functions if needed, WLOG we will assume that $R = 1$. We will also assume that $0 \in \mathcal{H}$ (i.e., the identically zero function is an element of $\mathcal{H}$).

To recap, let $(g_n, h_n) = \arg\min_{g\in\mathcal{G}, h\in\mathcal{H}} L_n(f)$ and $f_n = g_n + h_n$. We assume that the minimizers exist, but this is done only for the sake of convenience. Further, at this stage the uniqueness of the minimizers is unimportant.

Let us start with our assumptions on the loss function, ℓ.

Assumption 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) Lipschitzness: The loss function ℓ is Lipschitz with respect to both of its arguments over $\mathcal{Y}\times[-c,c]$ for any $c > 0$. In particular, we will denote the Lipschitz coefficient of ℓ over $\mathcal{Y}\times[-c,c]$ by $K_\ell(c)$: for any $y, y_1, y_2 \in \mathcal{Y}$ and $y', y_1', y_2' \in [-c,c]$, $|\ell(y,y_1') - \ell(y,y_2')| \le K_\ell(c)|y_1' - y_2'|$, and $|\ell(y_1,y') - \ell(y_2,y')| \le K_\ell(c)|y_1 - y_2|$.

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

Remark 1. The convexity assumption is standard, while the Lipschitz assumption is unrestrictive due to the boundedness of the ranges involved.

Remark 2. Unlike the first two assumptions, Assumption 1(iii), 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. One way of weakening this assumption would be to assume that for any distribution $P_{(X,Y)}$ of (X,Y) 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 1 (Quadratic Loss). In the case of quadratic loss, i.e., when $\ell(y,y') = \frac{1}{2}(y-y')^2$, $R_c^2 \le 2c + 2\Lambda^2$: Indeed, this follows from $\|f\|_n^2 \le \frac{1}{n}\sum_i 2(\mathbb{E}\left[(f(X_i) - Y_i)^2\,|X_{1:n}\right] + \mathbb{E}\left[Y_i^2\,|X_{1:n}\right]) \le 2\mathbb{E}\left[L_n(f)\,|X_{1:n}\right] + 2\mathbb{E}\left[Y_i^2\,|X_{1:n}\right]$ and our assumption that $|Y| \le \Lambda$.

Example 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 $R_c^2 \le c/\varepsilon$.

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 the previous section). If there are multiple minimizers, choose one.

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 $x \in \mathcal{X}, y \in \mathcal{Y}$, $f : \mathcal{X} \to \mathbb{R}$. The next assumption states that the loss function is locally "not flat":

Assumption 2 (Non-flat Loss). *Assume that there exists $\varepsilon > 0$ such that for any $h \in \mathcal{H}$ and vector $a \in [-\varepsilon, \varepsilon]^n \cap \mathrm{Im}(\Phi)$,*

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

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

Remark 4. It is easy to see that if for all $y \in \mathcal{Y}$, $y' \in \mathbb{R}$, $a \in [-\varepsilon, \varepsilon]$, $\ell(y, y' + a) - \ell(y, y') \ge \varepsilon a^2$ holds then Assumption 2 will also hold. This condition, although stronger than Assumption 2, may be easier to verify in some cases, e.g. L_p loss for $p \ge 3$.

Example 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 the minimizer of $\overline{L}_{h,n}(\cdot)$. Then $\overline{\theta}_{h,n} = \left(\Phi^\top\Phi\right)^+\Phi^\top(\mathbb{E}\left[Y_{1:n}|X_{1:n}\right] - h(X_{1:n}))$. A simple calculation (where we exploit that $a \in \mathrm{Im}(\Phi)$) shows that the assumption holds with equality and with $\varepsilon = 1$.

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

Definition 1. *For $\varepsilon > 0$, the ε-covering number $N(\varepsilon, H, d)$ of a set H equipped with a pseudo-metric d is the number of balls with radius ε measured with respect to d necessary to cover H. The ε-entropy of H is $H(\varepsilon, H, d) = \log N(\varepsilon, H, d)$.*

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 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 $\|f\|_n \le \|f\|_{\infty,n} \le \|f\|_\infty$. For the next two assumptions we introduce $G_{\lambda_{\min}}$ as the event when $\hat{\lambda}_{\min} \ge \lambda_{\min}/2$. We use $\|\cdot\|_{\infty,n}$ in our integrability assumption:

Assumption 3 (Integrable Entropy Numbers of $\mathcal{H}$). *There exists a (non-random) constant C_H such that, $\int_0^1 H^{1/2}(v, \mathcal{H}, \|\cdot\|_{\infty,n})\, dv < C_H$ holds a.s. on $G_{\lambda_{\min}}$.*

Remark 5. Assumption 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].

Assumption 4 (Lipschitzness of the Parametric Solution Path). *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 6. When $\overline{g}_{h,n}$ is uniquely defined, Assumption 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 4 (Quadratic loss). In the case of the quadratic loss, by Example 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 $G_{\lambda_{\min}}$, a simple calculation shows that

$$|\overline{g}_{h,n}(x) - \overline{g}_{h',n}(x)| \le \frac{2}{\lambda_{\min}}\|h' - h\|_{\infty,n}$$

where we used that $[P_X]$ a.e. $\|\phi(x)\|_2 \le 1$.

Results

Now, we are at the stage to state our main result. As suggested beforehand, we will state this result in terms of the Rademacher complexity of the function spaces of interest:

Definition 2. *Given a random sample $X_{1:n} = (X_1, ..., X_n)$ and a set $\mathcal{F}$ of real-valued functions with a common bounded range, we will denote the Rademacher complexity of $\mathcal{F}$ by*

$$R_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\sum_{i=1}^n \sigma_i f(X_i)\right],$$

where $\sigma_1, \ldots, \sigma_n$ are Rademacher random variables which take value from $\{-1, +1\}$ with equal probability and are independent of each other and $X_{1:n}$.

With this, we are ready to state our main result:

Theorem 1. *Let Assumptions 1 to 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_0, L_0, and $U \ge \|g^*\|_\infty$ such that for any $0 < \delta < 1$ and*

$$n \ge \max\left(16L_0^4, \frac{2}{\lambda_{\min}\log\left(\frac{\varepsilon}{2}\right)}\log\left(\frac{4\rho}{\delta}\right), \frac{4\log\left(\frac{4}{\delta}\right)}{C_0}\right),$$

with probability at least $1 - \delta$,

$$L(f_n) - L(f^*) \le 2\hat{K}_\ell\left\{R_n(\mathcal{H}) + R_n(\mathcal{G}(U))\right\} + \hat{K}_\ell(\Lambda + r + U)\sqrt{\frac{2\ln(4/\delta)}{n}}, \quad (4)$$

where $f_n = h_n + g_n$ is the minimizer of $L_n(\cdot)$ over $\mathcal{H} + \mathcal{G}$, $R_n(\mathcal{F})$ denotes the Rademacher complexity of $\mathcal{F}$ and $\hat{K}_\ell = K_\ell(r+U)$.

Remark 7. The actual value of U can be read out from the proof of Theorem 3 below (the value is inversely proportional to $\lambda_{\min}$ and depends on $(R_c)_c$ from the level-set assumption, the range of losses and r). The dependence on ρ and $\lambda_{\min}$ also show in the lower bound constraint of n. Further, the entropy integrability constraint for $\mathcal{H}$ controls the size of L_0 and C_0.

Remark 8. Kakade et al. [2009] gives bounds on the Rademacher complexity of various class of linear functions, which can be useful to bound $R_n(\mathcal{G}(U))$, while Bartlett and Mendelson [2002] provides several examples for bounding the Rademacher complexity for various classes of functions. With the normalization used here, the Rademacher complexity for linear function class will typically be of order $O(1/\sqrt{n})$. For further examples and connection to other measures of complexity, see Tewari and Bartlett [2013].

As explained earlier, the main difficulty in proving this result is that g_n is not penalized and hence showing that it is bounded requires substantial effort. As such, and also because we believe that the behavior of g_n may be of indepedent interest, we state this as a separate theorem. Once we know that g_n is bounded (with high probability), the proof of Theorem 1 is standard (this argument is presented in the next section). Before stating the result on the boundedness of g_n, we state an easy corollary that bounds the expected excess risk of the truncated predictor f_n^c defined by $f_n^c(x) = \max(\min(f(x), \Lambda), -\Lambda)$. Note that the truncation cannot increase the loss of f_n.

Corollary 2. *Let* $M(n) = \max\left(16L_0^4, \frac{2}{\lambda_{\min}\log\left(\frac{e}{2}\right)}\log(4n\rho), \frac{4\log(4n)}{C_0}\right)$ *and consider any* $n \geq M(n)$. *Then,*

$$\mathbb{E}\left[L(f_n^c) - L(f^*)\right] \leq 2\hat{K}_\ell\left\{R_n(\mathcal{H}) + R_n(\mathcal{G}(U))\right\} + \hat{K}_\ell(\Lambda + r + U)\left\{\sqrt{\frac{2\ln(4n)}{n}} + \frac{2}{n}\right\}.$$

Proof. It is not hard to see that the range of the loss $\ell(y, f(x))$ when $|f(x)| \leq U + r$ is $W = 2\hat{K}_\ell(\Lambda + r + U)$. Denote by E the event when (4) holds and let $B_n(n,\delta)$ be the bound on the right-hand side of (4). Take $\delta = 1/n$. Then, for $n \geq M(n)$, $\mathbb{E}\left[L(f_n^c) - L(f^*)\right] = \mathbb{E}\left[\mathbb{I}_E(L(f_n^c) - L(f^*))\right] + W\mathbb{P}(E^c) \leq B_n(n, 1/n) + W/n$, and the result follows by plugging in the definitions of B_n and W. □

The result guaranteeing that g_n is bounded with high probability is as follows:

Theorem 3. *Let Assumptions 1 to 4 hold. Then, there exist constants C_0, L_0, U such that for any $0 < \delta < 1$, and n such that $n \geq \frac{2}{\lambda_{\min}\log\left(\frac{e}{2}\right)}\log\left(\frac{2\rho}{\delta}\right)$, $n \geq 16L_0^4$ and $n \geq \frac{4\log\left(\frac{2}{\delta}\right)}{C_0}$, it holds that*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\|g_{h,n}\|_\infty \geq U\right) \leq \delta\,. \tag{5}$$

The rest of the paper is devoted to the proof of these two results. In the next section we show how Theorem 3 gives rise to Theorem 1. This is followed by the proof of Theorem 3.

The Proof of Theorem 1

In this section we assume that Theorem 3 holds true and based on this we prove Theorem 1. We start from the excess risk bound (3). Thus, our goal is to bound $L(f_n) - L_n(f_n)$ and $L(f^*) - L_n(f^*)$.

Let U be as in Theorem 3 and let E denote the event when $\sup_{h\in\mathcal{H}}\|g_{h,n}\|_\infty \leq U$. Define $\mathcal{G}(U) = \{g \in \mathcal{G} : \|g\|_\infty \leq U\}$ and $\mathcal{C} = \mathcal{H} + \mathcal{G}(U)$. On E, $f_n \in \mathcal{C}$, 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, $\|g_n\|_\infty \leq \sup_{h\in\mathcal{H}}\|g_{h,n}\|_\infty \leq U$ and so on E,

$$L(f_n) - L_n(f_n) \leq \sup_{f\in\mathcal{C}}|L(f) - L_n(f)| =: \Delta_n(\mathcal{C})\,.$$

Furthermore, by increasing U if necessary, we can always arrange for that $f^* = h^* + g^* \in \mathcal{C}$ (for this we may need to increase U so that $\|g^*\|_\infty \leq U$). Hence, by (3),

$$L(f_n) - L(f^*) \leq 2\Delta_n(\mathcal{C}) \text{ a.s. on } E.$$

and thus for any $z > 0$,

$$\begin{aligned}\mathbb{P}\left(L(f_n) - L(f^*) > z\right) &= \mathbb{P}\left(L(f_n) - L(f^*) > z, E^c\right) + \mathbb{P}\left(L(f_n) - L(f^*) > z, E\right)\\ &\leq \mathbb{P}(E^c) + \mathbb{P}\left(2\Delta_n(\mathcal{C}) > z, E\right)\\ &\leq \mathbb{P}(E^c) + \mathbb{P}\left(2\Delta_n(\mathcal{C}) > z\right)\,.\end{aligned} \tag{6}$$

Thus, it remains to bound $\Delta_n(\mathcal{C})$.

This is done by means of using two standard results. The first result bounds $\Delta_n(\mathcal{C})$ in terms of the Rademacher complexity of $\mathcal{C}$.

Theorem 4 (Tewari and Bartlett 2013, Section 3.2)**.** *Fix any n and let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample of size n over $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} \subset \mathbb{R}$. Let $\ell : \mathcal{Y} \times [-a, a] \to [\beta, \beta + b]$ be a loss that is c-Lipschitz in its second argument and let $\mathcal{C} \subset [-a, a]^{\mathcal{X}}$. Then, for any positive integer n and $0 < \delta < 1$, with probability $1 - \delta$,*

$$\sup_{f\in\mathcal{C}}|L(f) - L_n(f)| \leq 2cR_n(\mathcal{C}) + b\sqrt{\frac{\ln(2/\delta)}{2n}}\,,$$

where $L(f) = \mathbb{E}\left[\ell(Y, f(X))\right]$ and $L_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i))$.[4]

[4] Bartlett and Mendelson [2002] gives essentially this result, with slightly worse constants.

Applying this result to our setting gives

$$\Delta_n(\mathcal{C}) \le 2\hat{K}_\ell R_n(\mathcal{C}) + \hat{K}_\ell(\Lambda + r + U)\sqrt{\frac{2\ln(2/\delta)}{n}},$$

where $\hat{K}_\ell = K_\ell(r+U)$ with $K_\ell(\cdot)$ being the Lipchitz coefficient in Assumption 1 (ii). It remains to bound $R_n(\mathcal{C})$. By Part 7 of Theorem 12 of Bartlett and Mendelson [2002], $R_n(\mathcal{C}) \le R_n(\mathcal{H}) + R_n(\mathcal{G}(U))$.

Combining this with (6) and using

$$z_\delta = 2\hat{K}_\ell\{R_n(\mathcal{H}) + R_n(\mathcal{G}(U))\} + \hat{K}_\ell(\Lambda + r + U)\sqrt{\frac{2\ln(4/\delta)}{n}}$$

gives $\mathbb{P}(L(f_n) - L(f^*) > z_\delta) \le \mathbb{P}(E^c) + \frac{\delta}{2}$. Finally, by Theorem 3, $\mathbb{P}(E^c) \le \delta/2$ provided that $n \ge \frac{2}{\lambda_{\min}\log\left(\frac{e}{2}\right)}\log\left(\frac{4\rho}{\delta}\right)$, $n \ge 16L_0^4$ and $n \ge \frac{4\log\left(\frac{4}{\delta}\right)}{C_0}$, thus finishing the proof.

The Proof of Theorem 3

In this section we present the proof of Theorem 3, 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

$$g_{h,n}^2(x) \le \|\theta_{h,n}\|_2^2 \le \frac{\theta_{h,n}^\top \hat{G}\,\theta_{h,n}}{\hat{\lambda}_{\min}} = \frac{2\,\|g_{h,n}\|_n}{\lambda_{\min}}. \qquad (7)$$

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 5. *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

$$n \ge \frac{2}{\lambda_{\min}\log\left(\frac{e}{2}\right)}\log\left(\frac{\rho}{\delta}\right), \qquad (8)$$

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 (8), 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 omitted.

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

$$\|g_{h,n}\|_n \le \|g_{h,n} - \overline{g}_{h,n}\|_n + \|\overline{g}_{h,n}\|_n \qquad (9)$$

and develop uniform bound on the two terms on the right-hand side.

Lemma 6. *We have $\sup_{h\in\mathcal{H}} \|\overline{g}_{h,n}\|_n \le \bar{R}$, where $\bar{R} = R_{C_0} + r$ and $C_0 = \ell(0,0) + K_\ell(0)\,\Lambda + K_\ell(r)\,r$.*

The constant R_{C_0} that appears in the statement is defined in our "level-set assumption" (cf. Assumption 1(iii)).

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

By Assumption 1(iii), 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 $\|h + \overline{g}_{h,n}\|_n \le R_c$. 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 $\ell(Y_i, h(X_i))$. For this, we have $\ell(Y_i, h(X_i)) \le \ell(0,0) + K_\ell(0)\,\Lambda + K_\ell(r)r$, where we used that ℓ is $K_\ell(c)$-Lipschitz on $[-c,c]\times\mathcal{Y}$, $\mathcal{Y} = [-\Lambda, \Lambda]$, $Y_i \in \mathcal{Y}$, and $|h(X_i)| \le r$. Putting together the inequalities, we obtain that $\overline{L}_n(h + \overline{g}_{h,n}) \le \ell(0,0) + K_\ell(0)\,\Lambda + K_\ell(r)r \doteq c$ and thus $\|h + \overline{g}_{h,n}\|_n \le R_c$. □

Let us now consider bounding $\|g_{h,n} - \overline{g}_{h,n}\|_n$. 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 $\|g_{h,n} - \overline{g}_{h,n}\|_n$ on this event only. Note that by Lemma 5 (ii), $G_{\lambda_{\min}}$ holds with probability at least $1-\delta$.

Lemma 7. *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*

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \|g_{h,n} - \overline{g}_{h,n}\|_n \ge 1, G_{\lambda_{\min}}\Big) \le \exp\left(-\frac{C_0 n}{4}\right). \qquad (10)$$

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

Now, combining (7), (9) and Lemma 6 we get that on $G_{\lambda_{\min}}$,

$$\begin{aligned} G_{n,\infty} \doteq \sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty &\le \frac{2}{\lambda_{\min}} \sup_{h\in\mathcal{H}} \|g_{h,n}\|_n \\ &\le \frac{2}{\lambda_{\min}}\left(\overline{R} + \sup_{h\in\mathcal{H}} \|g_{h,n} - \overline{g}_{h,n}\|_n\right). \end{aligned} \qquad (11)$$

Since for any $A > 0$,

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

and by (11), $\mathbb{P}\left(G_{n,\infty} > A, G_{\lambda_{\min}}\right) \le \mathbb{P}\left(\frac{2}{\lambda_{\min}}\left(\overline{R} + \sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n\right) > A, G_{\lambda_{\min}}\right)$, choosing $A = \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)$, we see that $\mathbb{P}\left(G_{n,\infty} > \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)\right) \le \mathbb{P}\left(G^c_{\lambda_{\min}}\right) + \mathbb{P}\left(\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \ge 1, G_{\lambda_{\min}}\right)$. By Eq. (8) and Lemma 7, provided that $n \ge \frac{2}{\lambda_{\min}\log\left(\frac{\varepsilon}{2}\right)}\log\left(\frac{2\rho}{\delta}\right)$, $n \ge 16L_0^4$ and $n \ge \frac{4\log\left(\frac{2}{\delta}\right)}{C_0}$ we get that $\mathbb{P}\left(G_{n,\infty} > \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)\right) \le \delta$, which is the desired statement. In particular, we can choose $U = \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)$.

The Proof of Lemma 7

The proof follows the ideas from the paper of van de Geer [1990]. Lemma 7 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

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

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

$$\left\|g - \overline{g}_{h,n}\right\|_n = \frac{\|\omega_{g,h}\|_n}{1 - K\|\omega_{g,h}\|_n}. \tag{12}$$

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

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

Lemma 8 ("Basic Inequality"). *Let Assumption 2 hold. There exists a constant η, such that on the event $G_{\lambda_{\min}}$, for any $h \in \mathcal{H}$,*

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

The proof, which is omitted, 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 $\mathcal{V}_{h,n}(\omega) = \sqrt{n}\left(\Delta_{h,n}(\overline{g}_{h,n}) - \Delta_{h,n}(\overline{g}_{h,n} + \omega)\right)$ normalized by ω. In particular, it follows from Lemma 8 that for $L, \sigma > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \|\hat{\omega}_{h,n}\|_n \ge L\sigma\,, G_{\lambda_{\min}}\Big) \\
&\quad= \mathbb{P}\Big(\exists h \in \mathcal{H} \,:\, \|\hat{\omega}_{h,n}\|_n \ge L\sigma,\, \frac{\mathcal{V}_{h,n}(\hat{\omega}_{h,n})}{\|\hat{\omega}_{h,n}\|_n^2} \ge \eta\sqrt{n}\,, G_{\lambda_{\min}}\Big) \\
&\quad\le \mathbb{P}\left(\sup_{(g,h)\in\mathcal{G}\times\mathcal{H}:\|\omega_{g,h}\|_n \ge L\sigma} \frac{\mathcal{V}_{h,n}(\omega_{g,h})}{\|\omega_{g,h}\|_n^2} \ge \eta\sqrt{n}\,, G_{\lambda_{\min}}\right).
\end{aligned} \tag{13}$$

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 $(k = 1, \ldots, n)$, 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 $U_1, U_2, \ldots, U_n$ 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"

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

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 9. *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

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

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

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

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

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

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

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

Remark 9. 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 9 to our problem, we choose the spaces to be $V = \mathcal{H}$, $\Lambda = \cup_{h \in \mathcal{H}} \Lambda_h$, were $\Lambda_h = \Omega_{h,n}$. 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 9 we need to "condition" on $X_{1:n}$ when using this lemma.

To make our argument formal, let $(W, \mathcal{W}, \mathbb{P})$ be the probability space that holds our random variables. 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 are real-valued). Now, let $(\mathbb{P}_{x_{1:n}})_{x_{1:n} \in \mathcal{X}^n}$ 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] We will use Lemma 9 with the probability spaces $(W, \mathcal{W}, \mathbb{P}_{x_{1:n}})$ for $x_{1:n} \in \mathcal{X}^n$ fixed, the expectation operator corresponding to $\mathbb{P}_{x_{1:n}}$ will be denoted by $\mathbb{E}_{x_{1:n}}$, or $\mathbb{E}[\cdot | X_{1:n} = x_{1:n}]$, where the latter notation is justified by the definition of $\mathbb{P}_{x_{1:n}}$.

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

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

[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 $P_X(\{x_{1:n} : \mathbb{P}_{ux} \neq \hat{\mathbb{P}}_{x_{1:n}}\}) = 0$. For background on disintegration and conditioning, the reader is referred to Chang and Pollard [1997].

(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 $d_\Lambda(\omega, \tilde{\omega}) = \|\omega - \tilde{\omega}\|_n$)

$$\sup_{h \in \mathcal{H}} \sup_{\substack{\omega \in \Lambda_h: \\ d_\Lambda(\lambda_h, \omega) > L\sigma}} \frac{Z_n(h, \omega)}{d_\Lambda^2(\lambda_h, \omega)} = \sup_{h \in \mathcal{H}} \sup_{\substack{\omega \in \Omega_{h,n}: \\ \|\omega\|_n > L\sigma}} \frac{\mathcal{V}_{h,n}(\omega)}{\eta \|\omega\|_n^2} =: Q_n(L\sigma), \tag{14}$$

showing that the conclusion of the lemma suffices to bound the quantity of interest appearing in (13). One can then show that the conditions of Lemma 9 are satisfied for $[P_X]$ almost every $x_{1:n} \in \mathcal{X}^n$ such that $\lambda_{\min}(x_{1:n}) \doteq \lambda_{\min}(\Phi(x_{1:n})^\top \Phi(x_{1:n})) \geq \lambda_{\min}/2$ (details are omitted due to the lack of space).

Further, one can show that $\alpha_n(L, \sigma) \leq \frac{2C'}{\sqrt{n} L \sigma^2} \leq \frac{2C'}{L}$ provided that $\sqrt{n}\sigma^2 \geq 1$, where C' is a constant that is independent of $x_{1:n}, L, n, K, \sigma$ and we assumed that $\sigma \leq 1$. Therefore, L_0 and C_0 can be selected independently of $x_{1:n}$, K, n and σ.

Therefore, using (14) 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 $\lambda_{\min}(x_{1:n}) \geq \lambda_{\min}/2$, $\mathbb{P}_{x_{1:n}}\Big(Q_n(L\sigma) \geq \sqrt{n}\Big) \leq \exp(-C_0 L^2 \sigma^2 n)$. Now, by the definition of $\mathbb{P}_{x_{1:n}}$,

$$\mathbb{P}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) \exp(-C_0 L^2 \sigma^2 n).$$

Hence, by combining (12) and (13), using the definition of $Q_n(L\sigma)$ in (14) and choosing $L = L_0$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left\| g_{h,n} - \overline{g}_{h,n} \right\|_n \geq \frac{L_0 \sigma}{1 - K L_0 \sigma}, G_{\lambda_{\min}} \right)$$
$$\leq \mathbb{P}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) \leq \exp(-C_0 L_0^2 \sigma^2 n).$$

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

Conclusions and Future Work

While the present paper makes the first steps in analyzing the excess risk of semiparametric models, much work remains to be done: The current excess risk is slower than what is expected when using the squared (or similarly smooth) loss. By using the boundedness result with more advanced techniques (that exploit the curvature of losses), one should be able to prove faster rates. However, perhaps more interesting is to consider other losses, like the hinge loss, to which the current results cannot be applied. In particular, the hinge loss lacking any curvature seems to call for entirely new ideas.

References

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

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.

Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: Error analysis. *J. Mach. Learn. Res.*, 5: 1143–1175, December 2004.

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

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.

Alex Gittens and Joel 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.

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

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

S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS 21*, pages 793–800. MIT Press, 2009. URL http://books.nips.cc/nips21.html.

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.

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.

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

Alex J. Smola, Thilo T. Frieß, and Bernhard Schölkopf. Semiparametric support vector and linear programming machines, 1998.

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

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

James 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.