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

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

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

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

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 $Z_i = (X_i, Y_i)$ $(i = 1, \dots, n)$. 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:

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

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.

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 $X_{1:n} = (X_1, \dots, X_n)$, 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 $(f(X_1), \dots, f(X_n))^\top$ 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 $\overline{L}_{h,n}(g) = \mathbb{E}\left[L_n(h+g) \,|\, X_{1:n}\right]$.

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-*

[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.

rameters (β, Γ) *if*

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

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

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

(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 $c_1, c_2 \in [-c, c]$,

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

and such that $\mathbb{E}\left[\exp(|\beta K_\ell(Y, c)|^2)|\, X\right] \le \Gamma_c < \infty$ 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 $c \ge 0$, $R_c = \sup_{f \in \mathcal{F} : \mathbb{E}[L_n(f)|X_{1:n}] \le c} \|f\|_n$ is finite and independent of n.

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 $\mathbb{E}\left[\exp(|\beta K_\ell(Y, c)|^2)\right] = \mathbb{E}\left[\exp(\frac{1}{2}(|Y| + c)^2)\right] \le \mathbb{E}\left[\exp(|Y|^2)\right] + \exp(c^2)$.

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 $\ell(y, y') = \frac{1}{2}(y - y')^2$, $R_c^2 \le 4c + 8Q + 4s^2$ where $s = \|\hat{h}\|_\infty$. Indeed, this follows from

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

Then $\mathbb{E}\left[Y_i^2\,|X_i\right] \le 2\mathbb{E}\left[(Y_i - \hat{h}(X_i))^2\,|X_i\right] + 2\hat{h}^2(X_i) \le 4Q + 2s^2$. 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 $R_c^2 \le c/\varepsilon$.

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 $a \in [-\varepsilon, \varepsilon]^n \cap \mathrm{Im}(\Phi)$,*

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

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,

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

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

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

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

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*

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

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 $(X_1,Y_1),\ldots,(X_n,Y_n)$ 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) $\mathbb{E}\left[\hat\lambda_{\min}^{-1}(n)\right] = +\infty$.

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.