Estimation of Extreme Values and Associated Level Sets of a Regression Function via Selective Sampling

Stanislav Minsker SMINSKER@MATH.DUKE.EDU
Department of Mathematics, Duke University

Abstract

We propose a new method for estimating the locations and the value of an absolute maximum (minimum) of a function from the observations contaminated by random noise. Our goal is to solve the problem under minimal regularity and shape constraints. In particular, we do not assume differentiability of a function nor that its maximum is attained at a single point. We provide tight upper and lower bounds for the performance of proposed estimators. Our method is adaptive with respect to the unknown parameters of the problem over a large class of underlying distributions.

Keywords: Regression, optimization, level sets, selective sampling, active learning, multi-armed bandits.

1. Introduction

Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ be a random couple with unknown distribution P and let Π stand for the marginal distribution of X. In what follows, we will assume that the conditional expectation $\eta(x) := \mathbb{E}(Y|X = x)$ is well-defined. The main goal of this work is to investigate the problem of finding location(s) and the value $M(\eta) := \sup\{\eta(x),\ x \in \operatorname{supp}(\Pi)\}$ of the maximum of a regression function $\eta(x)$. In practice, the only source of information about η is the collection of noisy measurements $(X_i, Y_i),\ 1 \le i \le n$ sampled from P (so that $Y_i = \eta(X_i) + \xi_i$). In recent years, there has been a lot of interest in estimation techniques that can benefit from *adaptive design* (e.g., so-called *active learning* framework), as opposed to the algorithms that are designed to work with the iid samples. This interest is partially motivated by an observation that in some cases, the cost related to the process of collecting the data is associated with response variable Y while design points X are freely available.

Some advantages of adaptive design relative to our problem were understood long ago (Kiefer and Wolfowitz (1952), Blum (1954)). The majority of recent results on optimality of estimators were obtained under assumptions that the regression function is smooth (at least twice continuously differentiable) and that the maximum is attained at a unique point (see e.g. Polyak and Tsybakov (1990), Dippon (2003) and references therein); all these methods are based on stochastic optimization techniques going back to Robbins and Monro (1951) and the aforementioned assumptions are crucial to guarantee the success of estimation. Other references on the topic include Tang et al. (2011) and a recent paper by Belitser et al. (2012). While the two-stage estimation technique proposed in this work is different, it still requires same strong assumptions. Our interest in the problem is motivated by the fact that there is often no reason to believe either in uniqueness of the maximum or in smoothness of the regression function. We do not assume that the maximum is attained at a single point and will propose a method that constructs a set-valued estimator of the level set $L_M := \{x \in \mathrm{supp}(\Pi) : \ \eta(x) = M(\eta)\}$ and attains nearly optimal rates with respect to the Hausdorff distance (over certain classes). At the same time, proposed technique allows to estimate the value of $M(\eta)$. Moreover, our method is adaptive with respect to the underlying structure of the problem (such as smoothness).

It is necessary to mention that our topic of research is closely related to the "continuum-armed bandits" problem, see Kleinberg et al. (2008), Bubeck et al. (2011) and numerous references therein. In the latter framework, the regression function η is usually called the payoff function. While the case of Lipschitz-continuous payoff functions in Euclidean and general metric spaces is understood well, the questions related to *adaptation* to the unknown smoothness (or metric) are not answered completely, to the best of our knowledge. One of the main goals of this work is to make a step towards the solution of this problem by relating in to the *adaptive confidence bands* for η.

Our algorithm is partially motivated by the *active learning* literature in the framework of binary classification (in particular, Minsker (2012b)), which is also a level-set estimation problem by nature, so there is no wonder that similar ideas allow for improved results in our context. At the same time, the technique is similar in spirit to the *zooming algorithm* of Kleinberg et al. (2008) and the *hierarchical optimistic optimization* strategy of Bubeck et al. (2011).

The paper is organized as follows: next section formally introduces necessary notations, assumptions and briefly explains the main ideas behind our estimation technique. Then we proceed with the formal statements of main results which are complemented with proofs and relevant references.

2. Notations, assumptions and overview of results

For $x \in \mathbb{R}^d$, let $\|x\|_\infty = \max_{i\le d}|x_i|$ and $\|x\|_2 = \left(\sum_{i\le d} x_i^2\right)^{1/2}$. For any two bounded functions $f, g : \mathbb{R}^d \mapsto \mathbb{R}$ and $A \subset \mathbb{R}^d$ define $\|f-g\|_{\infty,A} = \sup\limits_{x\in A}|f(x)-g(x)|$.
Recall that the observations are sampled from the model $Y = \eta(X) + \xi$. We denote $M(\eta) := \sup\limits_{x\in\mathrm{supp}(\Pi)} \eta(x)$ and $L_M = \{x \in \mathrm{supp}(\Pi) : \ \eta(x) = M(\eta)\}$, where $\mathrm{supp}(\Pi)$ stands for the support of measure Π.

Assumption 1 *Assume that one of the following two conditions holds:*

1. *There exists* $0 < T < \infty$ *such that* $|Y| \le T$ P*-almost surely.*
2. *Random variable* ξ *is independent of* X *and has finite* ψ_1 *- norm, where* $\|\xi\|_{\psi_1} = \inf\{C > 0 : \mathbb{E}\exp(|\xi/C|) \le 2\}$.

Our framework for selective sampling is governed by the following rules:

1. Design points are sampled sequentially: X_k is sampled from the modified distribution $\hat{\Pi}_k$ that depends on $(X_1, Y_1), \dots, (X_{k-1}, Y_{k-1})$.
2. Y_k is sampled from the conditional distribution $P_{Y|X}(\cdot|X = x)$. Y_i's are conditionally independent given the features $X_i, \ i \le n$.

Next, we introduce our assumptions on the design distribution Π. We will also assume below that Π is known: indeed, since we measure performance only by the number of noisy function values Y_i, $\Pi(A)$ can be estimated to an arbitrary precision for any measurable set A.

Assumption 2 *Distribution Π is supported in the unit cube $[0,1]^d$ and is absolutely continuous with respect to Lebesgue measure with a density $p(x)$ such that*

$$0 < c_1 \le p(x) \le c_2 < \infty \quad \forall x \in [0,1]^d.$$

This assumption serves two main purposes:

1. It allows to construct the estimators of regression function which are close to η in sup-norm.

2. It is well-suited for the iterative structure of our algorithm.

More general classes of design distributions can be handled by our method, but we tried to avoid unnecessary technicalities if favor of emphasizing the main ideas.

In the definition below, $\lfloor\beta\rfloor$ stands for the largest integer which is strictly smaller than β (e.g., $\lfloor 1\rfloor = 0$).

Definition 1 *We say that $g : \mathbb{R}^d \mapsto \mathbb{R}$ belongs to the class $\Sigma(\beta, B_1, [0,1]^d)$ if g is $\lfloor\beta\rfloor$ times continuously differentiable and for all $x, x_1 \in [0,1]^d$ satisfies*

$$|g(x_1) - T_x(x_1)| \le B_1 \|x - x_1\|_\infty^\beta,$$

where T_x is the Taylor polynomial of degree $\lfloor\beta\rfloor$ of g at the point x.

Since our main goal is to design estimation methods that are targeted at non-smooth (in particular, non-differentiable) functions, we will mostly concentrate on the case $\beta \in (0,1]$. Extensions of our methods to higher order of smoothness are possible but can be suboptimal.

Next condition is similar to the well-known *margin assumption* (also called *Tsybakov low noise assumption* (Tsybakov (2004)) that effectively captures the complexity of the problem in the framework of binary classification.

Assumption 3 *There exist $K, \gamma > 0$ such that $\forall\ t > 0$*

$$\mathrm{Vol}\left(\{x : \ |\eta(x) - M(\eta)| \le t\}\right) \le K t^\gamma,$$

where $\mathrm{Vol}(A)$ stands for the Lebesgue measure of a set $A \subset \mathbb{R}^d$.

This condition naturally describes how "flat" or "spiky" the regression function is near its maximum. Intuitively, the larger value of γ is, the easier it is to identify the set L_M. However, larger values of γ yield smaller values of the smoothness β, while the rate of estimation depends on the interplay between these parameters, as shown below. The following fact describes the relationship between β, γ and d.

Lemma 2 *Assume that $\eta \in \Sigma(\beta, B, [0,1]^d)$ and that Assumption 3 is satisfied for $\gamma > 0$. Then*

$$\beta \min(1, \gamma) \le d.$$

Proof See the proof of the first part of Proposition 3.4 in Audibert and Tsybakov (2005). ■

For an integer $m \geq 1$, let $\mathcal{G}_m := \left\{ \left(\frac{k_1}{2^m}, \ldots, \frac{k_d}{2^m}\right),\ k_i = 1 \ldots 2^m,\ i = 1 \ldots d \right\}$ be the uniform grid on the unit cube $[0,1]^d$ with mesh size 2^{-m}. It naturally defines a partition into a set of 2^{dm} open cubes $R_i,\ i = 1 \ldots 2^{dm}$ with edges of length 2^{-m} and vertices in $\mathcal{G}_m$. Below, we consider the nested sequence of grids $\{\mathcal{G}_m,\ m \geq 1\}$ and corresponding dyadic partitions $\{\mathcal{H}_m,\ m \geq 1\}$ of the unit cube.

Given two nonnegative integers r and m, let

$$\mathcal{F}_m^r := \left\{ f = \sum_{i=1}^{2^{dm}} q_i(x_1, \ldots, x_d) I_{R_i} \right\}, \tag{1}$$

where $\mathcal{H}_m = \left\{ R_i,\ 1 \leq i \leq 2^{dm} \right\}$ is the aforementioned dyadic partition of the unit cube and $q_i(x_1, \ldots, x_d)$ are the polynomials in d variables of degree at most r. For example, when $r = 0$, $\mathcal{F}_m^0$ can be viewed as the linear span of first 2^{dm} Haar basis functions on $[0,1]^d$. Note that $\{\mathcal{F}_m^r,\ m \geq 1\}$ is a nested family. Let $\eta \in \Sigma(\beta, B, [0,1]^d)$ for $0 < \beta \leq r+1$. By $\bar{\eta}_m(x)$ we denote the $L_2(\Pi)$ - projection of regression function $\eta(x)$ onto $\mathcal{F}_m^r$. The following assumption is crucial for theoretical justification of our method:

Assumption 4 *Assume one of the following two conditions holds:*

1. *$\eta(x)$ belongs to $\mathcal{F}_{m_0}^r$ for some $m_0 \geq 1$;*
2. *There exists $B_2 := B_2(\eta, \Pi) > 0$ such that for all $m \geq 1$ the following holds true:*

$$\|\eta - \bar{\eta}_m\|_{\infty, \mathrm{supp}(\Pi)} \geq B_2 2^{-\beta m}.$$

Note that for $r = 0$ (which will be our main focus), only the second part of Assumption 4 is meaningful for continuos η (unless $\eta(x) = \mathrm{const}$). The intuition behind this assumption can be informally explained as follows: functions that satisfy Assumption 4 are the functions whose smoothness can be learned from the data.

Assume we know that $\eta \in \Sigma(\nu, B_1, [0,1]^d)$ for some $\nu \in (0,1)$. While one could use this information to design an algorithm that achieves optimal rates over the class $\Sigma(\nu, B_1, [0,1]^d)$ without imposing Assumption 4 (in the bandits framework this is done, for example, in Bubeck et al. (2011)), our goal is to make another step forward. Namely, if it happens that for some $\varepsilon > 0$, $\eta \in \Sigma(\nu + \varepsilon, B_1, [0,1]^d) \subset \Sigma(\nu, B_1, [0,1]^d)$, we want the algorithm to automatically utilize the additional smoothness and to achieve the rate of convergence which would be optimal over $\Sigma(\nu + \delta, B_1, [0,1]^d)$ rather than $\Sigma(\nu, B_1, [0,1]^d)$. We show that this is possible if Assumption 4 is satisfied. The following statement demonstrates that the class of functions satisfying Assumption 4 is sufficiently rich.

Lemma 3 *Assume one of the following conditions holds:*

1. *$\eta \in C^{r+1}\left([0,1]^d\right)$, the space of $(r+1)$–times continuously differentiable functions.*
2. *Let $f_u(t; x_0) := \eta(x_0 + tu),\ t \in \mathbb{R},\ u \in \mathbb{R}^d,\ \|u\|_2 = 1$. There exist $x_0 \in (0,1)^d$ and u such that $\lim\limits_{t \to 0} \left| \frac{f_u^{(\lfloor \beta \rfloor)}(t; x_0) - f_u^{(\lfloor \beta \rfloor)}(0; x_0)}{|t|^{\beta - \lfloor \beta \rfloor}} \right| = M > 0$.*

Then Assumption 4 is satisfied.

Proof See the proof of Propositions 2.5.7 and 2.5.8 in Minsker (2012a). ■

Definition 4 *Let $\beta, \gamma > 0$. We say that P belongs to $\mathcal{P}(\beta,\gamma)$ if and only if $\eta \in \Sigma(\beta, B_1, [0,1]^d)$ and Assumptions 1–4 are satisfied.*

2.1. Overview of estimation technique

Our main goal is to design a procedure that would allow to estimate L_M and $M(\eta)$ adaptively over $\bigcup\limits_{0<\nu\le\beta\le 1}\bigcup\limits_{\gamma>0}\mathcal{P}(\beta,\gamma)$ for some fixed ν. Adaptivity of proposed algorithm is the main improvement over previously available results (e.g., Bubeck et al. (2011), Kleinberg et al. (2008), albeit these works impose weaker assumptions). Additionally, we are interested in estimating the whole *level set* associated to the absolute maximum of η rather than just identifying its subset and the value of the maximum. Our approach is closely related to the methods developed in Minsker (2012b), Minsker (2012a) for the binary classification problem.

On each iteration, the algorithm attempts to reduce the search domain for the maximum of η. Reduction is based on sequentially refined estimators of the regression function. Assume that $\hat{A}_k$ is the set of possible maximums on step k (e.g., on step 1 $A_1 = [0,1]^d$). The algorithm constructs an estimator $\hat{\eta}_k$ such that $\|\eta - \hat{\eta}_k\|_{\infty,\hat{A}_k} \le \delta_k$ with high probability, where δ_k is decreasing. In turn, $\hat{\eta}_k$ is used to define $\hat{A}_{k+1} := \left\{x \in \hat{A}_k : \ |\hat{\eta}_k(x) - \hat{M}_k| \le 2\delta_k\right\}$, where $\hat{M}_k := \sup_{x\in\hat{A}_k}\hat{\eta}_k$. Improvement is achieved due to the fact that on each step, the algorithm only requests response Y for an observation $X \in \hat{A}_k$ and ignores $X \notin \hat{A}_k$, since $\hat{A}_k$ contains the relevant information about the maximum. Resulting estimator $\hat{L}_M$ produced by the algorithm has the property that $L_M \subseteq \hat{L}_M \subseteq \left\{x \in [0,1]^d : \ \eta(x) \ge M(\eta) - t_n\right\}$ with high probability, where $t_n \lesssim n^{-\frac{\beta}{2\beta+d-\beta\gamma}}$ (up to log-factors). In particular, if $\beta\gamma = d$, then for any $x \in \hat{L}_M$ we have $|\eta(x) - M(\eta)| \lesssim n^{-1/2}$, which is often the desired property for applications.

For provable performance guarantees, our algorithm requires the lower bound ν for regularity of η and a priori estimates on the constants B_1 and B_2 from Definition 1 and Assumption 4 respectively.

Complete description is given in Algorithm 1 below, followed by the performance guarantees of Theorem 14. See figure 1 below for graphical illustration.

3. Lower bounds

Recall that the Hausdorff distance between two non-empty subsets $A, B \subseteq [0,1]^d$ is defined as

$$d_H(A,B) = \max\left(\sup_{x\in A}\inf_{y\in B}\|x-y\|_2, \sup_{y\in B}\inf_{x\in A}\|x-y\|_2\right),$$

The main result of this section establishes the minimax lower bound for estimation of the set L_M via selective sampling:

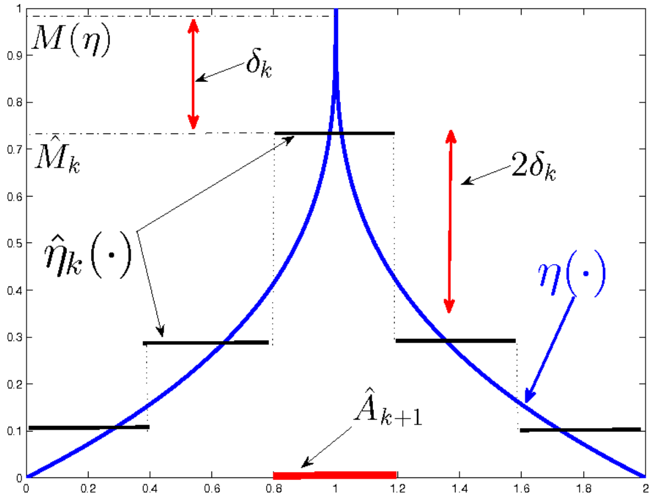

Figure 1: Estimation algorithm on step k

Theorem 5 *Let $\beta \in (0,1]$. There exists $C > 0$ such that for all n large enough and for any estimator $\hat{L}_M$ based on a sample of size n from P we have*

$$\sup_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{E}_P \, d_H\left(\hat{L}_M, L_M\right) \geq c n^{-\frac{1}{2\beta+d-\beta\gamma}}.$$

Remark 6 *Note that $d - \beta\gamma \geq 0$ for $0 < \beta \leq 1$.*

Proof The proof is based on Theorem 2.5 from Tsybakov (2009) and follows the relatively standard pattern by reducing the problem of estimating the minimax risk to the problem of hypotheses testing. Let $q = 2^l, \ l \geq 1$ and

$$G_q := \left\{ \left(\frac{2k_1 - 1}{2q}, \ldots, \frac{2k_d - 1}{2q} \right), \ k_i = 1 \ldots q, \ i = 1 \ldots d \right\}$$

be the grid on $[0,1]^d$. For $x \in [0,1]^d$, let $n_q(x) = \text{argmin } \{\|x - x_k\|_2 : \ x_k \in G_q\}$. If $n_q(x)$ is not unique, we choose a representative with the smallest $\|\cdot\|_2$ norm. The unit cube is partitioned with respect to G_q as follows: x_1, x_2 belong to the same subset if $n_q(x_1) = n_q(x_2)$. Let $' \succ '$ be some order on the elements of G_q such that $x \succ y$ implies $\|x\|_\infty \geq \|y\|_\infty$. Assume that the elements of the partition are enumerated with respect to the order of their centers induced by $' \succ '$: $[0,1]^d = \bigcup\limits_{i=1}^{q^d} R_i$. Fix $1 \leq m \leq q^d$ and let $S := \bigcup_{i=1}^m R_i$. Define $\mathcal{H}_m = \{P_\sigma : \sigma \in \{-1,1\}^m \setminus (-1,-1,\ldots,-1)\}$ to be the collection of probability distributions on $[0,1]^d \times \mathbb{R}$. Here, P_σ is the distributions of a couple (X,Y), where $Y = \eta_\sigma(X) + \varepsilon$, the marginal distribution of X is uniform on the unit cube (in particular, independent of σ), ε is standard normal and independent of X (so that the conditional distribution $P(Y|X=x)$ is normal with mean $\eta_\sigma(x)$ and variance 1). Let $\phi(x) := \left(1 - \|x\|_\infty^\beta\right) \exp\left(-\frac{1}{1-\|x\|_\infty^2}\right)$ and

$$\eta_\sigma(x) := \begin{cases} \sigma_i q^{-\beta} \phi(q[x - n_q(x)]), & x \in R_i, \ 1 \leq i \leq m, \\ -\left[\inf_{y \in S} \|x - y\|_\infty\right]^{d/\gamma}, & \text{else.} \end{cases}$$

Note that $M(\eta_\sigma) = q^{-\beta}$ for any $\sigma \in \mathcal{H}_m$. It is straightforward to check that η_σ satisfies smoothness requirements and second condition of Assumption 4 (by Lemma 3). We will check that Assumption 3 is satisfied whenever $mq^{-d} = O(q^{-\beta\gamma})$. Indeed, for $z \le q^{-\beta}$

$$\mathrm{Vol}\left\{x : |\eta_\sigma(x) - q^{-\beta}| \le z\right\} \le m\left(\frac{(2zq^{\beta})^{1/\beta}}{q}\right)^d = 2^d z^\gamma\, m z^{d/\beta-\gamma} \le K z^\gamma$$

given that $m \le C q^{d-\beta\gamma}$. For $z > q^{-\beta}$, Assumption 3 is also easily verified (by examining η_σ on S and $[0,1]^d \setminus S$ independently).

The proof proceeds by selecting a "well-separated" subset of $\mathcal{H}_m$. Let $\mathcal{H}' := \{P_{\sigma_0}, \dots, P_{\sigma_M}\}$, where $\sigma_0 = (1,1,\dots,1)$, be chosen such that collection $\{\sigma_0, \dots, \sigma_M\}$ satisfies the conditions of Gilbert-Varshamov bound (see Lemma 2.9 in Tsybakov (2009)). It remains to bound the Kullback-Leibler divergence $\mathrm{KL}(P_{\sigma,n} \| P_{\sigma_0,n})$, where $P_{\sigma,n}$ is the joint distribution of $(X_i, Y_i)_{i=1}^n$ under the hypothesis that the distribution of a couple (X,Y) is P_σ. In the case of adaptive design, this is done is detail in the proofs of Theorem 7 (Minsker (2012b)) or Theorem 3 (Polyak and Tsybakov (1990)) by factorizing $P_{\sigma,n}$ and using the fact that the $\mathrm{KL}\left(N(0,1)\|N(t,1)\right) = \frac{t^2}{2}$, where $N(t,1)$ is the normal distribution with mean t and variance 1; we skip the technical details for brevity and give the final bound:

$$\mathrm{KL}(P_{\sigma,n}\|P_{\sigma_0,n}) \le Cn \max_{x\in[0,1]^d} (\eta_\sigma(x) - \eta_{\sigma_0}(x))^2 \le Cnq^{-2\beta}.$$

Finally, setting $q = \lfloor C_1 N^{\frac{1}{2\beta+d-\beta\gamma}} \rfloor$ and $m = \lfloor C_2 q^{d-\beta\gamma} \rfloor$, it is easy to check that all conditions of Theorem 2.5 (Tsybakov (2009)) are satisfied for appropriate C_1, C_2. Since for any $\sigma_1, \sigma_2 \in \mathcal{H}'$ and corresponding level sets $L_M^i = \{x : \eta_{\sigma_i}(x) = \max_x \eta_{\sigma_i}(x)\}$, $i = 1,2$ we clearly have $d_H(L_M^1, L_M^2) \ge 2q$, then Theorem 2.5 in Tsybakov (2009) guarantees that $q = \lfloor C_1 n^{\frac{1}{2\beta+d-\beta\gamma}} \rfloor$ is the minimax lower bound for the risk, implying the result. ■

Remark 7 *Note that in the case $\beta\gamma = d$, which holds, for example, when the maximum is unique and $\eta(x) \simeq M(\eta) - \|x\|^\beta$ in a neighborhood of the maximum, the resulting rate $n^{-1/2\beta}$ is dimension-free. This result complements the well-known fact that the optimal rate for estimating the maximum of a smooth function is $n^{-\frac{\beta-1}{2\beta}}$, $\beta \ge 2$ (Theorem 3 in Polyak and Tsybakov (1990)). It is also interesting to compare our result with the optimal rate for estimating the mode of the univariate unimodal regression function, which is known to be $n^{-1/(2\beta+1)}$ for iid (hence, nonadaptive) design; here, β describes the rate of decay in vicinity of the mode (see Shoung and Zhang (2001)).*

Remark 8 *It is well-known that the minimax lower bound for estimating the value of the maximum $\sup\limits_{x\in[0,1]^d} \eta(x)$ over Hölder balls is $n^{-1/2}$. It will be shown in the following sections that our algorithm attains this rate up to logarithmic factors in the case $\beta\gamma = d$. Whenever $0 < \beta \le 1$ and $\beta\gamma < d$, we conjecture that the optimal rate is $n^{-\frac{\beta}{2\beta+d-\beta\gamma}}$. Let us mention that Theorem 2 in Auer et al. (2007) proves the lower bound of the form $n^{-\theta}$ for any $\theta < -\frac{\beta}{2\beta+1-\beta\gamma}$ (in the univariate case $d = 1$).*

4. Upper bounds

The main goal of this section is to give a rigorous description and analysis of the algorithm introduced in section 2.1.

4.1. Preliminary results

An important ingredient of our analysis is the bound for estimating $\eta(x)$ in sup-norm loss. To obtain such an estimate, we need to

1. Analyze approximation properties of classes $\mathcal{F}^r_m$ defined by (1);
2. Obtain the probabilistic bounds for estimation of elements of $\mathcal{F}^r_m$ from the noisy data.

We will concentrate on the case of piecewise-constant approximation $r = 0$ which is sufficient for our purposes. Let Proj_m be the $L_2(\Pi)$–projector on $\mathcal{F}^0_m$.

Lemma 9 *Assume $f \in \Sigma(\beta, B_1, [0,1]^d)$ for $0 < \beta \le 1$. Then*

$$\|f - \mathrm{Proj}_m f\|_{\infty,[0,1]^d} \le 2B_1 2^{-\beta m}.$$

Proof This immediately follows from the fact that for any dyadic cube R and $x \in R$ $f(x) - (\mathrm{Proj}_m f)(x) = \frac{1}{\Pi(R)} \int\limits_R (f(x) - f(y)) d\Pi(y)$. ■

Let $\mathcal{B}_m$ be the sigma-algebra generated by dyadic cubes R_j, $1 \le j \le 2^{dm}$ forming the partition of $[0,1]^d$. Given a nonempty set $A \in \mathcal{B}_m$, define $\hat{\Pi}_A(dx) := \Pi(dx|x \in A)$. Moreover, set $d_{m,A} := \dim\left(\mathcal{F}^0_m\big|_A\right)$. Next, we introduce an piecewise-constant estimator of the regression function on the set A. Let (X_i, Y_i), $i \le n$ be iid observations with $X_i \sim \hat{\Pi}_A(dx)$ and $m \ge 1$ be the resolution level. Define

$$\hat{\eta}_{m,A}(x) := \sum_{j: R_j \cap A \ne \emptyset} \frac{\sum_{i=1}^n Y_i I_{R_j}(X_i)}{n \hat{\Pi}_A(R_j)} I_{R_j}(x) \text{ for } x \in A \text{ and } 0 \text{ otherwise.} \tag{2}$$

This estimator is well-defined since we assumed that Π is known. Let $\bar{\eta}_m = \mathrm{Proj}_m \eta$ and note that for all $x \in A$, $\mathbb{E}\hat{\eta}_{m,A}(x) = \bar{\eta}_m(x)$.

We have the following concentration result:

Proposition 10 *Assume that the first condition of Assumption 1 holds. Then, for any $t > 0$,*

$$\Pr\left(\sup_{x \in A} |\hat{\eta}_{m,A}(x) - \bar{\eta}_m(x)| > C \cdot T \left[\sqrt{\frac{t\, 2^{dm}\Pi(A)}{n}} \vee 2^{dm}\Pi(A)\frac{t}{n}\right]\right) \le 2 d_{m,A} e^{-t}.$$

If the second condition of Assumption 1 holds, then

$$\Pr\left(\sup_{x \in A} |\hat{\eta}_{m,A}(x) - \bar{\eta}_m(x)| > C(\eta, \|\xi\|_{\psi_1}) \left[\sqrt{\frac{t\, 2^{dm}\Pi(A)}{n}} \vee dm 2^{dm}\Pi(A)\frac{t}{n}\right]\right) \le 2 d_{m,A} e^{-t}.$$

Proof The proof is based on the variants of Bernstein's inequality for bounded random variables and for random variables with bounded ψ_1-norms, combined with the union bound. See Appendix for details. ■

Remark 11 *Note that whenever* $2^{dm} \leq \frac{n}{\log^4 n}$ *and* $t \leq \log^3 n$, *bound of Proposition 10 becomes* $\Pr\left(\sup\limits_{x\in A} |\hat{\eta}_{m,A}(x) - \bar{\eta}_m(x)| > C\sqrt{\frac{t\,2^{dm}\Pi(A)}{n}}\right) \leq 2d_{m,A}e^{-t}$, *where* $C = C(T)$ *or* $C = C(\eta, \|\xi\|_{\psi_1})$.

4.2. Model selection

Below, we briefly describe the tools which are needed to make our learning algorithm adaptive with respect to the unknown smoothness β. It turns out that if Assumption 4 is satisfied, then information about smoothness can be captured from the data. Approach presented below was partially motivated by results of Giné and Nickl (2010) on adaptive confidence bands in density estimation. Given a sequence of finite dimensional subspaces $\mathcal{G}_m$(in our case, these are the piecewise–constant functions $\mathcal{F}^0_m$, possibly restricted to some subset of $[0,1]^d$), define the index set

$$\mathcal{J}(n) := \left\{m \in \mathbb{N} :\ 1 \leq \dim \mathcal{G}_m \leq \frac{n}{\log^4 n}\right\} \tag{3}$$

which is the set of all possible resolution levels of an estimator from $\mathcal{G}_m$ based on a sample of size n. For the model selection procedures described below, we will always assume that the index is chosen from the corresponding $\mathcal{J}(n)$.

Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from P, let $\left\{\hat{\eta}_m := \hat{\eta}_{m,[0,1]^d},\ m \in \mathcal{J}(n)\right\}$ be a collection of estimators of η on the unit cube defined by formula (2). Our goal is to choose the resolution level m in an optimal way using the given sample. Optimality is understood as a balance between the bias term coming from the piecewise-constant approximation and the random error coming from the use of noisy data. For $t > 1$, define

$$\hat{m} := \hat{m}(t, n) = \min\left\{m \in \mathcal{J}(n) :\ \forall l > m,\ l \in \mathcal{J}(n),\ \|\hat{\eta}_l - \hat{\eta}_m\|_\infty \leq K_1\sqrt{\frac{t\,2^{dl}l}{n}}\right\}. \tag{4}$$

We will compare $\hat{m}$ to the “optimal” resolution level $\bar{m}$ defined by

$$\bar{m} := \min\left\{m \in \mathcal{J}(n) : \|\eta - \bar{\eta}_m\|_\infty \leq K_2\sqrt{\frac{2^{dm}m}{n}}\right\}. \tag{5}$$

For $\bar{m}$, we immediately get the following:

Lemma 12 *If* $\eta \in \Sigma(B_1, [0,1]^d, \beta)$ *for* $0 < \beta \leq r + 1$, *then* $2^{\bar{m}} \leq C_1 \cdot \left(\frac{nB_1^2}{\log(nB_1^2)}\right)^{1/(2\beta+d)}$, *where* $C_1 = C_1(\Pi, d, r)$. *Moreover, if Assumption 4 is satisfied with a constant* B_2, *then* $2^{\bar{m}} \geq C_2 \cdot \left(\frac{nB_2^2}{\log(nB_2^2)}\right)^{1/(2\beta+d)}$.

We are ready to present the main result of this subsection:

Theorem 13 *Assume that $\eta \in \Sigma(\beta, B_1, [0,1]^d)$ and that Assumption 4 is satisfied with a constant B_2. Then there exists $t_0 = t_0(d) > 0$ and K_1 large enough such that for all $t \geq t_0$ we have*

$$\hat{m} \in \left(\bar{m} - \frac{1}{\beta}\left(\log_2 t + \log_2 \frac{B_1}{B_2} + h\right), \bar{m}\right]$$

with probability at least $1 - C2^{d\bar{m}} \log n \exp(-ct\bar{m})$, where h is some fixed positive number that depends on d, r, Π.

Proof See the proof of Theorem 2.5.6 in Minsker (2012a). ■

4.3. Estimation procedure: details and analysis

Complete description of estimation procedure is given in Algorithm 1 below; the details are mostly self-explanatory, and we will just make few clarifying comments. Note that on the first step of the algorithm, a sample of size $2N_0$ is divided into two equal parts. The first n_0 pairs denoted $S_{0,1}$ are used to define $\hat{m}_0$ and the rest (denoted $S_{0,2}$) are used to construct $\hat{\eta}_0$. Sample size n_k is chosen such that on every step, $2^{\hat{m}_k} \approx n_k^{1/(2\beta+d)}$ (this motivates the expression for τ_k).

We are ready to present the main result of this section. We will assume that ε is small enough so that $B_1 \leq \log^{1/2} \frac{1}{\varepsilon}$, $B_2 \geq \log^{-1/2} \frac{1}{\varepsilon}$, where $B_1,\ B_2$ are the constants from Definition 1 and Assumption 4 respectively (these can be replaced by any known bounds on B_1, B_2).

Theorem 14 *Assume that $P \in \mathcal{P}(\beta, \gamma)$ for $0 < \nu \leq \beta \leq 1$. Then with probability $\geq 1 - \alpha$ estimators $\hat{L}_M$ and $\hat{M}$ returned by Algorithm 1 satisfy*

$$L_M \subseteq \hat{L}_M \subseteq \left\{x \in [0,1]^d :\ \eta(x) \geq M(\eta) - 4\varepsilon\right\}, \tag{6}$$

$$|\hat{M} - M(\eta)| \leq \varepsilon, \tag{7}$$

while the total number of noisy function measurements requested by Algorithm 1 is

$$n \leq C\left(\frac{1}{\varepsilon}\right)^{\frac{2\beta+d-\beta\gamma}{\beta}} \log^p \frac{1}{\varepsilon\alpha}, \tag{8}$$

where $p \leq \left(\frac{4+2d}{4\nu}\right)^2 \left(1 + \frac{\beta\gamma}{2\beta+d-\beta\gamma}\right)$.

Remark 15 *Note that whenever $\eta(x)$ satisfies $M(\eta) - \eta(x) \geq B_3 \inf\limits_{y \in L_M} \|y - x\|_\infty^\beta$, then (6) implies that the estimator $\hat{L}_M$ produced by Algorithm 1 after requesting n noisy function values satisfies $d_H(\hat{L}_M, L_M) \leq C_3 n^{-\frac{1}{2\beta+d-\beta\gamma}} \mathrm{polylog}(n/\alpha)$ with probability $\geq 1 - \alpha$, which is, up to log-factors, the rate given by Theorem 5.*

```
input  : desired accuracy $\varepsilon$; confidence $\alpha$; minimal regularity $0 < \nu \le 1$
output: $\hat{L}_M$ – estimator of the level set $L_M$; $\hat{M}$ – estimator of $M(\eta)$.
$\omega := 2 + \frac{d}{2\nu}$
$k = 0,\ \hat{A}_0 := [0,1]^d$
$n_0 := \lfloor \varepsilon^{-\nu} \rfloor$
for $i = 1$ to $2n_0$ do
    sample i.i.d. $\left(X_i^{(0)}, Y_i^{(0)}\right)$ with $X_i^{(0)} \sim \Pi$
end
$S_{0,1} := \left\{ \left(X_i^{(0)}, Y_i^{(0)}\right),\ i \le n_0 \right\}, \quad S_{0,2} = \left\{ \left(X_i^{(0)}, Y_i^{(0)}\right),\ n_0 + 1 \le i \le 2N_0 \right\}$
$\hat{m}_0 := \hat{m}(s, n_0; S_{0,1})$                                   /* see equation (4)  */
$\hat{\eta}_0 := \hat{\eta}_{\hat{m}_0,[0,1]^d;S_{0,2}}$                                   /* see equation (2)  */
$\delta_0 := \tilde{D}(\log\frac{1}{\varepsilon\alpha})^{\omega\frac{\hat{m}_k}{\hat{m}_0}} \cdot \sqrt{\frac{2^{d\hat{m}_0}}{n_0}}$
while $\delta_k > \varepsilon$ do
    $\hat{M}_k = \max_{x \in \hat{A}_k} \hat{\eta}_k(x)$
    $k := k + 1$
    $\hat{A}_k := \left\{ x \in \hat{A}_{k-1} :\ \hat{\eta}_{k-1} \ge \hat{M}_{k-1} - 2\delta_{k-1} \right\}$          /* a new search domain */
    $\hat{m}_k := \hat{m}_{k-1} + 1$
    $\tau_k := \frac{\hat{m}_k}{\hat{m}_{k-1}}$
    $n_k := \lfloor n_{k-1}^{\tau_k} \rfloor$
    for $i = 1$ to $\lfloor n_k \cdot \Pi(\hat{A}_k) \rfloor$ do
        sample i.i.d. $\left(X_i^{(k)}, Y_i^{(k)}\right)$ with $X_i^{(k)} \sim \hat{\Pi}_k := \Pi(dx|x \in \hat{A}_k)$
        $S_k := \left\{ \left(X_i^{(k)}, Y_i^{(k)}\right),\ i \le \lfloor n_k \cdot \Pi(\hat{A}_k) \rfloor \right\}$
    end
    $\hat{\eta}_k := \hat{\eta}_{\hat{m}_k,\hat{A}_k}$                                   /* estimator based on $S_k$ */
    $\delta_k := \tilde{D}(\log\frac{1}{\varepsilon\alpha})^{\omega\frac{\hat{m}_k}{\hat{m}_0}} \cdot \sqrt{\frac{2^{d\hat{m}_k}}{n_k}}$                    /* bound on estimation error */
    $\hat{L}_M := \hat{A}_{k+1}, \quad \hat{M} := \hat{M}_k$   /* keeping track of the most recent estimators */
end
```

Algorithm 1: Learning the location and the value of $\max_{x \in [0,1]^d} \eta(x)$.

Proof The proof is conceptually simple but technical. Our main goal will be to construct high probability bounds for the size of the sets $\hat{A}_k$ defined by Algorithm 1. In turn, these bounds depend on the estimation errors δ_k. Suppose J is the number of steps performed by the algorithm before termination.

Let $n_k^{\text{act}} := \lfloor n_k \cdot \Pi(\hat{A}_k) \rfloor$ be the number of samples requested on k-th iteration of the algorithm. Claim: the following bounds hold uniformly for all $1 \le k \le J$ with probability at least $1 - \alpha$:

$$\|\eta - \hat{\eta}_k\|_{\infty, \hat{A}_k} \le \delta_k, \quad \delta_k \le C \left(\log \frac{1}{\varepsilon\alpha}\right)^{\omega\bar{\tau}} n_k^{-\beta/(2\beta+d)}, \tag{9}$$

$$\Pi(\hat{A}_k) \le C \left(\log \frac{1}{\varepsilon\alpha}\right)^{\gamma\omega\bar{\tau}} n_{k-1}^{-\beta\gamma/(2\beta+d)} \tag{10}$$

where $\omega = 2 + \frac{d}{2\nu}$ and $\bar{\tau} = 4 + \frac{2d}{\nu}$. Let $\mathcal{E}$ be the event on which (9),(10) hold.
Let us first assume that (9) has already been established and derive the result from it. Let $\bar{m}_0$ be the "optimal" resolution level for the corresponding sample of size n_0, see formula (5). First, we make a useful observation that, with high probability, numbers n_k grow geometrically: indeed, we have by the definition of $\hat{m}_k$

$$n_{k+1} = \lfloor n_k^{\hat{m}_{k+1}/\hat{m}_k} \rfloor \le n_k \cdot n_k^{1/\hat{m}_k} \le n_k \cdot \left(n_0^{\frac{\hat{m}_k}{\hat{m}_0}}\right)^{\frac{1}{\hat{m}_k}} = n_k \cdot n_0^{\frac{1}{\hat{m}_0}},$$

and by Theorem 13, if n_0 is sufficiently large, as guaranteed by our assumptions, $\frac{\log_2 n_0}{\bar{m}_0} \le \log_2 n_0^{1/\hat{m}_0} \le \frac{\log_2 n_0}{\frac{1}{2}\bar{m}_0}$ with probability $\ge 1 - \alpha$. Finally, Lemma 12 gives

$$\frac{1}{2\beta+d} \log n_0 + c \ge \bar{m}_0 \ge \frac{1}{2\beta+d}(\log n_0 - 2\log\log n_0) - c$$

which shows that $0 < C_1 \le \log_2 n_0^{1/\hat{m}_0} \le C_2$.
Next, inequality (9) implies, together with the previous observation, that the number of labels requested on step $k \ge 1$ satisfies $n_k^{\text{act}} = \lfloor n_k \Pi(\hat{A}_k) \rfloor \le C \cdot n_{k-1}^{\frac{2\beta+d-\beta\gamma}{2\beta+d}} \left(\log \frac{1}{\varepsilon\alpha}\right)^{\gamma\omega\bar{\tau}}$ with probability $\ge 1 - 2\alpha$. If n is the total number of labels requested by the algorithm, then, due to geometric growth,

$$n = \sum_{k=0}^{J} n_k^{\text{act}} \le C_3 \left(\log \frac{1}{\varepsilon\alpha}\right)^{\gamma\omega\bar{\tau}} \sum_{k=0}^{J} n_k^{\frac{2\beta+d-\beta\gamma}{2\beta+d}} \le C_4 \left(\log \frac{1}{\varepsilon\alpha}\right)^{\gamma\omega\bar{\tau}} n_{J-1}^{\frac{2\beta+d-\beta\gamma}{2\beta+d}}.$$

At the same time, we have $\delta_{J-1} > \varepsilon$ (otherwise algorithm would terminate on step $J-1$), hence by (9) we have $C\left(\log \frac{1}{\varepsilon\alpha}\right)^{\omega\bar{\tau}} n_{J-1}^{-\beta/(2\beta+d)} \ge \varepsilon$, and (8) follows by simple algebra.

To derive (6), note that on event $\mathcal{E}$

$$x \notin \hat{A}_{k+1} \implies \hat{\eta}_k(x) < \hat{M}_k - 2\delta_k \implies \eta(x) \le \hat{M}_k - 2\delta_k + |\eta(x) - \hat{\eta}_k(x)| \le$$
$$\le M(\eta) + \sup_{x \in \hat{A}_k} |(\eta - \hat{\eta}_k)(x)| - 2\delta_k + |\eta(x) - \hat{\eta}_k(x)| \le M(\eta)$$

where we used (9) in the last inequality. This gives $L_M \subseteq \hat{A}_k$ for all $1 \le k \le J$. On the other hand,

$$\begin{aligned} x \in \hat{A}_{k+1} \implies \hat{\eta}_k(x) \ge \hat{M}_k - 2\delta_k \implies \eta(x) \ge \hat{\eta}_k(x) - |\eta(x) - \hat{\eta}_k(x)| \ge \\ \ge M(\eta) - 2\delta_k - \sup_{x \in \hat{A}_k} |(\eta - \hat{\eta}_k)(x)| - |\eta(x) - \hat{\eta}_k(x)| \ge M(\eta) - 4\delta_k, \end{aligned} \tag{11}$$

hence on event $\mathcal{E}$

$$\hat{A}_{k+1} \subseteq \left\{ x \in [0,1]^d : \ \eta(x) \ge M(\eta) - 4\delta_k \right\}, \tag{12}$$

and (10) follows. It remains to show (9), (10). The main tools are given by Proposition 10 and Theorem 13. The proof of (9) consists of applying these results on each step of the algorithm combined with the union bound; (10) immediately follows from (9), (12) and Assumption 3. Detailed derivation is given in Appendix (also see the proof of Theorem 2.6.2 in Minsker (2012a) for a similar argument). ■

5. Concluding remarks

It is easy to see that our approach can be applied for estimation of arbitrary level sets of the form $\{x : \ \eta(x) = z\}$ for $z \in \mathrm{Range}(\eta)$, which might be useful in other applications.

While proposed algorithm possesses several nice properties (such as adaptivity), its performance is limited by the use of piecewise-constant approximation. Beyond Hölder smoothness $0 < \beta \le 1$, more complicated estimators are required to attain faster rates (e.g., classes defined by (1) for $r \ge 1$). Such extensions are possible (but not straightforward), and the resulting number of total noisy function evaluations required by Theorem 14 becomes $n \le C \left(\frac{1}{\varepsilon}\right)^{\frac{2\beta + d - (\beta \wedge 1)\gamma}{\beta}} \mathrm{polylog}(\frac{1}{\varepsilon\alpha})$ (compare to (8)). However, in many situations this bound is suboptimal due to the "zero-order" nature of the algorithm. We omit further details in this work to avoid associated technicalities.

Acknowledgments

I want to thank anonymous reviewers for their valuable comments that helped to identify some relevant references and improve the technical part of the paper. This research was partially supported by grant R01-ES-017436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH), NSF grants DMS-0650413 and CCF-0808847, and DOE contract 113054 G002745.

References

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Preprint (shorter version was published in Ann. Statist., 2007, Vol. 35(2))*, 2005. Available at: `http://imagine.enpc.fr/publications/papers/05preprint_AudTsy.pdf`.

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory*, pages 454–468. Springer, 2007.

E. Belitser, S. Ghosal, and H. van Zanten. Optimal two-stage procedures for estimating location and size of maximum of multivariate regression functions. *Ann. Statist.*, 40(6): 2850–2876, 2012.

J.R. Blum. Multidimensional stochastic approximation methods. *Ann. Statist.*, 25(4):737–744, 1954.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

J. Dippon. Accelerated randomized stochastic optimization. *Ann. Statist.*, 31(4):1260–1281, 2003.

E. Giné and R. Nickl. Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170, 2010.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Statist.*, 23(3):462–466, 1952.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 681–690. ACM, 2008.

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems.* Springer, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour.

S. Minsker. *Non-asymptotic bounds for prediction problems and density estimation.* PhD thesis, Georgia Institute of Technology, 2012a.

S. Minsker. Plug-in approach to active learning. *J. Mach. Learn. Res.*, pages 67–90, 2012b.

B. T. Polyak and A. B. Tsybakov. Optimal orders of accuracy for search algorithms of stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990. ISSN 0555-2923.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Statist.*, pages 400–407, 1951.

J.M. Shoung and C.H. Zhang. Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.*, 29(3):648–665, 2001.

R. Tang, M. Banerjee, and G. Michailidis. A two-stage hybrid procedure for estimating an inverse regression function. *Ann. Statist.*, 39(2):956–989, 2011.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32 (1):135–166, 2004.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996.

Appendix A. Proof of Proposition 10.

We will assume that $A = [0,1]^d$ for simplicity, since the same argument applies to any $A \in \mathcal{B}_m$. Let $R \in \mathbb{B}_m$ be a dyadic cube and define the random variable $Z := Y \cdot I\left\{X \in R\right\}$. To prove the first part of the proposition, note that $|Z| \le T$ P-a.s. and $\mathrm{Var}(Z) \le \mathbb{E}Z^2 \le T^2\Pi(R)$. It remains to apply Bernstein's inequality (see van der Vaart and Wellner (1996), Lemma 2.2.9) to $\frac{S_n}{n\Pi(R)}$, where $S_n := \sum_{j=1}^n (Z_j - \mathbb{E}Z_j)$. The union bound over all dyadic cubes R_j forming the partition of A, together with an observation that $\Pi(R_j) \ge c_1 2^{-dm}$ (by properties of the marginal distribution Π), concludes the proof.

The argument for the second part is slightly more complicated. We start by estimating $\|Z\|_{\psi_1}$. Note that the distribution P_Z of Z is a mixture $P_Z = (1-\Pi(R))\delta_0 + \Pi(R)P(Y|X \in R)$. If $Y_R \sim P(Y|X \in R)$, then

$$\mathbb{E}e^{D|Z|} - 1 = \Pi(R)\left(\mathbb{E}e^{D|Y_R|} - 1\right) = \Pi(R)\mathbb{E}\int\limits_0^{Y_R} De^{Ds}ds = \tag{13}$$

$$= \Pi(R)\int\limits_0^{\infty} \Pr\left(|Y_R| > s\right) De^{Ds}ds,$$

where we used Fubini's theorem in the last equality. Next, we estimate the moment generating function of Y_R as follows: for $t > 0$,

$$\mathbb{E}e^{tY_R} = \mathbb{E}\mathbb{E}\left(e^{t(Y_R - \eta(x) + \eta(x))}|X = x\right) = \mathbb{E}\left[e^{t\eta(X)}\mathbb{E}(e^{t\xi}|X = x)\right] \le e^{t\max_x|\eta(x)|}\mathbb{E}e^{t\xi}.$$

Since by definition of ψ_1-norm $\mathbb{E}e^{\frac{|\xi|}{\|\xi\|_{\psi_1}}} \le 2$, choosing $t := \|\xi\|_{\psi_1}^{-1}$ and setting $K(\eta) := \max_x|\eta(x)|$, we get $\mathbb{E}e^{Y_R\|\xi\|_{\psi_1}^{-1}} \le 2e^{K(\eta)\|\xi\|_{\psi_1}^{-1}}$. Furthermore, Chebyshev inequality applied to Y_R and $-Y_R$ yields $\Pr\left(|Y_R| > s\right) \le \inf\limits_{t>0} e^{-ts}(\mathbb{E}e^{tY_R} + \mathbb{E}e^{-tY_R}) \le 4e^{-s\|\xi\|_{\psi_1}^{-1}}e^{K(\eta)\|\xi\|_{\psi_1}^{-1}}$. Plugging this back into (13), one easily gets $\mathbb{E}e^{D|Z|} - 1 \le 4\Pi(R)e^{K(\eta)\|\xi\|_{\psi_1}^{-1}}\frac{D}{\|\xi\|_{\psi_1}^{-1} - D}$. The right-hand side of the last inequality is not greater than 1 for $D^{-1} \ge \|\xi\|_{\psi_1}(1 + 4\Pi(R)e^{K(\eta)\|\xi\|_{\psi_1}^{-1}})$, hence

$$\|Z\|_{\psi_1} \le \|\xi\|_{\psi_1}(1 + 4\Pi(R)e^{K(\eta)\|\xi\|_{\psi_1}^{-1}}).$$

Set $V(\eta, R) := \Pi(R)K(\eta) + \|\xi\|_{\psi_1}(1 + 4\Pi(R)e^{K(\eta)\|\xi\|_{\psi_1}^{-1}})$ and note that $\|Z - \mathbb{E}Z\|_{\psi_1} \le V(\eta, R)$. Finally, observe that $\mathbb{E}Z^2 \le 2\Pi(R)(\mathbb{E}\xi^2 + K^2(\eta))$. It remains to apply Bernstein's inequality for random variables with bounded ψ_1 - moments (Theorem 2.7 in Koltchinskii (2011)) to $S_n := \sum_{j=1}^n (Z_j - \mathbb{E}Z_j)$, which gives

$$\Pr\left(\frac{|S_n|}{n\Pi(R)} \ge C\left(\max(K(\eta), \mathbb{E}^{1/2}\xi^2)\sqrt{\frac{t}{n\Pi(R)}} \vee \frac{t}{n}\frac{V(\eta, R)}{\Pi(R)}\log\frac{V(\eta, R)}{\Pi(R)}\right)\right) \le e^{-t}.$$

The proof is concluded by applying the union bound similar to the first part of the proposition.

Appendix B. Proof of Theorem 14: remaining details

Let $\hat{\eta}_k$ be the estimator obtained on step k. For $k=0$, we have

$$\|\eta - \hat{\eta}_0\|_{\infty,\mathrm{supp}(\Pi)} \le \|\eta - \bar{\eta}_{\hat{m}_0}\|_{\infty,[0,1]^d} + \|\bar{\eta}_{\hat{m}_0} - \hat{\eta}_0\|_{\infty,[0,1]^d}.$$

By Proposition 10 (applied conditionally on $S_{0,1}$), with probability $\ge 1-\alpha$

$$\|\bar{\eta}_{\hat{m}_0} - \hat{\eta}_0\|_{\infty,\mathrm{supp}(\Pi)} \le C\log^{1/2}(1/\varepsilon\alpha)\sqrt{\frac{2^{d\hat{m}_0}}{n_0}}.$$

For the bias term $\|\eta - \bar{\eta}_{\hat{m}_0}\|_{\infty,\mathrm{supp}(\Pi)}$, by our assumptions on η (see Lemma 9),

$$\|\eta - \bar{\eta}_{\hat{m}_0}\|_{\infty,\mathrm{supp}(\Pi)} \le B_1 2^{-\beta\hat{m}_0}.$$

By Theorem 13, with probability $\ge 1-\alpha$

$$2^{-\beta\hat{m}_0} \le \frac{C}{\beta}\left(\log\frac{1}{\varepsilon\alpha}\right)^{1+\frac{d}{2\beta}}\sqrt{\frac{2^{d\hat{m}_0}\hat{m}_0}{n_0}}, \tag{14}$$

and by Theorem 13 and Lemma 12, with probability $\ge 1-\alpha$

$$\frac{2^{d\hat{m}_0}}{n_0} \le \frac{2^{d\bar{m}_0}}{n_0} \le C_1 n_0^{-2\beta/(2\beta+d)}, \tag{15}$$

so that, with probability $\ge 1-2\alpha$,

$$\begin{aligned}\|\eta - \hat{\eta}_0\|_{\infty,[0,1]^d} &\le C(\beta,d)\left(\log\frac{1}{\varepsilon\alpha}\right)^{3/2+\frac{d}{2\beta}}\sqrt{\frac{2^{d\hat{m}_0}\hat{m}_0}{n_0}} := \frac{\delta_0}{2} \le \qquad (16)\\ &\le C(\beta,d)\left(\log\frac{1}{\varepsilon\alpha}\right)^{2+\frac{d}{2\beta}} n_0^{-\frac{\beta}{2\beta+d}}.\end{aligned}$$

For $k\ge 1$, we have in a similar way

$$\|\eta - \hat{\eta}_k\|_{\infty,\hat{A}_k} \le \|\eta - \bar{\eta}_{\hat{m}_k}\|_{\infty,\hat{A}_k} + \|\bar{\eta}_{\hat{m}_k} - \hat{\eta}_k\|_{\infty,\hat{A}_k}.$$

By (15) and Proposition 10 applied for $A := \hat{A}_k$ and $n := n_k^{\mathrm{act}} = \lfloor n_k \cdot \Pi(\hat{A}_k)\rfloor$ (conditionally on $\bigcup\limits_{i=0}^{k-1} S_k$, where S_k is the subsample used by the Algorithm 1 on step k), with probability $\ge 1-\alpha$

$$\begin{aligned}\|\bar{\eta}_{\hat{m}_k} - \hat{\eta}_k\|_{\infty,\hat{A}_k} &\le C\log^{1/2}(1/\varepsilon\alpha)\sqrt{\frac{2^{d\hat{m}_k}}{n_k}} \le C\log^{1/2}(1/\varepsilon\alpha)\left(\sqrt{\frac{2^{d\hat{m}_0}}{n_0}}\right)^{\prod_{i=1}^k \tau_k} \le \\ &\le C\log^{1/2}(1/\varepsilon\alpha) n_k^{-\beta/(2\beta+d)}.\end{aligned} \tag{17}$$

Once again, for the bias term we have by (14)

$$\begin{aligned}\|\eta-\bar{\eta}_{\hat{m}_k}\|_{\infty,\hat{A}_k} &\le CB_1 2^{-\beta\hat{m}_k} = CB_1\left(2^{-\beta\hat{m}_0}\right)^{\prod_{i=1}^k\tau_i}\le\\ &\le C(\nu,d)B_1\left[\left(\log\frac{1}{\varepsilon\alpha}\right)^{1+\frac{d}{2\beta}}\sqrt{\frac{2^{d\hat{m}_0}\hat{m}_0}{n_0}}\right]^{\prod_{i=1}^k\tau_i}\le\\ &\le C(\nu,d)\left[\left(\log\frac{1}{\varepsilon\alpha}\right)^{2+\frac{d}{2\nu}}\right]^{\prod_{i=1}^k\tau_i}\sqrt{\frac{2^{d\hat{m}_k}}{n_k}} := \frac{\delta_k}{2}\le\\ &\le C(\nu,d)\left[\left(\log\frac{1}{\varepsilon\alpha}\right)^{2+\frac{d}{2\nu}}\right]^{\prod_{i=1}^k\tau_i} n_k^{-\beta/(2\beta+d)},\end{aligned}\tag{18}$$

which holds with probability $\ge 1-\alpha$ and gives together with (17) that

$$\|\eta-\hat{\eta}_k\|_{\infty,\hat{A}_k}\le\frac{\delta_k}{2}\le C(\nu,d)\left[\left(\log\frac{1}{\varepsilon\alpha}\right)^{2+\frac{d}{2\nu}}\right]^{\prod_{i=1}^k\tau_i} n_k^{-\beta/(2\beta+d)}\tag{19}$$

with probability $\ge 1-2\alpha$. Finally, it remains to note that for all $1\le k\le J$, with probability $\ge 1-2\alpha$,

$$\prod_{i=1}^k\tau_i\le\prod_{i=1}^J\tau_i\le 2\frac{2\nu+d}{\nu} := \bar{\tau}.\tag{20}$$

This follows from the observation that on step $J-1$, we have $\delta_{J-1}>\varepsilon$, hence

$$C\left[\left(\log\frac{1}{\varepsilon\alpha}\right)^{2+\frac{d}{2\nu}} n_0^{-\beta/(2\beta+d)}\right]^{\prod_{i=1}^{J-1}\tau_i}\ge\varepsilon.$$

Plugging in the value of $n_0=\lfloor\varepsilon^{-\nu}\rfloor$, we get the bound.

The union bound over all $0\le k\le J$ gives that, with probability $\ge 1-4(J+1)\alpha$, on every iteration we have

$$\|\eta-\hat{\eta}_k\|_{\infty,\hat{A}_k}\le\frac{\delta_k}{2}\le\bar{C}(\nu,\Pi)\left(\log\frac{1}{\varepsilon\alpha}\right)^{\bar{\tau}(2+\frac{d}{2\nu})} n_k^{-\beta/(2\beta+d)}.$$